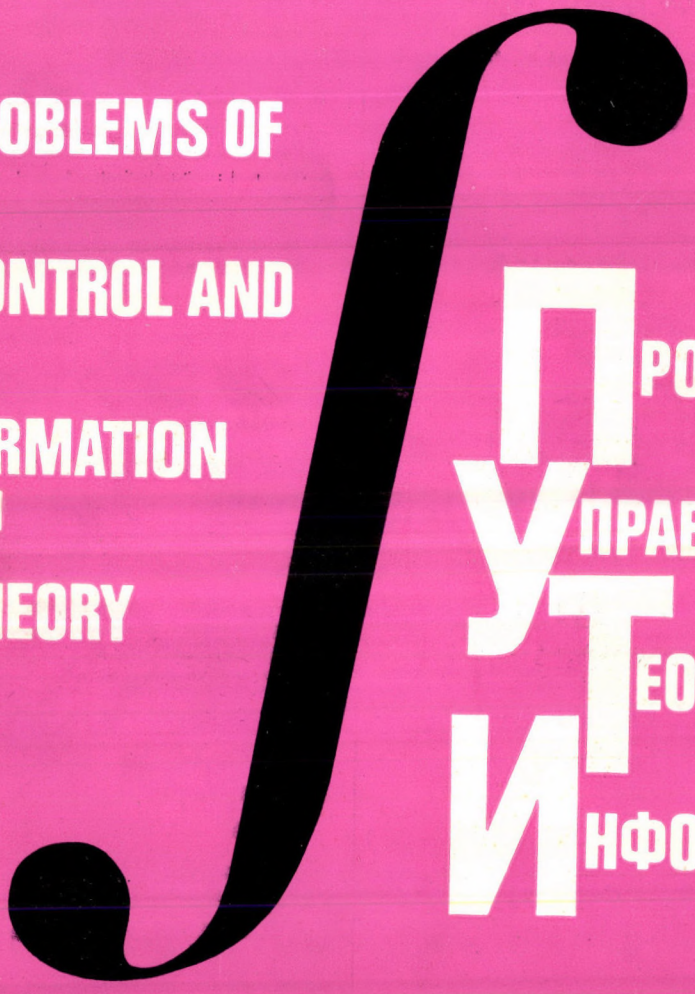
316.920
5
1976

VOL. TOM 5 * NUMBER HOMEP 1

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY



ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1976

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

is an international quarterly sponsored jointly by the Presidium of the Academy of Sciences of the USSR and of the Hungarian Academy of Sciences. It offers publicity for original papers and short communications of the following topics:

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

— theory and methods of adaptation, learning, identification and pattern recognition

— methods of information processing; application of digital computers in control and communication systems

— new physical principles in constructing technical devices for automation, control and information processing

The six issues published per year make up a volume of some 480 pp.

While this quarterly is mainly a publication forum of the research results achieved within the USSR and Hungary, also papers of international interest from other countries are welcome.

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

H-1389 Budapest 62, P.O.B. 149, Hungary

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

H-1363 Budapest P.O.B. 24, Hungary

Академия Наук СССР и Академия Наук ВНР совместно издают журнал

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал публикует материалы о новых результатах научных исследований, разработок и краткие сообщения о достижениях в следующих областях науки:

— общая теория процессов управления и теория систем
— теория автоматов
— теория информации
— исследование операций; теория сложных систем
— теория управления экономическими системами; модели систем
— теория и методы адаптации, обучения, идентификации и распознавания образов
— методы обработки информации; применение вычислительных машин в системах управления и передачи информации
— новые физические принципы создания технических средств автоматики, управления и передачи информации

Журнал быхогит шесть раз в год общим объёмом около 480 печатных страниц.

Журнал публикует статьи о научных достижениях обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

H-1389 Будапешт 62, п. о. 149, Венгрия

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии Наук Венгрии

H-1363 Будапешт п. о. 24, Венгрия

316.920

PROBLEMS OF CONTROL AND INFORMATION THEORY

Volume 5, 1976

PAPERS

Beneš, J.: Stochastic control of impulses on plane crystallographic group lattices. **5**, 1, pp. 13—29 13
Butkovskii, A. G., Mustafaev, M. I.: Fundamental finite control. **5**, 1, pp. 71—85 71
Donkov, D., Glushkov, L.: Simulation models of computing systems with two different memory structures. **5**, 3, pp. 239—246 239
Efimov, E. I.: Purposeful system behavior logic. **5**, 3, pp. 247—261 247
Emelyanov, S. V., Larichev, O. I.: A multicriteria approach to applied R and D planning: the case of qualitative criteria. **5**, 5—6, pp. 385—399 385
Frey, T., Gyürki, J.: Recursive state and parameter estimation in nonlinear stochastic systems, I. **5**, 1, pp. 39—49 39
Frigyes, A.: A special method of fast Fourier tpansform to determine the low and middle frequency characteristic of a control system **5**, 4, pp, 303—320 303
Fritz, J.: Two optimization methods in nearest neighbor pattern classification. **5**, 3, pp. 227—238 227
Fritz, J., Györfi, L.: On the minimization of classification error probability in statistical pattern recognition. **5**, 4, pp. 371—382 371
Giorgadze, A. Kh.: An approach to decomposition of probabilistic automata. **5**, 4, pp. 321—327 321
Goppa, V. D.: Decoding and diophantine approximations. **5**, 3, pp. 195—206 . 195
Grodzki, Z.: New class of the controlled shift-registers. I. **5**, 2, pp. 181—188 181
Grodzki, Z.: New class of the controlled shift-registers. II. **5**, 3, pp. 263—270 . 263
Györfi, L.: An upper bound of error probabilities for multihypotheses testing and its application in adaptive pattern recognition. **5**, 5—6, pp. 449—457 ... 449
Kárpáti, A., Szentai, E., Ipsits, I., Hermann, I.: Computerized reliability study of thyristorized continuous power supply systems. **5**, 5—6, pp. 459—471 459
Kazakov, I. E.: Analytical synthesis of the statistical quasi-optimal control of the nonlinear subject. **5**, 1, pp. 51—61 51
Krasovskii, A. A.: A generalization of a control system analytical design problem. **5**, 1, pp. 3—12 3
Krasovskii, N. N.: Game-theoretic control under incomplete phase-state information. **5**, 4, pp. 291—302 291
Lantos, B.: Necessary conditions for the optimality in abstract optimum control problems with nonscalar-valued performance criterion. **5**, 3, pp. 271—284 271
Lunderstädt, R.: Transportation systems with optimal parameters. **5**, 2, pp. 109—116 109
Masgras, V. V.: Connection between the weight distributions of a group code and its dual. **5**, 4, pp. 359—369 291
Mikhailov, F. A.: The method of generalized characteristic equation in the theory of linear time-varying systems. **5**, 2, pp. 149—165 149
Nevelson, M. B.: On stochastic approximation processes by sums of independent random variables. **5**, 2, pp. 117—133 117

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Petrov, B. N., Dobrushin, R. L., Pinsker, M. S., Ulanov, G. M., Ulyanov, S. V.: On some interrelations between the theories of information and control. **5,** 1, pp. 31—38 31
Pospelov, D. A., Cherezova, N. N., Shapiro, D. I.: Situation recognition in the pursuit problem. **5,** 5—6, pp. 421—436 421
Pupkov, K. A., Moroz, A. I.: Contribution to the theory of relay-linear control laws. **5,** 5—6, pp. 401—419 401
Raczyński, S.: A global optimization procedure for dynamic systems. **5,** 3, pp. 207—225 207
Rao, A. N. V., Tsokos, C. P.: On the existence and stability behavior of a stochastic integral equation in a Banach space. **5,** 1, pp. 87—95 87
Rejtő, L.: Pattern recognition in a Markovian case. **5,** 2, pp. 135—147 135
Romanov, V. G.: On method of moment dynamics. **5,** 5—6, pp. 437—448 437
Rózsa, P., Sinha, N. K.: Decoupling and pole placement in linear multivariable systems: A direct algebraic approach. **5,** 4, pp. 329—358 329
Schwarz, H.: Markov-parameter representation of linear time-variable systems with separable system functions. **5,** 1, pp. 63—70 63
Sultan, M. A.: A method of analysis of a limit cycling relay system. **5,** 2, pp. 167—180 167

AUTHORS

Benes, J. **5**, 1, pp. 13—29
Butkovskii, A. G. **5**, 1, pp. 71—85
Cherezova, N. N. **5**, 5—6, pp. 421—436
Dobrushin, R. L. **5**, 1, pp. 31—38
Donkov, D. **5**, 3, pp. 239—246
Efimov, E. I. **5**, 3, pp. 247—261
Emelyanov, S. V. **5**, 5—6, pp. 385—399
Frey, T. **5**, 1, pp. 39—49
Frigyes, A. **5**, 4, pp. 303—320
Fritz, J. **5**, 3, pp. 227—238
Giorgadze, A. Kh. **5**, 4, 331—327
Glushkov, L. **5**, 4, pp. 239—246
Goppa, V. D. **5**, 3, pp. 195—206
Grodzki, Z. **5**, 2, pp. 181—188; **5**, 3, pp. 263—270
Györfi, L. **5**, 4, pp. 371—382; **5**, 5—6, pp. 449—457
Gyürki, J. **5**, 1, pp. 39—49
Hermann, I. **5**, 5—6, pp. 459—471
Ipsits, I. **5**, 5—6, pp. 459—471
Kárpáti, A. **5**, 5—6, pp. 459—471
Kazakov, I. E. **5**, 1, pp. 51—61
Krasovskii, A. A. **5**, 1, pp. 3—12
Krasovskii, N. N. **5**, 4, pp. 291—302
Lantos, B. **5**, 3, pp. 271—284
Larichev, O. I. **5**, 5—6, pp. 385—399
Lunderstädt, R. **5**, 2, pp. 109—116
Masgras, V. V. **5**, 4, pp. 359—369
Mikhailov, F. A. **5**, 2, pp. 149—165
Mustafaev, M. I. **5**, 1, pp. 71—85
Moroz, A. I. **5**, 5—6, pp. 401—419
Nevelson, M. B. **5**, 2, pp. 117—133
Petrov, B. N. **5**, 1, pp. 31—38
Pinsker, M. S. **5**, 1, pp. 31—38
Pospelov, D. A. **5**, 5—6, pp. 421—436
Pupkov, K. A. **5**, 5—6, pp. 401—419
Raczyński, S. **5**, 3, pp. 207—225
Raibman, N. S. **5**, 1, pp. 97—105
Rao, A. N. V. **5**, 1, pp. 87—95
Rejtő, L. **5**, 2, pp. 134—147
Rodionova, N. I. **5**, 2, pp. 189—192
Romanov, V. G. **5**, 5—6, pp. 437—448
Rózsa, P. **5**, 4, pp. 329—357
Schwarz, H. **5**, 1, pp. 63—70
Shapiro, D. I. **5**, 5—6, pp. 421—436
Sinha, N. K. **5**, 4, pp. 329—357
Sultan, M. A. **5**, 2, pp. 167—180
Szentai, E. **5**, 5—6, pp. 459—471
Tsokos, C. P. **5**, 1, pp. 87—95
Ulanov, G. M. **5**, 1, pp. 31—38
Ulyanov, S. V. **5**, 1, pp. 31—38

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Том 5, 1976

СТАТЬИ

Бенеш Й.: Стохастическое управление импульсами в решетках плоских кристаллографических групп. **5**, 1, стр. 13—30 13
Бутковский А. Г., Мустафаев М. И.: Фундаментальные финитные управления. **5**, 1, стр. 71—85 71
Гиоргадзе А. Х.: Об одном подходе к задаче декомпозиции вероятностных автоматов. **5**, 4, стр. 321—328 321
Гоппа В. Д.: Декодирование и диофантовы приближения **5**, 3, стр. 195—206 195
Гродзки З.: Новый класс управляющих регистров сдвига. I. **5**, 2, стр. 181—188 181
Гродзки З.: Новый класс управляющих регистров сдвига. II. **5**, 3, стр. 263—270 263
Дерфи Л.: Верхняя граница вероятности ошибки при исследовании с несколькими гипотезами и применение результатов для адаптивного распознавания образов. **5**, 5—6, стр. 449—457 449
Донков Д., Глушков Л.: Модели вычислительных систем с двумя различными структурами памяти. **5**, 3, стр. 239—246 239
Емельянов С. В., Ларичев О. И.: Многокритериальный подход к проблеме планированная прикладных научных исследований и разработок при качественных критериях. **5**, 5—6, стр. 385—399 385
Ефимов Е. И.: Логика поведения целенаправленных систем. **5**, 3, стр. 247—261 247
Казаков И. Е.: Аналитический синтез статистически квазиоптимального управления нелинейным объектом. **5**, 1, стр. 51—61 51
Карпати А., Сентаи Е., Ипшич И., Херманн И.: Анализ надежности бесперебойного источника энергии на тиристорах с помощью вычислительной машины. **5**, 5—6, стр. 459—471 459
Красовский А. А.: Об одном обобщении аналитического конструирования систем управления. **5**, 1, стр. 3—12 3
Красовский Н. Н.: Игровое управление при неполной фазовой информации. **5**, 4, стр. 291—302 291
Лантои Б.: Необходимые условия оптимальности по нескалярному критерию в абстрактных проблемах оптимального управления. **5**, 3, стр. 271—284 271
Лундерштедт Р.: Транспортные системы автоматического управления с оптимальными параметрами. **5**, 2, стр. 109—116 109
Масграс В. В.: Связь между некоторым групповым кодом и весовыми распределениями его дуального кода. **5**, 4, стр. 359—369 359
Михайлов Ф. А.: Метод обобщенного характеристикого уравнения в теории нестационарных линейных систем. **5**, 2, стр. 149—165 149
Невельсон М. Б.: О приближении процессов стохастической аппроксимации суммами независимых случайных величин. **5**, 2, стр. 117—133 117
Петров Б. Н., Добрушин Р. Л., Пинскер М. С., Уланов Г. М., Ульянов С. В.: О некоторых взаимосвязях теории информации и теории управления. **5**, 1, стр. 31—38 ... 3
Поспелов Д. А., Черезова Н. Н., Шапиро Д. И.: Распознавание синтуаций в задаче о преследовании, **5**, 5—6, стр. 421—436 421
Пупков К. А., Мороз А. И.: К теории релейно-линейных законов управления. **5**, 5—6, стр. 401—419 401

Райбман Н. С.: Шестой Конгресс ИФАК а. **5**, 1, стр. 97—105 97
Рао А. Н. В., Цокос С. П.: К проблеме существования и устойчивости некоторого стохастического уравнения в банаховом пространстве. **5**, 1, стр. 87—95 87
Рачински С.: Метод глобальной оптимизации динамических систем. **5**, 3, стр. 207—225 207
Рейто Л.: Распознавание образов при марковских процессах. **5**, 2, стр. 135—147 ... 135
Родионова Н. И.: IV-ая Международная конференция по искусственному интеллекту. **5**, 2, стр. 189—192 189
Рожа П., Синха Н. К.: Разбиение линейных время-инвариантных многомерных систем. **5**, 4, стр. 329—358 329
Романов В. Г.: К вопросу о методе динамики моментов. **5**, 5—6, стр. 437—448 437
Сультан М. А.: Метод анализа предельного цикла релейной системы автоматического регулирования. **5**, 2, стр. 167—180 167
Фрей Т., Дюрки Й.: Рекуррентный алгоритм оценки состояния и параметров нелинейных стохастических систем. **5**, 1, стр. 39—49 39
Фридеш А.: Специальный метод ускоренного преобразования Фурье для определения областей низких и средних частот частотной характеристики системы автоматического управления. **5**, 4, стр. 303—320 303
Фриц Й.: Два способа оптимизации при использовании метода классификации по правилам «бижайших соседей». **5**, 3, стр. 227—238 227
Фриц Й., Дерфи Л.: О минимизации вероятности ошибки классификации в области статистического распознавания. **5**, 4, стр. 371—382 371
Шварц Х.: Параметрическое марковское представление сигналов линейных нестационарных систем с сепарируемой системной функцией. **5**, 1, стр. 63—70 63

АВТОРЫ

Бенеш Й. **5**, 1, стр. 13—29
Бутковский А. Г. **5**, стр. 71—85
Гиоградзе А. Х. **5**, 4, стр. 321—327
Глушков Л. **5**, 3, стр. 239—246
Гоппа В. Д. **5**, 3, стр. 195—206
Гродзки З. **5**, 2, стр. 181—188; **5**, 3, стр. 263—270
Дерфи Л. **5**, 4, стр. 371—382; **5**, 5—6, стр. 449—457
Добрушин Р. Л. **5**, 1, стр. 31—38
Донков Д. **5**, 3, стр. 239—246
Дюрки Й. **5**, 1, стр. 39—49
Емельянов С. В. **5**, 5—6, стр. 385—399
Ефимов Е. И. **5**, 3, стр. 247—261
Казаков И. Е. **5**, 1, стр. 51—61
Карпати А. **5**, 5—6, стр. 459—471
Красовский А. А. **5**, 1, стр. 3—12
Красовский Н. Н. **5**, 4, стр. 291—302
Лантош Б. **5**, 3, стр. 271—284
Ларичев О. И. **5**, 5—6, стр. 385—399
Масграс В. В. **5**, 4, стр. 359—369
Михайлов Ф. А. **5**, 2, стр. 149—165
Мороз А. И. **5**, 5—6, стр. 401—419
Мустафаев М. И. **5**, 1, стр. 71—85
Невельсон М. Б. **5**, 2, стр. 117—133
Петров Б. Н. **5**, 1, стр. 31—38
Пинскер М. С. **5**, 1, стр. 31—38
Поспелов Д. А. **5**, 5—6, стр. 421—436
Пупков К. А. **5**, 5—6, стр. 401—419
Райбман Н. С. **5**, 1, стр. 97—105
Рао А. Н. В. **5**, 1, стр. 87—95
Рачински С. **5**, 3, стр. 207—225
Рейто Л. **5**, 2, стр. 135—141
Родинова Н. И. **5**, 2, стр. 189—192
Рожа П. **5**, 4, стр. 329—357
Романов В. Г. **5**, 5—6, стр. 437—448
Сентаи Е. **5**, 5—6, стр. 459—471
Синха Н. К. **5**, 4, стр. 329—357
Султан М. А. **5**, 2, стр. 167—180
Уланов Г. М. **5**, 1, стр. стр. 31—38
Ульянов С. В. **5**, 1, стр. 31—38
Фрей Т. **5**, 1, стр. 39—49
Фридеш А. **5**, 4, стр. 303—320
Фриц Й. **5**, 3, стр. 227—238; **5**, 4, стр. 371—382
Цокос С. П. **5**, 1, стр. 87—95
Черезова Н. Н. **5**, 5—6, стр. 421—436
Шапиро Д. И. **5**, 5—6, стр. 421—436
Шварц, Х. **5**, 1, стр. 63—70

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)
F. CSÁKI (Budapest)

DEPUTY EDITORS

V. S. PUGACHEV (Moscow)
V. I. SIFOROV (Moscow)
S. CSIBI (Budapest)

COORDINATING EDITORS

M. A. GAVRILOV (Moscow)
L. KALMÁR (Szeged)

S. V. EMELYANOV (Moscow)
I. CSISZÁR (Budapest)

N. N. KRASOVSKII (Sverdlovsk)
A. PRÉKOPA (Budapest)

E. P. POPOV (Moscow)
L. VARGA (Budapest)

E. D. TERYAEV (Moscow)
J. KOCSIS (Budapest)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)
Ф. ЧАКИ (Будапешт)

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Ш. ЧИБИ (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

М. А. ГАВРИЛОВ (Москва)
Л. КАЛМАР (Сегед)

С. В. ЕМЕЛЬЯНОВ (Москва)
И. ЧИСАР (Будапешт)

Н. Н. КРАСОВСКИЙ (Свердловск)
А. ПРЕКОПА (Будапешт)

Е. П. ПОПОВ (Москва)
Л. ВАРГА (Будапешт)

Е. Д. ТЕРЯЕВ (Москва)
Я. КОЧИШ (Будапешт)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 5 (1), pp. 3—12 (1976)

ОБ ОДНОМ ОБОБЩЕНИИ ЗАДАЧИ АНАЛИТИЧЕСКОГО КОНСТРУИРОВАНИЯ СИСТЕМ УПРАВЛЕНИЯ

А. А. КРАСОВСКИЙ

(Москва)

(Поступила в редакцию 14 апреля 1975 г.)

Для объекта автоматического управления применена операторно-дифференциальная форма описания, в которой кроме производной по времени вектора главных координат фигурируют функционалы этого вектора и линейно-входящее управление. Показано, что на подобные объекты распространяется основная теорема аналитического конструирования по критерию обобщенной работы с тем отличием, что место функций (характеристик объекта) занимают функционалы. Для решения определяющего уравнения с частными производными и функционалами в качестве коэффициентов рекомендуется итерационный метод. Для линейного стационарного объекта, представленного в операторно-дифференциальной форме, указано общее точное решение задачи синтеза стационарного управления.

В обычной хорошо разработанной форме метода аналитического конструирования систем управления для многомерных многосвязных объектов существуют следующие трудности. Размерность систем уравнений, определяющих коэффициенты оптимальных управлений, весьма велика. Как правило, многие компоненты полного вектора состояния объекта не поддаются непосредственному измерению (неполная степень непосредственной наблюдаемости). Реализация оптимальных управлений с большим количеством составляющих сопряжена с трудностями. Все эти трудности имеют место и при применении критерия обобщенной работы [1], [2], хотя этот вариант аналитического конструирования наиболее эффективен. Для преодоления указанных трудностей в данной работе предлагается операторно-дифференциальная форма представления объекта, обуславливающая дифференциально-операторную форму всех последующих построений.

1. Нелинейный объект

Представим объект управления в виде

$$\dot{x}_i + F_i(x_1, \ldots, x_n) = \sum_{j=1}^{m} \Phi_{ij}(x_1, \ldots, x_n)\, u_j, \; i = 1, 2, \ldots, n, \qquad (1)$$

где F_i, Φ_{ij} — операторы, воздействующие на вектор $x = (x_1, \ldots, x_n)$ (функционалы), u_j — синтезируемые управления. Уравнения (1) лишь по внешнему виду напоминают дифференциальные уравнения в форме Коши, так как здесь вместо функций фигурируют функционалы. Функционалы F_i, Φ_{ij} могут быть интегральными, тогда (1) будет представлять собой систему интегро-дифференциальных уравнений. Функционалы F_i, Φ_{ij} могут иметь вид передаточных функций с входными величинами $x_1, \ldots, x_n$ (линейный вариант) или нелинейными функциями этих величин на входах (нелинейный вариант), и так да-

лее. В частном случае, когда функционалы F_i, Φ_{ij} вырождаются в функции, система (1) вырождается в дифференциальные уравнения в форме Коши. Вектор $x = (x_1, \ldots, x_n)$, именуемый ниже главным, можно рассматривать как проекцию полного вектора состояния объекта; размерность последнего может намного превышать n и даже быть бесконечной. Выбор главного вектора x подчиняется следующим условиям. Все компоненты $x_1, \ldots, x_n$ должны быть доступны для непосредственного измерения. Эти компоненты должны представлять собой наиболее важные координаты объекта, в которых можно сформулировать требования к качеству процессов управления.

Справедливо следующее положение, являющееся полным аналогом основной теоремы аналитического конструирования по критерию обобщенной работы [2]. Для объекта (1) оптимальными управлениями, минимизирующими функционал

$$I = V_3[x_1(t_2), \ldots, x_n(t_2)] + \int_{t_1}^{t_2} Q(x_1 \ldots, x_n)\, dt + \frac{1}{2}\int_{t_1}^{t_2} \sum_{j=1}^{m} \frac{u_j^2 + u_{j\text{оп}}^2}{k_j^2}\, dt, \quad (2)$$

являются управления

$$u_j = u_{j\text{оп}} = -k_j^2 \sum_{k=1}^{n} \varphi_{kj} \frac{\partial V}{\partial x_k}, \quad (3)$$

где V есть решение уравнения

$$\frac{\partial V}{\partial t} - \sum_{i=1}^{n} F_i \frac{\partial V}{\partial x_i} = -Q \quad (4)$$

при граничном условии $V_{t=t_2} = V_3$; V_3, Q, k_j — заданные функционал, функция и коэффициенты соответственно. Отличие от обычной формулировки теоремы заключается лишь в том, что V_3, F_i, Φ_{kj} в выражениях (1), (4) не функции, а операторы над вектором x. Доказательство может осуществляться различными путями, в том числе способом, полностью аналогичным использованному в [1], [2]. Действительно, интегрируя в пределах от t_1 до t_2 полную производную V, которая в силу уравнений (1) и соотношения (4) равна

$$\dot V = \frac{\partial V}{\partial t} + \sum_{i=1}^{n} \frac{\partial V}{\partial x_i}\left(-F_i + \sum_{j=1}^{m} \varphi_{ij} u_j\right) = -Q + \sum_{i=1}^{n}\sum_{j=1}^{m} \Phi_{ij} \frac{\partial V}{\partial x_i} u_j;$$

находим

$$I = V_{t=t_1} + \sum_{i=1}^{n}\sum_{j=1}^{m} \int_{t_1}^{t_2} \varphi_{ij} \frac{\partial V}{\partial x_i} u_i\, dt + \frac{1}{2}\sum_{j=1}^{m} \int_{t_1}^{t_2} \frac{u_j^2 + u_{j\text{оп}}^2}{k_j^2}\, dt =$$

$$= V_{t=t_1} + \frac{1}{2}\int_{t_1}^{t_2} \sum_{j=1}^{m} \frac{(u_{j\text{оп}} - u_j)^2}{k_j^2}\, dt.$$

Из этого выражения непосредственно видно, что величина I минимальна при $u_j = u_{j\text{оп}}$.

Для нахождения оптимальных управлений $u_{j\text{оп}}$ необходимо решить уравнение (4), которое в да нном случае является операторно-дифференциальным (в частных производных) уравнением. Для решения этой задачи естественно применить итерационный или операцио нныйметод решения линейных уравнений в частных производных [3], [4], причем как обычно целесообразно различать терминальную и нетерминальную (стационарную) задачи [2]. Для терминальной задачи с критерием (2) решение ищется в виде $V = V_1 + V_2 + \ldots$, где

$$\frac{\partial V_1}{\partial t} = -Q, \quad \frac{\partial V_{\nu+1}}{\partial t} = \sum_{i=1}^{m} F_i \frac{\partial V_\nu}{\partial x_i}, \quad \nu = 1, 2, \ldots.$$

причем

$$(V_1)_{t=t_2} = V_3, \quad (V_{\nu+1})_{j=t_2} = 0.$$

Если ряд $V_1 + V_2 + \ldots$ сходится равномерно вместе со своими первыми производными, то он сходится к решению уравнения (4). Это проверяется простым суммированием. Если ввести обозначения операторов

$$D = \frac{\partial}{\partial t}, \quad \frac{1}{D} = \int [\quad] dt,$$

$$P = \frac{1}{D} \sum_{i=1}^{n} F_i \frac{\partial}{\partial x_i},$$

то терминальное решение принимает вид

$$V = -(1 + P + P^2 + \ldots) \frac{Q}{D}. \tag{5}$$

Однако в большинстве случаев задача аналитического конструирования оптимальных управлений ставится как нетерминальная. Для получения соответствующего решения воспользуемся способом нестационарного функционала [2]. Зададим функцию Q в виде $Q = Q_* \exp(-\lambda t)$, где Q_* — положительно определенная функция x; λ — действительная постоянная величина. При этом V целесообразно представить в виде

$$V = V_* \exp(-\lambda t).$$

В соответствии с этим уравнение (4) уступает место следующему

$$\frac{\partial V_*}{\partial t} - \lambda V_* - \sum_{i=1}^{n} F_i \frac{\partial V}{\partial x_i} = -Q_*.$$

Для стационарного случая, когда V_* не зависит явным образом от времени, $\frac{\partial V_*}{\partial t} = 0$ и

$$\lambda V_* + \sum_{i=1}^{n} F_i \frac{\partial V_*}{\partial x_i} = Q_* .$$

Обозначая оператор

$$L = \sum_{i=1}^{n} F_i \frac{\partial}{\partial x_i} ,$$

итерационное решение представляем в виде

$$V_* = \frac{1}{\lambda}\left(1 - \frac{1}{\lambda} L + \frac{1}{\lambda^2} L^2 - \ldots\right) Q_* . \tag{6}$$

Оптимальные управления в соответствии с (3) будут иметь вид

$$u_j = - k_j^2 \sum_{k=1}^{n} \Phi_{kj} \frac{\partial V}{\partial x_k} = - k_j^2 \exp(-\lambda t) R_j V_* ,$$

где

$$R_j = \sum_{k=1}^{n} \Phi_{kj} \frac{\partial}{\partial x_k} .$$

Для того, чтобы оптимальные управления не зависели явным образом от времени, задаем

$$k_j^2 = k_{j*}^2 \exp(\lambda t) ,$$

где k_{j*} — постоянные величины.

Таким образом, можно сформулировать следующий результат. Если ряд (6) равномерно сходится вместе с его первыми производными, то нетерминальными оптимальными стационарными управлениями, минимизирующими функционал

$$I_* = V_* \exp(-\lambda t) + \int Q_* \exp(-\lambda t)\, dt + \frac{1}{2} \int \sum_{j=1}^{m} \frac{u_j^2 + u_{j\text{оп}}^2}{k_{j*}^2} \exp(-\lambda t)\, dt ,$$

для стационарного объекта являются

$$u_j = u_{j\text{оп}} = - \frac{k_{j*}^2}{\lambda} R_j \left(1 - \frac{1}{\lambda} L + \frac{1}{\lambda^2} L^2 - \ldots\right) Q_n . \tag{7}$$

Для убыстрения сходимости следует увеличивать λ. Однако необходимо иметь в виду, что при неограниченном увеличении λ эффективность критерия I^* снижается, так как при $\lambda \to \infty$ он из интегрального критерия превращается в оценку мгновенных значений координат и управлений.

2. Линейный стационарный объект. Квадратичный критерий

Уравнения линейного стационарного объекта записываются в виде

$$\dot{x}_i + \sum_{k=1}^{n} W_{ik}(D)\, x_k = \sum_{j=1}^{m} b_{ij}\, u_j, \qquad i = 1, 2, \ldots, n, \quad (8)$$

где $W_{ik}(D)$ — передаточные функции — линейные операторы; b_{ij} — постоянные коэффициенты.

Задается квадратичный критерий

$$Q_* = \frac{1}{2} \sum_{\nu, \mu=1}^{n} \beta_{\nu\mu}\, x_\nu\, x_\mu, \qquad \beta_{\nu\mu} = \beta_{\mu\nu},$$

где $\beta_{\nu\mu}$ — постоянные коэффициенты.

В данном случае

$$L = \sum_{i,k=1}^{n} [W_{ik}(D)\, x_k] \frac{\partial}{\partial x_i}, \qquad R_j = \sum_{k=1}^{n} b_{kj} \frac{\partial}{\partial x_k}.$$

Учитывая, что

$$\frac{\partial}{\partial x_\mu} \sum_{k=1}^{n} W_{ik}(D)\, x_k = W_{i\mu}(D)\, 1 = W_{i\mu}(0)$$

— установившееся значение оператора*, воздействующего на постоянную величину, последовательно находим

$$LQ_* = \sum_{i,k,j=1}^{n} \beta_{ij} [W_{ik}(D)\, x_k]\, x_j = x^T \beta W(D)\, x,$$

$$L^2 Q_* = x^T \beta W(0)\, W(D) x + [W(D)x]^T \beta W(D)\, x,$$

$$L^3 Q_* = [W(D)x]^T [\beta W(0)\, W(D)\, x + \beta W^2(0) x + 2W^T(0)\, \beta W(D)\, x].$$

Здесь $W = \|(D)\, W_{ik}(D)\|$ — квадратная матрица передаточных функций, $\beta = \|\beta_{ik}\|$ — квадратная матрица коэффициентов критерия, k_*^2 — диагональная матрица коэффициентов усиления k_{jn}^2, b — прямоугольная матрица коэффициентов b_{ij}, x — вектор-столбец. Символ «T» обозначает транспонирование. Подставляя эти выражения в формулу (7), получаем

$$u_{\text{оп}} = -\frac{1}{\lambda} k^2 b^T \left\{ \beta - \frac{1}{\lambda} [\beta W(D) + W^T(0)\, \beta] + \right.$$

* Если $W_{i\mu}(0) = \infty$ и рассматриваются нестационарные управления, то обозначение $W_{i\mu}(0)$ следует понимать как условное, эквивалентное $W_{i\mu}(D)1$.

$$+\frac{1}{\lambda^2}\left[\beta W(0)\,W(D)+2\,W^T(0)\,\beta W(D)+W^{T2}(0)\,\beta\right]- \tag{9}$$

$$-\frac{1}{\lambda^3}\left[\beta W(0)\,W(D)+3\,W^T(0)\,\beta W(0)W(D)+3W^{T2}(0)\,\beta W(D)+W^{T3}(0)\,\beta\right]+\ldots\Big\}x.$$

Если это выражение записать в виде

$$u_{\text{оп}}=-\frac{1}{\lambda}k_*^2\,b^T\left[A_1-\frac{1}{\lambda}A_2(D)+\frac{1}{\lambda^2}A_3(D)-\ldots\right]x, \tag{10}$$

то $A_1=\beta$, а каждый последующий оператор $A_\nu(D)$, ($\nu=2,3,\ldots$) получается из предыдущего на основе простого рекуррентного соотношения

$$A_\nu(D)=W^T(0)\,A_{\nu-1}(D)+A_{\nu-1}(0)\,W(D). \tag{11}$$

Легко проверить, что ряд в фигурных скобках формулы (9) или в квадратных скобках формулы (10), умноженный на λ^{-1}, есть итерационное решение уравнения

$$\lambda A(D)+W^T(0)\,A(D)+A(0)W\,(D)=\beta. \tag{12}$$

Матрицу $A(0)$ можно определить из уравнения (12) при $D=0$:

$$\lambda A(0)+W^T(0)\,A(0)+A(0)\,W(0)=\beta. \tag{13}$$

Заметим, что если объект представлен в традиционной форме $\dot{x}+a\,x=bu$, где a — матрица постоянныхкоэффициентов, то $W(D)=W(0)=0$ и уравнения (12), (13) обращаются в известное уравнение:

$$\lambda A+a^T A+A\,a=\beta.$$

Уравнения (12), (13), определяющие оператор $A(D)$ оптимального управления, можно получить также иным путем. Так, будем искать решение уравнения

$$\lambda V_*+\sum_{i=1}^{n}F_i\frac{\partial V_*}{\partial x_i}=\lambda V_*+\sum_{i,k=1}^{n}[W_{ik}(D)\,x_k]\frac{\partial V_*}{\partial x_k}=\frac{1}{2}\sum_{i,k=1}^{n}\beta_{i\,k}x_i x_k \tag{14}$$

в виде

$$V_*=\frac{1}{2}\sum_{i,k=1}^{n}A_{ik}(D)\,(x_i\,x_k).$$

Подставляя это выражение в (14), дифференцируя один раз по x, и приравнивая нулю оператор над одним и тем же вектором, получаем (12). Дифференцируя второй раз, находим (13). Итак, итоговый результат для линейного стационарного объекта можно сформулировать так.

Если дан объект

$$\dot{x}+W(D)\,x=bu, \tag{15}$$

то стационарным управлением, минимизирующим критерий

$$I_* = V_* \exp(-\lambda t) + \frac{1}{2}\int (x^T \beta x) \exp(-\lambda t)\, dt + \\ + \frac{1}{2}\int (u^T k_*^{-2} u + u_{оп}^T k_*^{-2} u_{оп}) \exp(-\lambda t)\, dt, \qquad (16)$$

является

$$u = u_{оп} = -k_*^2 b^T A(D) x, \qquad (17)$$

где оператор $A(D)$ определяется уравнениями (12), (13).

Заметим, что в принципе аналогичная задача может быть решена и традиционным путем. Действительно, уравнение (15) путем введения полного вектора состояния может быть приведено к системе дифференциальных уравнений первого порядка — форме Коши. Используя функционал того же вида (16), синтезируя оптимальное управление традиционным путем, и исключая из этого управления посредством передаточных функций все компоненты полного вектора состояния, кроме главных координат, получим управление в виде (17). Однако этот путь, как уже отмечалось, для многомерных объектов отличается громоздкостью. Кроме того, результаты двух сопоставляемых путей синтеза могут быть принципиально различными. Действительно, в минимизируемом функционале (16) мы назначаем лишь одну составляющую. Остальные составляющие получаются автоматически и могут быть различными при разных решениях.

В частных случаях соотношения (12), (13), (17) могут быть еще более развернуты. Так, если $n = 1$, т.е. имеется только одна главная координата, то уравнение объекта и выражения (12), (13), (17) приобретают скалярную форму:

$$\dot{x}_1 + W_{11}(D) x_1 = u_1,$$
$$\lambda A_{11}(D) + W_{11}(0) A_{11}(D) + A_{11}(0) W_{11}(D) = \beta_{11},$$
$$\lambda A_{11}(0) + 2W_{11}(0) A_{11}(0) = \beta_{11},$$
$$u_{1оп} = -k_1^2 A_{11}(D) x.$$

Отсюда получаем

$$u_{1оп} = -\frac{\beta_{11} k_1^2}{\lambda + W_{11}(0)} \left[1 - \frac{W_{11}(D)}{\lambda + 2W_{11}(0)}\right] x.$$

Общее решение в явном виде может быть также записано в случае, когда $W(0) = 0$. В этом случае согласно (12), (13)

$$A(D) = \frac{1}{\lambda}\left[\beta - \frac{1}{\lambda}\beta W(D)\right]$$

и

$$u_{\text{оп}} = -\frac{1}{\lambda} k_*^2 b^T \left[\beta - \frac{1}{\lambda}\beta W(D)\right] x .$$

Эта формула получается также из выражения (9), если в нем положить $W(0) = 0$.

Пример. Линеаризованные уравнения продольного движения самолета при постоянной тяге двигателя и пренебрежении изменением плотности воздуха с высотой записываются в виде [2]:

$$\begin{aligned}
&\Delta\dot V + a_x^V \Delta V + a_x^\alpha \Delta\alpha + a_x^\theta \Delta\theta = 0,\\
&\Delta\dot\theta + a_y^\alpha \Delta\alpha + a_y^V \Delta V + a_x^\theta \Delta\theta = 0,\\
&\Delta\dot\omega_z + a_{m_z}^{\omega_z}\Delta\omega_z + a_{m_z}^\alpha \Delta\alpha + a_{m_z}^V \Delta V = a_{m_z}^{\delta_b}\Delta\delta_b,\\
&\Delta\dot v - \Delta\omega_z = 0, \quad \Delta n_y = \frac{V_0}{T_\theta}\Delta\alpha, \quad \Delta\vartheta = \Delta\theta + \Delta\alpha,
\end{aligned} \tag{18}$$

где ΔV, $\Delta\theta$, $\Delta\omega_z$, $\Delta\vartheta$, $\Delta\alpha$, Δn_y, $\Delta\delta_B$ — приращения скорости полета, угла наклона траектории, угловой скорости тангажа, угла тангажа, угла атаки, нормального ускорения и отклонения руля высоты соответственно. Через «a» с соответствующими индексами и T_θ обозначены коэффициенты. Эти коэффициенты будем считать постоянными. Вводя обозначения

$$\Delta n_y = x_1, \quad \Delta n_y = x_2, \quad b_{11} = \frac{V_0 a_{m_z}^{\delta_b}}{T_\theta}, \quad \Delta\delta_b = u_1,$$

$$\begin{cases}
W_{11}(D) = a_{m_z}^{\omega_z} - \dfrac{a_{m_z}^V a_x^\alpha + (D + a_{m_z}^{\omega_z})\,[a_y^\alpha (D + a_x^V) - a_x^\alpha a_y^V]}{(D + a_x^V)(D + a_y^\theta) - a_y^\theta a_y^V},\\
W_{12}(D) = a_{m_z}^\alpha + a_{m_z}^V \dfrac{a_y^\alpha a_x^\theta - a_x^\alpha a_y^\theta}{(D + a_x^V)(D + a_y^\theta) - a_x^\theta a_y^V},
\end{cases} \tag{19}$$

преобразуем уравнения (18) к виду

$$\begin{aligned}
&\dot x_1 + W_{11}(D)\, x_1 + W_{12}(D)\, x_2 = b_{11} u_1,\\
&\dot x_2 - x_1 = 0,
\end{aligned}$$

или

$$\dot x + W(D)\, x = b u,$$

где

$$W(D) = \begin{Vmatrix} W_{11}(D) & W_{12}(D) \\ -1 & 0 \end{Vmatrix}, \quad b = \begin{Vmatrix} b_{11} \\ 0 \end{Vmatrix}.$$

Требуется синтезировать управление u_1, обеспечивающее оптимальный переходный процесс по перегрузке, точнее, по главным координатам $x_1 = \Delta n_y$, $x_2 = \Delta n_y$.

Задаем Q_* в виде

$$Q_* = \frac{1}{2}(\beta_{11} x_1^2 + \beta_{22} x_2^2).$$

Уравнения (12), (13) в раскрытой форме в данном случае имеют вид

$$\begin{cases} [\lambda + W_{11}(0)]\, A_{11}(D) - A_{12}(D) - A_{12}(0) = \beta_{11} - A_{11}(0)\, W_{11}(D), \\ [\lambda + W_{11}(0)]\, A_{12}(D) - A_{22}(D) = -A_{11}(0)\, W_{12}(D), \\ W_{12}(0)\, A_{12}(D) + \lambda A_{22}(D) = \beta_{22} - A_{21}(0)\, W_{12}(D), \end{cases} \tag{20}$$

$$\begin{cases} [\lambda + \sqrt{2}\, W_{11}(0)]\, A_{11}(0) - 2A_{12}(0) = \beta_{11}, \\ W_{12}(0)\, A_{11}(0) + [\lambda + W_{11}(0)]\, A_{12}(0) - A_{22}(0) = 0, \\ 2W_{12}(0)\, A_{12}(0) + \lambda A_{22}(0) = \beta_{22}. \end{cases} \tag{21}$$

Оптимальное управление согласно (17) имеет вид

$$u_1 = k_{1*}^2\, b_{11}\, [A_{11}(D)\, \Delta n_y + A_{12}(D)\, \Delta n_y].$$

Кривая 1 на рисунке соответствует переходному процессу по перегрузке при отсутствии управления u_1. Этот процесс вычислен при следующих значениях параметров

$a = 0{,}017$ сек$^{-1}$, $a_x^\alpha = 0{,}15$ м сек$^{-2}$, $a_x^\theta = 0{,}17$ м сек$^{-2}$,

$a_y^\alpha = -0{,}84$ сек$^{-1}$, $a_y^V = -0{,}003$ м сек$^{-1}$, $a_y^\theta = 0$,

$a_{m_z}^{\omega_z} = 0{,}66$ сек$^{-1}$, $a_{m_z}^\alpha = 55$ сек$^{-2}$, $a_{m_z}^V = -0{,}022$ м$^{-1}$ сек$^{-1}$,

$a_{m_z}^{\delta_b} = -22{,}5$ сек$^{-2}$, $V_0 = 575$ м сек$^{-1}$, $T_\theta = 1{,}7$ сек.

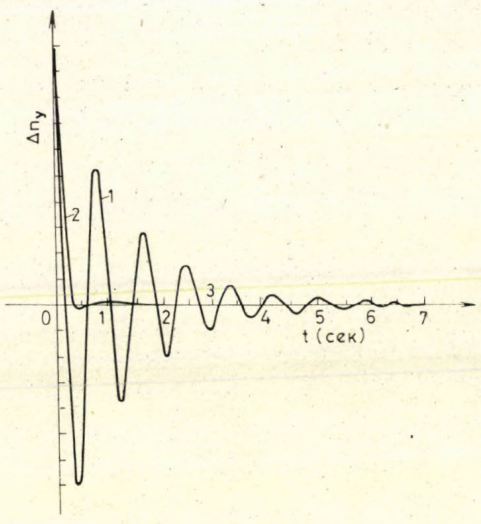


Рис. 1

При этих значениях параметров и $\lambda = -1$, $\beta_{11} = 10$, $\beta_{22} = 100$ по формулам (19)—(21) определено оптимальное управление:

$$u_1 = 3{,}3\,10^{-3}\left(\frac{0{,}506\,D^2 + 0{,}00277\,D + 0{,}00013}{D^2 + 0{,}017\,D + 0{,}00051}\,\Delta\dot{n}_y + \right.$$

$$\left. + \frac{1{,}33\,D^2 + 0{,}0226\,D + 0{,}0006}{D^2 + 0{,}017\,D + 0{,}00051}\,\Delta n\right) \approx$$

$$\approx 3{,}3\cdot 10^{-3}\left(\frac{0{,}506\,D + 0{,}0028}{D + 0{,}017}\,\Delta\dot{n}_y + \frac{1{,}33\,D + 0{,}023}{D + 0{,}017}\,\Delta n_y\right).$$

Переходный процесс в замкнутой системе, соответствующий этому управлению, представлен кривой 2.

Вычислительные преимущества предлагаемого «расширенного метода аналитического конструирования по критерию обобщенной работы» для многомерных многосвязных объектов весьма существенны.

Литература

1. *Красовский А. А.* Аналитическое конструирование контуров управления летательными аппаратами, «Машиностроение», 1969.
2. *Красовский А. А.* Системы автоматического управления полетом и их аналитическое конструирование, «Наука», 1973.
3. *Буков В. Н., Красовский А. А.* Операционный алгоритм оптимального управления, Автоматика и телемеханика, **10,** 1974.
4. *Красовский А. А., Буков В. Н.* Итерационное решение уравнения Колмогорова в статистической динамике непрерывных систем, ДАН, **216,** стр. 485, 1974.

A generalization of a control system analytical design problem

A. A. KRASOVSKII

(Moscow)

For an automatically controlled object an operator-differential form of a description is applied in which besides the derivative with respect to the time of the principal coordinate vector the functionals of this vector and the linear entering control appear. It is shown that the basic theorem of analytical design on the general work criterion is extended to similar objects, with such a difference that instead of functions (characteristics of the object) the functionals are shown. To solve a determining equation with partial derivatives and functionals as coefficients the iterative method is recommended. A general solution of a stationary control synthesis problem is shown for a linear stationary object represented in operator-differential form.

А. А. Красовский
Научный совет по технической кибернетике
СССР Москва В-333,
ул. Вавилова, 40

Problems of Control and Information Theory, Vol. 5 (1), pp. 13–29 (1976)

STOCHASTIC CONTROL OF IMPULSES ON PLANE CRYSTALLOGRAPHIC GROUP LATTICES

J. BENEŠ

(Prague)

(Received October 31, 1974)

Feedback control of intentionally oriented probabilistic spread of impulses on plane crystallographic group lattices is based here upon the information theoretical approach to the control of complexes. Epsilon-entropy and epsilon-capacity of the set of end points of the vector of command variables are used to estimate the variety of states of the controlled process when precision of control is prescribed. The spread of impulses on lattices with (in the order of) 100 nodes is modelled by Markov chains with sparse matrices with ten thousand elements, with exponents reaching often $n = 3001$, for typical representatives of the 17 main plane crystallographic groups existing. Comparative method is used to draw recommendations for the choice of the control. The investigation of the controlled spread of impulses may be useful for the automatic control of spatially extended complexes (nets of measuring and communication stations, adaptive random search processes, through-flow compartmental systems with iterative structure a. s. o.).

1. Introduction

The possibility to influence probabilities for the control of spatially extended complexes has been stressed already in 1963 at the 2nd Congress of IFAC in the author's paper: "On Systems with Automatic Control of Configuration". In the meantime, interest has grown in system theory in problems of the control of nets with many centres and in the use of graph theory, related to the theory of groups. Graphs are also used in the study of Markov processes. Recently the information theoretical approach to control [3] has brought results, sufficiently general and at the same time understandable and close to the needs of the designer. This and the availability of fast computers enable to proceed in the quantitative study of stochastic processes (as the spread of impulses) on lattices with many nodes in order to determine their optimal control.

2. Theoretical background

The scheme of the stochastic control of impulses on lattices can be considered as a case of feedback control of a complex (Fig. 1), consisting of centres linked together to form a lattice, the boundaries of which are N, S, E, W. Impulses from the generator GI are applied repeatedly to centre 1. Counters C1, C2, C3 measure the relative frequencies of impulses arriving to selected

output centres (here: 98, 99, 100). When these frequencies settle on practically stable values, their vector **X** is analyzed and the choice of the vector of control **U** is done in the control element and applied to the complex. The control variables are probabilities of transition in the directions of the lattice edges.

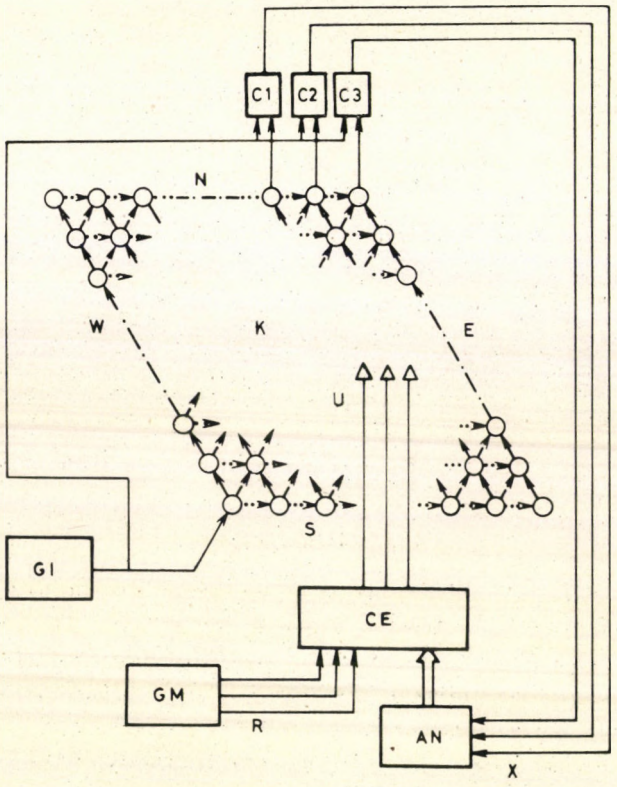


Fig. 1. The scheme of the feedback stochastic control of the spread of impulses. (K = complex, C1, C2, C3 = counters, GI = generator of impulses, AN = analyzer, CE = control element, GM = command element, N, S, E, W = line arrays of boundary elements of the complex)

The choice of control is done so as to obtain a vector **X** of the measured variables of the complex prescribed by the vector **R** of the command variables from the command element.

This scheme illustrates the stochastic control of the spread of impulses on a two-dimensional lattice, and it can be used also for modelling the spread of stuff in a complex of reservoirs (cells) with through-flow.

A problem which we analyze here is the possibility of selective control of the three relative frequencies measured on neighbouring centres. This prob-

lem can be related also to the selective orientation of the activity of centres (sensors) for search purposes at the boundaries N and E with the environment (artificial retina for search).

2.1. The feedback control

The control loop can be considered as a cascade of mappings from certain state spaces into subsequent ones [4]. Among them there is the important mapping from the state space of the measured variables of the complex into the state space of the analyzer. Using the information-theoretical approach [3] we shall be concerned with the choice of ε_{ut} — the utility threshold of resolution of the states of the complex by the analyzer through the measured variables of the complex. This threshold is defined in the Euclidean state space of the measured variables of the complex. With the assumption of the same dimensionality of this state space and of the Euclidean space of the states of the command element, we take $\varepsilon_{ut} = \varepsilon_{rv}$, where ε_{rv} is the threshold of resolution of the states of the command element by the control element through the command variables.

2.2. The basic inequality for the design of control

For quantitative design evaluations the following basic inequality [3], [4], based on the theoretical work [1] is considered:

$$H_{\varepsilon_s}(\mathfrak{R}) < \log_2 J < L_{\varepsilon_{rv}}(\mathfrak{R}), \tag{1}$$

where J is the number of states of the complex,

$\mathfrak{R}$ is the set of the end points of the vector of the command variables in the space of the states of the command element,

ε_S is the maximum admissible control deviation when reproducing the vector of the command variables by the vector of the measured variables of the complex,

$H_{\varepsilon_s}(\mathfrak{R})$ is the minimum ε_S-entropy of the set $\mathfrak{R}$ (in bits),

$L_{\varepsilon_{rv}}(\mathfrak{R})$ is the ε_{rv}-capacity of $\mathfrak{R}$ (in bits).

2.3. The modelling of the process of spread

For this aim the Markov chain model is used, described by the Chapman–Kolmogorov equation in matrix form

$$\mathbf{S}(n) = \mathbf{S}(0) \cdot \mathbf{p}_{jk}^{n} \tag{2}$$

where $\mathbf{S}(n)$ is the vector of the probabilities of the states at time n,

$\mathbf{S}(0)$ is the vector of the initial state probabilities,

$\mathbf{p}_{jk}$ is the matrix of transition probabilities from state j to state k.

2.4. Relation with physical through-flow models

There is an analogy between the processes of the spread of impulses (and of means) in a complex and the processes of the changes of concentration of a catalyzer or dying stuff in through-flow models with many reservoirs [4]. The assumptions are: 1. the through-flow is constant; 2. the total volume of all reservoirs is constant; 3. the flow is perfectly mixed in theoretical sense. This means that the transfer function of a reservoir is $F(p) = (p + I_z/V_z)^{-1}$, where I_z is the flow and V_z the volume. Then, the probability of a marked element remaining in the zth reservoir is $p_{zz} = \exp - P_z \Delta t$, where $P_z = I_z/V_z = \tau_z^{-1}$; 4. The state of the model is designated by the number of reservoir, where a hypothetical marked element is just present; 5. the transition between states can be modelled by a Markov chain.

2.5. The plane crystallographic groups

Wide use is done of the results of abstract crystallography. The number of infinite two-dimensional groups is 24, 7 containing translations in only one direction, so that 17 abstractly distinct are generally considered. As quoted in [2] all 17 of them were discovered empirically by the Moors in their decoration of the Alhambra at Granada; many of them also by the ancient Egyptians and the Chinese. For the denomination of these 17 groups we use the symbols by Hermann and Mauguin. There exist important subgroup relations among the 17 groups. Two of them, p4m, and p6m, taken together, contain all others as subgroups. According to [2], the group p4m contains the groups: p1, p2, pg, pm, cm, pgg, pmg, pmm, cmm, p4, p4g; the group p6m contains the groups p1, p2, pg, pm, cm, pgg, pmm, cmm, p3, p3m1, p31m, p6.

However these subgroup relations did not prove useful for control choice considerations (see point 3.3) and the different groups are to be investigated individually.

3.1. The method of procedure

For our investigation, use is made of Cayley diagrams, bearing in mind that the network of fundamental regions and the Cayley diagram form dual maps. In the graphs, corresponding to different complexes, the following conventions are introduced: 1. translations are indicated by arrows; 2. reflections are indicated by double arrows with opposite sense. This enables to pass directly from Cayley diagrams to the graphs of Markov processes on the different lattices. The appropriate boundary conditions for these graphs, depending upon the realization of the complex, are to be considered. The boundary states may be: 1. absorbing: a) with $p_{jj} \neq 0,1$; b) with $p_{jj} = 0$.

The general method consists in exciting one centre of the specific complex. Equation (2) is used to determine empirically by computation the probabilities of 3 specific neighbouring states of the complex. These correspond to the limit values to which the measured values of the relative frequencies on the counters of Fig. 1 would theoretically tend with the number of impulses from generator GI increasing. The influence of the chosen control upon the specific complex is expressed by the values of the transition probabilities in the matrix $\mathbf{p}_{jk}$ of Eq. (2). For different vectors of control and for n increasing, different idempotent stochastic matrices $\mathbf{p}_{jk}^{n}$ are obtained. These are used for mapping of points of the state space of the control element into points of the state space of the measured variables of the complex which can be identified (under the condition of the same number of dimensions) practically with the state space of the command element. Geometrically simple zones of this space are then analyzed using inequality (1) in order to meet by design the requirements placed upon the number of states of the complex to be considered when precision of control is prescribed by ε_s and $\varepsilon_{rv} = \varepsilon_{ut}$ is realizable by the analyzer. The choice of these figures is important both for the computer of the control element and for the analyzer.

3.2. The complexes investigated and results obtained

The complexes, based upon different abstract crystallographic groups are designated by key-words and are related to these groups as follows: PANDA PONY, PANEL to group p1, PAMPA, PAMIR to group pm, PEGAS to group pg, PAGING to group pgg, POKRM to group p4m, PETR, PETROL to group p3, PROSIM to group p6m.

To describe these various complexes and the graphs of the relevant Markov processes, these symbolic expressions are introduced: IS = input state; OAS = output absorbing state; AS = absorbing state; figures in brackets indicate states from which the transition is possible; states with $p_{jj} \neq 0,1$ are given after indication LOOP. Boundary states with $p_{jj} = 0$ are given after CUT; *N*, *S*, *W*, *E* are states on the relevant boundaries; MATRIX indicates type of the matrix of transition probabilities.

3.2.1. PANDA. Derived from group p1, generated by 3 translations in directions 0°, 60°, 120°. The fundamental region is a parallel-sided hexagon. The texture is in Fig. 2a. MATRIX (100, 100). IS = 1; OAS = 98, 99, 100; 98 (97, 87, 88), 99 (88, 89) 100 (89, 90); LOOP: other *N* states, *E* states; CUT: remaining *S* and *W* states.

The space of control variables is in Fig. 6a, the corresponding space of the measured state variables of the complex is shown in Fig. 6b. Typically,

for the control vector $\mathbf{U} = (p_{120}°; p_{60}°; p_{0}°) = (0.30625; 0.26875; 0.4250)$, we get already for $n = 121$ the vector $\mathbf{X} = (p^{\infty}_{98}, p^{\infty}_{99}, p^{\infty}_{100}) = (0.1822536460; 0.0960215945; 0.7217247595)$.

Consider the area of the quadrilateral 3–4–8–7, contained in the plane $x + y + z = 1$ of Fig. 6b, which is $S = 0.0175$. For $\varepsilon_{rv} = 0.004$ we estimate

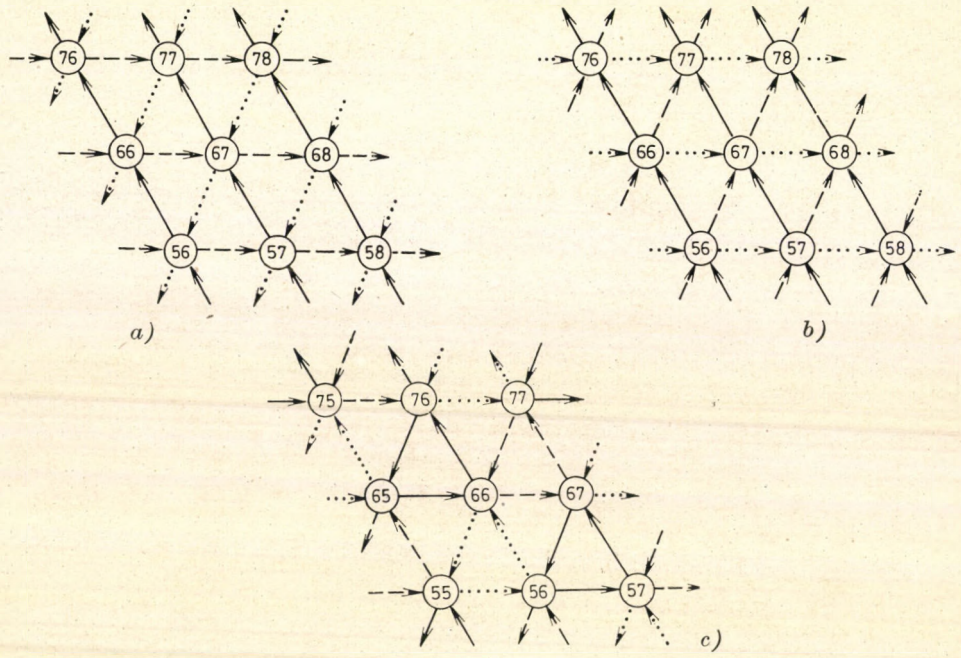


Fig. 2. (a) Texture of complexes PANDA and PANEL, derived from p1 (A — full line, B — dashed, C — dotted); (b) Texture of complex PONY, derived from group p1 (A — full line, B — dashed, C — dotted); (c) Texture of complex PETROL, derived from group p3 (A — full line, B — dashed, C — dotted)

$L_{\varepsilon_{rv}}(\mathfrak{R})$ as $\log_2$ of the quotient of S by ε^2_{rv}. It is approximatively $L_{\varepsilon_{rv}} = 10.094$ [bit]. For $\varepsilon_s = 0.008$ the ε_s-entropy is estimated as $\log_2$ of the quotient of S by $(2\varepsilon_s)^2$. It is $H_{\varepsilon_s} = 6.0874$ [bit]. We take the number of states $J = 6 \times 6 \times 6 = 216$, so that inequality (1) is satisfied ($6.0874 < 7.7548 < 10.094$).

3.2.2. PONY. Derived from group p1, generated by 3 translations in directions 0°, 120°, 240°. The fundamental region is a parallel-sided hexagon. This is one of the geometric varieties of the single abstract group. The texture is on Fig. 2b. MATRIX (100, 100). IS = 1; OAS = 98, 99, 100; 98 (97, 88), 99 (89), 100 (90); LOOP: all other N, S, E, W states including 1.

The spaces of end-points of $\mathbf{U}$ and of $\mathbf{X}$ are drawn in Fig. 8a. Typically, for $\mathbf{U} = (A; B; C) = (0.3; 0.3; 0.4)$ we get for $n = 3001$ the vector $\mathbf{X} \doteq (p^{\infty}_{98};$

p_{99}^{∞}; p_{100}^{∞}) = (0.5962340310; 0.1672522487; 0.2347598226), the sum of the elements of which is near to 1.

Consider the area of the triangle 2–4–5, which is $S = 0.08149$. For $\varepsilon_{rv} = 0.004$ we estimate the ε_{rv}-capacity as 12.314 [bit] and the ε_s-entropy for $\varepsilon_s = 0.008$ as 8.3128 [bit] from simple area consideration. We take $J = 9 \times 9 \times 9 = 729$, so that inequality $8.3128 < 9.5097 < 12.314$ is satisfied.

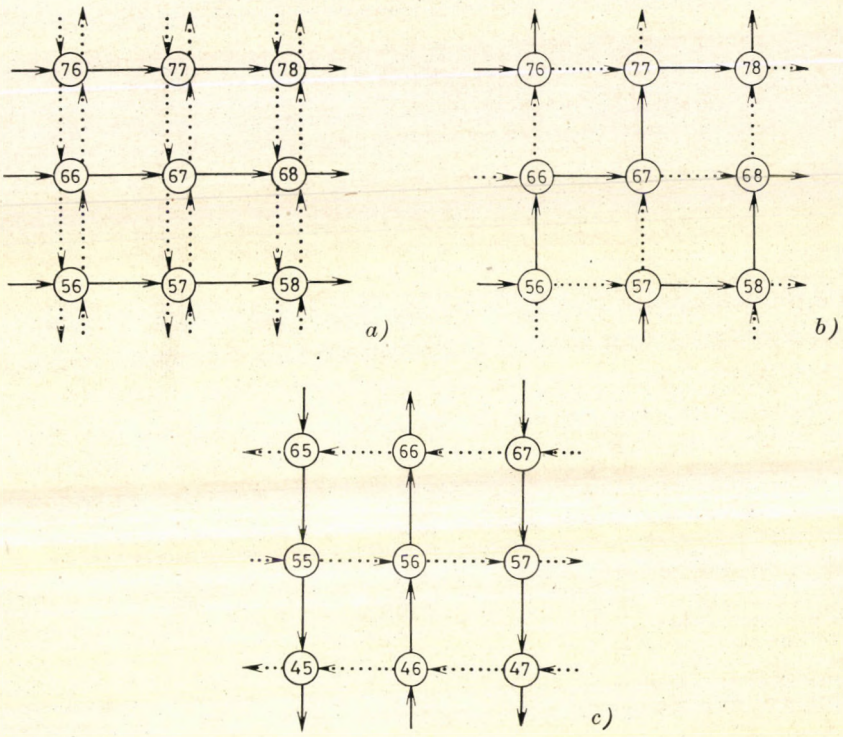


Fig. 3. (a) Texture of complexes PAMPA and PAMIR, derived from group pm (A — full line, $B/2$ — dotted); (b) Texture of complex PEGAS derived from group pg (A — full line, B — dotted); (c) Texture of complex PAGING derived from group pgg (A — full line, B — dotted)

3.2.3.PANEL. Derived from group p1 generated by 3 translations in directions 0°, 60°, 120°. PANEL differs from PANDA by number of absorbing boundary states. Texture is on Fig. 2a. MATRIX (100, 100). IS = 1; OAS = = 98, 99, 100; 98 (87, 88), 99 (88, 89), 100 (89); AS = all other N and E states. CUT: remaining W and S states.

The space of control is illustrated in Fig. 6a (the same as for PANDA). The space of the measured variables of the complex is in Fig. 7. Note, that due to absorbing states on boundaries the obtained mapping is not on a plane.

For same control vectors as in the case of PANDA we get non plane surface 3–4–7–8, which area is approximatively $S = 0.001698$. We even take it as a rectangle with sides 0.060 and 0.030 and estimate the $L_{\varepsilon_{rv}}(\mathfrak{R})$ as $\log_2 (16 \times 8) = 7.0000$, where $16 = (0.060 : 0.004) + 1$, $8 \doteq (0.030 : 0.004) + 1$. We estimate the ε_s-entropy for $\varepsilon_s = 0.008$ as $\log_2 (4 \times 3) = 3.5849$, where $4 \doteq (0.060 : 0.016) + 1$, $3 \doteq (0.030 : 0.016) + 1$. We take the number of states $J = 3 \times 3 \times 3 = 27$, so that inequality $3.5849 < 4.7548 < 7.0000$ is satisfied.

3.2.4. PAMPA. Derived from group pm generated by two reflections and one translation. Texture is on Fig. 3a. MATRIX (100, 100). IS = 1; OAS = 98, 99, 100; 98 (97, 88), 99 (89), 100 (90). LOOP: other N, E, S, states; CUT: remaining W states. The influence of the control variable (the probability A) is shown on Fig. 9. The convergence was practically good for some controls already at $n = 651$.

Investigations have been carried till $n = 3001$.

3.2.5. PAMIR. Derived from group pm generated by two reflections and one translation. PAMIR differs from PAMPA by number of absorbing boundary states. Texture is on Fig. 3a. MATRIX (99, 99). IS = 1; OAS = 97, 98, 99; 97 (87), 98 (88), 99 (89); AS = other N and E states; LOOP = remaining S states; CUT = then remaining W states. For results see point 3.3.6.

3.2.6. PEGAS. Derived from group pg generated by two parallel glide-reflections, the fundamental region of which is a rhombus. Texture is on Fig. 3b. MATRIX (100, 100). IS = 1; OAS = 98, 99, 100; 98 (97, 88), 99 (89), 100 (90). LOOP = other N and E states; CUT = remaining S and W states.

The space of the measured variables of the complex is in Fig. 10a, for different values of the probability A as one control variable. The other is the probability $B = 1 - A$. The dependence of the output variables upon A is shown in Fig. 10b, showing poor possibilities of control. The sets $\mathfrak{R}$ are points on the portion of curve in Fig. 10a from point for $A = 0.1$ to point for $A = 0.9$. We proceed to a rectification of this portion of the curve, which length is 0.346. For $\varepsilon_{rv} = 0.004$ we estimate the ε_{rv}-capacity directly from the quotient 0.346 : 0.004 (it must be an integer) as 6.4262. We estimate the ε_s-entropy directly from the quotient 0.346 : 0.016 as 4.3923. We take $J = 3 \times 3 \times 3$ $J = 3 \times 3 \times 3 = 27$, so that inequality $4.3923 < 4.7548 < 6.4262$ is satisfied.

3.2.7. PAGING. Derived from group pgg generated by two perpendicular glide-reflections. Texture is on Fig. 3c. MATRIX (99, 99). Two lowest rows of states 1–8 and 9–18; 10 (1, 9); 2 highest rows 79–88 and 89–97; 89 (80); IS = 9; OAS = 97, 98, 99; 97 (96, 88), 98 (93), 99 (95). LOOP = each second of the other states on perimeter. The state space of the complex is on Fig. 11a and the

output variables as functions of the control probability A (there is $B = 1 - A$) are on Fig. 11b, showing very good control possibilities. The portion of the curve on Fig. 11a has a length of 1.21, from which we estimate L for $\varepsilon_{rv} = 0.004$ directly by division as 8.2383 and similarly H for $\varepsilon_s = 0.008$ as 6.2288. We take $J = 6 \times 6 \times 6 = 216$, so as to have $6.2288 < 7.7548 < 8.2383$.

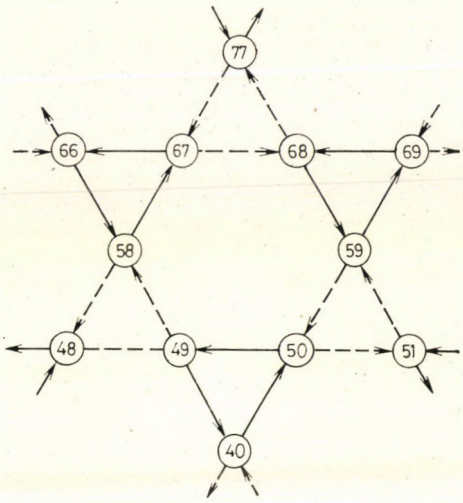


Fig. 4. Texture of complex PETR, derived from group p3 generated by two trigonal rotations (A — full line, B — dashed line)

3.2.8. POKRM. Derived from group p4m generated by reflections in the sides of a right-angled isoceles triangle. Texture is on Fig. 5a. MATRIX (77, 77). IS = 3, OAS = 75, 76, 77.

POKRM is little influenced by the choice of the control vector (of the probabilities A, B, C). Investigations of convergence were done till $n = 3001$ which proved appropriate.

3.2.9. PETR. Derived from group p3 generated by 2 trigonal rotations. Fundamental region is a rhombus, 1 angle of which is 60°. Texture is on Fig. 4. MATRIX (100, 100). Two lowest rows 1–5 and 6–17; 1 (7); 8 (1); IS = 6; OAS = 98, 99, 100; 98 (98), 99 (97), 100 (92); LOOP = all other states on perimeter including 6.

PETR is very little influenced by the choice of the probability A (while $B = 1 - A$). This has been checked for $n = 3001$. The convergence of the relative frequencies is relatively slower, and $n = 3001$ did not suffice in some cases of control.

3.2.10. PETROL. Derived from group p3 generated by 3 trigonal rotations. Texture is on Fig. 2c. MATRIX (100, 100). IS = 1; OAS = 98, 99, 100; 98 (97, 88), 99 (89), 100 (90); LOOP = E and S states and 91. Remaining states on N and W have bidirectional contact with boundary neighbours but 97 (96). Both the input and the output spaces of the complex are in Fig. 8b. Controls 2, 4 and 5 are the same as those for PONY (Fig. 8a). The area of the triangle 2, 4, 5 in the state space of the complex is $S = 0.001751$. For $\varepsilon_{rv} = 0.004$ we estimate the ε_{rv}-capacity as 6.76819 and the ε_s-entropy as 2.58495

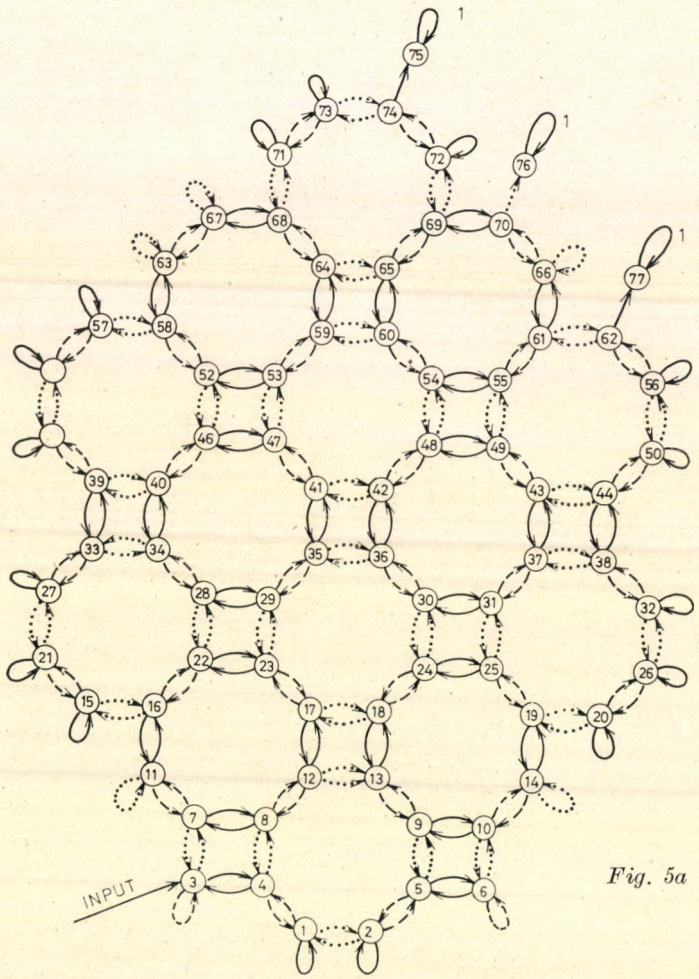


Fig. 5a

Fig. 5. (a) Graph of Markov chain for complex POKRM derived from group p4m. (Transition probabilities: A — full lines, B — dotted, C — dashed); (b) Graph of Markov chain for complex PROSIM derived from group p6m. (Transition probabilities: A — full line, B — dotted, C — dashed)

from area considerations. We take $J = 3 \times 3 \times 3 = 27$, so that $2.58495 < < 4.7548 < 6.76819$. The possibilities of control of PETROL have been investigated but they are limited. Typical data are: $\mathbf{U} = (0.6;\ 0.3;\ 0.1)$ and $\mathbf{X} = (0.3844254026;\ 0.0654069689;\ 0.5501658169)$ for $n = 3001$, showing good convergence already at about $n = 1081$.

3.2.11. PROSIM. Derived from group p6m generated by reflections in the sides of a bisected equilateral triangle. The fundamental region is a triangle with angles 30°, 60° and 90°. Each vertex of the graph of this group is common for a square, a triangle and a dodecagon. Texture on Fig. 5b. MATRIX (81, 81). IS = 5; OAS = 79, 80, 81.

PROSIM has been found very invariant to the choice of probabilities A, B, C as elements of the vector of control (at $n = 3001$). But it is possible to discern the influence of control vector upon the dynamics of the sum $(p_{79} + p_{80} + p_{81})$ taken as function of n. This is shown on Fig. 12a. Figure 12b

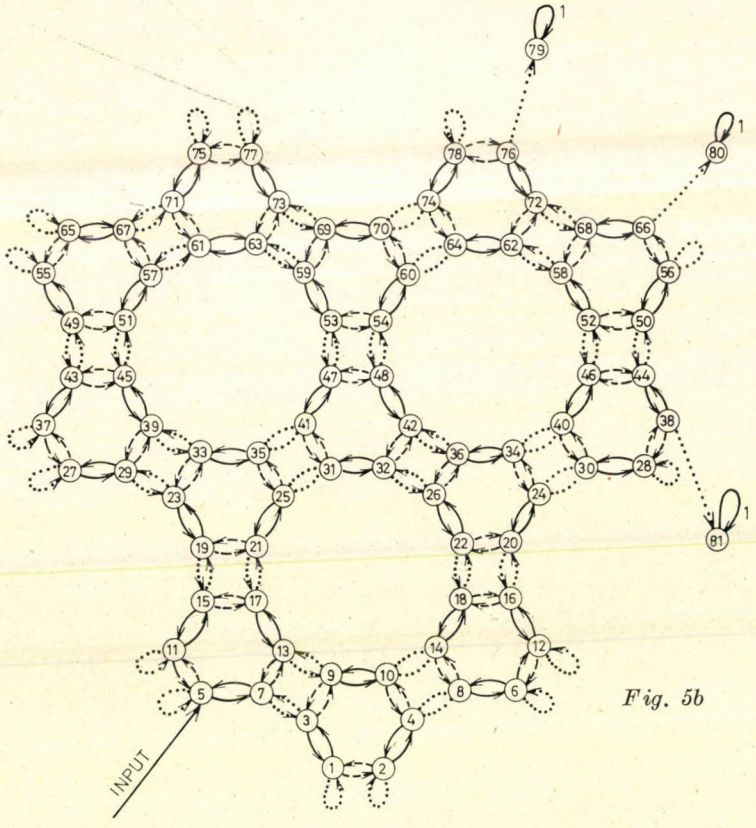


Fig. 5b

indicates as parameters values of this sum for $n = 501$ for different controls, 0.752 being a value near optimum.

3.3. Comparison of results and recommendations for control

The criterion for comparison is the possibility to control the orientation of the spread of impulses. Another criterion is the speed of convergence of the relative frequencies to stable values.

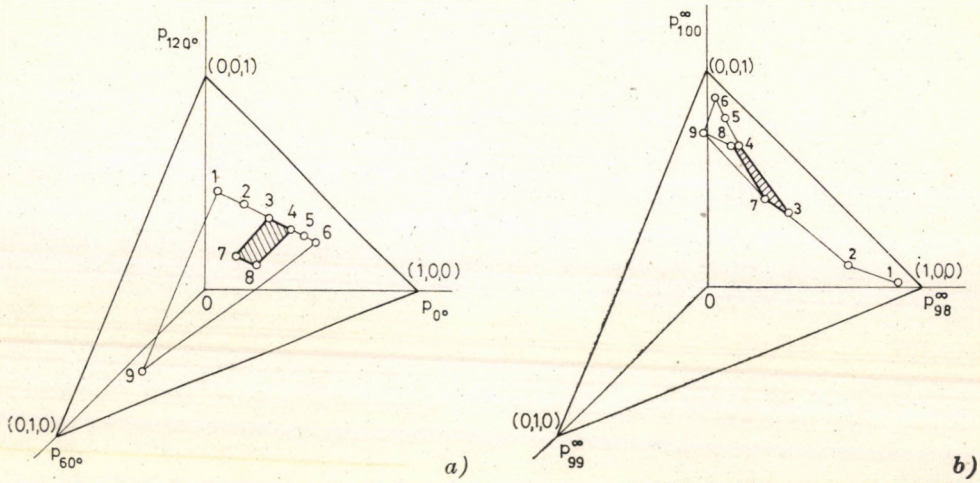


Fig. 6. (a) End-points of control vector for complexes PANDA (Fig. 6b) and PANEL (Fig. 7); (b) States of complex PANDA for controls on Fig. 6a

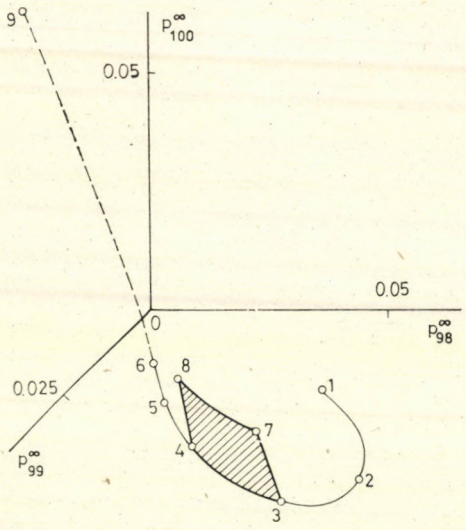


Fig. 7. States of complex PANEL for controls on Fig. 6a. (Note the scale indication on the axes)

3.3.1. Comparison of PANDA and PANEL. It is shown that the presence of absorbing states at the boundaries of PANEL limits the variety of states notably under otherwise similar conditions (thresholds ε_{rv} and ε_s). The effect of the same vectors of control stems from Figs 6a, 6b and 7.

3.3.2. Comparison of PONY and PETROL. Though the texture is similar (Fig. 2b and 2c), the possibilities of the control of PETROL are limited, those of PONY are better. In both cases the convergence of the relative frequencies

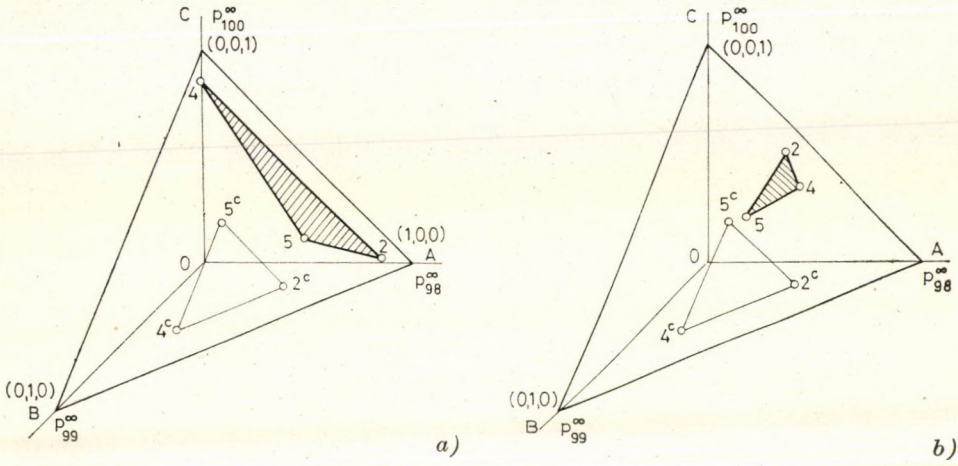


Fig. 8. (a) Mapping of end-points of control vector into states of complex PONY. (2^c, 4^c, 5^c are controls, 2, 4, 5 are states of complex); (b) Mapping of end points of control vector into states of complex PETROL (2^c, 4^c, 5^c are controls, 2, 4, 5 are states of complex)

was slow: for PONY for some controls $n = 2711$ or $n = 3001$ was not sufficient. For similar conditions and same controls (Figs 8a and 8b) the variety of states of PONY is much greater than that of PETROL.

3.3.3. Comparison of PETR and PETROL. Petr is practically invariant to control and has slow convergence of relative frequencies. PETROL is somewhat better from both these points of view and its possibilities of control are described in point 3.2.10 and Fig. 8b.

3.3.4. Comparison of PEGAS and PAGING. As follows from points 3.2.6 and 3.2.7 the variety of states of PAGING is much greater than that of PEGAS for otherwise similar conditions. As can be seen from Fig. 11b PAGING is particularly suitable from the point of view of control.

3.3.5. Comparison of POKRM and PROSIM. In both cases the influence of the control is small: in the case of PROSIM there is even practical invariance to control. Both have been investigated till $n = 3001$.

3.3.6. Comparison of PAMPA and PAMIR. In both cases it is advantageous to have small values of the probability A. This is shown on Fig. 9. There, owing to the influence of absorbing states of PAMIR, which otherwise has the same texture as PAMPA (Fig. 3a), the range of changes of PAMIR is smaller than that of PAMPA as only relatively few impulses arrive to the three output centres.

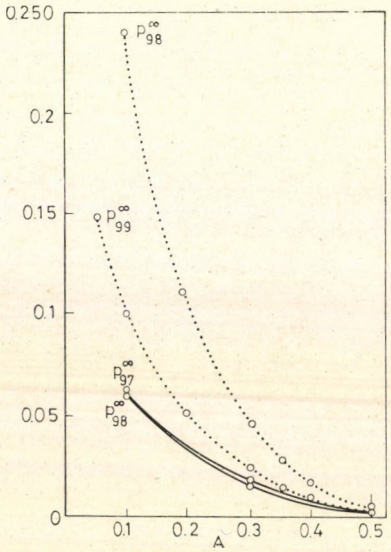


Fig. 9. Influence of control variable A upon output variables of complex PAMPA (dotted) and of complex PAMIR (full line)

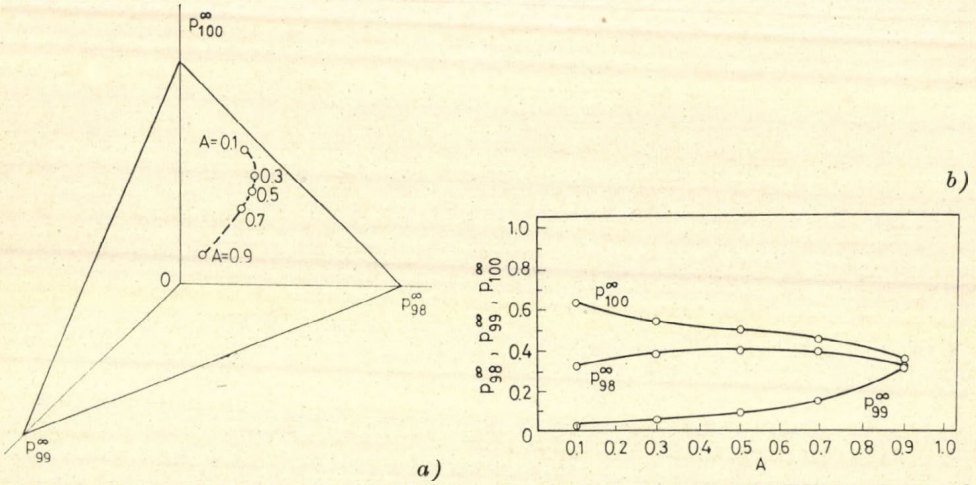


Fig. 10. (a) States of complex PEGAS for $A = 0.1$ till $A = 0.9$; (b) Effect of control upon output variables of complex PEGAS (Control variables are probabilities A and $B = 1-A$)

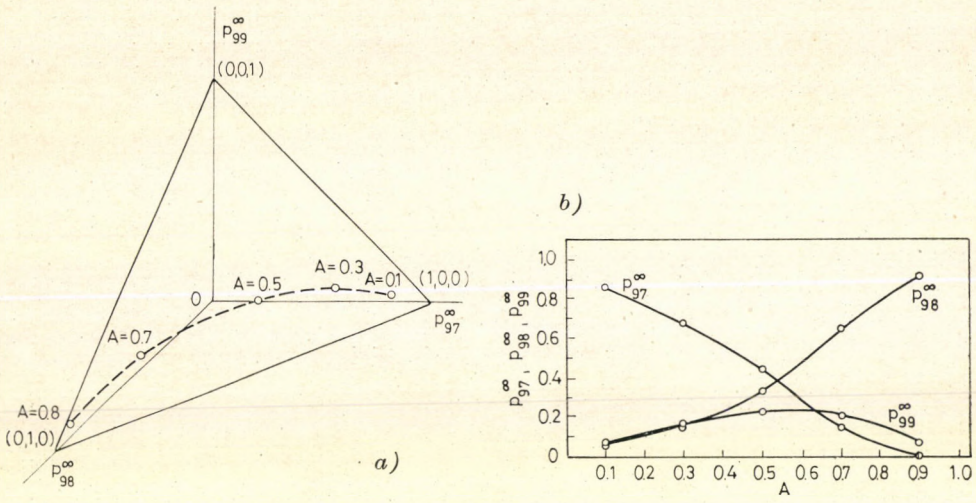


Fig. 11. (a) States of complex PAGING for $A = 0.1$ till $A = 0.8$; (b) Effect of control upon output variables of complex PAGING (Control variables are probabilities A and $B = 1 - A$)

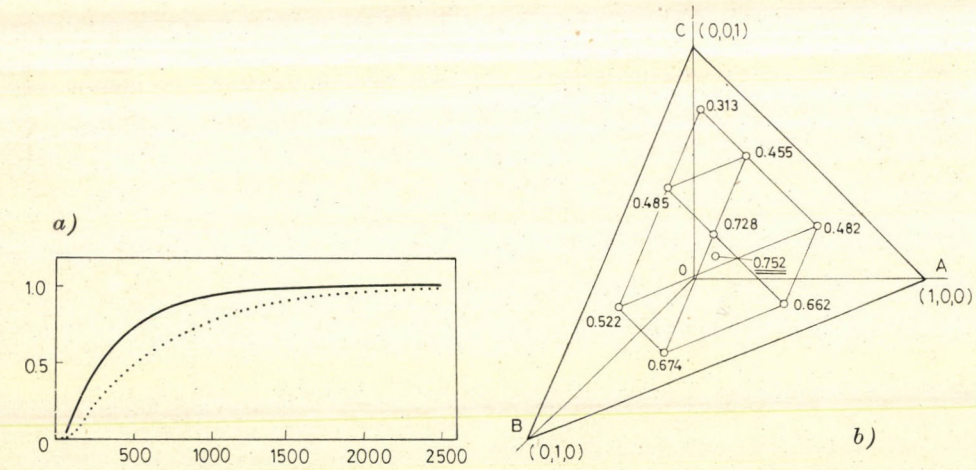


Fig. 12. (a) Dynamics of complex PROSIM. Influence of control upon the sum $(p_{79} + p_{80} + p_{81})$ as function of time

for $\mathbf{U} = (0.6;\ 0.1;\ 0.3)$ — dotted

for $\mathbf{U} = (0.3;\ 0.3;\ 0.4)$ — full line.

(b) Complex PROSIM at $n = 501$ for different controls. (Control vector $\mathbf{U} = (0.333;\ 0.333;\ 0.333)$ gives value of sum 0.752 near optimum)

4.0. Conclusions

As concluding remarks we can state, that:

1. The complex PAGING, derived from the group pgg has proved very suitable for the control of the spread of impulses,

2. Complexes with absorbing boundary centres have limited possible range of changes,

3. Complicated complexes with bidirectional contacts between states are invariant to control.

4. The method of procedure, described in point 3.1 is efficient and can be used for further quantitative and comparative investigations.

The selection of results and comparison of results show that some of the complexes investigated could be objects of automatic control. For quantitative evaluations only very simple ways of estimation of the ε-capacity and ε-entropy have been used here (in points 3.2.1, 3.2.2, 3.2.3, 3.2.6, 3.2.7, 3.2.10). For detailed theory, work [1] is fundamental.

The problem of invariance of the complex to the choice of the vector of control can be raised (as for PROSIM in point 3.2.11).

Further work can be directed towards the control of the spread of impulses on lattices based upon three-dimensional symmetry groups, especially upon crystallographic point groups related to space groups.

The spread of impulses on lattices is closely related, as has been stressed in point 2.4, with the spread of injected stuff in through-flow models. This opens potential applications for the modelling of the spread of catalysers or radioactive stuff (marked elements), and in pharmacodynamics through complex multicellular compartmental systems and models of tissues with homogeneous structure.

*

Acknowledgements. The author thanks J. Grim for efficient programming of computations carried out on an IBM 370/135 computer of the Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences.

References

1. *Kolmogorov, A. N., Tikhomirov, V. M.*, ε-entropia i ε-emkost mnozhestv v funktsionalnykh prostranstvakh. Uspekhi matem. nauk, **14**, 1959, *2*, pp. 3—86.
2. *Coxeter, H. S. M., Moser, W. O. J.*, Generators and relations for discrete groups. Ergebnisse d. Mathematik u. ihrer Grenzgebiete, *14*, 1957, Springer-Verlag, Berlin—Göttingen—Heidelberg.
3. *Petrov, B. N., Petrov, V. V., Ulanov, G. M., Ageiev, V. M., Zaporoshets, A. V., Kochubievskii, I. D., Mai, V. P., Uskov, A. S.*, Nachala informatsionnoi teorii upravleniya. In Tekhnicheskaya kibernetika 1968, 1969.
4. *Beneš, J.*, Teorie systémů (Řízení komplexů). (in Czech.) Academia, Praha 1974.

Стохастическое управление импульсами в решетках плоских кристаллографических групп

Й. БЕНЕШ

(Прага)

Управление с обратной связью намеренно ориентированного распространения импульсов в решетках плоских кристаллографических групп основано здесь на информационном подходе к управлению комплексами. ε-энтропия и ε-емкость множества конечных точек вектора задающих величин применяются для оценки разнообразия состояний управляемого процесса при заданной точности управления. Распространение импульсов в решетках с (порядка) 100 узлами моделируется марковскими цепями с редкими матрицами в десять тысяч элементов, с экспонентами достигающими часто $n = 3001$, для типичных представителей 17-ти существующих основных кристаллографических групп. Применяется метод сравнения для вывода рекомендаций для выбора управления. Исследование управляемого распространения импульсов можеть быть полезным для автоматического управления пространственно развернутыми комплексами (сетями измерительных станций и станций связи, самонастраивающимися процессами случайного поиска, проточными многоклеточными системами с однородной структурой и т. д.).

J. Beneš
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
180 76 Prague 8, Pod vodárenskou věží 4
Czechoslovakia

Problems of Control and Information Theory, Vol. 5 (1), pp. 31—38 (1976)

ON SOME INTERRELATIONS BETWEEN THE THEORIES OF INFORMATION AND CONTROL*

B. N. PETROV, R. L. DOBRUSHIN, M. S. PINSKER,
G. M. ULANOV, S. V. ULYANOV

(Moscow)

(Received June 28, 1975)

The notions of control and information are closely interrelated. Control requires information, and information is usually used for control purposes. Control maximizing the increment of information in the control system may be regarded as optimal (see, for example, bibliography in [1]).

Information theory (in the narrow sense of the word) is devoted to the problems involved into transmission of information through communication channels. In these problems coding and decoding methods may be varied and selected optimally. The very fact that communication channels are part and parcel of many control systems accounts for the position of the information theory within the framework of the control theory. The problem of interrelations between these theories at higher levels is much more profound.

Information transmission requirements led to the introduction of a new theoretical notion, amount of information, whose applicability reaches far beyond the scope of the above-mentioned problems. It is evident, however, that for such a manysided notion as information there cannot be a unique technique for measuring its amount applicable to all situations. Consider, for comparison, numerous techniques developed in physics for measuring the amount of matter. Yet, one may attempt to find some common features of various approaches to information measurement. In particular, the idea of measuring the amount of information as the difference between the uncertainty of the object under consideration prior to and after the arrival of information (this goes back to the founder of the information theory, C. Shannon [2]) seems to be very fruitful.

To put it more precisely, the amount of information contained in object θ about object $\tilde{\theta}$ is defined as

$$I(\theta, \tilde{\theta}) = H(\theta) - MH(\theta/\tilde{\theta}), \tag{1}$$

* This paper is a revised version of the presentation at the All-Union Seminar on Informational Processes (Vladivostok, 1972).

where $H(\theta)$ is the uncertainty of object θ, and $MH(\theta/\tilde{\theta})$ is that of θ when the results of observations of θ are known.

With such an approach, the choice of the information measure boils down to the choice of the uncertainty measure which may be done in many reasonable ways. For example, A. Kolmogorov's algorithmic approach to the measurement of the amount of information [3] may be mentioned, based on the introduction of the object uncertainty measure through the complexity of the algorithm describing this object. The statistical approach treating θ and $\tilde{\theta}$ as random variables with known probability distributions has numerous merits. With this approach, it suffices to introduce the uncertainty of one random variable θ, expressed in terms of its probability distribution in order to be able to treat in a natural way uncertainty $MH(\theta/\tilde{\theta})$ as an averaged (with respect to the distribution of the random variable θ) value of the uncertainty of the conditional probability distribution of θ under a given valueof $\tilde{\theta}$.

However, many different ways to introduce the uncertainty measure $H(\theta)$ are left here, each of them entailing its own choice of the amount of information. Besides Shannon's entropy-based method of uncertainty measurement that results in the conventional amount of information, it pays to isolate a classical method of the probability theory based on uncertainty measurements through variance random variable. Then the measure of the amount of information

$$I(\theta, \tilde{\theta}) = \mathfrak{D}\theta - M\mathfrak{D}(\theta/\tilde{\theta}) \tag{2}$$

is equal to

$$I(\theta, \tilde{\theta}) = \mathfrak{D}M(\theta/\tilde{\theta}) = M(M(\theta/\tilde{\theta}) - M\theta)^2, \tag{3}$$

where $M(\theta/\tilde{\theta})$ is the optimal mean-square estimate of θ on the basis of a given $\tilde{\theta}$. Hence, the amount of information $I(\theta, \tilde{\theta})$ characterizes in this case the mean-square difference between optimal estimation of θ under known and unknown values of $\tilde{\theta}$, i.e. the "degree" of estimation improvement due to our knowledge of the value of $\tilde{\theta}$. Other, non-quadratic loss functions may be taken.

Entropy defined for discrete random variable θ with values x_i, $i = \overline{1, m}$ as

$$H = -\sum_{i=1}^{m} p(x_i) \log p(x_i), \tag{4}$$

has two special features making it to stand out against a background of al possible measures of uncertainty. Firstly it is non-dimensional, i.e. independenl of the measurement scale. It is, therefore, insensitive to the meaning of infort mation, and this accounts for its successful application to information trans-

mission problems since difficulties involved into information transmission do not depend on the scales of transmitted variables. Note, however, that

$$H\varphi = \sum_{i=1}^{m} p(x_i)\,\varphi\big(p(x_i)\big) \tag{5}$$

may be used as a fairly natural non-dimensional characteristic of uncertainty where $\varphi(p(x_i))$ is an arbitrary positive function of $p(x_i) \geq 0$. Of all these functions, the power one, $\varphi(p(x)) = p^{\alpha+\beta-2}(x)/a(x)$, β, $a(x) > 0$ leads after the passage to the logarithmic scale and appropriate normalization, to entropy

$$H_\alpha^\beta = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{m}\frac{p(x_i)^{\alpha+\beta-1}}{a(x_i)}\right) \tag{6}$$

coinciding at $\beta = 1$, $a(x_i) = 1$, $i = \overline{1, m}$ with that of Rényi [4]:

$$H_\alpha = \frac{1}{1-\alpha}\log\left(\sum_{i=1}^{m} p(x_i)^\alpha\right). \tag{7}$$

Expressions (4) and (5) were used with success in the works of Soviet, Hungarian and Czechoslovak researchers.

Entropies (5) and (6) are very useful for error estimation in discriminating hypotheses. By taking $\beta = \beta_i$, $p(x) = p_j(x)$, $a(x) = p_k(x)$, $j, k = \overline{1, l}$ where $p_j(x)$ and $p_k(x)$ are probabilitites of x under the jth and kth hypotheses respectively, and by obtaining γ_{jk} denote the probability of taking the kth hypothesis when the jth one is true, obtain the following inequalitites

$$e^{(1-\alpha)H_\alpha^{\beta_i}} \leq \begin{cases} \gamma_{jk}^{1-\alpha}\gamma_{kk}^{\alpha+\beta-1} + (1-\gamma_{jk})^{1-\alpha}(1-\gamma_{kk})^{\alpha+\beta_j-1} \\ \sum_{s=1}^{l}\gamma_{js}^{1-\alpha}\gamma_{ks}^{\alpha+\beta_j-1} \\ \beta_j > 2;\ 1-\beta_j < \alpha < 2-\beta_j. \end{cases} \tag{8}$$

Another important feature of the entropy is the analytical simplicity of its properties due to the use of a logarithmic function in its definition. This fact is essential. In the conventional probability theory, variance is used as aquantitative measure of random variable uncertainty not because of some logical advantages, but owing to analytical conveniences brought about by the quadratic function.

The feasibility of axiomatic entropy definition, indicated already by Shannon, is based on the two above properties which outline the scope of possible applications of Shannon's entropy and the amount of information.

The major justification of the usefulness of the notion of entropy and that of based on its amount of information is that they enable one to express

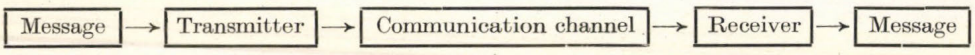
through them channel capacity, and message ε-entropy or entropy with a given level of fidelity, as indicated by Shannon. The latter variable (which is not as well-known as the capacity) is useful for continuous variables with infinite $H(\theta)$. It is defined by the following inequality

$$H_\varepsilon(\theta) = \inf \frac{1}{T} I(\theta, \tilde{\theta}), \tag{9}$$

where the lower bound is taken over pairs of random variables $(\theta, \tilde{\theta}) = \{\theta(t), \tilde{\theta}(t);\ 0 \leq t \leq T\}$ (T is transmission time) which differ from each other in some metric at most by ε; for example,

$$M(\theta(t) - \tilde{\theta}(t))^2 < \varepsilon^2. \tag{10}$$

According to Shannon's theorems, capacity and ε-entropy enable description of conditions where, with appropriate choice of coding and decoding, transmission is possible through the following communication system:



Transmission with error ε is possible when $H_\varepsilon(\theta)$ is less than channel capacity C and impossible if $H_\varepsilon(\theta) > C$.

In many situations, it is easier to determine and estimate the amount of information than special characteristics more appropriate to a given particular problem that can be later studied on the basis of the amount of information. For instance, in the problem of statistical evaluation of a signal taking values from $\theta_1, \ldots, \theta_m$ with probabilities $p(\theta_i)$, $i = \overline{1, m}$ under n independent observations ξ_k, $k = \overline{1, n}$, the amount of information is represented as [5]:

$$I\big(\theta, (\xi_1, \ldots, \xi_n)\big) = H - \frac{\lambda^n}{\sqrt{n}} 0(1), \quad n \to \infty \tag{11}$$

where H is the entropy defined in (4), and

$$\lambda = \max_{i \neq j} \min_{0 \leq \alpha \leq 1} \int_{-\infty}^{\infty} p(x/\theta_i)^\alpha p(x/\theta_j)^{1-\alpha}\, dx,$$

$p(x/\theta_j)$, $j = \overline{1, m}$ is conditional distribution of ξ_k provided that $\theta = \theta_j$. In the case of a continuous signal θ, and a noise ζ_i wich is independent and for n independent similar observations

$$\xi_i = \theta + \zeta_i,\ i = \overline{1, n}; \tag{12}$$

an amount of information about θ contained in an observation ξ_i for large n may be approximately represented as

$$I\big(\theta, (\xi(1), \ldots, \xi(n))\big) = \frac{1}{2}\ln(n\,\mathfrak{I}) - h(\theta) - \frac{1}{2}\ln(2\pi e), \tag{13}$$

where $\mathfrak{I}$ is Fischer's informational quantity of random variable ζ_i, and $h(\theta)$ is the differential entropy of random variable θ, i.e.

$$\mathfrak{I} = \int \left[\frac{p'(x)}{p(x)}\right]^2 p(x)\,dx, h(\theta) = -\int p_\theta(x)\log p_\theta(x)\,dx, \tag{13'}$$

where $p(x)$ and $p_\theta(x)$ are densities of distributions of ζ_i and θ, respectively. On the basis of this equality one can easily obtain bounds on moments of parameter θ.

In particular, for any estimate $\tilde{\theta} = \Psi(\xi_1, \ldots, \xi_n)$ of parameter θ such that

$$M(\theta - \tilde{\theta})^2 \leq \varepsilon^2 \tag{14}$$

amount of information must satisfy the inequality:

$$I(\theta, \tilde{\theta}) \geq H_\varepsilon(\theta). \tag{15}$$

It may be seen from (14) and (15) that

$$M(\theta - \tilde{\theta})^2 \geq [\varepsilon_\theta(I(\theta, \theta))]^2 \tag{16}$$

where the function $\varepsilon_\theta(H)$ is the inverse of $H_\varepsilon(\theta)$:

$$H_{\varepsilon_\theta(H)}(\theta) = H. \tag{17}$$

Indeed, assume that

$$M(\theta - \tilde{\theta})^2 = \varepsilon_1^2.$$

Then

$$I(\theta, \tilde{\theta}) < H_{\varepsilon_1}(\theta).$$

Consider the case of the squared error, $\varepsilon_\theta(H_{\varepsilon_1}(\theta)) = \varepsilon_1^2$. Since $H_\varepsilon(\theta)$ is a monotonic decreasing function of ε, $\varepsilon_\theta(H_{\varepsilon_1}(\theta))$ is decreasing as well, and therefore

$$\varepsilon_\theta(I(\theta, \tilde{\theta})) \leq \varepsilon_\theta(H_\varepsilon(\theta)) = \varepsilon_1^2 = M(\theta - \tilde{\theta})^2.$$

Thus, one can bound the mean-square error $M(\theta - \theta)^2$ from below by the known amount of information $I(\theta, \tilde{\theta})$. In a rather wide class of applications this estimate turns out to be asymptotically accurate for $\varepsilon \to 0$. This procedure, of getting estimates can be extended to more general situations, too.

Estimates of the increase in risk incurred because of reduction of observations may also be obtained through measurements of the amount of information.

It seems interesting to use the notion of ε-entropy to a wider class of control theory applications. However, some difficulties are met here. It is allowed in the usual formulation of information transmission problems that

message $\theta = \{\theta(t), -\infty < t < \infty\}$ will not be restored by the receiver instantly at the time of transmission, but after an arbitrarily large delay; at the same time, in the control theory it is essential that $\tilde{\theta}(t)$ be defined through $\theta(s)$ without anticipation, i.e. through its values for $s \leq t$, or even with forecast $\tau > 0$, i.e. through $\theta(s)$, $s \leq t - \tau$. Such a situation takes place, for example, in a problem where $\theta(t)$ is interpreted as a control input, $\tilde{\theta}(t)$ is regarded as a control variable, and input and output variables should be transformed so (or, in terms of the information theory, such coding and decoding should be used) that $\tilde{\theta}(t)$ meet imposed conditions.

Assume, for example, that $\tilde{\theta}(t)$ is reproducing signal $\theta(t)$ with mean-square error of type (10) provided that $\tilde{\theta}(t)$ depends only on the behaviour of $\theta(s)$ up to time t (restoration of the control input $\theta(t)$ without anticipation) or up to time $t - \tau$, $\tau > 0$ (restoration with forecast τ). To solve this new problem, one has to modify the notion of ε-entropy and to introduce the notion of ε-entropy without anticipation $H_\varepsilon^0(\theta)$ or with forecast $H_\varepsilon^\tau(\theta)$, $\tau > 0$. Under the mean-square fidelity criterion, $H_\varepsilon^0(\theta)$ and $H_\varepsilon^\tau(\theta)$ are defined as the lower bound of the amount of information computed under the constraint that $\theta(t)$ and $\tilde{\theta}(t)$ meet condition (10) and under an additional condition requiring that values of $\tilde{\theta}(t)$ are determined through those of $\theta(s)$ up to time t or $t - \tau$, respectively (i.e. for $r \leq t - s$, sequence of three random variables $\{\tilde{\theta}(s), s \leq t\}$, $\{\theta(r), r \leq t - r\}$ and $\{\theta(u), 0 \leq u \leq T\}$ form Markov chain for any t).

It turns out that if $\theta(t)$ is a Gaussian process, it is sufficient to consider only Gaussian pairs of processes $(\theta, \tilde{\theta})$ in order to determine ε-entropy $H_\varepsilon^\tau(\theta)$, $\tau \geq 0$ (like in the case of the usual ε-entropy).

For a stationary Markov—Gaussian process $\theta(t)$ with

$$M\theta(t)\,\theta(t+s) = Ae^{-\beta|s|};\ A, \beta > 0 \tag{18}$$

obtain, in particular, a simple design formula

$$H_\varepsilon^\tau(\theta) = \beta(A - \varepsilon^2)/(\varepsilon^2 - A(1 - e^{-\beta\tau}))$$
$$\tau \geq 0,\ A(1 - e^{-\beta\tau}) < \varepsilon^2 < A, \quad T \to \infty \tag{19}$$

and ε-entropy without anticipation is $\frac{\pi^2}{2}$ times as great as the usual entropy $H_\varepsilon(\theta)$ characterizing the situation where $\tilde{\theta}(t)$ is determined through the whole process $\theta(s)$, $-\infty < s < \infty$.

The new versions of the notion of ε-entropy lead to generalizations of the basic results in the information transmission theory. Thus, the inverse theorem takes place: if

$$H_\varepsilon^\tau(\theta) > C \tag{20}$$

no coding and decoding methods enable transmission through a channel where input signal $\eta(t)$ is determined from $\theta(s)$, $s \leq t - \tau$, and the message $\tilde{\theta}(t)$ is determined from the channel output $\tilde{\eta}(s)$ prior to time t such that the error of reproduction of the control signal is less than ε.

The direct theorem holds for rather general situations. Consider, for example, an important special case where control signal $\theta(t)$ is simulated as an arbitrary Gaussian stationary random process (with discrete or continuous time) and transmission is done through a channel with Gaussian white noise $\zeta(t)$ of power N with limited power P at the transmitter input, so that

$$\tilde{\eta}(t) = \eta(t) + \zeta(t),\ M\eta^2(t) \leq P,\ P > 0 \tag{21}$$

and there is a noiseless instantaneous feedback. In such a situation for

$$H_\varepsilon^\tau(\theta) < C,\ \tau > 0,\ \ T \to \infty \tag{22}$$

where $C = \dfrac{P}{2N}$ is the capacity of continuous-time channel (similar to $C = \dfrac{1}{2}\ln\left(1 + \dfrac{P}{N}\right)$ in the case of discrete time), one can determine linear transforms f and φ:

$$\eta(t) = f(\theta(s),\ \tilde{\eta}(r),\ s \leq t - \tau,\ r \leq t) \tag{23}$$

and

$$\tilde{\theta}(t) = \varphi(\tilde{\eta}(r),\ r \leq t) \tag{24}$$

so as to satisfy conditions (10) and (21).

The procedure of finding explicit expressions (23) and (24), and also $H_\varepsilon^\tau(\theta)$ for Gaussian processes $\theta(t)$ is similar to that of finding optimal filters in Kolmogorov—Wiener's problems of optimal linear filtration.

If $\theta(t)$ is a Markov—Gaussian process with correlation function (18), (23) and (24) may be rewritten as follows:

$$\eta(t) = \frac{\theta(t) - \tilde{\theta}(t)}{\mathfrak{D}\left(\theta(t) - \tilde{\theta}(t)\right)}\sqrt{P}; \tag{25}$$

and

$$\tilde{\theta}(t) = \sqrt{\frac{\beta(A - \varepsilon^2)}{N}} \int_{-\infty}^{t} e^{-\beta(s-t)}\, d\tilde{\eta}(s)$$

where ε^2 is determined from the equality

$$H_\varepsilon^\tau(\theta) = \frac{P}{2N}. \tag{27}$$

One can, thus, assert that the mutual influence of the control theory and the information transmission theory is fruitful. The problems arising in the control theory lead to new, original formulations of problems in the information theory, and their solutions are promising for the control theory.

References

1. *Petrov, B. N.*, Engineering cybernetics, 1968. Summary of science and technology. Engineering cybernetics, vol. 4. Summary of science and technology. Moscow, 1972 (in Russian).
2. *Shannon, C.*, A Mathematical Theory of Communications, Bell System Technical Journal, **27** (1948), No. 3, No. 4.
3. *Kolmogorov, A. N.*, Three approaches to definition of the notion of the "amount of information", Problemi Peredachi Informatsii, No. 1 (1965), 3 (in Russian).
4. *Rényi, A.*, On Measures of Entropy and Information. Proc. 4-th Berkeley Sympos. Math. Statist. and Probability, 1960, 1, 547—561.
5. *Vajda, I.*, On the Amount of Information Contained in a Sequence of Independent Observations, Kybernetika **4** (1970), 5, 306—323.

О некоторых взаимосвязях теории информации и теории управления

Б. Н. ПЕТРОВ, Р. Л. ДОБРУШИН, М. С. ПИНСКЕР, Г. М. УЛАНОВ, С. В. УЛЬЯНОВ

(Москва)

В статье обсуждаются вопросы взаимосвязи принципов теории информации и теории управления. Рассмотрены возможные подходы количественного измерения информации и построения мер количества информации. На конкретных примерах показана целесообразность и особенности применения информационных методов к классическим задачам различения гипотез, оценки пропускной способности каналов связи и др. Данные результаты являются основой использования информационных методов в теории управления. Показано взаимное влияние друг на друга принципов теории управления и теории передачи информации: проблемы теории управления приводят к новым оригинальным постановкам задачи в теории информации (определение эпсилон-энтропии без предвосхищения и др.), решения этих задач одновременно перспективны для применений в теории управления.

Б. Н. Петров
Академия Наук СССР
СССР, Москва В-71,
Ленинский проспект, 14

Problems of Control and Information Theory, Vol. 5 (1), pp. 39—49 (1976)

RECURSIVE STATE AND PARAMETER ESTIMATION IN NONLINEAR STOCHASTIC SYSTEMS. I

T. FREY, J. GYÜRKI

(Budapest)

(Received April 23, 1975)

1. Introduction

The available recursive parameter and state estimation procedures (first and second order conditional-mean-filtering algorithms [1]; maximum aposteriori filtering algorithms based on the invariant imbedding; a new method developed by us avoiding the use of invariant imbedding [2], etc.) present in the practice divergency properties. This phenomenon has been explained by data saturation and attacked by the artificial increase of the gain matrix or the plant noise. On the other hand if the system nonlinearities are significant, the source of the divergency phenomenon changes and is due to the applied incorrect approximations in the filtering procedures. The approximation and estimation errors will accumulate and — although the error variance, computed in the course of the filtering process will decrease as the estimation time interval increases — the actual error variance continually increases. The reason of this is that the computed minimum variance (MV) or maximum aposteriori (MAP) estimate becomes more and more inaccurate and the actual error variance related to the accurate MV or MAP estimate (which is computed in the course of the estimation process) becomes more and more unreliable and diverges from the true value.

In this first part we describe and analyze a method which is appropriate to modify the classical filtering procedures in such a way that the effect of the mentioned inaccuracies can be decreased to an arbitrarily small value and we give an estimate for the measure of the error propagation caused by the applied approximations for the nonlinearities of the system.

In Part II we present numerically attractive algorithms, offering new filtering and prediction techniques for systems described by nonlinear stochastic differential equations in form of plant and observation models. We will discuss the perturbation of stochastic differential equations and based on that we present exact error estimation formulae for the filtering and extrapolation procedures, applying different approximation principles.

In Part III we will separate the problems of state and parameter estimation in such a way, that for the latter one we propose a stochastic approximation

algorithm, controlled by the innovation process. In this way the dimension of the state space can significantly be decreased for combined estimation problems.

2. An outline of the methods

The basic idea of our first method is rather simple. In the classical procedures the nonlinearities are substituted by a first or second order polynomial in the vicinity of the estimated value, and the estimated value is also computed on this basis. This approach is sufficiently accurate if the state variable lies with almost probability one in that interval, where the second order polynomial, substituting the nonlinearity is sufficiently accurate. The quality of the results deteriorates if the probability that the state variable is outside of the acceptable approximation interval increases. To detect this fact it is practical to compare the computed error variance with the radius of that sphere where the applied approximation is better than a prescribed, sufficiently small ε value. When the variance becomes greater than this radius, the increase of the error can be expected.

A possibility in this case: let us substitute the system at the given time with a mixture of systems, characterized by a message and observation model which is identical with the original one but having different initial distribution. It means that the actual system agrees with a definite probability with the elements of the applied substituting mixture. The initial distributions of the new systems are declared so, that their variance is small, but their probabilities come from the fact that the initial distributions of the mixture have to agree with the given one at the initial time. In this way it can be guaranteed that for each subsystem the radius of the ε-approximation sphere is essentially greater than the variance of the state estimate. After this decomposition the evolution of each subsystem is traced by the applied approximation procedure. It is important to note however, that the innovation, supplied for each subsystem by the observation model gives information not only on the evolution of that particular subsystem, but it corrects its apriori probability, too.

It has to be mentioned here that the actual decomposition will be a simple task only at the true initial time, because the actual distribution of the initial state will be known at this point.

Later on the classical methods — and the procedure proposed by us — will produce only the conditional mean value and variance (related to the innovation) of the system state, but not its probability distribution. Consequently if the need for a decomposition would occur "in the course" of the estimation process we can choose from the following three possible procedures:

a) The probabilities which determine the decomposition will be chosen so, that they will reproduce the mean value and the variance;

b) We will return to the initial time and a new, better decomposition will be generated compared with the original one;

c) Based on the sufficiently accurate approximated solution of the conditional Fokker–Planck equation the actual conditional distribution function for the examined time t will be generated sufficiently accurately.

The first procedure from the preceding three possibilities needs relatively small amount of computation, but it is inaccurate: firstly it does not reduce the error accumulated in the course of the filtering process, secondly it is not an exact decomposition for the system, because the actual distribution of the mixture reflects the actual distribution of the state vector only up to its second moment.

The second procedure needs more computation, but it will reduce the accumulated error due to the better decomposition applied. On the other hand, it does not require a new decomposition in that time moment, when we "look back" to the critical time.

The third procedure needs the most computation but it reduces remarkably the accumulated errors.

3. The evolution of the mixture-plant

Let us consider first discrete-time plants, described by the message and observation model:

$$\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k), k) + \mathbf{G}(\mathbf{x}(k), k)\, \mathbf{w}(k) \tag{1}$$

$$\mathbf{z}(k) = \mathbf{h}(\mathbf{x}(k), k) + \mathbf{v}(k) \tag{2}$$

where

$\mathbf{x}(k)$ represents the N-dimensional state vector at the kth time moment,

$\mathbf{f}(.,.)$ is the deterministic evolution component of the plant,

$\mathbf{w}(k)$ represents the M-dimensional Gauss–Markov white noise, which is transformed into the actual plant noise by matrix $\mathbf{G}(.,.)$,

$\mathbf{z}(k)$ is the R-dimensional observation vector at the kth time moment which is related with the actual state of the plant and observation noise $\mathbf{v}$ characterized by Eq. (2),

$\mathbf{v}(k)$ represents the Gauss–Markov white observation noise, which is independent from $\mathbf{x}$, $\mathbf{w}$ and $\mathbf{z}$.

It is presupposed, that the mean value $\hat{\mathbf{x}}(0)$ and the variance matrix $\mathbf{V}_x(0)$ of the initial state $\mathbf{x}(0)$ is known, together with its distribution function. The variance matrices of the noises $\mathbf{v}$ and $\mathbf{w}$ are $\mathbf{V}_v(k)$ and $\mathbf{V}_w(k)$ and are known, too.

Let us decompose the initial distribution of $\mathbf{x}(0)$ as follows:

$$F(\mathbf{t}) = P(\mathbf{x}(0) < \mathbf{t}) = \sum_{i=1}^{S} P_i(0)\, F_i(\mathbf{t}) \tag{3}$$

where the distribution functions F_i characterize the distribution of the random variables $\mathbf{x}_i(0)$, which have finite mean $\hat{\mathbf{x}}_i(0)$ and variance $\mathbf{V}_{i,\mathbf{x}}(0)$. $P_i(0)$ are nonnegative numbers generating a discrete distribution (namely $\sum_i P_i(0) = 1$). Consequently it is true now that the following relations hold:

$$\hat{\mathbf{x}}(0) = \mathbf{E}\{\mathbf{x}(0)\} = E_i\{\mathbf{E}_{\mathbf{x}}\{\mathbf{x}_i(0) \mid i\}\} = \sum_i P_i(0)\, \hat{\mathbf{x}}_i(0) \tag{4}$$

$$\begin{aligned} \mathbf{V}_{\mathbf{x},0} &= E_i\, \mathbf{cov}\, \{\mathbf{x}_i \mid i\} + \mathbf{D}_i^2\, \mathbf{E}_{\mathbf{x}}\{\mathbf{x}_i \mid i\} = \\ &= \sum_i P_i(0)\, \mathbf{V}_{i,\mathbf{x}}(0) + \sum_i P_i(0)\, [\hat{\mathbf{x}}_i(0) - \hat{\mathbf{x}}(0)]\, [\hat{\mathbf{x}}_i(0) - \hat{\mathbf{x}}(0)]^* . \end{aligned} \tag{5}$$

It can be considered, that at the initial time the plant will be realized by the plant having $\mathbf{x}_i(0)$ as initial state with probability $P_i(0)$ and the evolution of each subsystem is governed by the message-observation model [1]–[2]. For a given index i the evolution of the ith plant can be traced by the usual approximation methods, developed for nonlinear discrete models (e.g. second-order minimum covariance-of-error estimation procedure [1]).

However the consecutive observations present information not only to trace the evolution of the different plants but they change the probability that the actual plant has been realized by the ith plant. This probability can be determined (approximately) by the Bayes rule. Let $\mathbf{X}_i(k)$ the series of state vectors for the ith plant (namely $\mathbf{X}_i(k) = \{\mathbf{x}_i(0), \mathbf{x}_i(1) \ldots \mathbf{x}_i(k)\}$) $\mathbf{Z}(k)$ the series of observations (namely $\mathbf{Z}(k) = \{\mathbf{z}(1), \ldots, \mathbf{z}(k)\}$) and $P_i(k \mid \mathbf{Z}(k))$ the aposteriori probability that after the kth observation the actual plant will be realized by the ith plant. Using the Bayes rule it can be written:

$$P_i(k+1 \mid \mathbf{Z}(k+1)) = \frac{1}{p(\mathbf{z}(k+1) \mid \mathbf{Z}(k))} \cdot P(\mathbf{z}(k+1) \mid i, \mathbf{Z}(k)) \cdot P_i(k \mid \mathbf{Z}(k)), \tag{6}$$

resp.

$$\vartheta_i(k+1) = \frac{P_i(k+1 \mid \mathbf{Z}(k+1))}{P_i(k \mid \mathbf{Z}(k))} = \frac{p(\mathbf{z}(k+1)\, i, \mathbf{Z}(k))}{\sum_j P_j(k \mid \mathbf{Z}(k)) \cdot p(\mathbf{Z}(k+1 \mid j, \mathbf{Z}(k))} . \tag{7}$$

Namely the recursive computation of $\vartheta_i(k+1)$ requires the computation of the following ratio:

$$p\big(\mathbf{z}(k+1)\,|\,i,\mathbf{Z}(k)\big)=\int_{-\infty}^{\infty}p\big(\mathbf{z}(k+1)\,|\,i,\mathbf{X}_i(k+1),\mathbf{Z}(k)\big)\cdot$$

$$\cdot p\big(\mathbf{X}_i(k+1)\,|\,i,\mathbf{Z}(k)\big)\,d\mathbf{X}_i(k+1)=$$

$$=\int_{-\infty}^{\infty}p\big(\mathbf{z}(k+1)\,|\,i,\mathbf{x}_i(k+1)\big)\cdot\frac{p\big(\mathbf{Z}(k)\,|\,i,\mathbf{X}_i(k+1)\big)\cdot p\big(\mathbf{x}_i(k+1)\,|\,i\big)}{p\big(\mathbf{Z}(k)\,|\,i\big)}\,d\mathbf{X}_i(k+1)=$$

$$=\frac{\int_{-\infty}^{\infty}p\big(\mathbf{z}(k+1)\,|\,i,\mathbf{x}_i(k+1)\big)\cdot p\big(\mathbf{Z}(k)\,|\,i,\mathbf{X}_i(k)\big)\cdot p\big(\mathbf{X}_i(k+1)\,|\,i\,d\mathbf{X}_i\big)(k+1)}{\int_{-\infty}^{\infty}p\big(\mathbf{Z}(k)\,|\,i,\mathbf{X}_i(k)\big)\cdot p\big(\mathbf{X}_i(k)\,|\,i\big)\,d\mathbf{X}_i(k)}=$$

$$=\frac{\int_{-\infty}^{\infty}p\big(\mathbf{Z}(k+1)\,|\,i,\mathbf{X}_i(k+1)\big)\cdot p\big(\mathbf{X}_i(k+1)\,|\,i\big)\,d\mathbf{X}_i(k+1)}{\int_{-\infty}^{\infty}p\big(\mathbf{Z}(k)\,|\,i,\mathbf{X}_i(k)\big)\cdot p\big(\mathbf{X}_i(k)\,|\,i\big)\,d\mathbf{X}_i(k)} \tag{8}$$

where the expressions in the numerator and the denominator are the well-known expressions of the maximum-likelihood-aposteriori (MLAP) state estimation [1], except that here the distribution of $\mathbf{x}_i(0)$ depends on the decomposition applied and in general is non-Gaussian. According to the well-known technique of the MLAP, the numerator of the $p\big(\mathbf{z}(k+1)\,|\,i,\,\mathbf{Z}(k)\big)$ can be written:

$$A\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2}\log\,\det\boldsymbol{\Omega}_{k+1}^{(i)}-\frac{1}{2}\|\,\mathbf{x}_i(k+1)-f\big(\mathbf{x}_i(k),k\big)\,\|^{2^{(i)-1}}_{\Omega_{k+1}}-\right.$$

$$\left.-\frac{1}{2}\|\,\mathbf{z}(k+1)-\mathbf{h}\big(\mathbf{x}_i(k+1),k+1\big)\,\|^{2}_{\mathbf{V}_V{}^{-1}}\right\}\cdot\int_{-\infty}^{\infty}p\big(\mathbf{Z}(k),\mathbf{X}_i(k)\,|\,i\big)\,d\mathbf{X}_i(k)\,d\mathbf{x}_i(k+1) \tag{9}$$

where

$$\boldsymbol{\Omega}_{k+1}^{(i)}=\mathbf{G}(\mathbf{x}_i(k+1),k+1)\,\mathbf{V}_W(k+1)\,\mathbf{G}^T\big(\mathbf{x}_i(k+1),k+1\big)\,.$$

To evaluate exactly the integral in (9), we would need the exact distribution of $\mathbf{X}_i(k+1)$, which is difficult to be calculated. Due to this, we apply the following approximation. We know the $\hat{\mathbf{x}}_i(k)$, $\hat{\mathbf{x}}_i(k+1)$, $\mathbf{V}_{i,x}(k)$, $\mathbf{V}_{i,x}(k+1)$ values, needed to evaluate expression (9) up to the effectivity of the accuracy of that algorithm, which has been applied to trace the evolution of the ith

plant. Based on these values $\mathbf{\Omega}_{k+1}^{(i)}$ can be substituted by the following expression:

$$\mathbf{\Omega}_{k+1}^{(i)} \cong \mathbf{G}(\hat{\mathbf{x}}_i(k+1), k+1)\,\mathbf{V}_W(k+1)\,\mathbf{G}^T(\hat{\mathbf{x}}_i(k+1), k+1)\,.$$

With this approximation $\mathbf{\Omega}_{k+1}^{(i)}$ will be independent from $\mathbf{x}_i(k+1)$ and the following expressions are also true:

$$\begin{aligned}\|\mathbf{z}(k+1) - \mathbf{h}\big(\mathbf{x}_i(k+1), k+1\big)\|^2_{\mathbf{V}_V(k+1)} = \big[\big(\mathbf{z}(k+1) - \mathbf{h}\big(\hat{\mathbf{x}}_i(k+1), k+1\big) + \\ + \mathbf{h}\big(\hat{\mathbf{x}}_i(k+1), k+1\big) - \mathbf{h}\big(\mathbf{x}_i(k+1), k+1\big)\big)\big]^T \mathbf{V}_w^{-1}(k+1)\,\big[\mathbf{z}(k+1) - \\ - \mathbf{h}\big(\hat{\mathbf{x}}_i(k+1), k+1\big)\big) + \mathbf{h}\big(\mathbf{x}_i(k+1), k+1\big) - \mathbf{h}\big(\mathbf{x}_i(k+1), k+1\big)\big] \cong \\ \cong \mathbf{I}_i^{(T)}(k+1)\,\mathbf{V}_V^{-1}\,\mathbf{I}_i(k+1) + 2\mathbf{I}_i^{(T)}(k+1)\,\mathbf{V}_V^{-1}\cdot \\ \cdot \frac{\partial \mathbf{h}[\hat{\mathbf{x}}_i(k+1), k+1]}{\partial \hat{\mathbf{x}}_i}\,[\hat{\mathbf{x}}_i(k+1) - \mathbf{x}_i(k+1)], \qquad (10)\end{aligned}$$

$$\begin{aligned}\|\mathbf{x}_i(k+1) - \mathbf{f}\big(\mathbf{x}_i(k), k\big)\|^2_{\mathbf{\Omega}_{k+1}^{(i)-1}} \sim \big(\mathbf{x}_i(k+1) - \mathbf{f}(\mathbf{x}_i(k), k\big)^T\,\mathbf{\Omega}_{k+1}^{-1}\big(\mathbf{x}_i(k+1) - \\ - \mathbf{f}(\mathbf{x}_i(k), k)\big) \sim [\mathbf{x}_i(k+1) - \mathbf{f}\big(\hat{\mathbf{x}}_i(k), k\big)]^T\,\hat{\mathbf{\Omega}}_{k+1}^{-1}[\mathbf{x}_i(k+1) - \mathbf{f}\big(\hat{\mathbf{x}}_i(k), k\big)] + \\ + 2[\mathbf{x}_i(k+1) - \mathbf{f}\big(\mathbf{x}_i(k), k\big)]^T \cdot \hat{\mathbf{\Omega}}_{k+1}^{(i)-1} \cdot \frac{\partial \mathbf{f}\big(\hat{\mathbf{x}}_i(k), k\big)}{\partial \hat{\mathbf{x}}_i} \cdot [\hat{\mathbf{x}}_i(k) - \mathbf{x}_i(k)] \qquad (11)\end{aligned}$$

where $\mathbf{I}_i(k+1)$ represents the so-called aposteriori innovation, calculated on the basis of $\mathbf{Z}(k+1)$ and is given by

$$\mathbf{I}_i(k+1) = \mathbf{z}_{k+1} - \mathbf{h}\big(\hat{\mathbf{x}}_i(k+1), k+1\big)\,.$$

Substituting expressions (10) and (11) into (9); resolving the sums in the exponents into multiplicative terms; applying second-order polynomial approximation for the exponential function, which is linear in $\mathbf{x}_i(k+1)$; applying first-order polynomial approximation for those terms which contain both $\mathbf{x}_i(k+1)$ and $\mathbf{x}_i(k)$, we get:

$$\begin{aligned}p\big(\mathbf{z}(k+1) \mid \mathbf{Z}(k)\big) = \exp\left\{-\frac{1}{2}\mathbf{I}_i^{(T)}(k+1)\,\mathbf{V}_V^{-1}(k)\,\mathbf{I}_i(k+1)\right\}\cdot \\ \cdot\left[1 + \frac{1}{2}\mathbf{I}_i^{(T)}(k+1)\,\mathbf{V}_V^{-1}\cdot\frac{\partial \mathbf{h}\big(\mathbf{x}_i(k+1), k+1\big)}{\partial \hat{\mathbf{x}}_i}\cdot\right. \qquad (12) \\ \left.\cdot\,\mathbf{V}_{i,\mathbf{x}}(k+1)\,\frac{\partial^T \mathbf{h}\big(\hat{\mathbf{x}}_i(k+1, k+1\big)}{\partial \hat{\mathbf{x}}_i}\,\mathbf{I}_i(k+1)\cdot\int_{-\infty}^{\infty} p\big(\mathbf{Z}(k), \mathbf{X}_i(k) \mid i\big)\,d\mathbf{x}_i(k)\right].\end{aligned}$$

Using the relations of

$$r_n = \frac{s_n}{s_{n+1}} = s_{n-1}\,\sigma_n(1+\delta_n); \quad r_{n-1} = \frac{s_{n-1}}{s_{n-2}} = s_{n-2}\,\sigma_{n-1}(1+\delta_{n-1})$$

we can reduce the following expressions:

$$\frac{r_n}{r_{n-1}} = \frac{s_{n-1}}{s_{n-2}} \cdot \frac{\sigma_n}{\sigma_{n-1}} \cdot \frac{1+\delta_n}{1+\delta_{n-1}} = r_{n-1} \frac{\sigma_n}{\sigma_{n-1}} \cdot \frac{1+\delta_n}{1+\delta_{n-1}} .$$

Applying these relations, expression (12) can be simplified as follows:

$$p\big(z(k+1) \mid i, \mathbf{Z}(k)\big) \cong p^2\big(\mathbf{z}(k) \mid i, \mathbf{Z}(k-1)\big) \cdot \exp\left\{\frac{1}{2} \mathbf{I}_i^{(T)}(k)\, \mathbf{V}_v^{-1}(k)\, \mathbf{I}_i(k) - \right.$$

$$\left. - \frac{1}{2} \mathbf{I}_i^{(T)}(k+1)\, \mathbf{V}_V^{-1}\, \mathbf{I}_i(k+1)\right\} \cdot \left[1 + \frac{1}{2} \mathbf{I}_i^{(T)}(k+1)\, \mathbf{V}_V^{-1} \frac{\partial \mathbf{h}\big(\hat{\mathbf{x}}_i(k+1), k+1\big)}{\partial \hat{\mathbf{x}}_i} \cdot \right.$$

$$\left. \cdot\, \mathbf{V}_{i,\mathbf{x}}(k+1) \frac{\partial^T \mathbf{h}\big(\hat{\mathbf{x}}_i(k+1), k+1\big)}{\partial \hat{\mathbf{x}}_i} \cdot \mathbf{I}_i(k+1)\right] \cdot \tag{13}$$

$$\cdot \left[1 + \frac{1}{2} \mathbf{I}_i^{(T)}(k)\, \mathbf{V}_{\bar{A}}^{-1}(k) \frac{\partial \mathbf{h}\big(\hat{\mathbf{x}}_i(k), k\big)}{\partial \hat{\mathbf{x}}_i} \cdot \mathbf{V}_{i,\mathbf{x}}(k) \cdot \frac{\partial^T \mathbf{h}\big(\hat{\mathbf{x}}_i(k), k\big)}{\partial \hat{\mathbf{x}}_i} \cdot \mathbf{I}_i(k)\right]^{-1} .$$

Expression (13) is valid only for $k > 0$. For $k = 0$ the denominator of $P(\mathbf{Z}(1) \mid i)$ is equal to l, namely

$$p\big(\mathbf{z}(1) \mid i\big) = \exp\left\{-\frac{1}{2} \mathbf{I}_i^{(T)}(1)\, \mathbf{V}_V^{-1}(1)\, \mathbf{I}_i(1)\right\} \cdot$$

$$\cdot \left[1 + \frac{1}{2} \mathbf{I}_i^{(T)}(1)\, \mathbf{V}_V^{-1}(1) \cdot \frac{\partial \mathbf{h}\big(\hat{\mathbf{x}}_i(1), 1\big)}{\partial \hat{\mathbf{x}}_i} \cdot \mathbf{V}_{i,\mathbf{x}}(1) \frac{\partial^T \mathbf{h}\big(\hat{\mathbf{x}}_i(1), 1\big)}{\partial \hat{\mathbf{x}}_i} \cdot \mathbf{I}_i(1)\right] . \tag{14}$$

Based on expressions (13) and (14) $p(\mathbf{z}(k+1) \mid i, \mathbf{Z}(k))$ can be computed recursively, then using expression (7) $\vartheta_i(k+1)$ and finally $P_i(k+1 \mid \mathbf{Z}(k+1))$ can also be calculated.

The estimate of the system's state vector and its variance — which characterizes its reliability — are given by

$$\hat{\mathbf{x}}(k) = \sum_i P_i(k \mid \mathbf{Z}(k))\, \hat{\mathbf{x}}_i(k) \tag{15}$$

$$\mathbf{V}\hat{\mathbf{x}}(k) = \sum_i P_i((k \mid \mathbf{Z}(k)) \cdot [\mathbf{V}_{i,\mathbf{x}}(k) + (\hat{\mathbf{x}}(k) - \hat{\mathbf{x}}_i(k))\, (\hat{\mathbf{x}}(k) - \hat{\mathbf{x}}_i(k))^T] . \tag{16}$$

In the second part of this work we will demonstrate the value of the proposed method by numerical examples.

Let us consider now the case of continuous plants, described by the following message-observation model:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}(t), t)\, dt + \mathbf{G}(\mathbf{x}(t), t)\, d\boldsymbol{\omega}(t) \tag{17}$$

$$\mathrm{d}\mathbf{y} = \mathbf{h}(\mathbf{x}(t), t)\, dt + d\boldsymbol{\nu}(t) . \tag{18}$$

We will explain here only the evolution of the probabilities $P_i(t)$, because otherwise there is no formal difference in the application of the mixture-principle. The Bayes rule can be applied only formally to develop the $P_i(t) = P(i \mid \mathbf{Y}(t))$ probabilities, since the $p(\mathbf{Y}(t+dt) \mid i, \mathbf{Y}(t))$ does not have a meaning directly.

Rewriting of expression (7) formally as

$$\vartheta_i(t, d\mathbf{y}) = \frac{p(d\mathbf{y} \mid i, \mathbf{Y})}{p(d\mathbf{y} \mid \mathbf{Y})} \tag{19}$$

results in a meaningful interpretation as follows: considering the Wiener process $\nu(t)$ in the $0 \leq \tau < t + dt$ interval, it will determine here a σ-algebra $\mathfrak{A}_{t+dt}$ and a corresponding Wiener measure over it. The $\mathbf{Y}(t)$ process is also measurable over $\mathfrak{A}_{t+dt}$, moreover it will generate an absolutely continuous measure P_i, resp. P_0 regarding the preceding Wiener measure. Here the index "i" refers to the measure generated by the ith element and index "0" refers to the measure induced by the mixture. These two measures are absolutely continuous for each other too, and ϑ_i means the derivative of P_i with regard to P_0 in that point of the space $\mathfrak{A}_{t+dt}$ which is determined by the observation $\mathbf{Y}(t)$. Due to the fact that having the usual regularity conditions for the derivatives (with regard to each other) of those measures which are considered over an algebra generated by a sufficiently fine decomposition will approximate ϑ_i sufficiently well [3, 4], and here the Bayes rule can be applied, the ideas which we explained for the discrete case can be used with a slight modification for the continuous case, too.

It can be written consequently

$$\vartheta_i(t, d\mathbf{y}) = p(dy \mid i, \mathbf{Y}) \cdot p^{-1}(d\mathbf{y} \mid \mathbf{Y}) =$$

$$= \frac{\int_{-\infty}^{\infty} p\big(d\mathbf{y} \mid i, \mathbf{x}_i(t), \mathbf{Y}(t)\big) \cdot p\big(\mathbf{x}_i(t) \mid i, \mathbf{Y}(t)\big)\, d\mathbf{x}_i}{\sum_j P_j(t) \int p\big(d\mathbf{y} \mid j, \mathbf{x}_j(t), \mathbf{Y}(t)\big) \cdot p\big(\mathbf{x}_j(t) \mid j, \mathbf{Y}(t)\big)\, d\mathbf{x}_j} . \tag{20}$$

Thus consider the following integral

$$\int P(d\mathbf{y} \mid i,\ \mathbf{x}_i(t), \mathbf{Y}(t)) \cdot p(\mathbf{x}_i(t) \mid,\ \mathbf{Y}(t))\, d\mathbf{x}_i \tag{21}$$

wich is needed to determine (20) and where

$$p\big(d\mathbf{y} \mid i, \mathbf{x}_i(t), \mathbf{Y}(t)\big) = p\big(d\mathbf{y} \mid i, \mathbf{x}_i(t)\big) =$$

$$= \exp\left\{ -\frac{1}{2dt} [d\mathbf{y} - \mathbf{h}\big(\mathbf{x}_i(t), t\big)\, dt]^T\, \mathbf{V}_V^{-1}(t) [d\mathbf{y} - \mathbf{h}\big(\mathbf{x}_i(t), t\big)\, dt] \right\}. \tag{22}$$

By completing the multiplications on expression (22) we can neglect the bilinear terms in $d\mathbf{y}$ (because it is independent from both $\mathbf{x}_i$ and i, we can eliminate the corresponding factor from the numerator and denominator by taking out it from the integral). Denoting by $\tilde{p}$ the truncated p we get:

$$\int p\big(d\mathbf{y} \mid i, \mathbf{x}_i(t), \mathbf{Y}(t)\big) \cdot p\big(\mathbf{x}_i(t) \mid i, \mathbf{Y}(t)\, d\mathbf{x}_i = A \int_{-\infty}^{\infty} \exp\Big[-\Big\{ d\mathbf{y}^T\, \mathbf{V}_V^{-1}(t)\, \mathbf{h}\big(\mathbf{x}_i(t), t\big) -$$

$$- \frac{1}{2}\, \mathbf{h}^T\big(\mathbf{x}_i(t), t\big)\, \mathbf{V}_V^{-1}(t) \cdot \mathbf{h}\big(\mathbf{x}_i(t), t\big)\, dt \Big\}\Big] \cdot p\big(\mathbf{x}_i(t) \mid i, \mathbf{Y}(t)\big)\, d\mathbf{x}_i\,. \tag{23}$$

Applying first-order approximation for the exponential term in dt (which is as second-order in $d\mathbf{y}$) we can get:

$$\exp\Big\{ - d\mathbf{y}^T\, \mathbf{V}_V^{-1}(t)\, \mathbf{h}\big(\mathbf{x}_i(t), t\big) + \frac{1}{2}\, \mathbf{h}^T\big(\mathbf{x}_i(t), t\big)\, \mathbf{V}_V^{-1} \mathbf{h}\big(\mathbf{x}_i(t), t\big)\, dt \Big\} \cong$$

$$\cong 1 + \frac{1}{2}\, \mathbf{h}^T\big(\mathbf{x}_i(t), t\big)\, \mathbf{V}_V^{-1} \mathbf{h}\big(\mathbf{x}_i(t), t\big)\, dt - \mathbf{h}^T\big(\mathbf{x}_i(t), t\big)\, \mathbf{V}_V^{-1}\, d\mathbf{y} +$$

$$+ \frac{1}{2}\, d\mathbf{y}^T \mathbf{V}_V^{-1} \mathbf{h}\big(\mathbf{x}_i(t), t\big)\, \mathbf{h}^T\big(\mathbf{x}_i(t), t\big)\, \mathbf{V}_V^{-1}\, d\mathbf{y} \cong 1 + \hat{\mathbf{h}}^T \mathbf{V}_V^{-1}\, \hat{\mathbf{h}} dt +$$

$$+ \frac{1}{2}\, \big(\mathbf{x}_i(t) - \hat{\mathbf{x}}_i(t)\big)^T \frac{\partial \hat{\mathbf{h}}^T}{\partial \hat{\mathbf{x}}_i}\, \mathbf{V}_V^{-1} \hat{\mathbf{h}}\, dt + \frac{1}{2}\, \hat{\mathbf{h}}^T \mathbf{V}_V^{-1} \frac{\partial \hat{\mathbf{h}}}{\partial \hat{\mathbf{x}}_i}\, \big(\mathbf{x}_i(t) - \mathbf{x}_i(t)\big)\, dt +$$

$$+ \frac{1}{2}\, \mathrm{tr}\Big(\mathbf{V}_V^{-1} \frac{\partial \mathbf{h}}{\partial \mathbf{x}_i}\, (\mathbf{x}_i - \mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_i)^T\, dt \Big) - \mathbf{h}^T \mathbf{V}_V^{-1}\, d\mathbf{y} - \big(\mathbf{x}_i(t) - \hat{\mathbf{x}}_i(t)\big)\, dt +$$

$$+ \frac{1}{2}\, \hat{\mathbf{h}}^T \mathbf{V}_V^{-1} d\mathbf{y} d\mathbf{y}^T\, \mathbf{V}_V^{-1} \hat{\mathbf{h}} + \frac{1}{2}\, (\mathbf{x}_i - \hat{\mathbf{x}}_i)\, \frac{\partial \hat{\mathbf{h}}^T}{\partial \hat{\mathbf{x}}_i}\, (\mathbf{x}_i - \mathbf{x}_i) +$$

$$+ \frac{1}{2}\, \mathrm{tr}\Big(\mathbf{V}_V^{-1} d\mathbf{y} d\mathbf{y}^T \mathbf{V}_V^{-1}\, \frac{\partial \hat{\mathbf{h}}}{\partial \hat{\mathbf{x}}_i}\, (\mathbf{x}_i - \hat{\mathbf{x}}_i)(\mathbf{x}_i - \hat{\mathbf{x}}_i)^T \frac{\partial \hat{\mathbf{h}}^T}{\partial \hat{\mathbf{x}}_i} \Big)$$

where the meaning of $\hat{\mathbf{h}}$ and $\dfrac{\partial \hat{h}}{\partial \hat{\mathbf{x}}_i}$ is that the corresponding matrices are evaluated at the $\hat{\mathbf{x}}_i$ coordinates. Using the well-known relation of

$$d\mathbf{y}\, d\mathbf{y}^T = \mathbf{V}_V\, dt + \mathbf{o}(dt)$$

after substitution and completing the integral we can get the following relation:

$$\int \tilde{p}\big(d\mathbf{y} \mid i, \mathbf{x}_i(t), \mathbf{Y}(t)\big) \cdot p\big(\mathbf{x}_i(t) \mid i, \mathbf{Y}(t)\big)\, d\mathbf{x}_i \cong 1 + \frac{1}{2}\hat{\mathbf{h}}^T \mathbf{V}_V^{-1} \hat{\mathbf{h}}\, dt +$$

$$+ \frac{1}{2}\operatorname{tr}\left(\mathbf{V}_V^{-1} \frac{\partial \hat{\mathbf{h}}}{\partial \hat{\mathbf{x}}_i} \mathbf{V}_{i,\mathbf{x}}(t) \cdot \frac{\partial \hat{\mathbf{h}}^T}{\partial \hat{\mathbf{x}}_i}\right) dt - \hat{\mathbf{h}}^T \mathbf{V}_V^{-1}\, d\mathbf{y} + \frac{1}{2}\hat{\mathbf{h}}^T \mathbf{V}_V^{-1} \hat{\mathbf{h}}\, dt +$$

$$+ \frac{1}{2}\operatorname{tr}\left(\mathbf{V}_V^{-1} \frac{\partial \hat{\mathbf{h}}}{\partial \hat{\mathbf{x}}_i} \mathbf{V}_{i,\mathbf{x}}(t) \cdot \frac{\partial \hat{\mathbf{h}}^T}{\partial \hat{\mathbf{x}}_i}\right) dt = 1 - \mathbf{I}_i^T(d\mathbf{y}, t)\, \mathbf{V}_V^{-1} \hat{\mathbf{h}} +$$

$$+ \operatorname{tr}\left(\mathbf{V}_V^{-1} \frac{\partial \hat{\mathbf{h}}}{\partial \mathbf{x}_i} \mathbf{V}_{i,\mathbf{x}}(t) \cdot \frac{\partial \hat{\mathbf{h}}^T}{\partial \mathbf{x}_i}\right) dt \tag{24}$$

where

$$\mathbf{I}_i(d\mathbf{y}, t) = d\mathbf{y}(t) - \mathbf{h}(\mathbf{x}_i(t), t)\, dt.$$

Substituting (24) into expression (20), keeping in mind the meaning of $\vartheta_i(t, d\mathbf{y})$, furthermore dividing by dt and taking the limit $dt \to 0$ we conclude in the following relation:

$$\frac{dP_i(t)}{dt} = \frac{\operatorname{tr}\left\{\mathbf{V}_V^{-1} \frac{\partial \hat{\mathbf{h}}_i(t)}{\partial \hat{\mathbf{x}}_i} \cdot \mathbf{V}_{i,x_i}(t) \frac{\partial \hat{\mathbf{h}}_i^T(t)}{\partial \hat{\mathbf{x}}_i}\right\} - \mathbf{I}_i^T(d\mathbf{y}, t)\, \mathbf{V}_V^{-1} \hat{\mathbf{h}}_i(t)}{\sum_j P_j(t)\left\{\operatorname{tr}\left[\mathbf{V}_V^{-1} \frac{\partial \hat{\mathbf{h}}_j^T(t)}{\partial \hat{\mathbf{x}}_j} \mathbf{V}_{j,\mathbf{x}})t\right) \frac{\partial \hat{\mathbf{h}}_j^T(t)}{\partial \mathbf{x}_j}\right] - \mathbf{I}_j^T(d\mathbf{y}, t)\, \mathbf{V}_V^{-1} \hat{\mathbf{h}}_j(t)\right\}}. \tag{25}$$

The accuracy of (25) is better than that of the discrete case. We shall present an illustrating example for this in the second part of our work.

References

1. *Sage, A. P., Melsa, J. L.*, Estimation Theory with Application to Communications and Control. McGraw Hill, New York, 1971.
2. *Frey, T.*, Investigations of Optimal Controllability and Observatibility of Some Stochastic Processes. Preprints of Stochastic Control Symposium. Budapest, September 25—27, 1974.
3. *Gichman, I. I., Skorohod, A. W.*, Stochastische Differentialgleichungen. Akademie-Verlag, Berling, 1971.
4. *Wong, E.*, Stochastic Processes in Information and Dynamical Systems. McGraw Hill, New York, 1971.

Рекуррентный алгоритм оценки состояния и параметров нелинейных стохастических систем

Т. ФРЕЙ, Й. ДЮРКИ

(Будапешт)

При использовании известных методов рекуррентной оценки параметров и состояния системы часто возникает проблема сходимости алгоритмов. Нарушение правил сходимости объясняется явлениями насыщения при сборе данных и может быть устранено путем увеличения матрицы усиления или возмущений. Если в системе присутствуют существенные нелинейности, то расходимость появляется под влиянием неточности приближения в алгоритмах фильтрации. Ошибки приближения и оценки суммируются и вариация действительной ошибки непрерывно растет. Причина данного явления заключается в том, что рассчитанная оценка с минимальной дисперсией становится все более и более неточной, а вариация действительной ошибки, относящаяся к оценке с минимальной дисперсией, будет все более отличаться от действительного значения.

Приводится анализ предложенной методики, приемлемой для усовершенствования классического метода фильтрации таким образом, что при ее использовании удается в значительной мере уменьшить вышеприведенные неточности.

T. Frey
J. Gyürki
Computer and Automation Institute
Hungarian Academy of Sciences
H-1502 Budapest, XI. Kende u. 13—17.

Problems of Control and Information Theory, Vol. 5 (1), pp. 51—61 (1976)

ANALYTICAL SYNTHESIS OF THE STATISTICAL QUASIOPTIMAL CONTROL OF THE NONLINEAR SUBJECT

I. E. KAZAKOV

(Moscow)

(Received November 3, 1974)

The procedure of the analytical synthesis of the quasioptimal control based on the quadratic criteria of the generalized work for the nonlinear subject with non-exact measure and random disturbances of general case is considered. The terminal and nonterminal problems are discussed.

The method of the analytical synthesis of the optimal control based on the criteria of the generalized work is developed due to the exact measures of phase coordinates and a particular types disturbances influencing on the nonlinear subject [1, 2]. In paper [3] the generalization of the method of synthesis of the control for the linear subject under random disturbances of the arbitrary type and the quadratic criteria of the generalized work and non-exact measures of the phase coordinates is proved. This result can be also expanded for the nonlinear subject of the control.

Let the subject of the control be given by the nonlinear vector equation

$$\dot{\mathbf{Y}} = \boldsymbol{\varphi}(\mathbf{Y}, t) + \mathbf{D}\mathbf{u} + \mathbf{u}_0 + \mathbf{V}, \quad \mathbf{Y}(t_0) = \mathbf{Y}_0\,, \tag{1}$$

where $\mathbf{Y}$ is the n-dimensional vector of the phase coordinates; $\boldsymbol{\varphi}$ — the vector nonlinear function, $\mathbf{D}(t)$ — the given $n \times r$ matrix and $r \leq n$, $\mathbf{u}$ is the r-dimensional control vector, $\mathbf{V}(t)$ is white Gaussian noise vector with the matrix of intensity $\mathbf{G}(t)$ and with zero the mathematical expectation; $\mathbf{u}_0$ — the deterministic vector function. The initial state of the subject is random with the given vector of the mathematical expectation $\mathbf{m}_0$ and the correlation matrix $\boldsymbol{\Theta}_0$. The required state vector of the control subject satisfies the equation

$$\dot{\mathbf{y}}^* = \boldsymbol{\varphi}(\mathbf{y}^*, t) + \mathbf{u}_0, \mathbf{y}^*(t_0) = 0\,. \tag{2}$$

In the particular cases the vector $\mathbf{u}_0 = 0$. Then the vector is identical zero ($\mathbf{y}^* = 0$). In the interval (t_0, t_k) the vector $\mathbf{Z}(t)$ of the dimension $m \leq n$ can be measured. Vector $\mathbf{Z}(t)$ is connected with vector $\mathbf{Y}(t)$ by the linear vector dependence

$$\mathbf{Z} = \mathbf{C}\mathbf{Y} + \mathbf{N}, \tag{3}$$

where $\mathbf{C}(t)$ is the given $m \times n$ matrix and $m \leq n$, $\mathbf{N}(t)$ is the white Gaussian

noise vector which is not connected with $\mathbf{V}(t)$ and $\mathbf{Y}_0$ with the matrix of intensity $\mathbf{Q}(t)$ and with the mathematical expectation $M[\mathbf{N}(t)] = 0$.

The problem is how to determine the control vector $\mathbf{u}$ which minimizes the quadratic functional of the generalized work

$$M[F] = M[\mathbf{X}^T(t_k)\,\boldsymbol{\Gamma}_k\mathbf{X}(t_k)] + \int_{t_0}^{t_k} M[\mathbf{X}^T(t)\,\mathbf{L}(t)\mathbf{X}(t) + \mathbf{u}^T(t)\mathrm{K}^{-1}\,\mathbf{u}(t) + \\ + \mathbf{X}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{D}(t)\,\mathbf{K}\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t)]\,\mathrm{d}t, \tag{4}$$

where $\mathbf{X}(t) = \mathbf{Y}(t) - \mathbf{y}^*(t)$ is the vector error, $\boldsymbol{\Gamma}_k$, $\mathbf{L}(t)$ are given symmetrical positively definite matrices, K — the given diagonal matrix with the positive coefficients, M — the operation of the mathematical expectation. The matrix $\boldsymbol{\Gamma}$ can be determined by the additional equation but it should be symmetrical positively definite and it should satisfy the condition

$$\boldsymbol{\Gamma}(t_k) = \boldsymbol{\Gamma}_k. \tag{5}$$

In [1, 2] two types of the terminal and nonterminal control problems are considered. The problem of minimizing of the functional (4) in the given interval $t_k - t_0$ is terminal or quasiterminal [2] as the transient process is also taken into consideration. The interval $t_k - t_0$ is arbitrary for the nonterminal problem and $t_k = t$ is current time, the matrix $\boldsymbol{\Gamma}_k$ is supposed to be equal to the matrix $\boldsymbol{\Gamma}$.

The exact analytical solution of the problem discussed here cannot be obtained, but one can approximately determine the quasioptimal control vector for the linearized subject. In general case while applying the statistical linearization of the nonlinear vector function $\boldsymbol{\varphi}$ regarding the vector of the mathematical expectation m_y^* and the correlation matrix $\boldsymbol{\Theta}_y^*$ which exist in the optimal control we have

$$\boldsymbol{\varphi}(\mathbf{Y}, t) = \mathbf{A}(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t)\,\mathbf{Y} + \boldsymbol{\varphi}_0'(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t), \tag{6}$$

where

$$\boldsymbol{\varphi}'(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t) = \boldsymbol{\varphi}_0(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t) - \mathbf{A}(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t)\mathbf{m}_y^*.$$

$\boldsymbol{\varphi}_0$ is the statistical vector characteristic of nonlinearity, $\mathbf{A}$ is the matrix of the statistical equivalent coefficients of the gain [4]. Taking into account Eq. (6) we can write the approximate linearized equation of the subject in the form

$$\dot{\mathbf{Y}} = \mathbf{AY} + \mathbf{Du} + \mathbf{u}_0 + \mathbf{V} + \boldsymbol{\varphi}_0', \mathbf{Y}(t_0) = \mathbf{Y}_0. \tag{7}$$

We obtain the approximate equation for the error $\mathbf{X}(t)$ by the subtraction of Eq. (7) from Eq. (2)

$$\dot{\mathbf{X}} = \mathbf{AX} + \mathbf{Du} + \mathbf{V} + \mathbf{f}(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, \mathbf{y}^*, t),\ \mathbf{X}(t_0) = \mathbf{Y}_0, \tag{8}$$

where vector $\mathbf{f}$ is

$$\mathbf{f}(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, \mathbf{y}^*, t) = \boldsymbol{\varphi}_0'(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t) - \boldsymbol{\varphi}(\mathbf{y}^*, t) + \mathbf{A}\mathbf{y}^*.$$

The vector $\mathbf{f}$ has small components and by ordinary linearization regarding $\mathbf{y}^*$ it is equal to zero.

The quasioptimal vector of the control $\mathbf{u}$ can be determined by the following theorem.

Theorem. The optimal vector $\mathbf{u}$ which minimizes the quadratic functional (4) for the nonlinear subject (1) approximately given by Eq. (7) during the measure of the vector (2) is the following:

$$\mathbf{u} = \mathbf{K}\mathbf{D}^T\boldsymbol{\Gamma}(\hat{\mathbf{Y}}(t) - \mathbf{y}^*(t)), \tag{9}$$

where $\hat{\mathbf{Y}}(t) = M[\mathbf{Y}(t) \,|\, \mathbf{Z}(\tau), t_0 \leq \tau \leq t] = M_z[\mathbf{Y}]$ is the conditional mathematical expectation.

It means the estimation of the state vector $\mathbf{Y}(t)$ during the measure of the vector $\mathbf{Z}(t)$ in the interval (t_0, t). The matrix $\boldsymbol{\Gamma}$ in the terminal control under the bordering conditions (5) is the solution of the equation

$$\dot{\boldsymbol{\Gamma}} + \mathbf{A}^T\boldsymbol{\Gamma} + \boldsymbol{\Gamma}\mathbf{A} = -\mathbf{L} \tag{10}$$

and in the nonterminal control is a particular solution of the equation according to the right part $\mathbf{L}(t)$.

To prove the theorem we change the functional (4). For this purpose we are to differentiate the quadratic form $\mathbf{X}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t)$ by the time

$$\frac{d}{dt}[\mathbf{X}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t)] = \mathbf{X}^T(t)\,\dot{\boldsymbol{\Gamma}}(t)\,\mathbf{X}(t) + 2\dot{\mathbf{X}}^T\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t).$$

Integrating this relation in the interval (t_0, t_k) we get

$$\mathbf{X}^T(t_k)\,\boldsymbol{\Gamma}(t_k)\,\mathbf{X}(t_k) = \mathbf{X}^T(t_0)\,\boldsymbol{\Gamma}(t_0)\,\mathbf{X}(t_0) + \int_{t_0}^{t_k} (\mathbf{X}^T(t)\,\dot{\boldsymbol{\Gamma}}(t)\,\mathbf{X}(t) + 2\dot{\mathbf{X}}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t))\,dt.$$

Taking into consideration Eq. (8) and condition (5) we write the latter expression in the form

$$\mathbf{X}^T(t_k)\,\boldsymbol{\Gamma}_k\mathbf{X}(t_k) = \mathbf{X}^T(t_0)\,\boldsymbol{\Gamma}(t_0)\,\mathbf{X}(t_0) + \int_{t_0}^{t_k} (\mathbf{X}^T(t)\,\dot{\boldsymbol{\Gamma}}(t)\,\mathbf{X}(t) +$$

$$+ 2\mathbf{X}^T(t)\,\mathbf{A}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t) + 2\mathbf{u}^T(t)\,\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t) + 2\mathbf{f}^T(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, \mathbf{y}^*, t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t) +$$

$$+ 2\mathbf{V}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t))\,dt.$$

Substituting the obtained expression in the functional (4) and taking into account that the matrix $\mathbf{\Gamma}$ satisfies Eq. (10), we get

$$M[F] = M[\mathbf{X}^T(t_0)\,\mathbf{\Gamma}(t_0)\,\mathbf{X}(t_0)] + \int_{t_0}^{t_k} M[2\mathbf{u}^T(t)\,\mathbf{D}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{X}(t) +$$

$$+ \mathbf{X}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{D}(t)\,\mathbf{K}\mathbf{D}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{X}(t) + \mathbf{u}^T(t)\,\mathbf{K}^{-1}\mathbf{u}(t) + 2\mathbf{f}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{X}(t) +$$

$$+ 2\mathbf{V}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{X}(t)]\,dt. \tag{11}$$

Let us calculate the last member of the functional (11) using the dependence of the vector of error $\mathbf{X}(t)$ on the input signal in the form

$$\mathbf{X}(t) = \int_{t_0}^{t} \mathbf{G}(t,\,\tau)\;(\mathbf{D}(\tau)\,\mathbf{u}(\tau) + \mathbf{f}(\tau) + \mathbf{V}(\tau))\,d\tau,$$

where $\mathbf{g}(t,\,\tau)$ is the matrix of the weight functions of Eq. (8). We write the last member of the functional (11) in the following form

$$M[\mathbf{V}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{X}(t)] = \int_{t_0}^{t} M[\mathbf{V}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{G}(t,\,\tau)\;(\mathbf{D}(\tau)\,\mathbf{u}(\tau) + \mathbf{f}(\tau) + \mathbf{V}(\tau))]\,d\tau.$$

The integrant expression in the r.h.s. of the obtained equality is changed with the application of the following relations for the vectors $\mathbf{H}$ and $\mathbf{W}$:

$$[\mathbf{H}^T\,\mathbf{W}] = \mathrm{tr}\,[\mathbf{H}\mathbf{W}^T].$$

After this change we have

$$M[\mathbf{V}^T(t)\,\mathbf{\Gamma}(t)\,\mathbf{X}(t)] = \mathrm{tr} \int_{t_0}^{t} M[\mathbf{V}(t)\,\mathbf{u}^T(\tau)]\,\mathbf{D}^T(\tau)\,\mathbf{g}(t,\,\tau)\,\mathbf{\Gamma}(t)\,d\tau +$$

$$+ \int_{t_0}^{t} M[^T\mathbf{V}(t)]\,\mathbf{\Gamma}(t)\,\mathbf{g}(t,\,\tau)\,\mathbf{f}(\tau)\,d\tau + \mathrm{tr} \int_{t_0}^{t} M[\mathbf{V}(t)\,\mathbf{V}^T(\tau)]\,\mathbf{g}^T(t,\,\tau)\,\mathbf{\Gamma}(t)\,d\tau.$$

Taking into consideration that $M[\mathbf{V}(t)] = \mathbf{0}$, $M[\mathbf{V}(t)\,\mathbf{V}^T(\tau)] = \mathbf{G}(t)\;\delta(t-\tau)$ and $\mathbf{g}^T(t,\,t) = \mathbf{I}(t)$ is the unique matrix, we get

$$\int_{t_0}^{t} M[\mathbf{V}^T(t)]\,\mathbf{\Gamma}(t)\,\mathbf{g}(t,\,\tau)\,\mathbf{f}(\tau) = 0,$$

$$\mathrm{tr} \int_{t_0}^{t} M[\mathbf{V}(t)\;\mathbf{V}^T(\tau)]\;\mathbf{g}(t,\,\tau)\,\mathbf{\Gamma}(t)\,d\tau = \frac{1}{2}\,\mathrm{tr}[G(t)\,\mathbf{\Gamma}(t)].$$

Due to the independence between $\mathbf{V}(t)$ and $\mathbf{u}^T(\tau)$ for all τ in the interval $t_0 \leq$ $\leq \tau \leq t-\varepsilon$ for any small ε we have

$$\operatorname{tr} \int_{t_0}^{t} M[\mathbf{V}(t)\,\mathbf{u}^T(\tau)]\,\mathbf{D}^T(\tau)\,\mathbf{g}^T(t,\tau)\,\boldsymbol{\Gamma}(t)\,d\tau = \operatorname{tr} \int_{t_0}^{t-\varepsilon} M[\mathbf{V}(t)\,\mathbf{u}^T(\tau)]\,.$$

$$\cdot\,\mathbf{D}^T(\tau)\,\mathbf{g}^T(t,\tau)\,\boldsymbol{\Gamma}(t)\,d\tau + \operatorname{tr} \int_{t-\varepsilon}^{t} M[\mathbf{V}(t)\,\mathbf{u}^T(\tau)]\,\mathbf{D}^T(\tau)\,\mathbf{g}^T(t,\tau)\,\boldsymbol{\Gamma}(t)\,d\tau =$$

$$= \varepsilon \operatorname{tr}\left(M[\mathbf{V}(t)\,\mathbf{u}^T(t)]\,\mathbf{D}^T(t)\,\mathbf{g}^T(t,t)\,\boldsymbol{\Gamma}(t)\right).$$

The above expression tends to zero when $\varepsilon \to 0$ because the components of the matrix $M[\mathbf{V}(t)\,\mathbf{u}^T(t)]$ are limited. Therefore

$$M[\mathbf{V}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t)] = \frac{1}{2}\operatorname{tr}[\mathbf{G}(t)\,\boldsymbol{\Gamma}(t)].$$

Finally the functional (11) takes the form

$$M[F] = M[\mathbf{X}^T(t_0)\,\boldsymbol{\Gamma}(t_0)\,\mathbf{X}(t_0)] + \int_{t_0}^{t_k} M\,[2\mathbf{u}^T(t)\,\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t) + \mathbf{u}^T(t)\,\mathbf{K}^{-1}\mathbf{u}(t) + \cdot$$

$$+ \mathbf{X}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{D}(t)\,\mathrm{K}\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t)]\,dt + \int_{t_0}^{t_k} 2\mathbf{f}^T(t)\,\boldsymbol{\Gamma}(t)\,(\mathbf{m}_y^*(t) - \mathbf{y}^*(t))\,dt +$$

$$+ \int_{t_0}^{t_k} \operatorname{tr}\,[\mathbf{G}(t)\,\boldsymbol{\Gamma}(t)]\,dt\,, \qquad (12)$$

where we approximately assume $M[\mathbf{X}(t)] \approx \mathbf{m}_y^*(t) - \mathbf{y}^*(t)$ for the optimal control.

The first and the two last members of the positively definite functional (12) are independent from the vector $\mathbf{u}$. Therefore the vector of the optimal control $\mathbf{u}$ should be found with the help of the minimization of the parts depending on it

$$\min_{\mathbf{u}} \{ \int_{t_0}^{t_k} M[2\mathbf{u}^T(t)\,\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t) + \mathbf{u}^T(t)\,\mathrm{K}^{-1}\mathbf{u}(t) +$$

$$+ \mathbf{X}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{D}(t)\,\mathbf{K}\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t)]\,dt\}\,.$$

While measuring the vector $\mathbf{Z}(t)$ connected with the state vector $\mathbf{Y}(t)$ for the minimum of the mathematical expectation of any functional of the state vector it is enough to minimize the conditional mathematical expectation of the functional for the measured realization of the vector $\mathbf{Z}(\tau)$, $(t_0 \leq \tau \leq t)$ [5]

$$\min_{\mathbf{u}} \{M\,[\int_{t_0}^{t_k} M_z[2\mathbf{u}^T(t)\,\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t) + \mathbf{u}^T(t)\,\mathrm{K}^{-1}\mathbf{u}(t) +$$

$$+ \mathbf{X}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{D}(t)\,\mathbf{K}\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t)]\,dt]\}\,.$$

Applying the operation of the conditional mathematical expectation and taking

into consideration that for the given measured realization $\mathbf{Z}(\tau)$ the control vector $\mathbf{u}$ is the deterministic function, we write

$$\min_{\mathbf{u}} \{ \int_{t_0}^{t_k} M[2\mathbf{u}^T(t)\,\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\hat{\mathbf{X}}(t) + \hat{\mathbf{X}}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{D}(t)\,\mathbf{K}\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\,\mathbf{X}(t) + \quad (13)$$

$$+ \mathbf{u}^T(t)\,\mathbf{K}^{-1}\mathbf{u}(t)]\,dt + \int_{t_0}^{t_k} M\,[(\mathbf{X} - \hat{\mathbf{X}})^T\,\boldsymbol{\Gamma}(t)\,\mathbf{D}(t)\,\mathbf{K}\mathbf{D}^T(t)\,\boldsymbol{\Gamma}(t)\;(\mathbf{X} - \hat{\mathbf{X}})]\,dt\}\,.$$

Functional (13) is positively definite and contains two positively definite parts. It reaches its smallest value zero when the control vector $\mathbf{u}$ has the form (9) because with such a control the integral function of the first member in formula (13) is identical with zero. Thus the theorem is proved.

To form the control in accordance with formulae (9) one should determine the matrix $\boldsymbol{\Gamma}$ and the estimator vector $\hat{\mathbf{Y}}(t)$. The matrix $\boldsymbol{\Gamma}$ is determined from Eq. (10) which contains the matrix $\mathbf{A}(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t)$ depending on the unknown vector $\mathbf{m}_y^*$ and the matrix $\boldsymbol{\Theta}_y^*$. To determine the vector $\mathbf{m}_y^*$ it is necessary to apply the operation of absolute mathematical expectation using formulae (9) to Eq. (7)

$$\dot{\mathbf{m}}_y^* = \mathbf{A}(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t)\,\mathbf{m}_y^* - \mathbf{D}\mathbf{K}\mathbf{D}^T\,\boldsymbol{\Gamma}\mathbf{m}_y^* + \boldsymbol{\varphi}_0', \qquad \mathbf{m}_y^*(t_0) = \mathbf{m}_0\,. \quad (14)$$

The equation for the centralized vector $\mathbf{Y}_0$ we get by the subtraction of Eq. (7) from Eq. (14)

$$\dot{\mathbf{Y}}^0 = \mathbf{A}(\mathbf{m}_y^*, \boldsymbol{\Theta}_y^*, t)\,\mathbf{Y}^0 - \mathbf{D}\mathbf{K}\mathbf{D}^T\,\boldsymbol{\Gamma}\hat{\mathbf{Y}}^0 + \mathbf{V}, \qquad \mathbf{Y}^0(t_0) = \mathbf{Y}_0^0. \quad (15)$$

While determining the absolute probability characteristics of the vector $\mathbf{m}_y^*$ and the correlation matrix $\boldsymbol{\Theta}_y^*$ the vector $\mathbf{Z}(t)$ should be considered random. That is why the vector $\hat{\mathbf{Y}}_0$ is also random. Using (15) let us form the equation for the matrix $\boldsymbol{\Theta}_y^*$ [4]

$$\left.\begin{aligned} &\boldsymbol{\Theta}_y^* = \mathbf{A}\boldsymbol{\Theta}_y^* + \boldsymbol{\Theta}_y^*\mathbf{A}^T - \mathbf{D}\mathbf{K}\mathbf{D}^T\,\boldsymbol{\Gamma}\boldsymbol{\Theta}_{\hat{y}}^* - \boldsymbol{\Theta}_{\hat{y}}^*(\mathbf{D}\mathbf{K}\mathbf{D}^T\boldsymbol{\Gamma})^T + \mathbf{G}, \;\; \boldsymbol{\Theta}_y^*(t_0) = \boldsymbol{\Theta}_0, \\ &\boldsymbol{\Theta}_{\hat{y}}^* = \boldsymbol{\Theta}_y^* - \mathbf{R}, \end{aligned}\right\} \quad (16)$$

where

$$\boldsymbol{\Theta}_y^* = M[\mathbf{Y}^0(t)\,\mathbf{Y}_0(t)], \;\; \mathbf{R} = M[(\hat{\mathbf{Y}}(t) - \hat{\mathbf{Y}}(t))\;(\mathbf{Y}(t) - \hat{\mathbf{Y}}(t)^T)^T].$$

Using Eqs. (7) and (3) we get the equation of the generalized Kalman–Bucy filter for the estimator vector

$$\dot{\hat{\mathbf{Y}}} = \mathbf{A}\hat{\mathbf{Y}} - \mathbf{D}\mathbf{K}\mathbf{D}^T\,\mathbf{Y}\hat{\boldsymbol{\Gamma}} + \mathbf{R}\mathbf{C}^T\mathbf{Q}^{-1}(\mathbf{Z} - \mathbf{C}\hat{\mathbf{Y}}) + \mathbf{u}_0 + \boldsymbol{\varphi}_0', \; \hat{\mathbf{Y}}(t_0) = \mathbf{m}_0\,, \quad (17)$$

$$\dot{\mathbf{R}} = \mathbf{A}\mathbf{R} + \mathbf{R}\mathbf{A}^T - \mathbf{R}\mathbf{C}^T\mathbf{Q}^{-1}\mathbf{C}\mathbf{R}, \quad \mathbf{R}(t_0) = \boldsymbol{\Theta}_0\,. \quad (18)$$

As a result of the cointegration of Eqs. (10), (14), (16) and (18) we get the matrix $\mathbf{\Gamma}$. The solution of this problem is difficult because the initial and terminal conditions are combined. The proper way of their solution is the method of consequent tests. The initial condition for $\mathbf{\Gamma}(t_0)$ in the moment $t = t_0$ will be taken arbitrarily. Then to carry out the given terminal condition one should take the initial condition for $\mathbf{\Gamma}(t_0)$. In the solution of this problem for the nonlinear subject there arise no restrictions for its stability. For the nonterminal task when the subject is stable the matrix $\mathbf{\Gamma}$ will be determined as a particular solution of Eq. (10) [2]. Here the system of Eqs. (10), (14), (16) and (18) should be solved under the following initial conditions:

$$\mathbf{\Gamma}(t_0) = 0, \ \mathbf{m}_y^*(t_0) = \mathbf{m}_0, \ \mathbf{\Theta}_y^*(t_0) = \mathbf{\Theta}_0, \ \mathbf{R}(t_0) = \mathbf{\Theta}_0 .$$

If the subject of control is stationary and stable then for the established regime for the infinite time of the observation and the control the abovementioned equations are turned into nondifferential ones of the following type:

$$\left.\begin{aligned}
&\boldsymbol{\varphi}_0(\mathbf{m}_y^*, \mathbf{\Theta}_y^*) - \mathbf{DKD}^T \mathbf{\Gamma m}_y^* = 0, \\
&\mathbf{A}^T(\mathbf{m}_y^*, \mathbf{\Theta}_y^*)\mathbf{\Gamma} + \mathbf{\Gamma A}(\mathbf{m}_y^*, \mathbf{\Theta}_y^*) = -\mathbf{L}, \\
&[\mathbf{A}(\mathbf{m}_y, \mathbf{\Theta}_y^*) - \mathbf{DKD}^T \mathbf{\Gamma}]\mathbf{\Theta}_y^* + \mathbf{\Theta}_y^*[\mathbf{A}(\mathbf{m}_y^*, \mathbf{\Theta}_y^*) - \mathbf{DKD}^T \mathbf{\Gamma}]^T + \\
&+ \mathbf{DKD}^T \mathbf{\Gamma R} + \mathbf{R}(\mathbf{DKD}^T \mathbf{\Gamma})^T = -\mathbf{G}, \\
&\mathbf{A}(\mathbf{m}_y^*, \mathbf{\Theta}_y^*)\mathbf{R} + \mathbf{RA}^T(\mathbf{m}_y^*, \mathbf{\Theta}_y^*) - \mathbf{RC}^T \mathbf{Q}^{-1} \mathbf{CR} = -\mathbf{G}.
\end{aligned}\right\} \tag{19}$$

The proved method of the solution of Eqs. (10), (14), (16) and (18) for the nonterminal control can be used also for the unstable subject if we preliminarily change the phase coordinates $Y_i(t)$ in accordance with the recommendations [2]

$$\hat{Y}_i = \chi(t) Y_i',$$

where $\chi(t) = \exp[\lambda t]$, $\lambda > 0$ and is greater then the characteristic value of the equation of the control subject.

As a result of the solution of Eqs. (10), (14), (16), (18) or (19) we can determine the matrices $\mathbf{A}$ and $\mathbf{\Gamma}$ in the linearized equation (7). Equation (17) is the equation of the filter for obtaining of the unbiased estimator vector $\hat{\mathbf{Y}}$ of the state vector $\mathbf{Y}$ by measuring the vector $\mathbf{Z}$.

Examples

Example 1. Nonlinear stable subject is given by the equation of the first order

$$\hat{Y} = \boldsymbol{\varphi}(Y, t) + du + V, \ Y(t_0) = Y_0,$$

where $V(t)$ is the white noise of the intensivenes G and $m_\gamma = 0$. The value Z connected with Y by the dependence of the type (3) is observed

$$Z = Y + N,$$

where $N(t)$ is the white noise of the intensiveness Q and $m_N = 0$. It is necessary to determine the quasioptimal terminal control $\mathbf{u}$ based on the quadratic criteria of the generalized work (4).

To solve this problem let us apply the statistical linearization of nonlinearity

$$\varphi(Y, t) = k_1(m_y^*, \Theta_y^*, t)\, Y + \varphi_0'.$$

In this case the quasioptimal control according to formulae (9) will be determined by the formulae

$$u = -kd\gamma_1 \hat{Y},$$

where γ_1 is determined from the equation of the type (11)

$$\dot{\gamma}_1 + 2k_1(m_y^*, \Theta_y^*, t)\, \gamma_1 + l = 0, \quad \gamma_1(t_k) = \gamma_k.$$

This equation needs to add the equations for m_y^*, Θ_y^*, R

$$m_y^* = \varphi_0(m_y^*, \Theta_y^*, t) - d^2k\gamma_1 m_y^*, \quad m_y^*(t_0) = m_0,$$

$$\Theta_y^* = 2(k_1(m_y^*, \Theta_y^*, t) - d^2k\gamma_1)\, \Theta_y^* + 2d^2k\gamma_1 R + G, \quad \Theta_y^*(t_0) = \Theta_0,$$

$$R = 2k_1(m_y^*, \theta_y^*, t)\, R - \frac{1}{Q} R^2 + G, \qquad R(t_0) = \theta_0.$$

Solving this equations, we find m_y^*, Θ_y^*, γ_1, k_1, R. The conditional estimator value Y of the value $\hat{Y}$ is determined by the equation of the filter

$$\dot{\hat{Y}} = (k_1 - d^2k\gamma_1)\, Y + \frac{R}{Q}(Z - \hat{Y}) + \varphi_0', \qquad \hat{Y}(t_0) = m_0.$$

Example 2. For the subject of the control discussed in Example 1 let us determine the quasioptimal nonterminal control for $m_0 = 0$ and $\varphi(Y) = -\operatorname{sgn} Y$.

In this case for the quasioptimal control we have

$$u = -kd\gamma_1 \hat{Y}$$

where γ_1 is determined as a particular solution of the equation

$$\dot{\gamma}_1 - 2k_1(\Theta_y^*)\, \gamma_1 = -l, \quad \gamma(t_0) = 0.$$

The solution of this equation has the following form:

$$\gamma_1 = l \mid 2k_1(\Theta_y^*).$$

The values Θ_y^*, R are determined from the equations

$$\dot{\Theta}_y^* = -2[k_1(\Theta_y^*) + d^2 k\gamma_1]\Theta_y^* + 2k\gamma_1 R + G, \; \Theta_y^*(t_0) = \Theta_0,$$

$$R = -2k_1(\theta_y^*)\, R - \frac{1}{Q} R^2 + G, \qquad R(t_0) = \theta_0,$$

$$k_1 = /2\sqrt{2\pi\theta_y^*}.$$

As a result for the control we have

$$u = -\frac{kdl}{2k_1(\Theta_y^*)}.$$

The estimation $\hat{Y}$ can be determined by the equation of the filter discussed in Example 1

$$\dot{\hat{Y}} = -\left(k_1 + d^2 \frac{kl}{2k_1}\right)\hat{Y} + \frac{R}{Q}(Z - \hat{Y}) \; . \qquad Y(t_0) = 0.$$

In Fig. 1 are shown the dependences $\Theta_y^*(t)$ for the following control and conditions:

$$k = 1, \; l = 1, \; d = 1, \; Q = 1, \; G = 1, \; \Theta_0 = 1,$$

1 — nonoptimal control for non-exact measurement

$$\theta_y^* = -\left(\frac{1{,}6}{\sqrt{\theta_y^*}} + 2\right)\theta_y^* + 2, \qquad \theta_y^*(t_0) = 1,$$

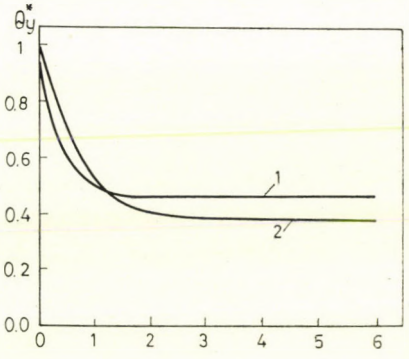


Fig. 1

2 — optimal control for non-exact measurement

$$\dot{\theta}_y^* = -\left(\frac{1{,}6}{\sqrt{\theta_y^*}} + 1{,}25\sqrt{\theta_y^*}\right)\theta_y^* + 1{,}25\sqrt{\theta_y^*}\,R + 1\,, \qquad \theta_y^*(t_0) = 1,$$

$$R = -\frac{1{,}6}{\sqrt{\theta_y^*}}\,R - R^2 + 1\,, \qquad R(t_0) = 1.$$

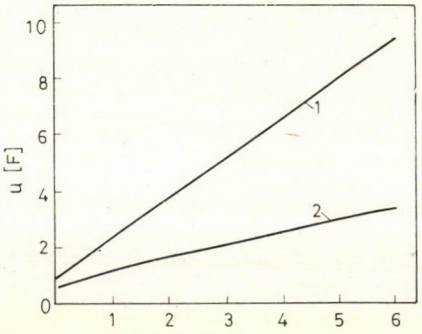


Fig. 2

In Fig. 2 are shown the value functional (4) for Example 2 and the following conditions: $t_0 = 0$, $t_k = t$, 1 — nonoptimal control

$$M[F] = \Theta_y^*(t) + 3\int_0^t \Theta_y^*(\tau)\,d\tau,$$

2 — optimal control

$$M[F] = 0.63\sqrt{\Theta_y^*(t)}\,\Theta_y^*(t) + \int_0^t [1 + 0.79\Theta_y^*(\tau) - 0.39R(\tau)]\,\Theta_y^*(\tau)\,\mathrm{d}\tau.$$

This example shows that with the help of chosen parameters of the functional we can get the best quality of the control.

References

1. *Krasovskii, A. A.*, Analytical design of the contours of flying machine control. Mashinostroenie, 1969.
2. *Krasovskii, A. A.*, Automatic control systems of flight and their analytical design. Nauka, 1973.
3. *Kazakov, I. E.*, Statistical synthesis of the control based on the quadratic criteria of the generalized work. Automatica and Telemekhanica. **10,** 1974.
4. *Kazakov, I. E.*, Statistical methods in control systems design. Mashinostroenie, 1969.
5. *Pugachev, V. S.*, Theory of random functions and application to control problems, 3rd edition, Fizmatgiz, 1962.

Аналитический синтез статистически квазиоптимального управления нелинейным объектом

И. Е. КАЗАКОВ

(Москва)

Предлагается процедура аналитического синтеза квазиоптимального управления при неточных измерениях вектора состояния и случайных возмущениях, действующих на нелинейный объект. Нелинейный объект описывается векторным уравнением (1). На интервале (t_0, t_k) наблюдается вектор $\mathbf{Z}$, связанный с вектором состояния уравнением (3). Требуется определить вектор квазиоптимального управления u на основе минимизации квадратического функционала обобщенной работы (4) относительно ошибки $\mathbf{X}(t) = \mathbf{Y}(t) - \mathbf{y}^*(t)$, где $\mathbf{Y}(t)$ — фактический, а $\mathbf{y}^*(t)$ — требуемый векторы состояния объекта. Представляют интерес терминальная (интервал $t_k - t_0$ задан) и нетерминальная (момент времени $t_k = t$ произволен) задачи.

Квазиоптимальный вектор управления определяется для статистически линеаризованного уравнения объекта (7) на основании следующей теоремы. Для нелинейного объекта (1), приближенно описываемого уравнением (7) при измерении вектора (2) оптимальным в смысле минимума квадратического функционала (4) является вектор управления вида (9)

$$\mathbf{u} = -\mathbf{K}\mathbf{X}^T \boldsymbol{\Gamma}(\hat{\mathbf{Y}}(t) - \mathbf{y}^*(t)),$$

где $\mathbf{Y}(t)$ — оценка вектора состояния при измерении вектора $\mathbf{Z}(\tau)$ на интервале (t_0, t) $y^*(t)$ — требуемый вектор состояния, определяемый уравнением (2), $\mathbf{K}$ и $\mathbf{D}$ — заданные матрицы. Матрица $\boldsymbol{\Gamma}$ при терминальном управлении является решением уравнения (10) при граничном условии (5), а при нетерминальном управлении — частным решением этого уравнения, соответствующим правой части.

Доказательство теоремы основывается на преобразовании функционала к виду, для которого непосредственной подстановкой выражения (9) показывается его минимальность и, следовательно, оптимальность управления.

Для формирования управления в соответствии с формулой (9) надо определить матрицу $\boldsymbol{\Gamma}$ и вектор оценки $\hat{\mathbf{Y}}(t)$. Матрица $\boldsymbol{\Gamma}$ может быть получена в результате совместного решения уравнений (10), (14), (16) (18). Сложность решения этой задачи при терминальном управлении состоит в наличии смешанных начальных и конечных условий в уравнениях. Стандартным приемом их решения является метод проб. Для нетерминальной задачи при устойчивом объекте матрица $\boldsymbol{\Gamma}$ определяется как решение тех же уравнений (10), (14), (16) (18), но в уравнении (10) следует принять нулевое начальное условие. В результате решения этих уравнений определяются вектор $\mathbf{m}_y^*$, матрицы θ_y^*, $\boldsymbol{\Gamma}$, а также матрица A, составленная из статистических коэффициентов усиления нелинейностей.

Для приближенной оценки вектора состояния $\hat{\mathbf{Y}}(t)$ можно применить обобщенный фильтр Калмана—Бьюси (17), который получается на основании использования линеаризованного уравнения (7).

И. Е. Казаков
Научный совет по технической кибернетике
СССР, Москва В-333,
ул. Вавилова, 40

Problems of Control and Information Theory, Vol. 5 (1), pp. 63—70 (1976)

MARKOV-PARAMETER REPRESENTATION OF LINEAR TIME-VARIABLE SYSTEMS WITH SEPARABLE SYSTEM FUNCTIONS

H. SCHWARZ

(Duisburg)

(Received March 17, 1975)

The well-known system representation of linear time-invariant systems by Markov parameters which has very successfully proved for digital computer applications is extended to a particular class of time-variable linear systems. The approach is based on the concept of separable system functions.

1. Introduction

The concept of system representation of linear time-invariant systems was published first by Ho [1, 2] in 1966. It is based on the fact that the functions relating to the input-output behaviour of the system — the weighting function in the time domain and the Laplace transformed weighting function in the frequency domain — are expanded in infinite power series respectively. In the time domain as well as in the domain of the Laplace transform the very same sequences of coefficients will be found, which are called Markov parameters [6]. By means of these coefficients one may construct purely algebraic models for describing the dynamic behaviour of linear systems, which are particularly well suitable for programming of digital computers. In recent years the Ho-algorithm of minimal realization, i.e. constructing of state space representations from input-output data of a linear system, based on the Markov-parameter representation was further studied and refined [3]. Further the efficiency for practical applications in control engineering for problems of adaptive control was established [4].

These experiences gave motivations for investigations in order to extend the concept of the Markov-parameter representation to an enlarged class of linear systems. The subject of this paper is to show that this extension is possible for the class of time-variable systems having separable system functions. In Section 2 some known essential results of the Markov-parameter representation are collected. Section 3 is devoted to the description of the concept of transfer functions for time-variable systems due to Zadeh. The main result of this paper, the Markov-parameter representation of time-variable systems with separable system, functions is given in Section 4. The conclusion in Section 5 points out some directions of further research work.

2. Markov parameter for time-invariant systems

We consider first a time-invariant system with one input and one output variable:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t), \quad \mathbf{x}_0 = \mathbf{x}(0) \tag{1}$$

$$y(t) = \mathbf{c}^T\mathbf{x}(t)$$

with $\mathbf{x} \in R_n$ (or R),the n, n system matrix $\mathbf{A}$, the n, 1 iput matrix $\mathbf{b}$ and the 1, n output matrix $\mathbf{c}^T$ respectively. The system (1) which is assumed to be completely controllable and observable has the transfer function:

$$G(s) = \mathbf{c}^T[\mathbf{1}s - \mathbf{A}]^{-1}\mathbf{b} = \frac{\sum_{i=1}^{n} b_i s^{n-i}}{s^n + \sum_{j=1}^{n} a_j s^{n-j}}, \tag{2}$$

where some or all b_i with exception of b_n may equal zero. In [1] or [6] one may find this important result:

Theorem 1. Every rational function given by (2) can be expanded into a power series representation

$$G(s) = \frac{M_1}{s} + \frac{M_2}{s^2} + \frac{M_3}{s^3} + \ldots = \sum_{i=1}^{\infty} M_i s^{-1}, \tag{3}$$

where the so-called Markov parameters M_i satisfy a recurrence relation:

$$M_i = \sum_{\nu=1}^{n} a_\nu M_{i-\nu}, \qquad \forall i \geq n \tag{4}$$

with the coefficients a_j of the denominator polynomial in (2).

If (3) is retransformed term by term into the time domain one yields a power series expansion of the related weighting function:

$$g(t) = M_1 + M_2 t + \frac{M_3 t^1}{2!} + \frac{M_4 t^4}{3!} + \ldots = \sum_{i=1}^{\infty} M_i \frac{t^{i-1}}{(i-1)!}, \qquad t \geq 0\,. \tag{5}$$

The solution of the state-space equation (1) for $\mathbf{x}_0 = \mathbf{0}$ enables a more detailed representation of the systems weighting function:

$$g(t) = \mathbf{c}^T e^{\mathbf{A}t}\mathbf{b} = \sum_{i=1}^{\infty} \mathbf{c}^T \mathbf{A}^{i-1}\mathbf{b} \frac{t^{i-1}}{(i-1)!}, \qquad t \geq 0 \tag{6}$$

and then compare the coefficients in (5) and (6) to obtain a relation between the Markov parameters M_i and the system realization $\{\mathbf{c}^T, \mathbf{A}, \mathbf{b}\}$:

$$M_i = \mathbf{c}^T\mathbf{A}^{i-1}\mathbf{b} \qquad i = 1, 2, \ldots\,. \tag{7}$$

Under the assumption, that the input signals $u(t)$ und the output signals $y(t)$ are defined for $t \in (0^+, \infty)$ and are arbitrary often differentiable we can expand these signals into Taylor series for $t = 0$:

$$\begin{aligned} u(t) &= u_0\delta(t) + u_1 + u_2 t + \ldots + u_i \frac{t^{i-1}}{(i-1)!} + \ldots \\ y(t) &= \quad\; y_1 + y_2 t + \ldots + y_i \frac{t^{i-1}}{(i-1)!} + \ldots, \end{aligned} \tag{8}$$

where $\delta(t)$ is the Dirac delta.

By means of the coefficients in (8) and the Markov parameters in (7) the behaviour of the signal transfer of (1) may be represented by a purely algebraic system model [1, 2, 3]:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \end{bmatrix} = \begin{bmatrix} M_1 & 0 & 0 & \ldots \\ M_2 & M_1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \\ M_i & M_{i-1} & M_{i-2} \ldots & \\ \vdots & \vdots & \vdots & \end{bmatrix} \cdot \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{i-1} \\ \vdots \end{bmatrix} + \begin{bmatrix} M_1 & M_2 & M_3 \ldots M_m \\ M_2 & M_3 & M_4 \ldots M_{m+1} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix} \beta . \tag{9}$$

In (9) the first term of the right side gives the relation between the input and output signals while the second term is related to the initial conditions for $t = 0$. If $\mathbf{x}_0 = \mathbf{0}$ the infinite dimensional vector β is zero, too. The continuous-time system (1) may be represented by an equivalent discrete-time system for which an algebraic system corresponding to (9) exists. These models then are extremely well suitable for "one-line" identification of an unknown system by its input-output data [3, 4].

3. Transfer functions for time-variable systems

Now a time-variable system with one input and one output variable is given:

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}(t)\,\mathbf{x}(t) + \mathbf{b}(t)\,u(t), \quad \mathbf{x}_0 = \mathbf{x}(t_0) \\ y(t) &= \mathbf{c}^T(t)\,\mathbf{x}(t), \end{aligned} \tag{10}$$

without loss of generality we may assume that the system matrix $\mathbf{A}(t)$ has phase-variable canonical form, so (10) may be written in a more detailed form:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \vdots \\ \dot{x}_n(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ -\alpha_n(t) & \cdot & \cdot & \cdot & & -\alpha_1(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ b(t) \end{bmatrix} u(t) \quad (10a)$$

$$y(t) = [c(t),\ 0 \dots 0]\,\mathbf{x}(t).$$

The homogeneous differential equation (10) with $u(t) = 0$ possesses the fundamental matrix:

$$\mathbf{\Phi}(t, \tau) = \mathbf{W}(t)\,\mathbf{W}^{-1}(\tau) \tag{11}$$

where $\mathbf{W}(t)$ is the Wronskian formed from n-linear independent solutions $\mathbf{x}_i(t)$ of the homogeneous equation (10). The impulse response $g(t, \tau)$ for a Dirac-delta impulse $u(t) = \delta(t - \tau)$ at time $t = \tau$ when the system has initial conditions $\mathbf{x}_0 = \mathbf{x}(\tau) = 0$ is given by [7, 8]:

$$g(t, \tau) = \mathbf{c}^T(t)\,\mathbf{\Phi}(t, \tau)\,\mathbf{b}(t) = \mathbf{c}^T(t)\,\mathbf{W}(t)\,\mathbf{W}^{-1}(\tau)\,\mathbf{b}(\tau) = \sum_{i=1}^{n} p_i(t)\,q_i(\tau). \tag{12}$$

The $p_i(t)$ are n-linear independent basis functions of this homogeneous differential equation related to (10):

$$L\big(y(t)\big) = a_0(t)\frac{d^n}{dt^n}y(t) + \dots + a_{n-1}(t)\frac{d}{dt}y(t) + a_n(t)\,y(t) = 0, \tag{13}$$

while the $q_i(t)$ are the basis functions of the adjoint homogeneous differential equation:

$$L^*\big(y^*(t)\big) = (-1)^n\frac{d^n}{dt^n}[a_0(t)\,y^*(t)] + \dots$$

It is proved [7, 8] that all impulse–response functions $g(t, \tau)$ of linear systems are separable according to (12), if these impulse response is due to a system with n-concentrated energy storages.

Zadeh [9] proposed a particular Laplace transform for weighting functions of time-variable systems called "system function":

$$G(s, \tau) = \int_{-\infty}^{t} g(t, \tau)\,e^{-s(t-\tau)}\,d\tau \tag{14}$$

which has for many problems of system analysis and synthesis analogous properties to the well-known transfer function $G(s)$ in (2) for time-invariant systems.

4. Markov parameters for separable system functions

A system function $G(s, \tau)$ is said to be separable if:

$$G(s, \tau) = \sum_{i=1}^{m} m_i(\tau)\, H_i(s). \tag{15}$$

Substituting the weighting function of the time-variable system (12) into (14), conditions may be found for the property of separability. It is proved [10] that for this the functions $q_i(\tau)$ in (12) must be linear combinations of exponentials:

$$q_i(\tau) = \sum_{k=1}^{m} d_{ik}\, \tau^{l_i}\, e^{\alpha_{ik}\tau}. \tag{16}$$

When this equation is combined with (12) and (14) and using the substitution $t - \tau = \eta$ we find:

$$G(s, t) = \sum_{i=1}^{n} p_i(t) \sum_{k=1}^{m} d_{ik}\, t^{l_i}\, e^{\alpha_{ik}t}\, \frac{1}{(s + \alpha_{ik})^{l_i}}; \tag{17}$$

and the related weighting function by Laplace transform into the time domain:

$$g(t, \tau) = \sum_{i=1}^{n} p_i(t) \sum_{k=1}^{m} d_{ik}\, t^{l_i}\, e^{\alpha_{ik}\tau}, \quad \forall t \geq \tau\,. \tag{18}$$

Comparison of (17) and (15) shows, that the $H_i(s)$ of the separable system functions are rational functions of the Laplace variable s. Systems with separable system functions therefore can be synthesized by appropriately interconnecting a linear time-invariant system with time-varying amplifiers (Fig. 1).

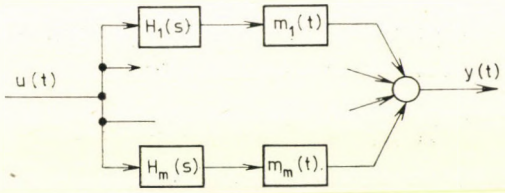


Fig. 1. Block-diagram representation of linear time-variable systems with separable system functions

The above considerations lead together with Theorem 1 to our result:

Theorem 2. Time-variable systems with separable system functions have Markov-parameter representations:

$$G(s, \tau) = \sum_{i=1}^{\infty} M_i(t)\, s^{-i} \tag{19}$$

and there is a polynomial

$$P(s) = s^{\pi} + P_1 s^{\pi-1} + \ldots + P_{\pi-1} s^1 + P_{\pi} \tag{20}$$

with coefficients p_i which give rise to the recurrence formula:

$$M_i = \sum_{\nu=1}^{\pi} P_\nu M_{i-\nu} \quad \forall_i \geq \pi. \tag{21}$$

Two essential items have to be pointed out for applications. First the system matrix **A** of time-variable systems with separable system functions are time-invariant and the elements of the input-matrix $\mathbf{b}(t)$ have to be linear combinations of exponentials according to (16). Further in general the polynomial $P(s)$ in (20) does not have the same degree as the characteristic polynomial of the (constant) system matrix **A** but often it will have a higher degree. The roots of $\det [\mathbf{1}s - A] = 0$ and those of $p(s)$ are different in general.

Example (a). The system with one input and one output

$$\dot{x}(t) = \lambda x(t) + te^{\beta t} u(t)$$

$$y(t) = c(t)x(t)$$

has the weighting function

$$g(t, \tau) = c(t)\, e^{\lambda(t-\tau)}\, \tau e^{\beta\tau} = c(t)\, e^{\lambda t} e^{-\lambda t}\, \tau e^{\beta\tau} = p(t)\, q(\tau), \qquad t \geq \tau.$$

With the substitution $t - \tau = \eta$ we get from (14):

$$G(s, t) = c(t) e^{\beta t} \left[\frac{t}{s - \lambda + \beta} - \frac{1}{(s - \lambda + \beta)^2} \right].$$

With the polynomial $p(s) = s^2 + p_1 s + p_2$; $p_1 = 2(\lambda - \beta)$; $p_2 = (\lambda + \beta)^2$ we find the Markov-parameter expansion of $G(s, t)$

$$G(s, t) = \frac{M_1(t)}{s} + \frac{M_2(t)}{s^2} + \ldots$$

$$M_1(t) = tc(t)\, e^{\beta t}; \quad M_2(t) = c(t)\, e^{\beta t} \quad [-1 - t(\lambda - \beta) - p_1 M(t);$$

$$M_3(t) = -p_1 M_2(t) - p_2 M_1(t), \ldots$$

Example (b). The system

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} e^{\beta_1 t} \\ e^{\beta_2 t} \end{bmatrix} u(t)$$

$$y(t) = [c_1(t),\ c_2(t)]\ \mathbf{x}(t)$$

has its weighting function:

$$g(t, \tau) = c_1(t)\, e^{\beta_1 \tau}\, e^{-\lambda_1 (t-\tau)} + c_2(t) e^{\beta_2 t}\, e^{-\lambda_2 (t-\tau)}$$

and its Markov-parameter representation is:

$$G(s, t) = \sum_{i=1}^{\infty} M_i s^{-1}$$

with

$$M_1(t) = c_1(t)\, e^{\beta_1 t} + c_2(t)\, e^{\beta_2 t}$$

$$M_2(t) = (\lambda_2 - \beta_2)\, c_1(t)\, e^{-\beta_1 t} + (\lambda_1 - \beta_1)\, c_2(t)\, e^{-\beta_2 t} - p_1 M_1(t);$$

$$M_3(t) = -p_2 M_1(t) - p_1 M_2(t);\ M_4 = \ldots$$

with $p_1 = \beta_1 + \beta_2 - \lambda_1 - \lambda_2;\ p_2 = (\lambda_1 - \beta_1)\ (\lambda_2 - \beta_2).$

5. Conclusion

In this paper it is shown, that the Markov-parameter representation of linear time-invariant systems proposed by Ho may be extended to a particular class of time-variable systems having separable system functions. Due to lack of space we considered single-variable systems only but the above-developed result is assignable to multivariable systems if the results of Gerth [3] are taken into account. An other research objective is to extend the results to establish a realization algorithm capable for the identification problem for time-variable systems.

*

Acknowledgement. The author would like to thank Prof. M. Thoma for some clarifying discussions.

References

1. *Ho, B. L.*, On Effective Construction of Linear State-variable Models from Input-Output Descriptions. Diss. Stanford, 1966.
2. *Schwarz, H.*, Mehrfachregelungen. Vol. 2. Berlin, Springer 1971.
3. *Gerth, W.*, Zur Identifikation und Minimalrealisierung von Mehrgrössenübertragungssystemen durch Markovparameter. Diss. TU Hannover, 1972.
4. *Hartmann, U.*, Ein Beitrag zum Entwurf digitaler, selbstadaptiver Flugregelsysteme. Diss. TU Hannover, 1974.
5. *Kaerger, W.*, Markovparameter linearer zeitvariabler Systeme. Dipl.-Arbeit, Inst. f. Regelungstechnik, TU Hannover, 1974.
6. *Gantmacher, F. R.*, Matrizenrechnung. Berlin, VEB Deutscher Verlag der Wissenschaften, 1958.
7. *D'angelo, H.*, Linear Time-Varying Systems: Analysis and Synthesis. Boston, Allyn and Bacon, 1970.

8. *Kaplan, W.*, Operational Methods and Linear Systems. Reading, Massachusetts, Addison—Wesley, 1962.
9. *Zadeh, L.*, Frequency Analysis of Variable Networks, IRE Proc. **38** (1950), pp. 291—299.
10. *Weiss, L.*, On System Functions with the Property of Separability, Int. J. Control **1** (1965), pp. 487—496.

Параметрическое марковское представление сигналов линейных нестационарных систем с сепарируемой системной функцией

Х. ШВАРЦ

(Дуисбург)

Приведены понятия сепарируемых системных функций. Обобщается системное описание линейных инвариантных систем, которое оказалось весьма эффективным при использовании ЭВМ в случае исследования автоматической системы с времязависящими параметрами одного частного класса. Методика расчета основана на применении сепарируемых системных функций. Показано, что предложенное приближение осуществимо в том случае, когда имеется возможность сконструировать сепарируемые системные функции.

Рассмотрены некоторые известные основные методы исследования таких систем на основе марковского представления.

Приведены передаточные функции систем автоматического управления на основе подхода Заде.

Главные результаты относительно сепарируемых системных функций приведены в заключительной части работы и указаны некоторые возможности дальнейших исследований.

H. Schwarz
Institute of Control Engineering
University of Duisburg
4100 Duisburg
Bismarckstr. 81
FRG

Problems of Control and Information Theory, Vol. 5 (1), pp. 71–85 (1976)

ФУНДАМЕНТАЛЬНЫЕ ФИНИТНЫЕ УПРАВЛЕНИЯ

А. Г. БУТКОВСКИЙ, М. И. МУСТАФАЕВ

(Москва, Баку)

(Поступила в редакцию 15 октября 1974 г.)

В работе вводится понятие фундаментального финитного управления (а также некоторые его модификации) для линейных управляемых систем с распределенными параметрами и устанавливаются аналитические формулы для представления финитных управлений соответствующего класса через фундаментальные финитные управления. Приводятся также формулы пересчета, связывающие различные модификации фундаментальных финитных управлений.

В теории линейных уравнений и, в частности, в теории линейных дифференциальных уравнений существует важное понятие *фундаментального решения* [1]. Грубо говоря, фундаментальным решением является такое решение, через которое может быть выражено любое другое решение уравнения с помощью некоторой линейной операции типа взятия линейной комбинации, свертки и т. д.

Например, рассмотрим линейное операторное уравнение

$$l(Q(x,\ t)) = f(x,t)\,, \tag{1}$$

где оператор l определяется некоторой задачей математической физики вида:

$$\sum_{k=0}^{m}\sum_{s=0}^{n} A_{ks}(x,t)\frac{\partial^{k+s}Q}{\partial t^k\,\partial x^s} = f(x,t),\quad x_0 < x < x_1,\ t > t_0\,, \tag{2}$$

с однородными граничными условиями:

1) однородными краевыми условиями

$$\sum_{i=0}^{1}\sum_{k=0}^{m-1}\sum_{s=0}^{n-1} a_{\nu iks}(t)\frac{\partial^{k+s}Q}{\partial t^k\,\partial x^s}\bigg|_{x=x_i} = 0,\quad t > t_0,\quad \nu = 1, 2, \ldots \leq n\,, \tag{3}$$

2) нулевыми начальными условиями

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = 0,\quad x_0 < x < x_1, k = 0, 1, \ldots, m-1\,. \tag{4}$$

Заметим, что неоднородные граничные условия, вообще говоря, сводятся к однородным путем соответствующего изменения функции $f(x, t)$ (см. п. 2).

Фундаментальное решение $\mathcal{C}(x, t; \xi, \tau)$ оператора l—это решение операторного уравнения (1), когда правая часть, т. е. функция $f(x, t)$, является δ-функцией, т. е.

$$f(x, t) = \delta(x - \xi)\ \delta(t - \tau). \tag{5}$$

Зная фундаментальное решение $\mathcal{C}(x, t; \xi, \tau)$ оператора l, легко получить его любое другое решение, соответствующее произвольной правой части $f(x, t)$ путем применения интегрального оператора с ядром $\mathcal{C}(x, t; \xi, \tau)$ [1]:

$$Q(x, t) = \int_{t_0}^{t} \int_{x_0}^{x_1} \mathcal{C}(x, t; \xi, \tau)\, f(\xi, \tau)\, d\xi\, d\tau \equiv l^{-1}\left(f(x, t)\right). \tag{6}$$

Фундаментальное решение $\mathcal{C}(x, t; \xi, \tau)$ часто называют функцией Грина оператора l, а возмущение вида (5) назовем *гриновским возмущением*.

Аналогичная ситуация имеет место, когда решается задача управляемости для некоторой линейной управляемой системы с сосредоточенными, дискретными и распределенными параметрами. По аналогии с понятием фундаментального решения здесь вводится понятие *фундаментального финитного управления* (ф. ф. у.) для линейных систем.

Под финитным управлением системы понимается перевод ее из данного произвольного состояния в состояние равновесия за заданное фиксированное время [2], [3].

Фундаментальным финитным управлением (ф. ф. у.) будем называть такое финитное управление, которое за фиксированное время $T = t_1 - t_0$ гасит *гриновское возмущение* (5).

Знание ф. ф. у. позволяет также легко получить любые другие управления, соответствующие произвольным возмущениям $f(x, t)$, путем применения некоторой простой операции композиции.

Перейдем к точным формулировкам.

Рассмотрим, например, линейные системы с распределенными параметрами. Аналогичное рассмотрение возможно также для линейных сосредоточенных, непрерывных и дискретных систем.

Итак, пусть состояние линейной управляемой системы в момент времени $t \geq t_0$ описывается уравнением

$$\sum_{k=0}^{m} \sum_{s=0}^{n} A_{vks}(x, t) \frac{\partial^{k+s} Q}{\partial t^k\, \partial x^s} = f(x, t) + B(x, t)\, U(x, t), \quad x_0 < x < x_1 \quad t > t_0, \tag{7}$$

с однородными граничными условиями:

1) однородными краевыми условиями

$$\sum_{i=0}^{1}\sum_{k=0}^{m-1}\sum_{s=0}^{n-1} a_{viks}(t)\left.\frac{\partial^{k+s}Q}{\partial t^k \partial x^s}\right|_{x=x_i} = 0, \quad t > t_0, \quad V = 1, 2, \ldots \leq n; \tag{8}$$

2) нулевыми начальными условиями

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m-1, \tag{9}$$

где функция $Q(x, t)$ характеризует состояние системы в точке x в момент времени t; $f(x, t)$ — возмущающий член; $B(x, t)$ — известная вектор-функция; $U(x, t)$ — векторное управляющее воздействие; а $B(x, t)\,U(x, t)$ — скалярное произведение.

Задача финитного управления состоит в нахождении такого векторного управления $U(x, t)$, при котором в некоторый момент времени $t_1 = t_0 + T$, $(T > 0)$ система переходит в состояние равновесия, т. е. в такое состояние, что выполняются условия успокоения системы

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m-1. \tag{10}$$

При этом предполагается, что возмущение $f(x, t)$ действует в некотором ограниченном интервале времени $t_0 \leq t \leq t_0 + T = t_1$.

На первый взгляд может показаться, что в данной постановке задача финитного управления тривиальна. Действительно, если для заданного возмущения $f(x, t)$ управление $U(x, t)$ выбрать так, чтобы выполнялось равенство

$$f(x, t) + B(x, t)\,U(x, t) = 0, x_0 \leq x \leq x_1\, t \geq t_0, \tag{10а}$$

то данная управляемая система при всех $t \geq t_0$ останется в положении равновесия в силу того, что все возмущения, входящие в правую часть основного уравнения (7), будут тождественно равны нулю. Однако, к сожалению, указанное уравнение далеко не всегда имеет решение относительно $U(x, t)$. Например, в случае, когда $U(x, t)$ и $B(x, t)$ есть скаляры, для существования решения уравнения (10а) необходимо и достаточно, чтобы

$$U(x, t) = -\frac{f(x, t)}{B(x, t)} \text{ для всех } x \in [x_0, x_1] \text{ и } t \geq t_0. \tag{10б}$$

Но интересные практические случаи соответствуют тому, что $B(x, t) = 0$ там, где $f(x, t) \neq 0$ и, следовательно, уравнение (10а) не имеет решения. Такая ситуация имеет место тогда, когда подмножество M тех точек (x, t), $x \in [x_0,$

$x_1]$, $t \geq t_0$, где $f(x, t) \neq 0$, не включается в подмножество N тех же точек (x, t) из полосы $x \in [x_0, x_1]$, $t \geq t_0$, где $B(x, t) \neq 0$.

Физически факт несуществования решения уравнения (10а) соответствует тому, что возмущающие воздействия (например, начальные условия) распределены в одном месте физического пространства, а управления — в другом месте пространства.

С такой ситуацией приходится иметь дело во многих практически важных задачах управления распределенными объектами.

Например, при нагреве массивного тела, занимающего некоторую область $\mathcal{D}$ физического пространства, возмущением является начальная температура, распределенная во всей области $\mathcal{D}$. В то же время управляющие воздействия (температура окружающей тело среды) распределены лишь на границе Γ этой области $\mathcal{D}$. Ясно, что пространственные области определения возмущения и управления, т. е. $\mathcal{D}$ и Γ, не совпадают друг с другом и, следовательно, уравнение (10а) не разрешимо.

Аналогичная ситуация имеет место и в другой важной задаче об успокоении распределенной в некоторой области пространства $\mathcal{D}$ колебательной системы с помощью управляющих воздействий, распределенных лишь на границе Γ области $\mathcal{D}$. Здесь, например, возмущениями могут быть, скажем, порывы ветра, действующие на весь объект в целом, а управление осуществляется с помощью некоторых «рулей», расположенных на границе Γ области $\mathcal{D}$.

В конце статьи приведен пример задачи финитного управления, где имеет место указанная выше ситуация: возмущение колебательной системы действует внутри пространственного отрезка $\mathcal{D} = [0, \pi]$, а управляющее воздействие сосредоточено в граничной точке $x = 0$ этого отрезка $\mathcal{D}$. Естественно, в этом случае уравнение (10а) не поможет найти финитное управление.

Постановку задачи финитного управления в форме однородных краевых и нулевых начальных условий будем называть задачей финитного управления в *стандартной форме*.

Задача финитного управления в форме с неоднородными краевыми условиями и ненулевыми начальными условиями сводится к задаче финитного управления в *стандартной форме* путем соответствующего выбора функции $f(x, t)$ и вектор функции $B(x, t)$. Пример сведения к стандартной форме задачи финитного управления рассматривается дальше (см. задачу (29)–(32)), а также в приложении 2.

Фундаментальное финитное управление, соответствующее гриновскому возмущению, ($f(x, t) = \delta(x-\xi)\,\delta(t-\tau)$, $x_0 < x_1$, $\xi < x_1$, $t_0 < t$, $\tau < t_1$) обозначим через $U_\delta(x, t; \xi, \tau)$.

Тогда, поскольку мы имеем дело с линейными управляемыми системами, на основании справедливого для них закона *линейности или принципа суперпозиции (при сложении внешних воздействий их результаты также складываются)*, зная ф. ф. у. $U_\delta(x, t; \xi, \tau)$, можно показать, что в случае произвольных возмущений $f(x, t)$, соответствующее финитное управление имеет вид:

$$U(x, t) = \int_{t_0}^{t_1}\int_{x_0}^{x_1} f(\xi, \tau)\, U_\delta(x, t\,; \xi, \tau)\, d\xi\, d\tau,\ x_0 < x < x_1,\ t_0 < t < t_1. \quad (11)$$

Выражение в правой части формулы (11) будем называть композицией двух функций $f(x, t)$ и $U_\sigma(x, t; \xi, \tau)$.

В самом деле, по предположению для ф. ф. у. $U_\delta(x, t; \xi, \tau)$ и соответствующего $Q_\delta(x, t\,; \xi, \tau)$ выполняются тождества:

$$\sum_{k=0}^{m}\sum_{s=0}^{n} A_{ks}(x, t)\,\frac{\partial^{k+s} Q_\delta(x, t\,; \xi, \tau)}{\partial t^k\, \partial x^s} \equiv \delta(x-\xi)\,\delta(t-\tau) + B(x, t)\, U_\delta(x, t\,; \xi, \tau), \quad (12)$$

$$\sum_{i=0}^{1}\sum_{k=0}^{m-1}\sum_{s=0}^{n-1} a_{\nu iks}(t)\,\frac{\partial^{k+s} Q_\delta(x, t; \xi, \tau)}{\partial t^k\, \partial x^s}\bigg|_{x=x_i} \equiv 0, \quad \nu = 1, 2, \ldots, \leq n, \quad (13)$$

$$\left(\frac{\partial^k Q_\delta(x, t; \xi, \tau)}{\partial t^k}\right)_{t=t_0} \equiv 0, \quad k = 0, 1, \ldots, m-1, \quad (14)$$

$$\left(\frac{\partial^k Q_\delta(x, t; \xi, \tau)}{\partial t^k}\right)_{t=t_0+T=t_1} \equiv 0, \quad k = 0, 1, \ldots, m-1. \quad (15)$$

Умножая обе части тождеств (12)—(15) на $f(\xi, \tau)$, а затем интегрируя по ξ и τ в соответствующих пределах и пользуясь формулой

$$f(x, t) = \int_{t_0}^{t_1}\int_{x_0}^{x_1} f(\xi, \tau)\ \delta(x-\xi)\ \delta(t-\tau)\, d\xi\, d\tau, \quad (16)$$

получим, что управление $U(x, t)$, которое за промежуток времени $t_1 - t_0 = T$ переводит рассматриваемую систему в состояние равновесия (см. (10) имеет вид (11).

Иногда практически бывает удобнее находить финитное управление от возмущений, взятых не в виде δ-функций, а в виде производных от δ-функций.

Поэтому целесообразно несколько обобщить введенное выше понятие ф. ф. у.

Пусть возмущение задается с помощью «производных» от δ-функции в виде:

$$f(x, t) = \delta_x^{(r)}(x - \xi)\, \delta_t^{(q)}(t - \tau),\ x_0 < x_1,\ \xi < x_1,\ t_0 < t,\ \tau < t_1, \qquad (17)$$

где $\delta_x^{(r)}(x-\xi)$ — есть «r-ая производная» от δ-функции ($r = 0, \pm 1, \pm 2, \ldots$), причем отрицательным r соответствует $|r|$-ая первообразная от δ-функции по x, а $\delta_t^{(q)}(t-\tau)$ — есть q-ая производная по t от δ-функции ($q = 0, 1, \ldots$).

Возмущение вида (17) назовем *гриновским возмущением типа* (r,q). Финитное управление, гасящее гриновское возмущение типа (r, q) назовем фундаментальным финитным управлением (ф. ф. у.) типа (r, q) и обозначим через $U_{\delta(r,q)}(x, t; \xi, \tau)$.

Следует заметить, что в данной работе не рассматривается задача вычисления ф. ф. у. Дело в том, что ф. ф. у., как и всякое ф. у., может быть найдено с помощью методов и приемов, рассмотренных ранее в [2], [3].

Зная ф. ф. у. типа (r, q), на основании принципа суперпозиции можно показать, что финитное управление соответствующее произвольным возмущениям $f(x, t)$, имеет вид:

$$U(x, t) = \int_{t_0}^{t_1}\int_{x_0}^{x_1} f^{[-r-q]}(\xi, \tau)\, U_{\delta(r,q)}(x, t; \xi, \tau)\, d\xi\, d\tau,\ x_0 < x < x_1,\ t_0 < t < t_1, \qquad (18)$$

где $f^{[-r-q]}(\xi, \tau)$ обозначает результат интегрирования $f(\xi, \tau)$ последовательно r раз по ξ в пределах от x_0 до ξ и раз по τ в пределах от t_0 до τ (при $r < 0$ интегрирование по ξ заменяется дифференцированием $|r|$ раз).

Действительно, заменяя в (12) $\delta(x-\xi)\,\delta(t-\tau)$ на $\delta_x^{(r)}(x-\xi)\,\delta_t^{(q)}(t-\tau)$, и, следовательно, $U_\delta(x, t; \xi, \tau)$ на $U_{\delta(z,q)}(x, t; \xi, \tau)$, а затем умножая соответствующие тождества (12)—(15) на $f(\xi, \tau)$ и интегрируя по ξ и τ с учетом формулы

$$f(x, t) = \int_{t_0}^{t_1}\int_{x_0}^{x_2} f^{[-r-q]}(\xi, \tau)\, \delta_x^{(r)}(x - \xi)\, \delta_x^{(q)}(t - \tau)\, d\xi\, d\tau, \qquad (19)$$

придем к (18).

Заметим, что ф. ф. у. типа $(r + s, q + k)$ может быть получено от ф. ф. у. типа (r,q) посредством s-кратного и k-кратного дифференцирования соответственно по ξ и по τ и умножением на $(-1)^{s+k}$ (если $s < 0$ и $k < 0$, то соответствующие дифференцирования заменяются интегрированием, причем для соблюдения финитности управления должно быть $|k| \leq q$ в случае $k < 0$).

Справедлива формула:

$$U_{\delta(r+s,q+k)}(x, t; \xi, \tau) = (-1)^{s+k} \frac{\partial^{s+k} U_{\delta(r,q)}(x, t; \xi, \tau)}{\partial \xi^s\, \partial \tau^k} \qquad (20)$$

в указанном выше смысле.

Действительно, по определению ф. ф. у. типа $(r+s, q+k)$ $U_{\delta(s+r,q+k)}(x,t;\xi,\tau)$ соответствует возмущению:

$$f(x,t) = \delta_x^{(r+s)}(x-\xi)\,\delta_t^{(q+k)}(t-\tau). \tag{20a}$$

Согласно формуле (18) имеем:

$$U_{\delta(r+s,q+k)}(x,t;\xi,\tau) =$$

$$= \int_{t_0}^{t_1}\int_{x_0}^{x_1} [\delta_{\xi_1}^{(r+s)}(\xi_1-\xi)\,\delta_{\tau_1}^{(q+k)}(\tau_1-\tau)]^{[-r-q]}\cdot U_{\delta(r,q)}(x,t;\xi,\tau_1)\,d\xi,d\tau_1 =$$

$$= \int_{t_0}^{t_1}\int_{x_0}^{x_1} \delta_{\xi_1}^{(s)}(\xi_1-\xi)\,\delta_{\tau_1}^{(k)}(\tau_1-\tau)\,U_{\delta(r,q)}(x,t;\xi_1,\tau_1)\,d\xi,\,d\tau_1 =$$

$$= (-1)^{s+k}\frac{\partial^{s+k}U_{\delta(r,q)}(x,t;\xi,\tau)}{\partial\xi^s\,\partial\tau^k},$$

что и требовалось доказать.

В случае автономных во времени систем и по предположению, что $B(x,t) = B(x)$, можно получить формулу

$$U_{\delta(r,q+k)}(x,t;\xi,\tau) = \frac{\partial^k U_{\delta(r,q)}(x,t;\xi,\tau)}{\partial t^k}, \tag{21}$$

а в случае системы с постоянными коэффициентами и $B(x,t) = B$, где B — постоянный известный вектор, справедлива формула

$$U_{\delta(r+s,q+k)}(x,t;\xi,\tau) = \frac{\partial^{s+k}U_{\delta(r,q)}(x,t;\xi,\tau)}{\partial x^s\,\partial t^k}. \tag{21a}$$

Формулы (20), (21), (21a) дают возможность выразить ф. ф. у. одного типа через ф. ф. у. другого типа, в частности, типа (0, 0), т. е. через $U_\delta(x,t;\xi,\tau)$.

Рассмотрим случай, когда внешние возмущения действуют только в начальный момент.

Пусть, например, управляемая система описывается уравнением:

$$\sum_{k=0}^{m}\sum_{s=0}^{n} A_{ks}(x,t)\frac{\partial^{k+s}Q}{\partial t^k\,\partial x^s} = B(x,t)\cdot U(x,t),\quad x_0<x<x_1,\quad t_0<t<t_1, \tag{22}$$

с однородными краевыми условиями

$$\sum_{i=0}^{1}\sum_{k=0}^{m-1}\sum_{s=0}^{n-1} a_{\nu iks}(t)\frac{\partial^{k+s}Q}{\partial t^k\,\partial x^s}\bigg|_{x=x_i} = 0,\quad t_0<t<t_1,\quad \nu=1,2,\ldots,\leq n, \tag{23}$$

и пусть в начальный момент времени $t = t_0$ имеется некоторое возбужденное состояние данной системы, описываемое начальными условиями вида:

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = Q_k(x), \quad x_0 < x < x_1 \quad k = 0, 1, \ldots, m-1. \tag{24}$$

Требуется перевести систему (22), (23) из начального возбужденного состояния (24) в состояние равновесия

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m-1 \tag{25}$$

за конечное время $t_1 - t_0 = T$ посредством специального управляющего воздействия $U(x, t)$.

Финитное управление, соответствующее нулевым начальным условиям за исключением p-го начального условия, взятого в виде δ-функции, т. е. начальным условиям вида

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = \delta_{kp} \cdot \delta(x-\xi), \quad x_0 < x, \ \xi < x_1, \quad k, p = 0, 1, \ldots, m-1, \tag{26}$$

где δ_{kp} — символ Кронеккера, — назовем частным фундаментальным финитным управлением (ч. ф. ф. у.) порядка p и обозначим через $U_{\delta p}(x, t; \xi)$.

Далее финитное управление, гасящее начальное возмущение вида

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = \delta_{kp} \cdot Q_k(x), \quad x_0 < x < x, \quad k, p = 0, 1, \ldots, m-1, \tag{27}$$

обозначим через $U_p(x, t; Q_p)$.

Тогда на основании принципа суперпозиции легко убедиться, что финитное управление, соответствующее произвольным начальным условиям (24), имеет вид:

$$U(x, t) = \sum_{p=0}^{m-1} U_p(x, t; Q_p) = \sum_{p=0}^{m-1} \int_{x_0}^{x_1} Q_p(\xi)\, U_{\delta p}(x, t; \xi)\, d\xi,$$
$$x_0 < x < x_1, \quad t_0 < t < t_1. \tag{28}$$

Совокупность $\{U_{\delta p}(x, t; \xi)\}$, которую можно упорядочить в вектор, назовем *вектором фундаментального финитного управления*. Тогда формулу (28) символически можно записать в виде

$$U(x, t) = Q^0(\xi) * U_f(x, t; \xi), \quad x_0 < x < x_1, \ t_0 < t < t_1,$$

где знак * обозначает соответствующий интеграл по ξ от скалярного произведения, $Q^0(x) = (Q^0(x), \ldots, Q_{m-1}(x))$ — вектор начальных условий, а $U_f(x, t; \xi) = (U_{\delta,0}(x, t; \xi), \ldots, U_{\delta,m-1}(x, t; \xi))$ — вектор фундаментальных финитных управлений.

Аналогично тому, как это делалось выше с правой частью уравнения (7), (см. (17)), в начальные условия (26) можно ввести производные от δ-функции, т. е. вместо (26) рассмотреть условия вида

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} \delta_{kp}\, \delta_x^{(r_k)}(x-\xi),\ x_0 < x,\quad \xi < x_1,\quad k,\ p = 0,\ 1,\ \ldots,\ m-1, \tag{28a}$$

где $r_k = 0,\ \pm 1,\ \pm 2,\ \ldots$ ($r_k < 0$ соответствует $|r_k|$-ой первообразной от δ-функции).

Финитное управление, соответствующее начальным условиям (28а) назовем частным фундаментальным финитным управлением порядка p ранга r_p и обозначим через $U_{\delta p}^{r_p}(x,\ t;\ \xi)$. Доказывается, что управление $U(x,\ t)$, гасящее начальные возмущения (24), за промежуток времени $T = t_1 - t_0$ представляется в виде:

$$U(x,\ t) = \sum_{p=0}^{m-1} \int_{x_0}^{x_1} Q_p^{[-r_p]}(\xi)\, U_{\delta p}^{r_p}(x,\ t\ ;\ \xi)\, d\xi,\quad x_0 < x < x_1,\quad t_0 < t < t_1, \tag{29}$$

где $Q_p^{[-r_p]}(\xi)$ при $r_p > 0$ обозначает интегрирование по ξ последовательно r_p раз в пределах от x_0 до ξ, а при $r_p < 0$ дифференцирование по ξ $|r_p|$ раз.

Аналогично (20) доказывается справедливость формулы перехода:

$$U_{\delta p}^{r_p+s}(x,\ t\ ;\ \xi) = (-1)^s \frac{\partial^s U_{\delta p}^{r_p}(x,\ t\ ;\ \xi)}{\partial \xi^s},$$

$$p = 0,\ 1,\ \ldots,\ m-1\ ;\ r_p,\ s = 0,\ \pm 1,\ \pm 2,\ \ldots\ . \tag{30}$$

Можно установить связь между ч. ф. ф. у. различных порядков, и, следовательно, между двумя произвольными ч. ф. ф. у. разных порядков и разных рангов. Наиболее просто эта связь выявляется в случае систем вида:

$$\frac{\partial^m Q}{\partial t^m} + \sum_{k=0}^{m-1} \sum_{s=0}^{n} A_{ks}(x,\ t)\, \frac{\partial^{k+s} Q}{\partial t^k\, \partial x^s} = B(x,t)\, U(x,t),\ x_0 < x_1\ t_0 < t < t_1, \tag{31}$$

$$\sum_{i=0}^{1} \sum_{s=0}^{n-1} a_{vis}(t)\, \frac{\partial^s Q}{\partial x^s}\bigg|_{x=x_i} = 0,\quad t_0 < t < t_1,\quad v = 1,\ 2,\ \ldots,\ < n, \tag{32}$$

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = Q_k(x),\ x_0 < x < x_1,\quad k = 0,\ 1,\ \ldots,\ m-1, \tag{33}$$

с конечными условиями

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0,\quad x_0 < x < x_1,\quad k = 0,\ 1,\ \ldots,\ m-1. \tag{34}$$

Поставленную задачу финитного управления сведем к задаче финитного управления в стандартной форме. При этом продолжим функцию $Q(x, t)$ нулем на $t < t_0$ по формуле

$$Q^*(x, t) = Q(x, t) \cdot 1(t - t_0), \tag{35}$$

где $1(t-t_0)$ — единичная функция

$$1(t - t_0) = \begin{cases} 1 \text{ при } t > t_0 \\ 0 \text{ при } t < t_0. \end{cases} \tag{36}$$

Тогда, пользуясь формулами

$$\frac{\partial^{k+s} Q^*}{\partial t^k \partial x^s} = \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} 1(t - t_0) + \sum_{p=0}^{k-1} \left.\frac{\partial^{p+s} Q}{\partial t^p \partial x^s}\right|_{t=t_0} \cdot \delta_t^{(k-p-1)}(t - t_0), \tag{37}$$

$$A_{ks}(x, t) \frac{\partial^{k+s} Q^*}{\partial t^k \partial x^s} = A_{ks}(x, t) \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} 1(t - t_0) +$$

$$+ \sum_{p=0}^{k-1} \left.\frac{\partial^{p+s} Q}{\partial t^p \partial x^s}\right|_{t=t_0} \sum_{j=0}^{k-p-1} (-1)^j C_{k-p-1}^j \left.\frac{\partial^j A_{ks}(x, t)}{\partial t^j}\right|_{t=t_0} \cdot \delta_t^{(k-p-j-1)}(t - t_0), \tag{38}$$

из уравнений (31)—(34), заключаем, что

$$\frac{\partial^m Q}{\partial t^m} + \sum_{k=0}^{m-1} \sum_{s=0}^{n} A_{ks}(x, t) \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} = \sum_{p=0}^{m-1} Q_p(x) \delta_t^{(m-p-1)}(t - t_0) +$$

$$+ \sum_{k=1}^{m-1} \sum_{s=0}^{n} \sum_{p=0}^{k-1} Q_p^{(s)}(x) \sum_{j=0}^{k-p-1} (-1)^j C_{k-p-1}^j \left.\frac{\partial^j A_{ks}(x, t)}{\partial t^j}\right|_{t=t_0} \cdot \delta_t^{(k-p-j-1)}(t - t_0) +$$

$$+ B(x, t)\, U(x, t), \quad x_0 < x < x_1, \quad t_0 + \leq t < t_1, \tag{39}$$

$$\sum_{i=0}^{1} \sum_{s=0}^{n-1} a_{\nu is}(t) \left.\frac{\partial^s Q}{\partial x^s}\right|_{x=x_i} = 0, \quad t_0 + < t < t_1, \nu = 1, 2, \ldots, \leq n, \tag{40}$$

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0-} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m - 1, \tag{41}$$

и

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m - 1. \tag{42}$$

Теперь, последовательно подставляя в (33) и (39)

$$Q_k(x) = \delta_{k, m-1} \cdot \delta(x - \xi), \; Q_k(x) = \delta_{k, m-r} \cdot \delta(x - \xi), \ldots, Q_k(x) = \delta_{k,0} \cdot \delta(x - \xi)$$

и учитывая линейность системы, придем к соотношениям:

$$U_{\delta, m-1}(x, t\,;\, \xi) = U_{\delta(0,0)}(x, t;\, \xi, t) = U_\delta(x, t_0\,;\, \xi, \tau),$$

$$U_{\delta,m-q}(x,t\,;\,\xi) = U_{\delta(0,q-1)}(x,t\,;\,\xi,t_0) + \sum_{k=m-q+1}^{m-1}\sum_{s=0}^{n}\sum_{r=0}^{s}\sum_{j=0}^{k-m+q-1} \times$$

$$\times\,(-1)^{j+r}\,C_s^r\,C_{k-m+q-1}^{j}\,\frac{\partial^{j+r}A_{ks}(x,t)}{\partial t^j\,\partial x^s}\bigg|_{x=\xi,\,t=t_0} U_{\delta(s-r,\,k-j-m+q-1)}\,(x,t;\xi,t_6), \quad (43)$$

$$q = 2, 3, \ldots, m.$$

В частности, когда система автономна во времени, и $B(x, t)$ также не зависит от времени, то, учитывая (20) и (21), из (43) получим

$$U_{\delta,m-1}(x,t;\xi) = U_{\delta}(x,t\,;\,\xi,t_0)\,,$$

$$U_{\delta,m-q}(x,t\,;\xi) = \left[\frac{\partial^{q-1}}{\partial t^{q-1}} + \sum_{k=m-q+1}^{m-1}\sum_{s=0}^{n}\sum_{r=0}^{s}(-1)^s\,A^{(r)}_{m-q+1,\,s}(\xi)\,\frac{\partial^{(k-m+q-1)+(s-r)}}{\partial t^{k-m+q-1}\,\partial x^{s-r}}\right]\times$$

$$\times\,U_{\delta}(x,t\,;\,\xi,t_0)\,, \qquad q = 1, 2, \ldots, m. \quad (44)$$

Наконец, когда в автономной системе равны нулю тождественно все $A_{ks}(x)$, за исключением коэффициентов вида $A_{0s}(x)$ и $A_{k0}(x)$, и если $B(x, t)$ также не зависит от t, то формулы (44) принимают более простой вид

$$U_{\delta,m-q}(x,t\,;\,\xi) = \left[\frac{\partial^{q-1}}{\partial t^{q-1}} + A_{m-1,0}(\xi)\,\frac{\partial^{q-2}}{\partial t^{q-1}} + \ldots + A_{m-q+1,0}(\xi)\right]\times$$

$$\times\,U_{\delta,m-1}(x,\,t\,;\,\xi), \qquad q = 1, 2, \ldots, m. \quad (44a)$$

Заметим, что нет принципиальных трудностей для обобщения понятия ф. ф. у. на случай, когда состояние системы Q и управление U являются векторами, подчиненными векторным уравнениям.

Заметим еще, что введенные выше понятия допускают непосредственное обобщение и на многомерный случай. Если оператор l в уравнении вида (1) определяется некоторой задачей, аналогичной задаче (22)—(24), где вместо интервала (x_0, x_1) фигурирует некоторая область $\mathfrak{D}$ n-мерного евклидова пространства, под ч. ф.,ф. у. p-го порядка и векторного ранга $r_p = (r_{p_1}, r_{p_2}, \ldots, r_{p_n})$ будем понимать ф. ф. у., соответствующее начальным условиям вида:

$$\frac{\partial^k Q}{\partial t^k}(x,t_0) = \delta_{k,p}\,\delta_{x_1}^{(r_{k1})}(x_1-\xi_1)\ldots\,\delta_{x_n}^{(r_{kn})}(x_n-\xi_n)\,, \quad (45)$$

где $k,\ p = 0, 1, \ldots, m-1$; $r_j = 0, \pm1, \pm2, \ldots$; $j = 1, 2, \ldots, n$.

В этом случае интеграл в (29) следует брать по $\mathfrak{D}$, а $Q^{[-rp]}(\xi)$ нужно понимать как частную производную (соответственно частотный интеграл, если какое-то $r_j > 0$) от $Q_p(\xi)$ порядка $r_p = (r_{p_1}, r_{p_2}, \ldots, r_{p_n})$, соответственно по $\xi_1, \xi_2, \ldots, \xi_n$.

Соответствующее обобщение формулы (30) на этот случай имеет вид:

$$U^{r_p+s}_{\delta p}(x, t\,;\,\xi) = (-1)^{\bar{s}} \frac{\partial^s U^{r_p}_{\delta p}(x, t\,;\,\xi)}{\partial \xi_1^{s_1} \ldots \partial \xi_n^{s_n}}, \tag{46}$$

где $s = (s_1, \ldots, s_n)$; $\bar{s} = s_1 + \ldots + s_n$.

(В случае $s_j < 0$ подразумевается интегрирование по ξ_j порядка $|s_j|$ в соответствующей области).

Приложение 1

Пример 1. Найдем финитное управление (ф. у.) $U(t)$, $0 \leq t \leq T$, для волнового уравнения

$$\frac{\partial^2 Q}{\partial t^2} = \frac{\partial^2 Q}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0, \tag{47}$$

с краевыми условиями

$$Q(0, t) = U(t), \qquad Q(\pi, t) = 0, \qquad t > 0 \tag{48}$$

и произвольными начальными условиями

$$Q(x, 0) = Q_0(x), \qquad \frac{\partial}{\partial t} Q(x, 0) = Q_1(x), \quad 0 < x < \pi. \tag{49}$$

Сначала найдем ч. ф. ф. у. порядка 1 ранга 1, т. е. финитное управление $U^1_{\delta,1}(t, \xi)$, соответствующее начальным условиям,

$$Q(x, 0) = 0, \quad \frac{\partial}{\partial t} Q(x, 0) = \delta'_x(x - \xi), \quad 0 < x, \xi < \pi. \tag{50}$$

Используя метод характеристик (метод распространяющихся волн Даламбера), нетрудно найти, что минимальное T в этой системе равно 2π (для самого «неблагоприятного» ξ из $(0, \pi)$) и

$$U^1_{\delta,1}(t, \xi) = \frac{1}{2}[\delta(t - \xi) + \delta(t + \xi - 2\pi)], \quad 0 < \xi < \pi, \; 0 \leq t \leq 2\pi. \tag{51}$$

Но формуле (30) при $p = 1$, $r_1 = 1$ и $s = -1$ получим

$$U^0_{\delta,1}(t, \xi) = U_{\delta,1}(t, \xi) = -\frac{1}{2} \int_0^{\xi} [\delta(t - \eta) + \delta(t + \eta - 2\pi)]\, d\eta =$$

$$= \frac{1}{2}[1(t - \xi) - 1(t + \xi - 2\pi)], \quad 0 < \xi < \pi, \quad 0 \leq t \leq 2\pi. \tag{52}$$

Теперь по формуле (44а) при $m = 2, q = 2$, $A_{1,0}(\xi) = 0$ и $A_{0,0}(\xi) = 0$, находим

$$U_{\delta,0}(t, \xi) = \frac{\partial}{\partial t} U_{\delta,1}(t, \xi) = \frac{1}{2}[\delta(t-\xi) - \delta(t+\xi-2\pi)], \quad 0 < \xi < \pi, \; 0 \leq t \leq 2\pi. \tag{53}$$

Таким образом, вектор фундаментальных финитных управлений имеет вид:

$$U_f(t, \xi) = \left\| \frac{\delta(t-\xi) - \delta(t+\xi-2\pi)}{2}, \; \frac{1(t-\xi) - 1(t+\xi-2\pi}{2} \right\| \tag{54}$$

следовательно, финитное управление соответствующее произвольным начальным условиям (49) в соответствии с формулами (21) и (44а) равно

$$\begin{aligned} U(t) = Q^0(\xi) * U_f(t, \xi) &= \sum_{p=0}^{1} \int_0^{\pi} Q_p(\xi) \, U_{\delta,p}(t, \xi) \, d\xi = \\ &= \frac{1}{2} \int_0^{\pi} Q_0(\xi) \, [\delta(t-\xi) - \delta(t+\xi-2\pi)] \, d\xi + \\ &+ \frac{1}{2} \int_0^{\pi} Q_1(\xi) \, [1(t-\xi) - 1(t+\xi-2\pi)] \, d\xi = \end{aligned} \tag{55}$$

$$= \frac{1}{2} \begin{cases} Q_0(t) + \int_0^t Q_1(\xi) \, d\xi, & \text{при} \quad 0 < t < \pi, \\ -Q_0(2\pi - t) + \int_0^{\pi} Q_1(\xi) \, d\xi - \int_{\pi}^{t} Q_1(2\pi - \xi) \, d\xi, & \text{при} \quad \pi < t < 2\pi, \\ 0, & \text{при} \quad t \leq 0, \; t \geq 2\pi \end{cases}$$

отсюда, считая, что $Q_0(x) = Q_1(x) \equiv 0$, при $x \in (x_0, x_1)$,

$$U(t) = \frac{1}{2}[Q_6(t) - Q_6(2\pi - t)] + \int_0^t [Q_1(\xi) - Q_1(2\pi - \xi)] \, d\xi, \quad 0 \leq t \leq 2\pi. \tag{56}$$

Приложение 2

Рассмотрим задачу финитного управления

$$\frac{\partial^2 Q}{\partial t^2} = \frac{\partial^2 Q}{\partial x^2} + C(x) \, Q + f(x, t), \quad x_0 < x < x_1, \quad t > t_6 \tag{57}$$

$$\left.\begin{aligned} \alpha_0 Q(x_0, t) + \beta_0 Q'_x(x_0, t) &= U_0(t), \quad t > t_0, \\ \alpha_1 Q(x_1, t) + \beta_1 Q'_0(x_1, t) &= U_1(t), \quad t > t_0, \end{aligned}\right\} \tag{58}$$

$$Q(x, t_0) = Q_0(x), \quad \frac{\partial}{\partial t} Q(x, t_0) = Q_1(x), \quad x_0 < x < x_1, \tag{59}$$

и

$$Q(x, t_1) = 0, \qquad \frac{\partial}{\partial t} Q(x, t_1) = 0, \quad x_0 < x < x_1. \tag{60}$$

Предположим, что существует функция Грина $\mathcal{C}(x, \xi, t-\tau)$, с помощью которой решение задачи (57)—(59) можно представить в виде:

$$Q(x,t) = \int_{t_0}^{t} \left[\mathcal{C}(x,\ \xi,\ t-\tau) \frac{\partial Q(\xi,\ \tau)}{\partial \xi} - Q(\xi,\ \tau) \frac{\partial \mathcal{C}(x,\ \xi,\ t-\tau)}{\partial \xi} \right]_{\xi=x_0}^{\xi=x_1} d\tau +$$

$$+ \int_{x_0}^{x_1} Q_0(\xi)\, \mathcal{C}_t(x,\ \xi,\ t-t_0)\, d\xi + \int_{x_0}^{x_1} Q_1(\xi)\, \mathcal{C}(x,\ \xi,\ t-t_0)\, d\xi +$$

$$+ \int_{t_0}^{t} \int_{x_0}^{x_1} f(\xi, \tau)\, \mathcal{C}(x, \xi, t-\tau)\, d\xi\, d\tau, \quad x_0 < x < x_1, \quad t > t_0. \tag{61}$$

Исходя из этой формулы, задачу (57)—(60) можно свести к следующей стандартной задаче финитного управления:

$$\frac{\partial^2 Q}{\partial t^2} = \frac{\partial^2 Q}{\partial x^2} + C(x)\, Q + Q_0(x)\ \delta_t'(t-t_0) + Q_1(x)\, \delta(t-t_0) + f(x,t) +$$

$$+ b_0(x)\, u_0(t) + b_1(x)\, u_1(t), \quad x_0 < x < x_1,\ t_0 + \leq t \leq t_1, \tag{62}$$

$$\left.\begin{aligned} &\alpha_0\, Q(x_0, t) + \beta_0\, Q_x'(x_0, t) = 0, \quad t_0 + \leq t < t_1, \\ &\alpha_1\, Q(x_1, t) + \beta_1\, Q_x'(x_1, t) = 0, \quad t_0 + \leq t < t_1, \end{aligned}\right\} \tag{63}$$

$$Q(x, t_0 -) = 0, \ \frac{\partial}{\partial t} Q(x, t_0 -) = 0, \quad x_0 < x < x_1, \tag{64}$$

$$Q(x, t,) \quad = 0, \ \frac{\partial}{\partial t} Q(x, t_1) = 0, \qquad x_0 < x < x_1, \tag{65}$$

$$b_0(x) = \begin{cases} -\dfrac{1}{\beta_0}\, \delta(x - x_0) & \text{при} \quad \beta_0 \neq 0 \\ \dfrac{1}{\alpha_0}\, \delta_x'(x - x_0) & \text{при} \quad \beta_0 = 0 \end{cases} \tag{66}$$

$$b_1(x) = \begin{cases} \dfrac{1}{\beta_1}\, \delta(x - x_1) & \text{при} \quad \beta_1 \neq 0 \\ -\dfrac{1}{\alpha_1}\, \delta_x'(x - x_1) & \text{при} \quad \beta_1 = 0. \end{cases} \tag{67}$$

Литература

1. *Тихонов А. И., Самарский А. А.* Уравнение математической физики. «Наука», 1966.
2. *Бутковский А. Г.* Финитное управление и управляемость в распределенных системах. Докл. АН СССР, **191,** 6, 1970.
3. *Фельдбаум А. А., Бутковский А. Г.* Методы теории автоматического управления. «Знание», 1971.

Fundamental finite control

A. G. BUTKOVSKII, M. I. MUSTAFAEV

(Moscow, Baku)

For distributed-parameter systems whose motion is described by linear partial differential equations, the problem of finite control is discussed. An important notion of fundamental finite control is introduced which is, by definition, the finite control for disturbances of the form of δ-functions, i.e. Green-type disturbances. This notion is similar to that of linear equation fundamental solution.

With the fundamental finite control known, through some integration operation one can obtain the finite control for any type of disturbances which may occur in a given controllable system: initial condition disturbances, disturbances in the right side of the basic differential equation, and disturbances in the right side of the boundary conditions.

Apart from the notion of the fundamental finite control, notions of partial fundamental finite controls are introduced which are, by definition, finite controls for disturbances in the initial distributions having the form of derivatives of δ-function.

Formulae are presented enabling recalculation of one kind of fundamental finite control into another, thus making flexible use of physical considerations when looking for fundamental finite control. An example is presented of finding the fundamental finite control for a distributed system described by the wave equation with control involved into the boundary condition.

А. Г. Бутковский,
Институт проблем управления
СССР, Москва В-485,
Профсоюзная ул., 81

Problems of Control and Information Theory, Vol. 5 (1), pp. 87–95 (1976)

ON THE EXISTENCE AND STABILITY BEHAVIOR OF A STOCHASTIC INTEGRAL EQUATION IN A BANACH SPACE*

A. N. V. RAO, C. P. TSOKOS

(Tampa)

(Received April 23, 1975)

1. Introduction

Random or stochastic integral equations are currently being studied by many scientists and mathematicians. Such equations play a very important role in model building of various problems in life sciences and engineering among others. For an up to date development of the theory and applications of random integral equations we refer the reader to A. T. Bharucha-Reid [2], C. P. Tsokos and W. J. Padgett [8].

Our main purpose in the present study is to investigate random integral equation in Banach space of the form

$$x(t\,;\omega) = h(t\,;\omega) + \int_S K(t,s\,;\omega)\, f\big(s, x(s\,;\omega)\;;\;\omega\big)\, d\mu(s) \tag{1.1}$$

where

(i) S is a locally compact metric space, with metric d and μ is a complete σ-finite measure defined on the σ-algebra of Borel subsets of S;

(ii) $\omega \in \Omega$ is the supporting set of a complete probability measure space (Ω, F, P);

(iii) $x(t\,;\omega)$ is *the unknown random function* defined on $S \times \Omega$ with values in X, a separable Banach space with norm $\|\cdot\|$;

(iv) $h(t\,;\omega)$, is a map from $S \times \Omega$ into X;

(v) $K(t,s\,;\omega)$, *the stochastic kernel*, is a map from $S \times S \times \Omega$ into the space of bounded linear operators from X into X;

(vi) $f(t, x;\ \omega)$ is a map from $S \times X \times \Omega$ into X.

The integral in (1.1) is to be understood as a Bochner integral.

We shall be concerned with existence, uniqueness and stability behavior of the random solutions of the integral equation (1.1). This equation is a generalization of the recent results of Padgett [6], [7] and Ahmed and Teo [1].

* This research was supported by the United States Air Force, Air Force Office of Scientefic Research, under Grant No. AFOSR-74-2711.

2. Preliminaries

We shall assume that S is the union of a countable family of compact sets $\{C_n\}$, with $\mu(C_n) < \infty$, such that $C_1 \subset C_2 \subset C_3 \ldots$ and that every compact set in S is contained in some $C_k \in \{C_n\}$ [6]. From these assumptions, it is clear that S is separable.

Let $L_2(\Omega, X)$ denote the space of all functions, $y(\omega)$, from Ω into X such that

$$\int_{\Omega} ||y(\omega)||^2 \, dP < \infty,$$

where $||\cdot||$ denotes the norm in X. It is known ([2], p. 23) that $L_2(\Omega, X)$ is a Banach space with norm $||\cdot||_{L_2}$, defined by

$$||y(\omega)||_{L_2} = \{\int_{\Omega} ||y(\omega)||^2 \, dP\}^{1/2},$$

$$y(\omega) \in L_2(\Omega, X).$$

Definition 2.1. By C, we shall mean the space of all continuous functions $x(t; \omega)$ from S into $L_2(\Omega, X)$.

We shall induce a topology on C by means of the following countable family of semi-norms.

$$||x(t\,;\omega)||_n = \sup_{t \in C_n} ||x(t\,;\omega)||_{L_2}.$$

It is known [9] that this topology is metrizable and the resulting metric space is a Fre'chet space.

$B(X)$ will denote the space of all bounded linear operators from X into X. Then $B(X)$ is a Banach space, with norm $||\cdot||_B$ given by

$$||T||_B = \sup_{||x|| \leq 1} ||Tx||, \; T \in B(X), \; x \in X.$$

Let $L(t; \omega)$ be an almost separably valued measurable map from $S \times \Omega$ into $B(X)$ such that

$$g(t) \, ||x|| \leq ||L(t; \omega)\, x||, \; x \in X, \; w \in \Omega$$

where $g(t) > 0$ is a real continuous function on S. Then, it is clear ([4] p. 42) that $L^{-1}(t; \omega)$ exists and is a measurable map from $S \times \Omega$ into $B(X)$ ([2], p. 76). Also, since L is almost separably valued, so is L^{-1}. Hence, $||L(t; \omega)||$ and $||L^{-1}(t; \omega)||$ are measurable (see [2], p. 77).

Definition 2.2. C_L will denote the space of all continuous functions $x(t; \omega) \in C$ such that for each $x(t; \omega) \in C_L$, there is a $\Lambda > 0$, such that

$$||L(t; \omega)\, x(t; \omega)||_{L_2} \leq \Lambda.$$

It is easy to show that C_L is a Banach space with norm $||\cdot||_{C_L}$ defined by

$$||x(t;\omega)||_{C_L} = \sup_{t\in S} ||L(t;\omega)\,x(t;\omega)||_{L_2}.$$

Definition 2.3. If $L(t;\omega) = I$, the identity operator, then C_L will be denoted by C_1.

Definition 2.4. By a random solution of Eq. (1.1), we will mean a function $x(t;\omega) \in C$, which satisfies the equation P — a.e.

Let $L_\infty(\Omega, B)$ denote the space of all maps, $T(\omega)$, from Ω into $B(X)$ such that

$$P - \operatorname*{ess\,sup}_{\omega\in\Omega} ||T(\omega)||_B < \infty.$$

For $T(\omega) \in L_\infty(\Omega, B)$, define

$$|||T(\omega)|||_B = P - \operatorname*{ess\,sup}_{\omega\in\Omega} ||T(\omega)||_B.$$

We list the following conditions regarding the behavior of the random kernel $K(t, s; \omega)$:

(a) $K(t, s; \omega)$ is an almost separably valued map from $S\times S\times\Omega$ into $B(X.)$

(b) $K(t, s; \omega)\,L^{-1}(s; \omega)$ is continuous as a map from $S\times S$ into $L_\infty(\Omega, B)$.

(c) $\int_S |||K(t, s; \omega)\,L^{-1}(s; \omega)|||\,ds$ is a continuous function of t such that $\int_S |||K(t, s; \omega)\,L^{-1}(s; \omega)|||\,ds \le A(t)$ where $A(t) \ge 0$ is continuous on S.

Remark. Assumption (b) implies that $K(t, s; \omega)L^{-1}(s; \omega)$ is measurable (see [5] pp. 156—159).

Since K and L^{-1} are almost separably valued, it follows that ([2] p. 73) $|K(t, s; \omega)\,L^{-1}(s; \omega)||$ is measurable.

We also note that K and L are uniform random operators ([2], p. 73).

We shall now define the operator T on C_L by

$$(Tx)(t;\ \omega) = \int_S K(t, s; \omega)\,x(s; \omega)\,ds. \tag{2.1}$$

Let $t \in S$ and $x(s; \omega) \in C_L$. Then

$$\int_S \Big\{\int_\Omega ||K(t, s; \omega)\,x(s; \omega)||\,dP\Big\}\,d\mu(s) \le \int_S \Big\{\int_\Omega |||K(t, s; \omega)\,L^{-1}(s; \omega)|||$$

$$||L(s; \omega)\,x(s; \omega)||\,dP\Big\}\,d\mu(s) \le \int_S |||K(t, s; \omega)\,L^{-1}(s; \omega)|||$$

$$\Big\{\int_\Omega ||L(s; \omega)\,x(s; \omega)||^2\,dP\Big\}^{\frac{1}{2}}\,d\mu(s) \le$$

$$\le ||x(t; \omega)||_{C_L} \int_S |||K(t, s; \omega)\,L^{-1}(s; \omega)|||\,d\mu(s) \le ||x(t; \omega)||_{C_L}\,A(t). \tag{2.2}$$

Hence, by Tonelli's theorem $||K(t, s; \omega)\, x(s; \omega)||$ is μ-integrable on S, a.s., for each $t \in S$. Thus, the Bochner integral in (2.1) is well defined.

We shall now show that for each $t \in S$, $(Tx)(t; \omega) \in L_2(\Omega, X)$. It follows from (2.2) and from Fubini's theorem for Bochner integrals ([4], p. 84) that $(Tx)(t; \omega)$ is measurable in ω for each $t \in S$. Moreover,

$$||\int_S K(t, s; \omega)\, x(s; \omega)\, d\mu(s)||_{L_2} \leq ||\int_S |||K(t, s; \omega)\, L^{-1}(s; \omega)|||$$

$$||L(s; \omega)x(s; \omega)||\, d\mu(s)||_{L_2} \leq \int_S |||K(t, s; \omega)\, L^{-1}(s; \omega)|||$$

$$||L(s; \omega)\, x(s; \omega)||_{L_2}\, d\mu(s) \leq ||x(t; \omega)||_{C_L} \int_S |||K(t, s; \omega)\, L^{-1}(s; \omega)|||\, d\mu(s) \leq$$

$$\leq ||x(t; \omega)||_{C_L}\, A(t)\,. \tag{2.3}$$

This proves the claim that $(Tx)(t; \omega)$ is in $L_2(\Omega, X)$ for each $t \in S$. It is easy to show that $(Tx)(t; \omega)$ is continuous as a map from S onto $L_2(\Omega, X)$. Thus, T is a map from C_L into C.

Definition 2.5. Let B_1 and B_2 be Banach spaces such that $B_1 \subset B_2$. Then B_1 is said to be stronger than B_2 if every sequence which converges in the topology of B_1 converges in the topology of B_2.

Definition. 2.6. Let B and D be Banach spaces. The pair (B, D) is said to be admissible with respect to the operator T if $TB \subset D$.

We shall now state a lemma which will be basic for our main results. The proof can be found in ([3], p. 25).

Lemma 2.1. Let T be a continuous linear operator from C_L into C. Let $B \subset C_L$ and $D \subset C$ be Banach spaces stronger than C_L and C, respectively. If (B, D) is admissible with respect to T, than T is continuous from B into D.

We shall denote

$$M = ||T|| = \sup_{||x||_B \leq 1} \frac{||Tx||_D}{||x||_B}\,. \tag{2.4}$$

3. Existence and uniqueness of random solutions

We shall now present some results concerning the existence and uniqueness of random solutions to the random integral equation (1.1).

Theorem 3.1. Consider the stochastic integral equation (1.1) subject to the following conditions:

(i) $K(t, s; \omega)$ satisfies conditions (a), (b) and (c);

(ii) B and D are Banach spaces stronger than C_L and C, respectively, such that (B, D) is admissible with respect to the operator T defined by (2.1);

(iii) Let

$$Q_p = \{x(t;\omega) : x(t;\omega) \in D,\ ||x(t;\omega)||_D \leq p\},$$

where $p > 0$ is a constant. For $x(t;\omega) \in Q_p$,

$$||f(t, x(t;\omega);\omega) - f(t, y(t;\omega);\omega)||_B \leq \lambda\, ||x(t;\omega) - y(t;\omega)||_D$$

for $x(t;\omega)$, $y(t;\omega) \in Q_p$ and λ, a constant.

(iv) $h(t;\omega) \in D$.

Then, there exists a random solution to Eq. (1.1) in Q_p provided $\lambda M < 1$ and

$$||h(t;\omega)||_D + M\, ||f(t,0;\omega)||_B \leq p(1 - \lambda M),$$

where M is defined in (2.4).

Proof. Define the operator U from Q_p into D by

$$(Ux)(t;\omega) = h(t;\omega) + \int_S K(t,s;\omega)\, f(s, x(s;\omega);\omega)\, d\mu(s)\,. \tag{3.1}$$

From Lemma (2.1) and the Lipschitz condition on f, we obtain

$$||(Ux)(t;\omega)||_D \leq ||h(t;\omega)||_D + M\, ||f(t, x(t;\omega);\omega)||_B \leq$$

$$\leq ||h(t;\omega)||_D + M\, ||f(t,0;\omega)||_B + M\lambda\, ||x(t;\omega)||_D \leq p\,.$$

Hence,

$$UQ_p \subset Q_p.$$

Now let $x(t;\omega)$, $y(t;\omega) \in Q_p$. Then

$$||(Ux)(t;\omega) = (Uy)(t;\omega)||_D \leq M\, ||f(t, x(t;\omega);\omega) - f(t, y(t;\omega;\omega))||_B \leq$$

$$\leq M\lambda\, ||x(t;w) - y(t;w)||_D\,.$$

Since by hypothesis $M\lambda < 1$, U is a contraction on Q_p. By Banach's fixed point theorem there exists a unique random solution in Q_p to Eq. (1.1).

In the following theorem, we shall choose $L^{-1}(t;\omega) = r(t;\omega)$, where $r(t;\omega)$ is a positive real valued function on $S \times \Omega$, continuous in t and measurable in ω. We shall further assume that there is a continuous real valued function $g(t)$ on S such that

$$0 < g(t) \leq r(t;\omega)\,.$$

With $L^{-1}(t;\omega) = r(t;\omega)$, we shall denote the space C_L by C_r.

Theorem 3.2. Consider the integral equation (1.1) under the following conditions:

(i) $K(t, s; \omega)$ satisfies conditions (a), (b) and (c) with $L^{-1}(t; w) = r(t; w)$ and $A(t) = A$, a constant;

(ii) $f(t, x; \omega)$ is a map from $S \times X \times \Omega$ into X such that f is continuous in t and x and measurable in ω and satisfies the condition

$$||f(t, x; \omega) - f(t, y; \omega)|| \leq \lambda\, r(t; \omega)||\, x - y||\,,$$

for $x, y \in X$;

(iii) $f(t, 0; \omega) \in C_r$ and $h(t, \omega) \in C_1$.

Then there exists a unique random solution to (1.1) such that

$$||x(t; \omega)||_{C_1} \leq \frac{||h(t; \omega)||_{C_1} + A||f(t, 0; \omega)||_{C_r}}{(1 - \lambda A)}$$

provided $\lambda A < 1$.

Proof. From Eq. (2.3) and assumption (i) of the theorem, it is clear that (C_r, C_1) is admissible with respect to the operator T (defined in (2.1)) with $M = A$. Also, if $x(t; \omega) \in C_1$, then by (ii) we have

$$||f(t, x(t; \omega); \omega)\, r^{-1}(t; \omega)||_{L_2} \leq \lambda\, ||x(t; \omega)||_{L_2} + ||f(t, 0; \omega) r^{-1}(t; \omega)||_{L_2}$$

which implies

$$||f(t, x(t; \omega); \omega)||_{C_r} \leq \lambda\, ||x(t; \omega)||_{C_1} + ||f(t, 0; \omega)||_{C_r}.$$

The above inequality shows that $f(t, x(t; \omega); \omega) \in C_r$ whenever $x(t; \omega) \in C_1$.

It is easy to show that

$$||f(t, x(t; \omega); \omega) - f(t, y(t; \omega); \omega)||_{C_r} \leq \lambda\, ||x(t; \omega) - y(t; \omega)||_{C_1}$$

for $x(t; \omega)$, $y(t; \omega) \in C_1$. If we now choose

$$\varrho = \frac{||h(t; \omega)||_{C_1} + A||f(t, 0; \omega)||_{C_r}}{(1 - \lambda A)}$$

it is easily verified that all the conditions of Theorem 3.1 are met with $B = C_r$ and $D = C_1$. Hence, there exists a unique solution to (1.1) in C_1.

We state the following corrolary as it is useful in applications.

Corrolary 3.1. Assume that the function $r(t; \omega)$ in Theorem 3.2 is P — ess. bounded uniformly in t. Suppose that all the conditions of Theorem 3.2 hold, except (i), which is replaced by

(i′) $K(t, s; \omega)$ satisfies conditions (a), (b) and (c) with $L^{-1}(t, \omega)$ replaced by I, the identity operator on X.

Then the conclusion of Theorem 3.2 holds.

Proof. It is an easy matter to verify that (i′) together with the P — ess. boundedness of $r(t;\omega)$ implies assumption (i) of Theorem 3.2

4. Stability behavior

Throughout this section we shall assume that the function $f(t,x;\omega)$ appearing in Eq. (1.1) satisfies the conditions $f(t,0;\omega)=0$.

Definition 4.1. The system (1.1) is said to be *bounded input bounded output* (b. i. b. o.) *stable in the mean* if

(i) the solution $x(t;\omega)$ is in C_1 whenever $h(t;\omega)$ is in C_1;

(ii) let $x_1(t;\omega)$ and $x_2(t;\omega)$ be, respectively, the two solutions corresponding to the stochastic inputs $h_1(t;\omega)$ and $h_2(t;\omega)$. Then, for any $\varepsilon>0$ there is a $\delta>0$ such that

$$||h_1(t;\omega)-h_2(t;\omega)||_{C_1}\leq\delta$$

implies

$$||x_1(t;\omega)-x_2(t;\omega)||_{C_1}\leq\varepsilon\,.$$

In the following definition, we shall assume that $S=R^+=[0,\infty)$ and μ the Lebesque measure.

Definition 4.2. The system (1.1) is said to *exponentially stable* in the mean, if $||h(t;\omega)||_{L_2}\leq\alpha\,e^{-\delta_1 t}$ implies that $||x(t;\omega)||_{L_2}\leq\gamma\,e^{-\delta_1 t}$, where $x(t;\omega)$ is a solution of (1.1) and γ and δ_1 are positive constants.

We shall now prove two theorems concerning the stability of the system (1.1).

Theorem 4.1. Suppose that. (1.1) satisfies the conditions of Theorem 3.2. Then the system (1.1) is b. i. b. o. stable.

Proof. Noting that $f(t,0;\omega)=0$, we see from Theorem 3.2 that

$$||x(t;\omega)||_{C_1}\leq\frac{||h(t;\omega)||_{C_1}}{1-\lambda M}\,. \tag{4.1}$$

Equation (4.1) implies that $x(t;\omega)\in C_1$ if $h(t;\omega)\in C_1$.

Let $x_1(t;\omega)$ and $x_2(t;\omega)$ be the two solutions of (1.1) corresponding to the inputs $h_1(t;\omega)$ and $h_2(t;\omega)$. Then

$$||x_1(t;\omega)-x_2(t;\omega)||_{L_2}\leq||h_1(t;\omega)-h_2(t;\omega)||_{L_2}+$$

$$+\int_S|||K(t,s;\omega)\,r(s;\omega)|||\,\lambda\left\|\frac{x(s;w)-x_2(s;w)}{r(s;\omega)}\right\|_{L_2}d\mu(s)$$

$$\leq ||h_1(t;\omega) - h_2(t;\omega)||_{L_2} + \lambda A\, ||x_1(t;\omega) - x_2(t;\omega)||_{C_1}. \tag{4.2}$$

It is clear that Eq. (4.2) implies that

$$\|x_1(t;\omega) - x_2(t;\omega)\|_{C_1} \leq \frac{\|h_1(t;\omega) - h(t_2;\omega)\|_{C_1}}{(1-\lambda A)}. \tag{4.3}$$

Equations (4.1) and (4.3) together imply the b. i. b. o. stability in the mean of the system (1.1).

Theorem 4.2. Let the integral equation (1.1) satisfy the following conditions:

(i) same as (i) of Theorem (3.2) with $C(\omega)$ replacing $r(t;\omega)$, where $C(\omega) > 0$ is a measurable function on Ω. In addition assume that

$$\int_0^\infty |||K(t,s;\omega)\, C(\omega)|||\, e^{-\delta s} ds \leq A e^{-\delta_1 t},$$

for some $\delta > 0$ and $0 < \delta_1 \leq \delta$;

(ii) same as (ii) of Theorem (3.2) with $C(\omega)$ replacing $r(t;\omega)$.

Then the system (1.1) is exponentially stable in the mean provided $\lambda A < 1$.

Proof. Set $L_1^{-1}(t;\omega) = C(\omega)\, e^{-\delta t}$ and $L_2^{-1}(t;\omega) = e^{-\delta_1 t}$.

We will show that the pair (C_{L_1}, C_{L_2}) is admissible with respect to the operator T defined in (2.1). Let $x(t;\omega) \in C_{L_1}$. Then

$$\|Tx\|_{L_2} \leq \int_0^\infty |||K(t,s;\omega)\, C(\omega)|||\, e^{-\delta s} \left\| \frac{x(t;\omega)}{C(\omega)\, e^{-\delta s}} \right\|_{L_2} d\mu(s) \leq$$

$$\leq \|x(t;\omega)\|_{C_{L_1}} \int_0^\infty |||K(t,s;\omega)\, C(\omega)|||\, e^{-\delta s}\, d\mu(s) \leq$$

$$\leq A \exp\{-\delta_1 t\}\, \|x(t;w)\|_{C_{L_1}}. \tag{4.4}$$

Thus,

$$||Tx(t;\omega)||_{C_{L_2}} \leq A\, ||x(t;\omega)||_{C_{L_1}}.$$

which proves the admissibility of (C_{L_1}, C_{L_2}) with respect to T.

Also, it is easy to check that $f(t,x(t;\omega);\omega) \in C_{L_1}$ whenever $x(t;\omega) \in C_{L_2}$ and

$$||f(t,x(t;\omega);\omega) - f(t,y(t;\omega);\omega)||_{C_{L_1}} \leq \lambda\, ||x(t;\omega) - y(t;\omega)||_{C_{L_2}},$$

for $x(t;\omega),\ y(t;\omega) \in C_{L_2}$.

Therefore, all the conditions of Theorem 3.1 are satisfied with $B = C_{L_1}$ and $D = C_{L_2}$ and

$$p = ||h(t;\omega)||_{C_{L_2}} \ (1 - \lambda A)^{-1}.$$

Hence, there is a unique solution to (1.1) in C_{L_2} i.e. a solution $x(t;\omega)$ such that $||x(t;\omega)||\, L_2 \leq \gamma e^{-\delta_1 t}$, which proves the theorem.

References

1. *Ahmed, N. V., Teo, K. L.*, On the Stability of a Class of Nonlinear Stochastic Systems, Information and Control, **20,** *3*, April, 1972, pp. 276—293.
2. *Bharucha-Reid, A. T.*, Random Integral Equations. Academic Press, New York, October, 1973.
3. *Corduneanu, C.*, Integral Equations and Stability of Feedback Systems. Academic Press, New York, December, 1973.
4. *Hille, E., Phillips, R. S.*, Functional Analysis and Semi Groups, American Mathematical Society, Providence, R. I., 1957.
5. *Gikhman, I. I., Skorokhod, A. V.*, Introduction to the Theory of Random Processes, W. B. Saunders, Comp., Philadelphia, Pa., 1969.
6. *Padgett, W. J.*, On a Nonlinear Stochastic Integral Equations of the Hammerstein Type, Proceedings of the American Math. Society, **38,** *3*, May, 1973, pp. 625—631.
7. *Padgett, W. J.*, On a Random Volterra Integral Equation, Mathematical System Theory, **7,** *2*, 1973, pp. 164—169.
8. *Tsokos, C. P., Padgett, W. J.*, Random Integral Equations with Applications in Life Sciences and Engineering, Academic Press, New York, March, 1974.
9. *Yosida, K.*, Functional Analysis, Academic Press, New York, New York, 1965.

К проблеме существования и устойчивости некоторого стохастического уравнения в банаховом пространстве

А. Н. В. РАО, С. П ЦОКОС

(Тампа)

В статье рассматривается стохастическое интегральное уравнение типа

$$x(t;\omega) = h(t;\omega) + \int_S K(t,s;\omega)\, f\big(s, x(s;\omega);\omega\big)\, d\mu(s)\,.$$

Приведены условия существования, однозначности и устойчивости стохастического решения данного интегрального уравнения.

A. N. V. Rao
C. P. Tsokos
Department of Mathematics
University of South Florida
Tampa, Florida 33620
USA

THE SIXTH CONGRESS OF IFAC

The sixth triennial congress of the International Federation of Automatic Control (IFAC) was held in Boston/Cambridge, Massachusetts, U.S.A., August 24—30, 1975. It was devoted to the theme of "Control Technology in the Service of Man". Host organization was the American Automatic Control Council representing the American Institute of Aeronautics and Astronautics, the American Institute of Chemical Engineers, the American Society of Mechanical Engineers, the Institute of Electrical and Electronics Engineers, the Institute of Traffic Engineers, the Instrument Society of America, the Society for Industrial and Applied Mathematics, and the Technical Association of the Pulp and Paper Industry. Technical sessions of the Congress were held at the Massachusetts Institute of Technology in Cambridge, world-renowned for its important contributions to the automatic control theory and the computer technology.

Like the earlier Congresses in Moscow (1960), Basel (1963), London (1966), Warsaw (1969) and Paris (1972), this one was a "large scale system" consisting of quite a few subsystems each of them being a separate theoretical or practical area of control developing according to its own laws and rules subordinated to the general requirements of the "system", the control theory. Performance of the system may be characterized by the following facts: 1350 scientists from 35 countries participated at the Congress; along with 8 plenary sessions, there were 63 sessions where 347 papers were read; there were 14 round table discussions, 9 technical visits to laboratories and enterprises involved into production and application of computers and control systems. The sessions and papers presented there were distributed between the committees as listed in the Table:

The mere listing of sessions and discussions gives some general idea of the Congress' scope. Theoretical papers were presented at the sessions on linear and nonlinear distributed parameter systems, differential games, optimal and stochastic systems, optimal and adaptive systems, algorithmic methods

Committee	Sessions	Papers
Theory	17	93
Applications	16	86
Systems Engineering	12	66
Space	5	30
Computers	5	29
Components and Instruments	4	21
Economics and Management Systems	2	11
Social Effects of Automation	1	6
Education	1	6

in control, optimization methods, digital adaptive systems, stability, model reference adaptive systems, and computational methods.

The Applications Committee papers were grouped into sessions on identification, power plants (hydro, nuclear and conventional steam, energy transmission), control in chemical, metallurgical, mining, pulp and paper, and electronic component industries, and national support of advanced developments in computers and industrial automation.

The System Engineering Committee's sessions were devoted to large scale systems, complex and hierarchical systems, information systems, man in systems, medical and health care systems, modelling of biological systems, transport systems, water resources, environmental systems, and system reliability.

At the sessions of the Space Committee papers were presented to control of manned orbiting and unmanned space vehichles, auxiliary control devices for space, theoretical investigations in the design and application of space mechanisms, on-board and launch support control methods and computers.

Sessions of the Computers Committee were devoted to the experience in computer control, digital systems, micro- and minicomputers in control systems, software for process control, testing and reliability in computer control.

The Economics and Management Systems Committee held sessions on urban, regional and national planning, and on control and management of human activity systems. A wide variety of problems in the modern control theory was discussed at round tables (organizers in parentheses).

1. "The Effect of Modern Control Theory on Engineering Practice" (R.E. Larson, U.S.A., and K. J. Åstrom, Sweden).

2. "Adaptive Control Today" (I.D. Landau, France, and H. Kaufman, U.S.A.).

3. "Fuzzy Automata and Decision Making" (L.A. Zadeh, U.S.A., and M. M. Gupta, Canada).

4. "Test Case for the Comparison of Identification Methods" (R. Iserman, F.R.G.).

5. "Modern Control Techniques in the Metal Industry" (D. Bristol, U.S.A.).

6. "Large Scale Systems Theory and Decentralized Control" (M. Athans, U.S.A.).

7. "Hardware-Software Tradeoffs in Real Time Applications" (J.D.N. VanWyk, Republic of S. Africa).

8. "Distributed Computer Control Systems" (T.M. Stout, U.S.A.).

9. "Continuing Education in Automatic Control" (P. Eykhoff, Netherlands).

10. "International Aspects in Control Education" (S. J. Kahne, U.S.A.).

11. "Aspects of World Dynamics" (J.F. Coals, England).

12. "Status of Control Terminology and Standards" (H.L. Mason, U.S.A.).

13. "Symbolics of the Sequence Control Systems" (A.E. Hekala, Finland).

14. "Productivity and Its Changes Done by Man" (E.L. Scott, U.S.A.).

Large activity of participants and thorough discussions of each of the topics should be mentioned. The discussions contributed to a better understanding, convergence of viewpoints, developments of common approaches to further advanced studies in this field.

According to the program, morning and afternoon sessions were preceded by 50-minutes plenary sessions where survey papers were presented. The paper by J. W. Forrester (U.S.A.) "World Dynamics" was devoted to the presentation of the fundamentals of national economics modelling. An approach developed at the Massachusetts Institute of Technology involves models of socie-economic systems based on the method of dynamics of averages. The developed model has sectors of industry (at least 14), personnel dynamics, demography, finances and household. For each sector, a broad spectrum of indexes is simulated characterizing its status and development dynamics. For instance, the industrial sector may model production of food, consumer goods, construction, capital investments, agriculture, resources, energy production, transportation, science, state service, military operations, etc. It is anticipated that on the basis of this modelling, in a year or two a model of the USA national economy would be developed, this being a major progress in the large scale systems control. The author believes that the developed model

would enable computer simulation of socio-economical processes going on in USA and study of effects of decisions made at various levels of economic control.

In his paper on "Future of Control" H. M. Rosenbrock (England) surveys the status of the control theory and technology and analyses their possible contributions to the society in future. Since 1945, the development of the control theory has two periods: the period of the "classical" theory (years 1945–1960) and that of "modern" theory (since 1960), the term "modern" theory referring often to formulations and methods allowing algorithmic solution. The gap between the theory and the practice which appeared by the end of the sixties, is attributed by the author to the attempt to reduce engineering problems to mathematical ones having algorithmic solutions. Further, difficulties are analysed involved into actual system design and due to formulation of the final aim, great number of various constraints, limited information about physical systems, and decisions options depending on the accepted tradeoff. This leads the author to the conclusion that the algorithmic, purely mathematical approach cannot give desired results, that the engineering design is an art, rather than a strict science, implying that this is a higher level. Therefore, development of mathematical synthesis methods is less promising than accumulation of data about the basic properties of systems with various structures. On the basis of the analysis, the author concludes that the success in the design of modern systems lies in man (designer)-computer interaction. He discusses two possible ways: 1) man-machine interaction making the most of their possibilities, and 2) division and algorithmization of design on the basis of knowledge and experience of designers where design is reduced to a sequence of simple decisions. The author regards the first way as more effective, cheap and bringing greater satisfaction, he calls not to lose faith in human capabilities, to make computer to analyse various situations and to leave the final decisions to man.

"Systems Engineering Methodology" was the theme of a survey by P. Faure (France) The author separates the following periods in the history of the systems engineering:

— before 1930 when it was an art, rather than an engineering discipline;

— 1930 to 1950, a period of the classical control theory and of parallelly developing operations research when application areas of these theories were disjoint;

— 1950 to 1970, a period of computers, of closer contacts between the automatic control theory and operations research, and of the advent of analytical methods of design of automatic controls; and

— since 1970, the period of complete recognition of the systems engineering approach.

According to the author's opinion, totality of the following four approaches defines the systems engineering methodology for system design:

1) experimental method of successive trials;

2) successive trials based on system modelling;

3) analytical method of successive trials which is characteristic of the classical automatic control theory; and

4) analytical and optimal design.

The author emphasises also the importance of the systems engineering methodology in curriculum of studies.

The joint paper by K. S. Bibby (England), I. M. Makarov (U.S.S.R.), F. Margulis (Austria), J. E. Pijnsdorp (The Netherlands) and R. M. J. Withers (England) "Man's Role in Control Systems" gives a survey of current works on sociological and psychological aspects of using man in control systems. The paper is accompanied by a detailed bibliography of the latest 54 publications in this field and by two appendices. The first appendix gives a detailed description of particular designs of automatic control systems as regards the functions of human operator. A detailed design sequence is presented along with a multi-criteria estimate showing to what extent the psycho-physiological requirements of human operator are satisfied. The second appendix contains the final statement of the members of the IFAC working group on social effects of automation adopted in Federal Republic of Germany in 1974.

Various scales and tables presented and analysed in the paper enable classification of the work done by human operator in control systems with various degrees of mechanization and automation. It is emphasised that the widely spread attitude toward the role of human operator in automatic control systems as vanishing is delusive.

Great attention is paid to the man-computer dialogue during observations of a process, distribution of functions between them under various situations, and to rational information representation to the operator. It is indicated that the modern control theory should develop with allowance for practical needs. To this end, not only control in closed loop, but interaction of multiple closed control loops of one control plant coupled by human operator, his role in control systems under different modes of operation, reasonable level of human operator's acquaintance with computerized control algorithms, etc. should be also studied. Major stages of man-machine system design are de-

scribed, those elements of the system are indicated which are to be selected and which influence essentially human operator functions, forms of his dialogue with the computer and the control plant, etc. Problems arising in introduction of such systems are also discussed from the viewpoint of human operator.

Contributions of specialists on control, sociologists, psychologists, and trade union workers to comprehensive rational organization of control systems and the form of human activity in them are outlined. As the authors indicate, the modern control theory allows for the human factor insufficiently, although it is of prime importance.

In a paper devoted to "National Economic Management" L. R. Klein discusses the technical aspects of design and application of large economic control systems. Econometric problems are described from the viewpoint of specialists on systems engineering and automatic control. Consideration is given to the general form of a set of equations describing a large scale economic system. There are two versions of these equations: one of them is the group of political control instruments (budget, taxes, control of trade, credit regulations), the other one is the group of environmental actions upon economics (earthquakes, amount of precipitation, plague). A number of constraints on the set of equations (limited number of variables, linearization of equations, introduction of statistical disturbances, etc.) is described enabling solution of the set on present-day computers. The paper emphasises not only the development and study of a national economy model, but its practical applications to various economic problems, including that of economic index forecasting for various time intervals. Common points in approaches and problems of econometrics and the control theory are demonstrated. Major difficulties involved into national economy model are analysed, and urgent problems are pinpointed whose solution would result in more precise design and deeper analysis of the large scale economic models. These problems are shown to be urgent for the control theory as well.

D. E. Okhotsimskii (U.S.S.R.) and M. Vukobratovic (Yugoslavia) in their paper on "Control of Legged Locomotion Robots" classify them into two groups, with and without internal stability. Robots with six legs having multi-level hierarchical control system are presented as an example of first-group robots. This system manifests some features of artificial intelligence. The second group was exemplified by a two-legged moving system. For such artificial two-legged "walkers", a new synthesis approach is suggested to control algorithms for small and great system disturbances.

Although the paper sums up the results of only three years' work, discussion is fairly detailed and makes use of conventional nonlinear differential equations.

For six-legged walkers general robot configuration, problems in kinematics control, dynamics of six-legged walkers, movement control system structure, kinematics of separate parts of the mechanism, block-diagram of control process modelling, stabilization systems and experimental results are considered.

For two-legged walkers, control at various levels is discussed (level of normal dynamics, and that of adaptation). Synthesis of two-legged walkers and stabilization against disturbances by means of inclining the upper part of the body in various directions are considered. Movement of numerous types of surfaces has been studied.

The chair of the plenary session devoted to the day of cosmonautics was occupied by Academician B. N. Petrov (U.S.S.R.). Pilot-cosmonaut of U.S.S.R. N. N. Rukavishnikov was among the speakers.

The paper by P. L. Reichetz (F.R.G.) "Hospital and Health Care Systems" gives a detailed description of problems arising in medical service systems. The major difficulties here are high dimensionality, and also the problem of appropriate criterion. The paper discusses various indexes: expected time of life, infantine death rate, speed of dying, speed of birth, mean time for treatment, mortality rate, illness rate, etc.

The paper was mostly descriptive; however, it is of interest, especially because it outlines a particular medical system in Hannover.

W. A. Hall shows in his paper that during millions of years men tried to control the environment. Men have a unique feature of being able to construct conceptual models of the major characteristics of their environment. The man also has a unique ability to find out physical actions directed toward modification of these characteristics with the aim of creating a situation optimal for himself.

To attain desired changes man used three basic concepts; namely, the concept of prediction model, the concept of corresponding resources and deliberate actions making use of them and oriented toward positive changes predicted by the model. Prediction model enables approximate determination of the environment response to possible actions. Moreover, other models may be used to establish how one can make use of resources in order to make the environment to "behave" optimally.

Some shortcomings of modelling are discussed which present more precise analysis of social systems.

Models are classified by the author as mental (direct human perception of the situation), descriptive (in the form of oral or written description), descriptively qualitative, and mathematical.

Among shortcomings of the popular now mathematical modelling the author sees to great number of factors which should be taken into consideration, most important factors being often omitted because they do not yield to mathematical analysis. Another shortcoming is the necessity to consider a number of hierarchies of descriptive and qualitative "substitutes" prior to development of mathematical model. Neglect of the lower levels of multi-level social objects is a most dangerous mistake involved into modelling of social control systems.

A third shortcoming closely related to the second one, is the unability of the mathematical model to maximize or minimize more than one aspect of a social system, although it is essentially of multi-aspect. Author's reasoning is supported by a number of particular examples of controlling action upon the environment, including all those changes in the status of the universe which are important for the majority of social objects disregarded by the so called "optimal control". The author believes that the honest consideration of the today's status of "optimal control" policy with the aid of scientific mathematical modelling and allowance for the limited and non-exact nature of optimality criteria would make it clear that if we do not learn to integrate descriptive and estimating models with mathematical ones, amount and intensity of harmful effects of our control actions upon the environment may in the future grow dramatically as compared with the past periods.

Limited volume of this publication prevents exposition of numerous papers presented. The above list of sessions evidences greater percentage of papers on applications, large and hierarchical systems, computers and space. It is due primarily to the fact that IFAC closely follows changes in the area of control, it is leader and influences, to a large measure, the level of control theory and practice. Papers read at the Congress generalize the major results and directions of studies in the field of control all over the world. These results are demonstrations that theoretical studies become wider and deeper; number of applications grows as well as the number of scientific results which are presented as computer algorithms; and analysis and design methods for large scale integrated systems with hierarchical structure are being developed. Introduction of adaptive systems, integrated systems, control systems for auxiliary plants should be mentioned primarily among the practical results. I would like to emphasize the great scope of works on the theory and practice of control during recent years. IFAC President J. C. Lozier said at the opening ceremony that in spite of the rigorous selection of papers, the program committee has still more than 350 acceptable papers among those that were rejected.

It is the merit of the Organizing Committee that the Congress worked smoothly, all the planned events took place, the working atmosphere was fine, and the information service was perfect. All these are undoubtedly strenuous for the organizers.

Elective organs of IFAC had their meetings during the Congress. Dr. U. A. Luoto representing Finland, the host country of the next, Seventh IFAC Congress, was elected President.

N. S. Raibman

BOOK REVIEW

N. AHMED and K. R. RAO: Orthogonal Transform for Digital Signal Processing (Springer Verlag, 1975, 263 pp)

The orthogonal transforms play an important role in the field of digital data processing. Their application may result in a significant reduction of the required computer memory.

A transform of a sampled analog signal (or sequence) consisting of N points results in another sequence of identical size. If among these we neglect the points of relatively small value we can get after the execution of the inverse transformation a good approximation of the original sequence. This procedure is called data compression, and it is very useful if the storage of the sequence is necessary.

The book consists of 10 chapters.

Chapter 1 is a short introduction with general remarks and terminology.

Chapter 2 reviews the Fourier methods of signal representation which play an important role in engineering. It gives the connection between the Fourier series and Fourier transformation. Finally, some aspects of convolution and correlation are discussed.

Chapter 3 deals with the Fourier representation of sequences by means of the Discrete Fourier Transform (DFT). The discrete analogs of convolution and correlation and their corresponding theorems are developed and the spectral properties of e.g. amplitude, phase and power spectra are also described.

The transformation with the aid of a digital computer may be executed easily. The fast algorithm of the Fourier transform (FFT) increases its efficiency further. The main objective of Chapter 4 is to develop FFT. The algorithm significantly reduces the number of arithmetic operations and the memory required to compute DFT (or its inverse).

Chapter 5 introduces a class of nonsinusoidal orthogonal functions as Rademacher, Haar and Walsh–Hadamard functions. These functions consist of square or rectangular waves. The functions of these sets are distinguished by a parameter called sequency or generalized frequency. The main advantage of the transforms based on these functions as compared with the Fourier transform is the use of real operations among the points.

Chapter 6 is devoted to the study of the Walsh–Hadamard transform (WHT), which is perhaps the best known nonsinusoidal orthogonal transformation. The transform needs only addition and subtraction among the points. Fast algorithms for the computation of the WHT (FWHT) are developed and the notation of Walsh spectra is introduced.

Chapter 7 introduces some miscellaneous orthogonal transforms. These are the generalized, Haar, slant and the discrete cosine transform. The WHT and the DFT are two special cases of the generalized transform. The motivation for studying the Haar, slant and discrete cosine transform is that they are useful for certain applications, some of which are discussed in Chapter 9. Their mathematical description is based on the matrix factorization.

The first application studied in Chapter 8 concerns the classical signal processing technique known as Wiener filtering. It is shown that orthogonal transforms may be used to extend Wiener filtering to the processing of discrete signals with emphasis on the reduction of computational requirements. Orthogonal transforms may also be used for applications in the area of image processing.

As it was mentioned before an important application of the orthogonal transforms is data compression. Chapter 9 deals with this topic. The different transforms may be compared by some error criteria, most often by the mean square error criterion. The data

compression is illustrated by examples pertaining to the processing of electrocardiographic and image data.

In Chapter 10 the possibility of applying the orthogonal transforms for pattern recognition is studied, some simple classification algorithms and their implementations are shown. The results prove that orthogonal transform processing enables substantial dimensionality reduction with only a small percentage decrease in correct classification. It is easy to note the close connection between dimensionality reduction and the principle of data compression.

The book contains many references after each chapter for those who are interested in further study. It gives also FORTRAN programs for the execution of the transformations. Many problems are given after each chapter to bring the thoughts presented close to the reader.

The book is perhaps the first summarizing the knowledge in this field. Besides the theory it shows also practical applications and induces for problem solving. This book is recommended to workers in the control engineering and communication field as a source of ideas and methods.

Ruth Bars and *László Kónya*

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY, VOL. 5 (1)
ENGLISH TRANSLATION OF THE COMMUNICATIONS IN RUSSIAN

A GENERALIZATION OF A CONTROL SYSTEM ANALYTICAL DESIGN PROBLEM

A. A. KRASOVSKII
(Moscow)

The control plant is described in the operator-differential form involving, along with the derivative of the main coordinate vector with respect to time, functionals of the vector and control included in the linear form. The basic theorem of analytical design with respect to the generalized work criterion is shown to be applicable to such plants with the only difference that functionals are substituted for functions (plant characteristics). The iteration method is recommended to solve the defining partial differential equation with functionals as coefficients of the derivatives. For the linear stationary plant represented in the operator-differential form, a general precise solution is found for stationary control synthesis. The major merits of the method are better demonstrated in the case of multidimensional multivariate plants.

The following difficulties are involved into the conventional, well developed form of the control system analytical design method. Dimensionality of the equation set defining coefficients of optimal controls is large.

As a rule, many components of the complete state vector are not measurable directly (incomplete direct observability). Difficulties involved into the implementation of optimal controls with large number of components occur also if the generalized work criterion is applied [1, 2], although this version of the analytical design is the most effective one. To overcome the difficulties, the present paper proposes an operator-differential form of control plant representation dictating differential-operator form of all subsequent constructions.

1. Nonlinear plant

Represent the control plant as

$$\dot{x}_i + F_i(x_1, \ldots, x_n) = \sum_{j=1}^{m} \varphi_{ij}(x_1, \ldots, x_n)\, u_j\,, \qquad (1)$$

$$(i = 1, 2, \ldots, n),$$

where F_i, φ_{ij} are operators affecting vector $x = (x_1, \ldots, x_n)$ (functionals), u_j are controls to be synthesized. Equation (1) reminds the differential Cauchy equation only in appearance as they involve functionals rather than functions. Functionals F_i, φ_{ij} may be integral, (1) being in this case in the form of a set of integro-differential equations. Functionals F_i, φ_{ij} may be in the form of transfer functions with linear input variables $x_1, \ldots, x_n$ (linear version) or nonlinear input functions of these variables (nonlinear version), etc. In the

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

particular case where functionals F_i, φ_{ij} degenerate into functions, set (1) degenerates into differential equations in the Cauchy form. Vector $x = (x_1, \ldots, x_n)$ referred to below as the main vector may be regarded as a projection the complete plant state vector whose dimensionality may be much more than n or even infinite. The choice of the main vector is dictated by the following conditions. All components $x_1, \ldots, x_n$ should be directly measurable most important plant coordinates in terms of which requirements to the quality of control may be formulated.

The following statement is true which is a complete analog of the basic theorem of analytical design with respect to the generalized work criterion [2]: for (1), the optimal controls minimizing functional

$$I = V_3[x_1(t_2), \ldots, x_n(t_2)] + \int_{t_1}^{t_2} Q(x_1, \ldots, x_n)\, dt + \frac{1}{2} \int_{t_1}^{t_2} \sum_{j=1}^{m} \frac{u_j^2 + u_{j\text{on}}^2}{k_j^2}\, dt, \qquad (2)$$

are as follows

$$u_j = u_{j\text{on}} = k_j^2 \sum_{k=1}^{n} \varphi_{kj} \frac{\partial V}{\partial x_k}, \qquad (3)$$

where V is solution of equation

$$\frac{\partial V}{\partial t} - \sum_{i=1}^{n} F_i \frac{\partial V}{\partial x_i} = -Q \qquad (4)$$

under boundary condition $V_{t=t_2} = V_3$; V_3, Q and k_j are given functional, function and coefficients, respectively. The only difference from the conventional formulation of the theorem is that V_3, F_i, φ_{kj} in (1) and (4) are operators over vector x rather than over functions. The proof may follow different ways including one completely similar to that used in [1, 2]. Indeed, by integrating from t_1 to t_2 complete derivative V which is as follows

$$\dot{V} = \frac{\partial V}{\partial t} + \sum_{i=1}^{n} \frac{\partial V}{\partial x_i} \left(-F_i + \sum_{j=1}^{m} \varphi_{ij} u_j \right) = -Q + \sum_{i=1}^{n} \sum_{j=1}^{m} \varphi_{ij} \frac{\partial V}{\partial x_i} u_j,$$

in virtue of (1) and (4), obtain that

$$I = V_{t=t_1} + \sum_{i=1}^{n} \sum_{j=1}^{m} \int_{t_1}^{t_2} \varphi_{ij} \frac{\partial V}{\partial x_i} u_j\, dt + \frac{1}{2} \sum_{j=1}^{m} \int_{t_1}^{t_2} \frac{u_j^2 + u_{j\text{on}}^2}{k_j^2}\, dt =$$

$$= V_{t=t_1} + \frac{1}{2} \int_{t_1}^{t_2} \sum_{j=1}^{m} \frac{(u_{j\text{on}} - u_j)^2}{k_j^2}\, dt.$$

It may be directly seen from this expression that I is minimal for $u_j = u_{j\,\mathrm{on}}$.

To find optimal controls $u_{j\,\mathrm{on}}$, it is necessary to solve (4) which in this case is operator-(partial)-differential equation. It is natural to tackle this problem by means of iterative or operational method for solving linear partial differential equations [3, 4]. In doing so, it is expedient, as usual, to distinguish terminal and nonterminal (stationary) problems [2]. For terminal problem with criterion (2), solution is looked for in the form of $V = V_1 + V_2 \ldots$, where

$$\frac{\partial V}{\partial t} = -Q, \quad \frac{\partial V_{\nu+1}}{\partial t} = \sum_{i=1}^{m} F_j \frac{\partial V_\nu}{{}_i\partial x_i}, \qquad \nu = 1, 2, \ldots.$$

and

$$(V_1)_{t=t_2} = V_3, \quad (V_{\nu+1})_{t=t_2} = 0.$$

If series $V_1 = V_2 + \ldots$ converges uniformly to its first derivatives, it converges to the solution of (4), that may be verified by simple summing. With the notation of operators as below

$$D = \frac{\partial}{\partial t}, \quad \frac{1}{D} = \int [\ ]\, dt, \quad P = \frac{1}{D} \sum_{i=1}^{n} F_i \frac{\partial}{\partial x_i},$$

terminal control assumes the following form

$$V = -(1 + P + P^2 + \ldots) \frac{Q}{D}. \tag{5}$$

In the majority of cases, however, the problem of analytical design of optimal controls is formulated as nonterminal one. To obtain the corresponding solution, make use of the nonstationary functional method [2]. Give function Q as $Q = Q_* \exp(-\lambda t)$, where Q is a positively defined function of x; λ is a real constant. It is expedient in this case to represent V as

$$V = V_* \exp(-\lambda, t).$$

Equation (4) is substituted by the following one:

$$\frac{\partial V_*}{\partial t} - \lambda V_* \sum_{i=1}^{n} F \frac{\partial x_i}{\partial V} = -Q_*.$$

For the stationary case where V_* is not explicitly time-dependent

$$\frac{\partial V_*}{\partial t} = 0 \quad \text{and} \quad \lambda V_* + \sum_{i=1}^{n} F_i \frac{\partial V_*}{\partial x_i} = -Q_*.$$

Denote operator

$$L = \sum_{i=1}^{n} F_i \frac{\partial}{\partial x_i},$$

and represent iterative solution as

$$V_* = \frac{1}{\lambda}\left(1 - \frac{1}{\lambda}L + \frac{1}{\lambda^2}L^2 - \ldots\right) Q_*. \tag{6}$$

In virtue of (3), the optimal controls will be as follows

$$u_j = -k_j^2 \sum_{k=1}^{n} \varphi_{kj} \frac{\partial V}{\partial x_k} = k^2 \exp(-\lambda t)\, R_j\, V_*,$$

where

$$R_j = \sum_{k=1}^{n} \varphi_{kj} \frac{\partial}{\partial x_k}.$$

In order that the optimal controls do not depend explicitly on time, take that

$$k_j^2 = k_{j*}^2 \exp(\lambda t),$$

where k_{j*} are constants.

The following result may be, thus, formulated: if series (6) uniformly converges to its first derivatives, then controls

$$u_j = u_{j\,\text{on}} = -\frac{k_{j*}^2}{\lambda} R_j \left(1 - \frac{2}{\lambda}L + \frac{1}{\lambda^2}L^2 - \ldots\right) Q_* \tag{7}$$

are, for the stationary control plant, nonterminal optimal stationary controls minimizing functional

$$I_* = V_* \exp(-\lambda t) + \int Q_* \exp(-\lambda t)\, dt + \frac{1}{2}\int_{j=1}^{m} \frac{u_j^2 + u_{j\,\text{on}}^2}{k_{j*}^2} \exp(-\lambda t)\, dt.$$

To enhance the rate of convergence, one has to increase λ. It should be, however, kept in mind that with the unlimited increase of λ, effectiveness of criterion I_* decreases because for $\lambda \to \infty$ it becomes an estimate of instantaneous values of coordinates and controls instead of being an integral criterion.

2. Linear stationary plant. Quadratic criteria

For the linear stationary plant equations are as follows:

$$\dot{x}_i + \sum_{k=1}^{n} W_{ik}(D)\, x_k = \sum_{j=1}^{m} b_{ij}\, u_j, \quad i = 1, 2, \ldots, n, \tag{8}$$

where $W_{ik}(D)$ are transfer functions/linear operators, and b_{ij} are constant coefficients.

The following quadratic criterion is given

$$Q_* = \frac{1}{2}\sum_{\nu\mu=1}^{n} \beta_{\nu\mu}\, x_\nu x_\mu, \quad \beta_{\nu\mu} = \beta_{\mu\nu},$$

where $\beta_{\nu\mu}$ are constant coefficients. Here

$$L = \sum_{i,k=1}^{n} [W_{ik}(D)\, x_k] \frac{\partial}{\partial x_i} \qquad R_j = \sum_{k=1}^{n} b_{kj} \frac{\partial}{\partial x_k}.$$

Taking into account that

$$\frac{\partial}{\partial x_\mu} \sum_{k=1}^{n} W_{ik}(D)\, x_k = W_{i\mu}(D)\, 1 = W_{i\mu}(0)$$

is the stable value of the operator* affecting the constant, find consequtively

$$LQ_* = \sum_{i,k,j=1}^{n} \beta_{ij} [W_{ik}(D)\, x_k]\, x_j = x^{\mathrm{T}} \beta W(D)\, x,$$

$$L^2 Q_* = x^{\mathrm{T}} \beta W(0)\, W(D)\, x + [W(D)\, x]^{\mathrm{T}}\, \beta W(D)\, x,$$

$$L^3 Q_* = [W(D)\, x]^{\mathrm{T}}\, [\beta W(0)\, W(D)\, x + \beta W^2(0)\, x + 2W^{\mathrm{T}}(0)\, \beta W(D)\, x],$$

. . . . . .

Here $W(D) = \| W_{ik}(D) \|$ is a quadratic matrix of transfer functions; $\beta = \| \beta_{ik} \|$ is a quadratic matrix of criterion coefficients; k is a diagonal matrix of gains k_{j*}^2; b is a rectangular matrix of coefficients b_{ij}; and x is a column vector. Symbol "T" stands for transposition. By substituting these notations into (7) obtain

$$\begin{aligned} u_{\text{оп}} = -\frac{1}{\lambda} k^2 b^{\mathrm{T}} \Big\{ & \beta - \frac{1}{\lambda} [\beta W(D) + W^{\mathrm{T}}(0)\, \beta] + \\ & + \frac{1}{\lambda^2} [\beta W(0)\, W(D) + 2W^{\mathrm{T}}(0)\, \beta W(D) + W^{\mathrm{T}2}(0)\, \beta] - \\ & - \frac{2}{\lambda^3} [\beta W(0)\, W(D) + 3W^{\mathrm{T}}(0)\, \beta W(0)\, W(D) + \\ & + 3W^{\mathrm{T}2}(0)\, \beta W(0) + W^{\mathrm{T}3}(0)\, \beta] + \ldots \Big\} x. \end{aligned} \tag{9}$$

If this equation is written as

$$u_{\text{оп}} = -\frac{1}{\lambda} k_*^2 b^{\mathrm{T}} \left[A_1 - \frac{1}{\lambda} A_2(D) + \frac{1}{\lambda^2} A_3(D) - \ldots \right] x, \tag{10}$$

then $A_1 = \beta$, and each subsequent operator $A_\nu(D)$ ($\nu = 2, 3, \ldots$) may be obtained from its predecessor through a simple recurrent formula

$$A_\nu(D) = W^{\mathrm{T}}(0)\, A_{\nu-1}(D) + A_{\nu-1}(0)\, W(D). \tag{11}$$

* If $W_{i\mu}(0) = \infty$ and consideration is given to nonstationary controls, notation $W_{i\mu}(0)$ should be understood as a conditional one, equivalent to $W_{i\mu}(D)\, 1$.

The series inclosed in braces in (9) or in square brackets in (10) may be readily shown to be an iterative solution of the following equation

$$\lambda A(D) + W^{\mathrm{T}}(0)\, A(D) + A(0)\, W(D) = \beta. \tag{12}$$

if multiplied by λ^{-1}.

Matrix $A(0)$ may be determined from (12) for $D = 0$:

$$\lambda A(0) + W^{\mathrm{T}}(0)\, A(0) + A(0)\, W(0) = \beta. \tag{13}$$

Note that if the control plant is represented in the traditional form of $\dot{x} + ax = bu$ where a is the matrix of constant coefficients, then $W(D) = W(0) = a$ and (12) and (13) turn into the well-known equation:

$$\lambda A + a^{\mathrm{T}} A + Aa = \beta.$$

Equations (12) and (13) defining optimal control operator $A(D)$ may be determined differently. We shall look for a solution of equation

$$\lambda V_n + \sum_{i=1}^{n} F_i \frac{\partial V_*}{\partial x_i} = \lambda V_* + \sum_{i,k=1}^{n} [W_{ik}(D) x_k] \frac{\partial V_*}{\partial x_k} = \frac{1}{2} \sum_{i,k=1}^{n} \beta_{ik} x_i x_k \tag{14}$$

in the form of

$$V_* = \frac{1}{2} \sum_{i,k=1}^{n} A_{ik}(D)\, (x_i x_k).$$

By substituting this expression into (14), differentiating one time with respect to x and equating to zero the operator over the same vector, obtain (12). Second differentiation results in (13). Thus, the following result may be formulated for the linear stationary control plant; if a control plant is given

$$\dot{x} + W(D)\, x = bu, \tag{15}$$

then stationary control minimizing criterion

$$\begin{aligned} I_* = V_* \exp(-\lambda t) + \frac{1}{2} \int (x^{\mathrm{T}} \beta x) \exp(-\lambda t)\, dt + \\ + \frac{1}{2} \int (u^{\mathrm{T}} k_*^{-2} u + u_{\mathrm{on}}^{\mathrm{T}} k_*^{-2} u_{\mathrm{on}}) \exp(-\lambda t)\, dt \end{aligned} \tag{16}$$

is as follows

$$u = u_{\mathrm{on}} = -k_*^2\, b^{\mathrm{T}} A(D)\, x, \tag{17}$$

where operator $A(D)$ is defined by Eqs (12) and (13).

Note that a similar problem might be solved, in principle, in the traditional manner. Indeed, Eq. (15) may be usually reduced to a set of first-order differential equations (the Cauchy form) by introducing the complete state

vector. By making use of functional (16) which is of the same form, synthesizing the optimal control in the traditional manner and excluding from the control by means of transfer functions all complete state vector components save the main coordinates, obtain control in the form of (17). However, this way, as it was noted above, is too cumbersome for multidimensional plants. Moreover, the results of the two synthesis methods may be basically different. Indeed, in the minimized functional (16) only one component is prescribed, the rest of them are obtained automatically and may differ in different solutions.

In particular cases, relations (12), (13) and (17) may be even more developed. For example, if $n = 1$, i.e. if there is only one main coordinate, the plant equations and expressions (12), (13) and (17) take the scalar form

$$\dot{x}_1 + W_{11}(D)\, x_1 = u_1,$$

$$\lambda A_{11}(D) + W_{11}(0)\, A_{11}(D) + A_{11}(0)\, W_{11}(D) = \beta_{11},$$

$$\lambda A_{11}(0) + 2W_{11}(0)\, A_{11}(0) = \beta_{11},$$

$$u_{1\text{оп}} = -k_1^2 A_{11}(D)\, x.$$

Hence, obtain

$$u_{1\text{оп}} = -\frac{\beta_{11} k_1^2}{1 + W_{11}(0)} \left[1 - \frac{W_{11}(D)}{1 + 2W_{11}(0)}\right] x.$$

In the explicit form, the general solution may be written also for the case where $W(0) = 0$. In this case according to (12) and (13)

$$A\left(D = \frac{1}{\lambda}\left[\beta - \frac{1}{\lambda}\beta W(D)\right]\right)$$

and

$$u_{\text{оп}} = -\frac{1}{\lambda} k_*^2 b^{\mathrm{T}} \left[\beta - \frac{1}{\lambda}\beta W(D)\right] x.$$

This formula results also from (9) if we take that $W(0) = 0$.

Example. The linearized equations of the longitudinal motion of an aircraft are (under constant engine thrust and disregarding the changes of air density with altitude) as follows [2]:

$$\Delta\dot{V} + a_x^V \Delta V + a_x^\alpha \Delta\alpha + a_x^\theta \Delta\theta = 0,$$

$$\Delta\dot{\theta} + a_y^\alpha \Delta\alpha + a_y^V \Delta V + a_y^\theta \Delta\theta = 0,$$

$$\Delta\dot{\omega}_z + a_{mz}^{\omega z} \Delta\omega_z\, a_{mz}^\alpha \Delta\alpha + a_{mz}^V \Delta V = a_{mz}^{\delta b} \Delta_b^\delta,$$

$$\Delta\vartheta - \Delta\omega_z = 0,\ \Delta n_y = \frac{V_0}{T_\theta}\Delta\alpha,\ \Delta\vartheta = \Delta\theta + \Delta\alpha.$$

where ΔV, $\Delta\theta$, $\Delta\omega_z$, $\Delta\vartheta$, $\Delta\alpha$, Δn_y, $\Delta\delta_b$ are increments of speed, trajectory angle, pitch angular speed, pitch angle, angle of incidence, normal acceleration and elevator displacement, respectively. "a" and "T_0" with appropriate indexes stand for coefficients which will be assumed constant. Introduce notations

$$\Delta\dot{n}_y = x_1,\ \Delta n_y = x_2;\ b_{11} = \frac{V_0\, a_{mz}^{\delta b}}{T_0},\ \Delta\delta_b = u_1$$

$$W_{11}(D) = a_{mz}^{\omega z} - \frac{a_{mz}^{V} a_x^{\alpha} + (D + a_{mz}^{\omega z})\,[a_y^{\alpha}(D + a_x^{V}) - a_x^{\alpha} a_y^{V}]}{(D + a_x^{\nu})\,(D + a_y^{\theta}) - a_x^{\theta} a_y^{\nu}}, \tag{19}$$

$$W_{12}(D) = a_{mz}^{\alpha} + a_{mz}^{V}\,\frac{a_y^{\alpha} a_x^{\theta} - a_x^{\alpha} a_y^{\theta}}{(D + a_x^{V})\,(D + a_y^{\theta}) - a_x^{\theta} a_y^{V}}$$

and transform (18) as follows

$$\dot{x}_1 + W_{11}(D)\, x_1 + W_{12}(D)\, x_2 = b_{11} u_1,$$

$$\dot{x}_2 - x_1 = 0$$

or

$$\dot{x} + W(D)\, x = bu,$$

where

$$W(D) = \left\| \begin{matrix} W_{11}(D) & W_{12}(D) \\ -1 & 0 \end{matrix} \right\|,\quad b = \left\| \begin{matrix} b_{11} \\ 0 \end{matrix} \right\|.$$

Optimal control u_1 is to be synthesized providing optimal transient conditions with respect to g-load or, to put it more precisely with respect to the main coordinates $x_1 = \Delta\dot{n}_y$, $x_2 = \Delta n_y$. Q is given in the following form

$$Q_* = \frac{1}{2}\,(\beta_{11} x_1^2 + \beta_{22} x_2^2).$$

In the developed form, (12) and (13) are as follows:

$$\begin{cases} [\lambda + W_{11}(0)\, A_{11}(D) - A_{12}(D) - A_{12}(0) = \beta_{11} - A_{11}(0)\, W(D), \\ [\lambda + W_{11}(0)]\, A_{12}(D) - A_{22}(D) = -A_{11}(0)\, W_{12}(D), \\ W_{12}(0)\, A_{12}(D) + \lambda A_{22}(D) = \beta_{22} - A_{21}(0)\, W_{12}(D), \end{cases} \tag{20}$$

$$\begin{cases} [\lambda + \sqrt{2}\, W_{11}(0)]\, A_{11}(0) - 2A_{12}(0) = \beta_{11}, \\ W_{12}(0)\, A_{11}(0) + [\lambda + W_{11}(0)]\, A_{12}(0) - A_{22}(0) = 0, \\ 2W_{12}(0)\, A_{12}(0) + \lambda A_{22}(0) = \beta_{22}. \end{cases} \tag{21}$$

With regard to (17), the optimal control has the following form

$$U_1 = U_{1\text{оп}} = -k_{1*}^2\, b_{11}[A_{11}(D)\, \Delta\dot{n}_y + A_{12}(D)\, \Delta n_y].$$

Line 1 in the figure corresponds to the transient regime with respect to g-load without control u. The process was computed for the following values of parameters

$$a_x^V = 0.017 \text{ sec}^{-1},\ a_x^\alpha = 0.15 \text{ sec}^{-2},\ a_x^\theta = 0.17 \text{ msec}^{-2},$$

$$a_y^\alpha = -0.84 \text{ sec}^{-1},\ a_y^V = -0.003 \text{ m}^{-1},\ a_y^\theta = 0,$$

$$a_{mz}^{mz} = 0.66 \text{ sec}^{-1},\ a_{mz}^\alpha = 55 \text{ sec}^{-1},\ a_{mz}^V = 0.022 \text{ m}^{-1}\text{sec}^{-1},$$

$$a_{mz}^{\delta b} = -22.5 \text{ sec}^{-2},\ V_0 = 575 \text{ msec}^{-1},\ T_\theta = 1.7 \text{ sec}.$$

For these values and for $\lambda = -1$, $\beta_{11} = 10$, $\beta_{22} = 100$ the optimal control was determined by formulae (19), (20) and (21):

$$u_1 = 3.310^{-3}\left(\frac{0.506\,D^2 + 0.00277\,D + 0.00013}{D^2 + 0.017\,D + 0.00051}\,\Delta\dot{n}_y + \right.$$

$$\left. + \frac{1.33\,D^2 + 0.0226\,D + 0.0006}{D^2 + 0.017\,D + 0.00051}\,\Delta n_y\right) \approx$$

$$\approx 3.310^{-3}\left(\frac{0.506\,D + 0.0028}{D + 0.017}\,\Delta\dot{n}_y + \frac{1.33\,D + 0.023}{D + 0.017}\right)\Delta n_y.$$

The transient regime in the closed-loop system corresponding to this control is represented by line 2.

The computational merits of the proposed "expanded method of analytical design with respect to the generalized work criterion" are essential for multidimensional multivariate control plants.

References

1. *Krasovskii, A. A.*, Analytical Design of Flying Vehicle Control Loops, Mashinostroenie Publ., 1969 (in Russian).
2. *Krasovskii, A. A.*, Automatic Flight Control Systems and Their Analytical Design, Nauka Publ., 1973 (in Russian).
3. *Bukov, V. N., Krasovskii, A. A.*, "An Operational Algorithm of Optimal Control", Avtomatika i Telemekhanika, **10,** 1974 (in Russian).
4. *Krasovskii, A. A., Bukov, V. N.*, "Iterative Solution of the Kolmogorov Equation in the Statistical Dynamics of Continuous Systems", Dokl. Akad. Nauk SSSR. **216,** p. 485, 1974 (in Russian).

FUNDAMENTAL FINITE CONTROL

A. G. BUTKOVSKII, M. I. MUSTAFAEV
(Moscow, Baku)

The notion of fundamental finite control and some of its modifications are introduced for linear controllable systems with distributed parameters. Analytical formulae are derived for representation of finite controls belonging to the corresponding class through fundamental finite controls. Recalculation formulae are presented relating various modifications of fundamental finite controls.

The theory of linear equations, and, in particular, that of linear differential equations, has a very important notion of *fundamental solution* [1]. Roughly speaking, a solution through which any other solution may be expressed by means of some linear operation (like linear combination, convolution, etc.) is fundamental.

Consider, for example, linear operator equation

$$l(Q(x,\ t)) = f(x,\ t) \tag{1}$$

where operator l is defined by some problem of the mathematical physics having the following form:

$$\sum_{k=0}^{m} \sum_{s=0}^{n} A_{ks}(x, t) \frac{\partial^{k+s} Q}{\partial t^k\, \partial x^s} = f(x, t),\ x_0 < x < x_1,\ t > t_0 \tag{2}$$

and uniform boundary conditions:

1. uniform boundary conditions

$$\sum_{i=0}^{1} \sum_{k=0}^{m-1} \sum_{s=0}^{n-1} a_{Viks}^{(t)} \frac{\partial^{k+s} Q}{\partial t\, \partial x^s}\bigg|_{x=x_i} = 0,\ t > t_0,\ V = 1,2, \ldots \leq n \tag{3}$$

2. zero initial conditions:

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = 0,\ x_0 < x < x_1,\ k = 0,1, \ldots, m-1\,. \tag{4}$$

Note that, generally speaking, the non-uniform boundary conditions resolve themselves into uniform ones by means of appropriate modification of function $f(x,\ t)$ (see Appendix 2).

Fundamental solution $\mathcal{C}(x,\ t;\ \xi,\ \tau)$ of operator l is the solution of operator equation (1) with the right side (i.e. function $f(x,\ t)$) being δ-function, i.e.

$$f(x,\ t) = \delta(x - \xi)\,\delta(t - \tau). \tag{5}$$

Knowing fundamental solution $\mathcal{C}(x,\ t;\ \xi,\ \tau)$ of operator l, one may easily obtain any other solution corresponding to arbitrary right side $f(x,\ t)$ by applying an integral operator with kernel $\mathcal{C}(x,\ t;\ \xi,\ \tau)$:

$$Q(x,t) = \int_{t_0}^{t}\int_{x_0}^{x} \mathcal{C}(x,t;\,\xi,\tau)\,f(\xi,\tau)\,d\xi\,d\tau = l^{-1}(f(x,t))\,. \tag{6}$$

Fundamental solution $\mathcal{C}(x,\ t;\ \xi,\ \tau)$ is often referred to as Green's function of operator l, and disturbance of type (5) is referred to as *Green's disturbance.*

A similar situation arises in solving the controllability problem for some linear controllable system with lumped, discrete and distributed parameters. By analogy to the notion of fundamental solution, a notion of *fundamental finite control* (FFC) is introduced here for linear systems.

Taking a system from a given arbitrary state into the equilibrium state during a given fixed time is understood as finite control [2, 3].

Such a finite control which during fixed time $T = t_1 - t_0$ damps *Green's disturbance* (5) is referred to as *fundamental finite control* (FFC).

Knowing FFC, one can easily obtain through convolution any other controls corresponding to arbitrary disturbances $f(x,\ t)$. Give now exact formulations.

Consider, for instance, linear distributed-parameter systems. Similar considerations may be applied to linear lumped, continuous and discrete systems.

Let the state of a linear controlled system be described at time $t \geqslant t_0$ by equation:

$$\sum_{k=0}^{m}\sum_{s=0}^{n} A_{viks}(x,\ t)\frac{\partial^{k+s} Q}{\partial t^k\,\partial x^s} = f(x,t) + B(x,t)\,U(x,t);\ x_0 < x < x_1;\ t > t_0 \tag{7}$$

and uniform boundary conditions:

1. uniform boundary conditions:

$$\sum_{i=0}^{1}\sum_{k=0}^{m-1}\sum_{s=0}^{n-1} a_{viks}(t)\,\frac{\partial^{k+s} Q}{\partial t^k\,\partial x^s}\bigg|_{x=x_i} = 0,\ t > t_0,\ V = 1,2,\ldots \leqslant n; \tag{8}$$

2. zero initial conditions:

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = 0;\ x_0 < x < x_1;\ k = 0,1,\ldots,\ m-1, \tag{9}$$

where function $Q(x, t)$ characterizes the state of the system in point x at time t; $f(x, t)$ is a disturbing term; $B(x, t)$ is a known vector-function; $U(x, t)$ is a vector control action; and $B(x, t)U(x, t)$ is a scalar product.

The objective of the finite control is to find a vector control $U(x, t)$ such that at some time $t_1 = t_0 + T (T > 0)$ the system comes to the equilibrium state, i.e. the state where the system stabilization conditions are met:

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0, \; x_0 < x < x_1, \; k = 0,1, \ldots, \; m-1. \tag{10}$$

In doing so, it is assumed that disturbance $f(x, t)$ acts within a limited time interval $t_0 \leq t \leq t_0 + T = t_1$.

At first sight, this formulation of the finite control might seem trivial. Indeed, if control $U(x, t)$ for a given disturbance $f(x, t)$ is taken so that the following equality

$$f(x, t) + B(x, t)\, U(x, t) = 0, \; x_0 \leq x \leq x_1, \; t \geq t_0 \tag{10a}$$

holds, the given controlled system would be in the equilibrium state under all $t \geq t_0$ in virtue of the fact that all disturbances involved into the right side of (7) would be identical to zero. Unfortunately, this equation is far from always having solution with respect to $U(x, t)$. For example, in the case where $U(x, t)$ and $B(x, t)$ are scalars, it is necessary and sufficient for solution of (10a) to exist that

$$U(x, t) = -\frac{f(x, t)}{B(x, t)} \quad \text{for all} \quad x \in [x_0, x_1], \; t \geq t_0. \tag{10b}$$

But practically interesting cases correspond to $B(x, t) = 0$ when $f(x, t) \neq 0$ and, consequently, equation (10a) has no solution. Such a situation takes place where subset M of the points (x_0, t), $x \in [x, x_1]$, $t \geq t_0$ where $f(x, t) \neq 0$ is not included in subset N of the same points (x, t) from the band $x \in [x_0, x_1]$, $t \geq 0$ where $B(x, t) = 0$.

Physically, the fact that there is no solution to (10a) means that disturbing actions (say, initial conditions) are distributed over one area of the physical space, and control is in another.

One meets with such a situation in many practically important problems of distributed object control.

For example, when heating a massive body occupying some domain $\mathfrak{D}$ of the physical space, it is the initial temperature distributed over all domain $\mathfrak{D}$ that is the disturbance. At the same time, control actions (environmental temperature) are distributed only over the boundary Γ of domain $\mathfrak{D}$. It is clear that the spatial domains of definition of the disturbance and the control (i.e. $\mathfrak{D}$ and Γ) do not coincide, and, hence, equation (10a) is not solvable.

A similar situation arises in another important problem of damping an oscillatory system distributed over some spatial domain $\mathfrak{D}$ by means of control actions distributed only over boundary Γ of domain $\mathfrak{D}$. Here, for instance, gusts of wind might be disturbances acting upon an object as a whole, and control might be exercised by some "controls" situated in boundary Γ of domain $\mathfrak{D}$.

At the end of this paper, an example of finite control is presented where the above situation occurs, disturbances of an oscillatory system act within a spatial interval $\mathfrak{D} = [0, \pi]$, and control action is concentrated in the boundary point $x = 0$ of interval $\mathfrak{D}$. Naturally, (10a) is unable to find a finite control for this case.

Formulation of the finite control problem in the form of uniform boundary and zero initial conditions will be referred to as *the standard form* of the finite control problem.

By appropriate choice of function $f(x)$ and vector function $B(x, t)$, the problem of finite control with nonuniform boundary conditions and nonzero initial conditions may be reduced to *the standard form.*

Below, an example is presented of reducing a finite control problem to the standard form (see problems (29)–(32) and Appendix 2).

Denote by $U_\delta(x, t; \xi, \tau)$ the fundamental finite control, i.e. a finite control corresponding to Green's disturbances ($f(x, t) = \delta(x - \xi)\, \delta(t - \tau)$, $x_0 < x,\ \xi < x_1,\ t_0 < t,\ \tau < t_1$). Then, since we deal with linear controllable systems and since the *law of linearity|the superposition principle (if external actions make a sum, their results make a sum as well)* holds for them, one can show that if FFC $U_\delta(x, t; \xi, \tau)$ is known the corresponding finite control under arbitray disturbance is as follows:

$$U(x,t) = \int_{t_0}^{t_1} \int_{x_0}^{x_1} U\delta(x,t;\xi,\tau)\, f(,\tau)\, d\xi d\tau, \quad x_0 < x < x_1;\ t_0 < t < t_1. \tag{11}$$

Indeed, for FFC $U\delta(\varkappa, t; \xi, \tau)$ and corresponding $Q_\delta(\varkappa, t; \xi, \tau)$ the following identities hold by assumption:

$$\sum_{k=0}^{m} \sum_{s=0}^{n} A_{ks}(x,t)\, \frac{\partial^{k+s} Q_\delta(x,t;\xi,\tau)}{\partial t^k\, \partial x^s} \equiv \delta(x-\xi)\, \delta(t-\tau) + B(x,t)\, U(x,t;\xi,\tau)$$

$$x_0 < x, \xi < x_1, \quad t_0 < t, \quad \tau < t_1. \tag{12}$$

$$\sum_{i=0}^{1} \sum_{k=0}^{m-1} \sum_{s=0}^{n-1} a_{V\,iks}(t)\, \frac{\partial^{k+s} Q_\delta(x,t;\xi,\tau)}{\partial t^k\, \partial x^s}\bigg|_{x=x_i} \equiv 0, \tag{13}$$

$$x_0 < x, \xi < x_1,\ t_0 < t,\ \tau < t,\ V = 1, 2, \ldots \leq n.$$

$$\left(\frac{\partial^k Q_\delta(x, t; \xi, \tau)}{\partial t^k}\right)_{t=t_0} \equiv 0, \tag{14}$$

$$x_0 < x,\ \xi < x_1\ t_0 < \tau < t_1,\ k = 0, 1, \ldots, m-1$$

$$\left(\frac{\partial^k Q_\delta(x, t; \xi, \tau)}{\partial t^k}\right)_{t=t_0+T=t_1} \equiv 0, \tag{15}$$

$$x_0 < x,\ \xi < x_1,\ t_0 < \tau < t_1,\ k = 0, 1, \ldots, m-1.$$

By multiplying both sides of identities (12)–(15) by $f(\xi, \tau)$ and then integrating over ξ and τ within appropriate limits and using formula

$$f(x, t) = \int_{t_0}^{t_1} \int_{x_0}^{x_1} f(\xi, \tau)\, \delta(x - \xi)\, \delta(t - \tau)\, d\xi\, d\tau \tag{16}$$

obtain that control $U(x, t)$ bringing the system under consideration into the equilibrium state (see (10)) during time interval $t_1 - t_0 = T$ has the form of (11).

Practically, sometimes it is more convenient to find finite control for disturbances taken as derivatives of δ-function rather than δ-functions themselves.

It is appropriate, therefore, to somewhat generalize the notion of FFC introduced above.

Let the disturbance be given by "derivatives" of δ-function in the form of

$$f(x, t) = \delta_x^r(x - \xi)\, \delta_t^{(q)}(t - \tau),\ x_0 < x_1,\ \xi < x_1,\ t_0 < t,\ \tau < t_1 \tag{17}$$

where $\delta_x^{(r)}(x - \xi)$ is "rth derivative" of δ-function ($r = 0, \pm 1, \pm 2, \ldots$) with $|\, r\,|$-th primitive of δ-function with respect to x corresponding to negative r; and $\delta_t^{(q)}(t - \tau)$ is qth derivative of δ-function with respect to t $q = 0,\ 1,\ 2\ldots$).

Disturbance of type (17) will be referred to as *Green's disturbance of type* (r, q). The finite control damping Green's (r, q) disturbance will be referred to as fundamental finite control (FFC) of type (r, q) and denoted as $U_{\delta(r,q)}(x, t; \xi, \tau)$.

It should be noted that the present paper does not consider computation of FFC because, like any FC, FFC may be found by methods and techniques considered earlier in [2, 3].

Knowing FFC of type (r, q) and using the superposition principle, one can show that the finite control corresponding to arbitrary disturbances $f(x, t)$ has the form of

$$U(x, t) = \int_{t_0}^{t_1} \int_{x_0}^{x_1} f_{(\xi, \tau)}^{(-r-q)}\, U_{\delta(r,q)}(x, t; \xi, \tau)\, d\xi\, d\tau,\ x_0 < x < x_1,\ t_0 < t < t_1 \tag{18}$$

where $f^{[-r-q]}(\xi, \tau)$ stands for the result of integrating $f(\xi, \tau)$ sequentially r times over ξ from x_0 to ξ and q times over τ from to t_0 to τ (under $r < 0$ integration over ξ is substituted by differentiation ($| r |$ times).).

Indeed, by substituting in (12) $\delta_x^{(r)}(x - \xi)\, \delta_t^{(q)}(t - \tau)$ for $\delta(x - \xi)\, \delta(t - \tau)$ and, consequently, $U_{\delta(r,q)}(x, t; \xi, \tau)$ for $U_\delta(x, t; \xi, \tau)$ and then multiplying corresponding identities (12)–(15) by $f^{[-r-q)}(\xi, \tau)$ and integrating over ξ and τ with regard to formula

$$f(x, t) = \int_{t_0}^{t_1} \int_{x_0}^{x_1} f^{[-r-q]}(\xi, \tau) \delta_x^{(r)}(x - \xi)\, \delta_t^{(q)}(t - r)\, d\xi\, d\tau \tag{19}$$

obtain (18).

Note, that FFC of type $(r + s;\ q + k)$ may be derived from FFC of type $(r,\ q)$ by differentiating s and k times, respectively, with respect to ξ and τ and multiplying by $(-1)^{s+k}$ (if $s < 0$ and $k < 0$, corresponding differentiations are substituted by integrations, and to preserve the finiteness of control for $k < 0$, it is required that $| k | \leq q$).

The following formula

$$U_{\delta(r+s,q+k)}(x, t; \xi, \tau) = (-1)^{s+k} \frac{\partial^{s+k} U_{\delta(r,q)}(x, t; \xi, \tau)}{\partial \xi^s\, \partial \tau^k} \tag{20}$$

is true in the sense indicated above.

Indeed, by definition, FFC of type $(r + s;\ q + k)$ corresponds to disturbance

$$f(x, t) = \delta_x^{(r+s)}(x - \xi)\, \delta_t^{(q+k)}(t - \tau). \tag{20a}$$

Accordingly to (18), obtain

$$U_{\delta(r+s,q+k)}(x, t; \xi, \tau) = \int_{t_0}^{t_1} \int_{x_0}^{x_1} \delta_{\tau_1}^{(r+s)}(\xi_1 + \xi)\, \delta_{\tau_1}^{(q+k)}(\tau_1 - \tau)\} \, U_{\delta(r,q)}^{[-r-q]}(x, t; \xi_1 \tau_1)\, d\xi_1 d\tau_1 =$$

$$= \int_{t_0}^{t_1} \int_{x_0}^{x_1} \delta_{\xi_1}^{(s)}(\xi_1 - \xi)\, \delta_{\tau_1}^{(k)}(\tau_1 - \tau)\, U_{\delta(r,q)}(x, t; \xi_1, \tau_1)\, d\xi_1\, d\tau_1 =$$

$$= (-1)^{s+k} \frac{\partial^{s+k} U_{\delta(r,q)}(x, t; \xi, \tau)}{\partial \xi^s\, \partial \tau^k}.$$

For the case of time-autonomous systems and assuming that $B(x, t) = B(x)$, the following formula may be derived

$$U_{\delta(r,q+k)}(x, t; \xi, \tau) = \frac{\partial^k U_{\delta(r,q)}(x, t; \xi, \tau)}{\partial t^k} \tag{21}$$

and in the case of a system with constant coefficients and with $B(x, t) = B$ (where B is a constant known vector) the following formula holds:

$$U_{\delta(r+s,q+k)}(x, t; \xi, \tau) = \frac{\partial^{s+k} U_{\delta(r,q)}(x, t; \xi, \tau)}{\partial x^s \, \partial t^k}. \tag{21a}$$

Formulae (20), (21) and (21a) enable expressing one FFC through FFC of another type, type (0,0) in particular, i.e. through $U_\delta(x, t; \xi, \tau)$.

Consider a case where external disturbances act at the initial moment only.

For example, let a controllable system be described by equation

$$\sum_{k=0}^{m} \sum_{s=0}^{n} A_{ks}(x, t) \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} = B(x, t) U(x, t), \; x_0 < x < x_1, \; t_0 < t < t_1 \tag{22}$$

with uniform boundary conditions

$$\sum_{i=0}^{1} \sum_{k=0}^{m-1} \sum_{s=0}^{n-1} a_{Viks}(t) \frac{\partial^{k+s} Q}{\partial t^k \partial x^s}\bigg|_{x=x_i} = 0, \; t_0 < t < t_1, \; V = 1, 2, \ldots \leq n \tag{23}$$

and let the given system be at the initial moment $t = t_0$ in some excited state described by the following initial conditions

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = Q_k(x), \; x_0 < x < x_1, \; k = 0, 1, \ldots, m-1. \tag{24}$$

The system (22)–(23)) is to be taken from initial excited state (24) to the equilibrium one

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0, \; x_0 < x < x_1, \; k = 0, 1, \ldots, m-1 \tag{25}$$

during finite time $T = t_1 - t_0$ by means of a special controlling change of action $U(x, t)$.

A finite control corresponding to zero initial conditions (with the exception of pth initial condition taken as δ-function), i.e. to initial conditions of the following form:

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = \delta_{kp} \delta(x - \xi), \; x_0 < x, \; \xi < x_1, \; k, q = 0, 1, \ldots, m-1 \tag{26}$$

where δ_{kp} is Kronecker's symbol, will be referred to as partial fundamental finite control (PFFC) of order p and denoted as $U_{\delta p}(x, t; \xi)$.

Further, a finite control damping initial disturbance of type

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = \delta_{kp} Q_k(x), \; x_0 < x < x_1, \; k, p = 0, 1, \ldots, m-1 \tag{27}$$

will be denoted as $U_p(x, t; Q_p)$.

Then, using the superposition principle, one can easily verify that the finite control corresponding to arbitrary initial conditions (24) is as follows:

$$U(x, t) = \sum_{p=0}^{m-1} U_p(x, t; Q_p) = \sum_{p=0}^{n-1} \int_{x_0}^{x_1} Q_p(\xi)\, U_{\delta p}(x, t; \xi)\, d\xi; \quad x_0 < x < x_1; \quad t_0 < t < t_1. \tag{28}$$

Totality $\{U_{\delta p}(x, t; \xi)\}$ which may be ordered into a vector will be referred to as *fundamental finite control vector.*

Then (28) may be symbolically rewritten as

$$U(x, t) = Q^0(\xi) * U_f(x, t; \xi), \; x_0 < x < x_1, \; t_0 < t < t_1$$

where sign $*$ stands for an appropriate integral over ξ of scalar product; $Q^0(\xi) = (Q_0(x), \ldots, Q_{m-1}(x))$ is initial conditions vector; and $U_f(x, t; \xi) = (U_{\delta 0}(x, t; \xi), \ldots, U_{\delta, m-1}(x, t; \xi))$ is vector of fundamental finite controls.

Similar to what was done above with the right side of control (7) (see (17)), one can introduce into initial conditions (26) derivatives of δ-function, i.e. instead of (26), the following conditions may be considered:

$$\left(\frac{\partial^x Q}{\partial t^x}\right) = \delta_{kp}\, \delta_x^{(r_k)}(x - \xi), \; x_0 < x < x_1, \; k, p = 0, 1, \ldots, m - 1 \tag{28a}$$

where $r_k = 0, \pm 1, \pm 2 \ldots$ ($r_k < 0$ corresponds to $| r |$-th primitive of δ-function).

A finite control corresponding to conditions (28a) will be referred to as partial fundamental finite control of order p and rank r_p and denoted as $U_{\delta p}^{r_p}(x, t; \xi)$, $(\mathrm{PFFC})_p^{r_p}$.

Control $U(x, t)$ damping initial disturbances (24) during time interval $T = t_1 - t_0$ is proved to be representable as

$$U(x, t) = \sum_{p=0}^{m-1} \int_{x_0}^{x_1} Q_p^{[-r_p]}(\xi)\, U_{\delta_p}^{r_p}(x, t; \xi)\, d\xi; \quad x_0 < x < x_1, \; t_0 < t < t_1 \tag{29}$$

where $Q_p^{[-r_p]}(\xi)$ denotes for $r_p > 0$ integration over ξ successively r_p times from x_0 to ξ, and for $r_p < 0$ differentiation with respect to ξ taken $| r_p |$ times.

Correctness of transition formula

$$U_{\delta_p}^{r_p+s}(x, t; \xi) = (-1)^s \frac{\partial^s U_{\delta_p}^{r_p}(x, t; \xi)}{\partial \xi^s} r_p, \; s = 0, \pm 1, \pm 2, \ldots \tag{30}$$

$$p = 0, 1, \ldots, m - 1$$

is proved analogously to (20).

A relation may be established between PFFC's of different orders and, consequently, between two arbitrary PFFC's of different orders and ranks. This relation is most easily revealed in the case of the following systems:

$$\frac{\partial^m Q}{\partial t^m} + \sum_{k=0}^{m-1} \sum_{s=0}^{n} A_{ks}(x,t) \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} = B(x,t)\, U(x,t), \quad x_0 < x < x_1 \qquad (31)$$

$$t_0 < t < t_1 .$$

$$\sum_{i=0}^{\mathrm{i}} \sum_{s=0}^{n-1} a_{Vis}(t) \frac{\partial^s Q}{\partial x^s}\bigg|_{x=x_i} = 0, \quad t_0 < t < t_1, \quad V = 1, 2, \ldots, \leq n, \qquad (32)$$

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0} = Q_k(x), \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m-1 \qquad (33)$$

with initial conditions

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m-1. \qquad (34)$$

Reduce the posed problem of finite control to that of standard form finite control. To this end, continue function $Q(x,\ t)$ over $t < t_0$ and make it equal to zero by means of formula

$$Q^*(x,t) = Q(x,t) \cdot 1(t - t_0) \qquad (35)$$

where $1(t - t_0)$ is a unit function

$$1(t - t_0) = \begin{cases} 1 & \text{for} \quad t > t_0 \\ 0 & \text{for} \quad t < t_0 . \end{cases} \qquad (36)$$

Then, making use of formulae

$$\frac{\partial^{k+s} Q^*}{\partial t^k \partial x^s} = \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} 1(t - t_0) + \sum_{p=0}^{k-1} \frac{\partial^{p+s} Q}{\partial t^p \partial x^s}\bigg|_{t=t_0} \delta_t^{(k-p-1)}(t - t_0), \qquad (37)$$

$$A_{ks}(x,t) \frac{\partial^{k+s} Q^*}{\partial t^k \partial x^s} = A_{ks}(x,t) \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} 1(t - t_0) +$$

$$+ \sum_{p=0}^{k-1} \frac{\partial^{p+s} Q}{\partial t^p \partial x^s}\bigg|_{t=t_0^+} \times \sum_{j=0}^{k-p-1} (-1)^j C_{k-p-1}^j \frac{\partial^j A_{ks}(x,t)}{\partial t^j}\bigg|_{t=t_0^+} \cdot \delta_t^{(k-p-j-1)}(t - t_0) \qquad (38)$$

conclude from equations (31), (32), (33) and (34) that

$$\frac{\partial^m Q}{\partial t^m} + \sum_{k=0}^{m-1} \sum_{s=0}^{n} A_{ks}(x,t) \frac{\partial^{k+s} Q}{\partial t^k \partial x^s} = \sum_{p=0}^{m-1} Q_p(x)\, \delta_t^{(m-p-1)}(t - t_0) +$$

$$+ \sum_{k=1}^{m-1} \sum_{s=0}^{h} \sum_{p=0}^{k-1} Q_p^{(s)}(x) \times \sum_{j=0}^{k-p-1} (-1)^j C_{k-p-1}^j \frac{\partial^j A_{ks}(x,t)}{\partial t^j}\bigg|_{t=t_0^+} \delta_t^{(k-p-j-1)}(t - t_0) +$$

$$+ B(x,t)\,U(x,t), \quad x_0 < x < x_1, \quad t_0^+ \leq t < t_1, \tag{39}$$

$$\sum_{i=0}^{1}\sum_{s=0}^{n-1} a_{Vis}(t)\,\frac{\partial^s Q}{\partial x^s}\bigg|_{x=x_i} = 0, \quad t_0^+ \leq t < t_1, \quad V = 1, 2, \ldots, \leq n\,. \tag{40}$$

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0^-} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m-1 \tag{41}$$

and

$$\left(\frac{\partial^k Q}{\partial t^k}\right)_{t=t_0+T=t_1} = 0, \quad x_0 < x < x_1, \quad k = 0, 1, \ldots, m-1. \tag{42}$$

Now, by substituting successively into (33) and (39)

$$Q_k(x) = \delta_{k,m-1}\,\delta(x-\xi),\ Q_k(x) = \delta_{k,m-2}\cdot\delta(x-\xi), \ldots,\ Q_k(x) = \delta_{k,0}\cdot\delta(x-\xi)$$

and allowing for linearity of the system, obtain the following relations:

$$U_{\delta,m-1}(x,t;\xi) = U_{\delta(0,0)}(x,t;\xi,t_0^+) = U_\delta(x,t;\xi,t_0^+),$$

$$U_{\delta,m-q}(x,t;\xi) = U_{\delta(0,q-1)}(x,t;\xi,t_0^+) + \sum_{k=m-q+1}^{m-1}\sum_{s=0}^{n}\sum_{r=0}^{s}\sum_{j=0}^{k-m+q+1}(-1)^{j+r}C_s^r C_{k-m+q-1}^j \times$$

$$\times \frac{\partial^{j+r}A_{ks}(x,t)}{\partial t^j\,\partial x^s}\bigg|_{\substack{x=\xi\\ t=t_0^+}} U_{\delta(s-r,k-j-m+q-1)}(x,t;\xi,t_0^+),\ q = 2,3,\ldots,m. \tag{43}$$

In particular, if the system is time-autonomous and $B(x,t)$ is also time-independent, then allowing for (20) and (21) obtain from (43)

$$U_{\delta,m-1}(x,t;\xi) = U_\delta(x,t;\xi,t_0^+),$$

$$U_{\delta,m-1}(x,t;\xi) = \left[\frac{\partial^{q-1}}{\partial t^{q-1}} + \sum_{k=m-q+1}^{m-1}\sum_{s=0}^{n}\sum_{r=0}^{s}(-1)^s C_s^r A_{m-q+1,s}^{(r)}(\xi)\,\frac{\partial^{(k-m+q-1)+(s-r)}}{\partial t^{k-m+q-1}\,\partial\xi^{s-r}}\right] \times$$

$$\times\, U_\delta(x,t;\xi,t_0^+), \quad q = 1,2,\ldots,m. \tag{44}$$

At last, when in an autonomous system all $A_{ks}(x,t)$ are identical to zero, aside from coefficients of the form $A_{0s}(x)$ and $A_{k0}(x)$, and if $B(x,t)$ is also independent of t, formula (44) takes a simpler form:

$$U_{\delta(m-q)}(x,t;\xi) = \left[\frac{\partial^{q-1}}{\partial t^{q-1}} + A_{m-1,0}(\xi)\,\frac{\partial^{q-2}}{\partial t^{q-2}} + \ldots + A_{m-q+1,0}(\xi)\right] U_{\delta,m-1}(x,t,\xi)\,,$$

$$q = 1,2,\ldots,m. \tag{44a}$$

Note that the generalization of FFC to the case where system state Q and control U are vectors subordinated to vector equations does not involve any principal difficulties.

Note also that notions introduced above enable direct generalization to the multidimensional case. If operator l in an equation of type (1) is defined by

some problems similar to problems (22)–(24) where some domain $\mathfrak{D}$ of n-dimensional Euclidean space is used instead of interval (x_0, x_1), FFC corresponding to the following initial conditions

$$\frac{\partial^k Q}{\partial t^k}(x, t_0) = \delta_{k,p} \cdot \delta_{x_1}^{(r_{k_1})}(x_1 - \xi_1) \ldots \delta_{x_n}^{(r_{k_n})}(x_n - \xi_n) \tag{45}$$

where k, $p = 0, 1, \ldots, m-1$; $r_{p_j} = 0, \pm 1, \pm 2, \ldots$; $j = 1, 2, \ldots, n$, will be understood as PFFC of pth order and vector rank $r_p = (r_{p_1}, r_{p_2}, \ldots, r_{p_n})$. In this case, the integral in (29) should be taken over $\mathfrak{D}$, and $Q_p^{[-rp]}(\xi)$ should be understood as a partial derivative (or partial integral if some $r_{pj} > 0$) of $Q_p(\xi)$ having order $r_p = (r_{p_1}, \ldots, r_{p_n})$ and taken with respect to $\xi_1, \ldots, \xi_n$.

The corresponding generalization of (30) to this case has the form of:

$$U_{\delta_p}^{r+s}(x, t; \xi) = (-1)^{\bar{s}} \frac{\partial^{\bar{s}} U_{\delta p}^{r}(x, t; \xi)}{\partial \xi_1^{s_1} \ldots \partial \xi_n^{s_n}} \tag{46}$$

where $s = (s_1, \ldots s_n)$; $\bar{s} = s_1 + s_2 + \ldots + s_n$.

(In the case of $s_j < 0$, integration over ξ_j of order $|s_j|$ within a corresponding domain is implied.)

Appendix I

Example 1. Find finite control (FC) $U(t)$, $0 \leq t \leq T$, for wave equation

$$\frac{\partial^2 Q}{\partial t^2} = \frac{\partial^2 Q}{\partial x^2}, \quad 0 < x < \pi, \quad t > 0 \tag{47}$$

with boundary conditions

$$Q(0, t) = U(t), \quad Q(\pi, t) = 0, \quad t > 0 \tag{48}$$

and arbitrary initial conditions

$$Q(x, 0) = Q_0(x), \frac{\partial}{\partial t} Q(x, 0) = Q_1(x), \quad 0 < x < \pi. \tag{49}$$

Find first PFFC of order 1 and rank 1, i.e. finite control $U_{\delta_1}^1(t, \xi)$ corresponding to the following initial conditions:

$$Q(x, 0) = 0, \frac{\partial}{\partial t} Q(x, 0) = \delta_x'(x - \xi), \quad 0 < x, \quad \xi < \pi \tag{50}$$

Using the method of characteristics (the D'Alembert's method of propagating waves) one can readily find that minimal T in this system is 2π (for the most "unfavorable" ξ from $(0, \pi)$) and that

$$U_{\delta_1}^1(t, \xi) = \frac{1}{2}[\delta(t - \xi) + \delta(t + \xi - 2\pi)], \quad 0 < \xi < \pi; \quad 0 \leq t \leq 2\pi. \tag{51}$$

For $p = 1$, $r_1 = 1$ and $s = 1$ obtain through (30) that

$$U^0_{\delta,1}(t, \xi) = U_{\delta,1}(t, \xi) = -\frac{1}{2}\int_0^{\xi} [\delta(t - \eta) + \delta(t + \eta - 2\pi)]\, d\eta = \\ = \frac{1}{2}[1(t - \xi) - 1(t + \xi - 2\pi)], \quad 0 < \xi < \pi, \quad 0 \leq t \leq 2\pi. \tag{52}$$

Now, for $m = 2$, $q = 2$, $A_{10}(\xi) = 0$ and $A_{00}(\xi) = 0$ obtain through formula (44′) that

$$U_{\delta,0}(t, \xi) = \frac{\partial}{\partial t} U_{\delta,1}(t, \xi) = \frac{1}{2}[\delta(t - \xi) - \delta(t + \xi - 2\pi)], \\ 0 < \xi < \pi, \quad 0, \leq t \leq 2\pi. \tag{53}$$

Thus, the vector of fundamental finite controls has the following form:

$$U_f(t, \xi) = \left(\frac{\delta(t - \xi) - \delta(t + \xi - 2\pi)}{2}, \frac{1(t - \xi) - 1(t + \xi - 2\pi)}{2}\right), \tag{54}$$

and, consequently, the finite control corresponding to arbitrary initial condition (49) is, accordingly to (21) and (44a), as follows:

$$U(t) = Q^0(\xi) * U_f(t, \xi) = \sum_{p=0}^{1} \int_0^{\pi} Q_p(\xi)\, U_{\delta p}(t, \xi)\, d\xi = \\ = \frac{1}{2}\int_0^{\pi} Q_0(\xi)\, [\delta(t - \xi) - \delta(t + \xi - 2\pi)]\, d\xi + \\ + \frac{1}{2}\int_0^{\pi} Q_1(\xi)\, [1(t - \xi) - 1(t + \xi - 2\pi)]\, d\xi = \\ = \frac{1}{2}\begin{cases} Q_0(t) + \int_0^t Q_1(\xi)\, d\xi & \text{for} \quad 0 < t < \pi \\ -Q_0(2\pi - t) + \int_0^{\pi} Q_1(\xi)\, d\xi - \int_0^{\pi} Q_1(2\pi - \xi)\, d\xi & \text{for} \quad \pi < t < 2\pi \\ 0 & \text{for} \quad t < 0,\ t > 2\pi \end{cases} \tag{55}$$

hence, assuming that $Q_0(x) = Q_1(x) \equiv 0$ obtain for $x \in (0, \pi)$

$$U(t) = \frac{1}{2}[Q_0(t) - Q_0(2\pi - t)] + \frac{1}{2}\int_0^t [Q_1(\xi) - Q_1(2\pi - \xi)]\, d\xi, \; 0 \leq t \leq 2\pi. \tag{56}$$

Appendix 2

Consider problem of finite control

$$\frac{\partial^2 Q}{\partial t^2} = \frac{\partial^2 Q}{\partial x^2} + c(x)\, R + f(x, t),\ x_0 < x < x_1,\ t > t_0, \tag{57}$$

$$\left.\begin{aligned} \alpha_0 Q(x_0, t) + \beta_0 Q'_x(x_0, t) &= U_0(t),\ t > t_0, \\ \alpha_1 Q(x_1, t) + \beta_1 Q'_x(x_1, t) &= U_1(t),\ t > t_0, \end{aligned}\right\} \tag{58}$$

$$Q(x, t_0) = Q_0(x), \frac{\partial}{\partial t} Q(x, t_0) = Q_1(x),\ x_0 < x < x_1, \tag{59}$$

and

$$Q(x, t)\,|_{t=t_0+T=t_1} = 0, \frac{\partial}{\partial t} Q(x, t)\,|_{t=t_0+T=t_1} = 0,\ x_0 < x < x_1. \tag{60}$$

Assume that there is a Green function $\mathcal{C}(x, \xi, t - r)$ enabling the following representation of problems (57)–(59):

$$\begin{aligned} Q(x, t) = &\int_{t_0}^{t} \left[\mathcal{C}(x, \xi, t - \tau) \frac{\partial Q(\xi, \tau)}{\partial \xi} - Q(\xi, \tau) \frac{\partial \mathcal{C}(x, \xi, t - \tau)}{\partial \xi} \right]_{\xi = x_0}^{\xi = x_1} d\tau + \\ &+ \int_{x_0}^{x_1} Q_0(\xi)\, \mathcal{C}_t(x, \xi, t - t_0)\, d\xi + \int_{x}^{x_1} Q_1(\xi)\, \mathcal{C}(x, \xi, t - t_0)\, d\xi + \\ &+ \int_{t_0}^{t} \int_{x_0}^{x} f(\xi, \tau)\, \mathcal{C}(x, \xi, t - \tau)\, d\xi\, d\tau. \end{aligned} \tag{61}$$

On the basis of this formula, problems (57)–(59) may be reduced to the following finite control problem:

$$\begin{aligned} \frac{\partial_2 Q}{\partial t^2} = \frac{\partial^2 Q}{\partial x^2} &+ c(x)\, Q + Q_0(x)\, \delta'_t(t - t_0) + Q_1(x)\, \delta(t - t_0) + f(x, t) + \\ &+ b_0(x)\, U_0(t) + b_1(x)\, U_1(t),\quad x_0 < x < x_1,\quad t_0 + \leq t < t_1, \end{aligned} \tag{62}$$

$$\left.\begin{aligned} \alpha_0 Q(x_0, t) + \beta_0 Q'_x(x_0, t) &= 0,\ t_0^+ + \leq t < t_1 \\ \alpha_1 Q(x_1, t) + \beta_1 Q'_x(x_1, t) &= 0,\ t_0^+ + \leq t < t_1 \end{aligned}\right\}, \tag{63}$$

$$Q(x, t_0^-) = 0,\ \frac{\partial}{\partial t} Q(x, t_0^-) = 0,\quad x_0 < x < x_1, \tag{64}$$

$$Q(x, t_1) = 0, \frac{\partial}{\partial t} Q(x, t_1) = 0, \quad x_0 < x < x_1, \tag{65}$$

where

$$b_0(x) = \begin{cases} -\dfrac{1}{\beta_0}\delta(x - x_0) & \text{for} \quad \beta_0 \neq 0 \\ \dfrac{1}{\alpha_0}\delta'_x(x - x_0) & \text{for} \quad \beta_0 = 0, \end{cases} \tag{66}$$

$$b_1(x) = \begin{cases} \dfrac{1}{\beta_1}\delta(x - x_1) & \text{for} \quad \beta_1 \neq 0 \\ -\dfrac{1}{\alpha_1}\delta'_x(x - x_1) & \text{for} \quad \beta_1 = 0. \end{cases} \tag{67}$$

References

1. *Tikhonov, A. I., Samarskii, A. A.*, Equations of the Mathematical Physics, Nauka Publ., 1966 (in Russian).
2. *Butkovskii, A. G.*, Finite Control and Controllability in Distributed Systems. Dokl. AN SSSR, **191**, *6*, 1970 (in Russian).
3. *Feldbaum, A. A., Butkovskii, A. G.*, Methods of the Automatic Control Theory, Znanie Publ., 1971 (in Russian).

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

H–1111 Budapest XI., Goldmann György tér 3, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4–5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которых должен содержать рисунки, таблицы и литературу) направляются

Е. Д. Теряеву — Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский Проспект, 14, Москва В-71, СССР

или

Я. Кочишу — Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета
Н–1111 Будапешт XI, площадь Голдманн Д., 3,
Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре.

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4–5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру .

Статье должна предшествовать аннотация объемом 50–100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме — на русском языке, если статья написана на английском, и наоборот — должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем резюме не должен превышать 10–15% объема статьи.

Авторам высылается корректура статьи, которую они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Красовский А. А. Об одном обобщении задачи аналитического конструирования систем управления (*Krasovskii, A. A.*: A generalization of a control system analytical design problem) 3

Beneš, J.: Stochastic control of impulses on plane crystallographic group lattices (*Бенеш Й.* Стохастическое управление импульсами в решетках плоских кристаллографических групп) 13

Petrov, B. N., Dobrushin, R. L., Pinsker, M. S., Ulanov, G. M., Ulyanov, S. V.: On some interrelations between the theories of information and control (*Петров Б. Н., Добрушин Р. Л., Пинскер М. С., Уланов Г. М., Ульянов С. В.* О некоторых взаимосвязях теории информации и теории управления) 31

Frey, T., Gyürki, J.: Recursive state and parameter estimation in nonlinear stochastic systems. I (*Фрей Т., Дюрки Й.* Рекуррентный алгоритм оценки состояния и параметров нелинейных стохастических систем) 39

Kazakov, I. E.: Analytical synthesis of the statistical quasi-optimal control of the nonlinear subject (*Казаков И. Е.* Аналитический синтез статистически квазиоптимального управления нелинейным объектом) 51

Schwarz, H.: Markov-parameter representation of linear time-variable systems with separable system functions (*Шварц Х.* Параметрическое марковское представление сигналов линейных нестационарных систем с сепарируемой системной функцией) 63

Бутковский А. Г., Мустафаев М. И. Фундаментальные финитные управления (*Butkovskii, A. G. Mustafaev, M. I.*: Fundamental finite control) 71

Rao, A. N. V., Tsokos, C. P.: On the existence and stability behavior of a stochastic integral equation in a Banach space (*Рао А. Н. В., Цокос С. П.* К проблеме существования и устойчивости некоторого стохастическово уравнения в банаховом пространстве) 87

Raibman, N. S.: The Sixth Congress of IFAC 97

Book review 107

316.920

VOL. 5 * NUMBER 2
TOM ТОМ НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1976

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

is an international quarterly sponsored jointly by the Presidium of the Academy of Sciences of the USSR and of the Hungarian Academy of Sciences. It offers publicity for original papers and short communications of the following topics:

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

— theory and methods of adaptation, learning, identification and pattern recognition

— methods of information processing; application of digital computers in control and communication systems

— new physical principles in constructing technical devices for automation, control and information processing

The six issues published per year make up a volume of some 480 pp.

While this quarterly is mainly a publication forum of the research results achieved within the USSR and Hungary, also papers of international interest from other countries are welcome.

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

H-1389 Budapest 62, P.O.B. 149' Hungary

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

H-1363 Budapest P.O.B. 24, Hungary

Академия наук СССР и Академия наук ВНР совместно издают журнал

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал публикует материалы о новых результатах научных исследований, разработок и краткие сообщения о достижениях в следующих областях науки:

— общая теория процессов управления и теория систем
— теория автоматов
— теория информации
— исследование операций; теория сложных систем
— теория управления экономическими системами; модели систем
— теория и методы адаптации, обучения, идентификации и распознавания образов
— методы обработки информации; применение вычислительных машин в системах управления и передачи информации
— новые физические принципы создания технических средств автоматики, управления и передачи информации

Журнал быхогит шесть раз в год общим объёмом 480 печатных страниц.

Журнал публикует статьи о научных достижениях обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

H-1389 Будапешт 62, п. о. 149, Венгрия

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии наук Венгрии

H-1363 Будапешт п. о. 24, Венгрия

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)
F. CSÁKI (Budapest)

DEPUTY EDITORS

V. S. PUGACHEV (Moscow)
V. I. SIFOROV (Moscow)
S. CSIBI (Budapest)

COORDINATING EDITORS

M. A. GAVRILOV (Moscow)
L. KALMÁR (Szeged)

S. V. EMELYANOV (Moscow)
I. CSISZÁR (Budapest)

N. N. KRASOVSKII (Sverdlovsk)
A. PRÉKOPA (Budapest)

E. P. POPOV (Moscow)
L. VARGA (Budapest)

E. D. TERYAEV (Moscow)
J. KOCSIS (Budapest)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)
Ф. ЧАКИ (Будапешт)

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Ш. ЧИБИ (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

М. А. ГАВРИЛОВ (Москва)
Л. КАЛМАР (Сегед)

С. В. ЕМЕЛЬЯНОВ (Москва)
И. ЧИСАР (Будапешт)

Н. Н. КРАСОВСКИЙ (Свердловск)
А. ПРЕКОПА (Будапешт)

Е. П. ПОПОВ (Москва)
Л. ВАРГА (Будапешт)

Е. Д. ТЕРЯЕВ (Москва)
Я. КОЧИШ (Будапешт)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES

BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 5 (2), pp. 109—116 (1976)

TRANSPORTATION SYSTEMS WITH OPTIMAL PARAMETERS

R. LUNDERSTÄDT

(Hamburg)

(Received March 25, 1975)

Optimization problems with ordinary differential equations and transportation equations (partial differential equations of first order) as system dynamics are considered. The system equations depend on control functions and on parameters. For the parameters necessary optimization conditions are derived by means of the calculus of variations in the formulation of Hamilton. The theoretical results are explained in an example.

1. Introduction

After the theoretical investigations of optimization problems for systems with lumped parameters have reached a certain closing, the interest is directed to systems with distributed parameters. The so-called transportation systems have here an intermediate position. These are systems described by partial differential equations of first order. A comprehensive part of the optimization theory developed for systems with lumped parameters can be transformed directly or after analogous extensions to transportation systems. This is based on the fact that the transportation equation can be described by ordinary differential equations using the method of characteristics and it is therefore solvable if the system is linear.

First systematic investigations of transportation systems in connection with optimizations problems can be found in [1] and [2]. Here, especially, it is shown that the transportation equation is suitable for a generalized description of systems with time lag. Necessary optimization conditions for transportation systems are given in [3]. With these conditions in [4] a general solution is carried out for a combined system with quadratic performance criteria existing of a set of ordinary differential equations and a set of transportation equations coupled by the boundary conditions. In [4] lumped and distributed controls are used. As optimization procedure in [1]–[4] the maximum principle is employed. In extension of the results of [3] and [4] in this paper the problem is discussed that in the system equations in addition to control functions also time constant parameters occur, the destination of which is task of the optimization problem, too. Results are presented already in [6]

for systems with only ordinary differential equations. The generalization of the results of [6] for transportation systems is shortly announced in [7] and [8]. Here, now a systematic investigation is given.

2. Problem

The system of ordinary differential equations is considered

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}\big(\mathbf{x}, \mathbf{z}, \mathbf{u}(t), \boldsymbol{\alpha}, t\big), \qquad t > t_0 \tag{1a}$$

and the transportation system

$$\frac{\partial \mathbf{y}}{\partial t} = v(t, s) \cdot \frac{\partial \mathbf{y}}{\partial s} + \mathbf{h}\big(\mathbf{y}, \mathbf{w}(t, s), \boldsymbol{\beta}, \boldsymbol{\gamma}(s), t, s\big), \quad \begin{matrix} t > t_0, \\ s_0 < s < L. \end{matrix} \tag{1b}$$

In (1) $\mathbf{x}(t)$ and $\mathbf{y}(t, s)$ are n-dimensional state variables. The controls are given by $\mathbf{u}(t)$ and $\mathbf{w}(t, s)$. They are q- and r-dimensional and constrained by $\mathbf{u} \in U$ and $\mathbf{w} \in W$. The unconstrained, time-constant parameters are denoted by $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}(s)$. The parameters have the appropriate dimensions d_i ($i = \alpha, \beta, \gamma$). The independent variables are the time t and the space coordinate s. It is assumed that the functions $\mathbf{f}$ and $\mathbf{h}$ fulfil the conditions of continuity and differentiability given in [1]. The scalar function $v(t, s)$ in (1b) is the transportation velocity. It is always $v(t, s) > 0$. To (1) the initial conditions

$$\begin{aligned} \mathbf{x}(t_0) &= \mathbf{x}_0, \\ \mathbf{y}(t_0, s) &= \mathbf{y}_0(s) \end{aligned} \tag{2}$$

and the boundary conditions

$$\begin{aligned} \mathbf{y}(t, s_0) &= \mathbf{x}(t), \\ \mathbf{y}(t, L) &= \mathbf{z}(t) \end{aligned} \tag{3}$$

belong. Final conditions are not explicitly prescribed because these do not influence the further derivations and for the solution of (1) they are not necessary. In (1) the controls $\mathbf{u}(t)$ and $\mathbf{w}(t, s)$ and the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}(s)$ are now to be destinated such that the general performance index

$$J = \int_{t_0}^{T} f_0(\mathbf{x}, \mathbf{z}, \mathbf{u}, \boldsymbol{\alpha}, t)\, dt + \int_{t_0}^{T} \int_{s_0}^{L} h_0(\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\gamma}, t, s) \cdot ds \cdot dt \tag{4}$$

is minimal. Solving the formulated optimization problem by means of the maximum principle [5] then the optimal controls $\overset{*}{\mathbf{u}}(t)$ and $\overset{*}{\mathbf{w}}(t, s)$ are held from necessary conditions given in [3]. About the optimal parameters $\overset{*}{\boldsymbol{\alpha}}, \overset{*}{\boldsymbol{\beta}}$ and $\overset{*}{\boldsymbol{\gamma}}(s)$ in [3] is said nothing. The destination of the optimal parameters is worked out here in the next section.

3. Optimal parameters

Necessary conditions for optimal controls and optimal parameters can be held from the first variation of (4). For this the systems (1a) and (1b) are at first formally extended by the additional coordinates

$$\frac{d\boldsymbol{\alpha}}{dt} = \mathbf{0}\,, \tag{5a}$$

$$\begin{cases} \dfrac{\partial \boldsymbol{\beta}}{\partial t} = \mathbf{0}\,, \\ \dfrac{\partial \boldsymbol{\gamma}}{\partial t} = \mathbf{0}\,. \end{cases} \tag{5b}$$

The Hamiltonians for the extended systems (1a) and (1b) are then given by

$$\begin{aligned} \hat{H}_x &= H_x + \boldsymbol{\pi}_\alpha^T \cdot \left(\frac{d\boldsymbol{\alpha}}{dt}\right), \\ \hat{H}_y &= H_y + \boldsymbol{\pi}_\beta^T \cdot \left(\frac{\partial \boldsymbol{\beta}}{\partial t}\right) + \boldsymbol{\pi}_\gamma^T \cdot \left(\frac{\partial \boldsymbol{\gamma}}{\partial t}\right). \end{aligned} \tag{6}$$

In (6) H_x and H_y are the Hamiltonians of the initial systems (1a) and (1b)

$$\begin{aligned} H_x &= -f_0 + \psi^T \cdot \mathbf{f}\,, \\ H_y &= -h_0 + \boldsymbol{\Phi}^T \cdot \left(-v(t,s) \cdot \frac{\partial \mathbf{y}}{\partial s} + \mathbf{h}\right) \end{aligned} \tag{7}$$

and ψ, $\boldsymbol{\Phi}$, $\boldsymbol{\pi}_\alpha$, $\boldsymbol{\pi}_\beta$, $\boldsymbol{\pi}_\gamma$ are the adjoint variables belonging to $\mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$. With (6) and (7) the performance index (4) can be written in a new form

$$\begin{aligned} J = &- \int_{t_0}^{T} \left[H_x - \psi^T \cdot \mathbf{f} - \boldsymbol{\pi}_\alpha^T \cdot \left(\frac{d\boldsymbol{\alpha}}{dt}\right)\right] \cdot dt - \\ &- \int_{t_0}^{T} \int_{s_0}^{L} \left[H_y - \boldsymbol{\Phi}^T \cdot \left(-v(t,s) \cdot \frac{\partial \mathbf{y}}{\partial s} + \mathbf{h}\right) - \boldsymbol{\pi}_\beta^T \cdot \left(\frac{\partial \boldsymbol{\beta}}{\partial t}\right) - \boldsymbol{\pi}_\gamma^T \cdot \left(\frac{\partial \boldsymbol{\gamma}}{\partial t}\right)\right] \cdot ds \cdot dt\,. \end{aligned} \tag{8}$$

The first variation of (8) reads as

$$\delta J = \delta J_0 - \int_{t_0}^{T} \left[\left(\frac{\partial H_x}{\partial \boldsymbol{\alpha}}\right)^T \cdot \delta\boldsymbol{\alpha} - \delta\left(\boldsymbol{\pi}_\alpha^T \cdot \frac{d\boldsymbol{\alpha}}{dt}\right)\right] \cdot dt -$$

$$-\int_{t_0}^{T}\int_{s_0}^{L}\left[\left(\frac{\partial H_y}{\partial \boldsymbol{\beta}}\right)^T \cdot \delta\boldsymbol{\beta} - \delta\left(\boldsymbol{\pi}_\beta^T \cdot \frac{\partial \boldsymbol{\beta}}{\partial t}\right)\right] \cdot ds \cdot dt - \tag{9}$$

$$-\int_{t_0}^{T}\int_{s_0}^{L}\left[\left(\frac{\partial H_y}{\partial \boldsymbol{\gamma}}\right) \cdot \delta\boldsymbol{\gamma} - \delta\left(\boldsymbol{\pi}_\gamma^T \cdot \frac{\partial \boldsymbol{\gamma}}{\partial t}\right)\right] \cdot ds \cdot dt\,,$$

where in the strong variation δJ_0 variations of $\mathbf{x}$, $\mathbf{z}$,ψ, $\mathbf{u}$, $\mathbf{y}$, $\partial\mathbf{y}/\partial s$, $\boldsymbol{\Phi}$, and $\mathbf{w}$ not interested here are included. Necessary condition for a minimum of (4) is now the demand that the first variation δJ vanishes. Because in (9) the variations of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are indepedent of each other and especially independent of the variations of the other variables $\mathbf{x}$, $\mathbf{z}$, ψ, $\mathbf{u}$, $\mathbf{y}$, $\partial\mathbf{y}/\partial s$, $\boldsymbol{\Phi}$ and $\mathbf{w}$ the variation δJ_0 as well as each of the three integrals in (9) must vanish. The demand

$$\delta J_0 = 0 \tag{10}$$

leads to the necessary conditions given in [3].

From the further demands one gets

$$I_1 = \int_{t_0}^{T}\left[\left(\frac{\partial H_x}{\partial \boldsymbol{\alpha}}\right)^T \cdot \delta\boldsymbol{\alpha} - \delta\left(\boldsymbol{\pi}_\alpha^T \cdot \frac{d\boldsymbol{\alpha}}{dt}\right)\right] \cdot dt = 0\,,$$

$$I_2 = \int_{t_0}^{T}\int_{s_0}^{L}\left[\left(\frac{\partial H_y}{\partial \boldsymbol{\beta}}\right)^T \cdot \delta\boldsymbol{\beta} - \delta\left(\boldsymbol{\pi}_\beta^T \cdot \frac{\partial \boldsymbol{\beta}}{\partial t}\right)\right] \cdot ds \cdot dt = 0\,, \tag{11}$$

$$I_3 = \int_{t_0}^{T}\int_{s_0}^{L}\left[\left(\frac{\partial H_y}{\partial \boldsymbol{\gamma}}\right)^T \cdot \delta\boldsymbol{\gamma} - \delta\left(\boldsymbol{\pi}_\gamma^T \cdot \frac{\partial \boldsymbol{\gamma}}{\partial t}\right)\right] \cdot ds \cdot dt = 0\,.$$

In (11) at first the integral I_1 is considered. If one solves the second part of this integral by partial integration gets after a simple vector algebraic transformation

$$I_1 = -\,\delta\boldsymbol{\alpha}^T \cdot \int_{t_0}^{T}\left[\frac{\partial H_x}{\partial \boldsymbol{\alpha}} + \frac{d\boldsymbol{\pi}_\alpha}{dt}\right] \cdot dt + \int_{t_0}^{T} \delta\boldsymbol{\pi}_\alpha^T \cdot \frac{d\boldsymbol{\alpha}}{dt} \cdot dt + \delta\boldsymbol{\alpha}^T \cdot \boldsymbol{\pi}_\alpha \Big|_{t=t_0}^{t=T}\,. \tag{12}$$

In (12) the second integral vanishes because of (5a) and the third term is equal to zero because of the transversality condition [3]. It remains

$$\delta\boldsymbol{\alpha}^T \cdot \int_{t_0}^{T}\left[\frac{\partial H_x}{\partial \boldsymbol{\alpha}} + \frac{d\boldsymbol{\pi}_\alpha}{dt}\right] \cdot dt = 0\,. \tag{13}$$

As the variation $\delta\boldsymbol{\alpha}$ is in general not equal to zero it follows from (13)

$$\boldsymbol{\pi}_{\alpha}(T) - \boldsymbol{\pi}_{\alpha}(t_0) + \int_{t_0}^{T} \frac{\partial H_x}{\partial \boldsymbol{\alpha}} \cdot dt = \mathbf{0}\,, \tag{14}$$

which leads with the transversality condition [3] to the final result

$$\int_{t_0}^{T} \frac{\partial H_x}{\partial \boldsymbol{\alpha}} \cdot dt = \mathbf{0}\,. \tag{15}$$

The demand (15) is the necessary condition for the destination of the optimal parameter $\overset{*}{\boldsymbol{\alpha}}$. Equation (15) is firstly given in [6].

Investigating now in (11) the integral I_2, one gets here also by partial integration of the second term and under consideration of the first equation (5b) and the transversality condition [3] an expression equivalent to (13)

$$\delta\boldsymbol{\beta}^T \cdot \int_{t_0}^{T} \int_{s_0}^{L} \left[\frac{\partial H_y}{\partial \boldsymbol{\beta}} + \frac{\partial \boldsymbol{\pi}_{\beta}}{\partial t}\right] \cdot ds \cdot dt = 0\,. \tag{16}$$

As analogous to (13) in (16) $\delta\boldsymbol{\beta}$ is in general not equal to zero one gets from the transversality condition [3] directly

$$\int_{t_0}^{T} \int_{s_0}^{L} \frac{\partial H_y}{\partial \boldsymbol{\beta}} \cdot ds \cdot dt = \mathbf{0}\,. \tag{17}$$

Equation (17) is a necessary condition, too; here for destination of the optimal parameter $\overset{*}{\boldsymbol{\beta}}$.

Finally the integral I_3 in (11) is investigated. With the same considerations as for I_1 and I_2 one gets at first the intermediate result

$$\int_{s_0}^{L} \delta\boldsymbol{\gamma}^T(s) \cdot \int_{t_0}^{T} \left[\frac{\partial H_y}{\partial \boldsymbol{\gamma}} + \frac{\partial \boldsymbol{\pi}_{\gamma}}{\partial t}\right] \cdot dt \cdot ds = \mathbf{0}\,. \tag{18}$$

Because the parameter $\boldsymbol{\gamma}$ depends on the space variable s the variation $\delta\boldsymbol{\gamma}$ cannot removed in front of the integral over ds. As $\delta\boldsymbol{\gamma}$ is in general not equal to zero under consideration of the transversality condition [3] from (18) the simple result

$$\int_{t_0}^{T} \frac{\partial H_y}{\partial \boldsymbol{\gamma}} \cdot dt = \mathbf{0} \tag{19}$$

is derived. The necessary condition for the destination of the optimal parameter $\overset{*}{\boldsymbol{\gamma}}$ is consequently identical with condition (15) for the optimal parameter $\overset{*}{\boldsymbol{\alpha}}$.

4. Example

For illustration of the derived conditions (15), (17) and (19) a simple example which is important for practical investigations is treated.

Given the linear time independent control system

$$\frac{d\mathbf{x}}{dt} = \mathbf{A} \cdot \mathbf{x} + \mathbf{B} \cdot \mathbf{z} + \mathbf{C} \cdot \mathbf{u} + \mathbf{D} \cdot \boldsymbol{\alpha} \tag{20a}$$

with the linear transportation equation

$$\frac{\partial \mathbf{y}}{\partial t} = -v_0 \cdot \frac{\partial \mathbf{y}}{\partial s} + \mathbf{E} \cdot \mathbf{w} + \mathbf{F} \cdot \boldsymbol{\beta} + \mathbf{G} \cdot \boldsymbol{\gamma} . \tag{20b}$$

Equations (20) have the general initial and boundary conditions (3) and (4). From the matrices $\mathbf{A}, \ldots, \mathbf{G}$ it is assumed that they are of suitable dimension to the vectors $\mathbf{x}, \ldots, \boldsymbol{\gamma}$. The transportation velocity is constant. It is denoted by v_0. In (20) the controls $\mathbf{u}(t)$ and $\mathbf{w}(t, s)$ and the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\boldsymbol{\gamma}(s)$ are destinated in an optimal way such that the quadratic performance index

$$J = \frac{1}{2} \cdot \mathbf{x}^T(T) \cdot \mathbf{K} \cdot \mathbf{x}(T) + \frac{1}{2} \cdot \int_0^T [\mathbf{u}^T \cdot \mathbf{R} \cdot \mathbf{u} + \boldsymbol{\alpha}^T \cdot \mathbf{M} \cdot \boldsymbol{\alpha}] \cdot dt +$$

$$+ \frac{1}{2} \cdot \int_0^T \int_0^L [\mathbf{w}^T \cdot \mathbf{Q} \cdot \mathbf{w} + \boldsymbol{\beta}^T \cdot \mathbf{N} \cdot \boldsymbol{\beta} + \boldsymbol{\gamma}^T \cdot \mathbf{P} \cdot \boldsymbol{\gamma}] \cdot ds \cdot dt \tag{21}$$

is minimal [4]. In (21) all matrices are constant and symmetric and of suitable dimension to the vectors $\mathbf{x}, \ldots, \boldsymbol{\gamma}$. The matrices $\mathbf{M}, \mathbf{N}, \mathbf{P}, \mathbf{R}, \mathbf{Q}$ are furthermore positive definite. For the optimal controls $\overset{*}{\mathbf{u}}(t)$ and $\overset{*}{\mathbf{w}}(t, s)$ one gets from the maximum principle [5] directly [3].

$$\begin{aligned} \mathbf{u}(t) &= \mathbf{R}^{-1} \cdot \mathbf{C}^T \cdot \boldsymbol{\psi}(t) , \\ \mathbf{w}(t, s) &= \mathbf{Q}^{-1} \cdot \mathbf{E}^T \cdot \boldsymbol{\Phi}(t, s) . \end{aligned} \tag{22}$$

In (22) $\boldsymbol{\psi}(t)$ and $\boldsymbol{\Phi}(t, s)$ are again the adjoint vectors belonging to $\mathbf{x}(t)$ and $\mathbf{y}(t, s)$.

With the Hamiltonian

$$H_x = -\frac{1}{2} \cdot (\mathbf{u}^T \cdot \mathbf{R} \cdot \mathbf{u} + \boldsymbol{\alpha}^T \cdot \mathbf{M} \cdot \boldsymbol{\alpha}) + \boldsymbol{\psi}^T \cdot (\mathbf{A} \cdot \mathbf{x} + \mathbf{B} \cdot \mathbf{z} + \mathbf{C} \cdot \mathbf{u} + \mathbf{D} \cdot \boldsymbol{\alpha}) \tag{23}$$

it follows from (15)

$$\int_0^T [-\mathbf{N}\cdot\overset{*}{\boldsymbol{\alpha}} + \mathbf{D}^T\cdot\boldsymbol{\psi}(t)]\cdot dt = \mathbf{0}\,, \tag{24}$$

which leads by integration of the first term of (24) to the optimal parameter

$$\overset{*}{\boldsymbol{\alpha}} = \frac{1}{T}\cdot\mathbf{N}^{-1}\cdot\mathbf{D}^T\cdot\int_0^T \boldsymbol{\psi}(t)\cdot dt\,. \tag{25}$$

The Hamiltonian for the transportation system reads as

$$\begin{aligned} H_y = &-\frac{1}{2}\cdot(\mathbf{w}^T\cdot\mathbf{Q}\cdot\mathbf{w} + \boldsymbol{\beta}^T\cdot\mathbf{N}\cdot\boldsymbol{\beta} + \boldsymbol{\gamma}^T\cdot\mathbf{P}\cdot\boldsymbol{\gamma}) + \\ &+\boldsymbol{\Phi}^T\cdot\left(-v_0\cdot\frac{\partial\mathbf{y}}{\partial t} + \mathbf{E}\cdot\mathbf{w} + \mathbf{F}\cdot\boldsymbol{\beta} + \mathbf{G}\cdot\boldsymbol{\gamma}\right). \end{aligned} \tag{26}$$

So one gets from (17) and (26)

$$\int_0^T\int_0^L [-\mathbf{N}\cdot\overset{*}{\boldsymbol{\beta}} + \mathbf{F}^T\cdot\boldsymbol{\Phi}(t,s)]\cdot ds\cdot dt = \mathbf{0}. \tag{27}$$

By integration of the first term in (27) over dt and ds for the optimal parameter $\overset{*}{\boldsymbol{\beta}}$ the expression

$$\overset{*}{\boldsymbol{\beta}} = \frac{1}{T\cdot L}\cdot\mathbf{N}^{-1}\cdot\mathbf{F}^T\cdot\int_0^T\int_0^L \boldsymbol{\Phi}(t,s)\cdot ds\cdot dt \tag{28}$$

is held. Finally it follows with (26) from (19)

$$\int_0^T [-\mathbf{P}\cdot\overset{*}{\boldsymbol{\gamma}}(s) + \mathbf{G}^T\cdot\boldsymbol{\Phi}(t,s)]\cdot dt = \mathbf{0}\,, \tag{29}$$

which leads in the same way as for $\overset{*}{\boldsymbol{\alpha}}$ and $\overset{*}{\boldsymbol{\beta}}$ to the optimal parameter

$$\overset{*}{\boldsymbol{\gamma}}(s) = \frac{1}{T}\cdot\mathbf{P}^{-1}\cdot\mathbf{G}^T\cdot\int_0^T \boldsymbol{\Phi}(t,s)\cdot dt\,. \tag{30}$$

Comparing the results (25), (28) and (30) especially with the optimal controls (22), it can be seen that the optimal parameters are the integral mean values of the controls if the weighting matrices **M**, **N**, **Q** and **R** and the system matrices **C**, **D**, **E**, **F** and **G** are chosen in a suitable way. This fact is as well interesting as important.

References

1. *Hiratsuka, S., Ichikawa, A.*, Optimal control of systems with transportation lags. IEEE Transactions on Automatic Control, **AC-14** (1969) 237–247.
2. *Shimemura, E.*, Optimal problems in control system with time-delay. Memoirs of the School of Science & Engineering, Waseda Univ. *33* (1969) 11–20.
3. *Gwinner, K., Lunderstädt, R.*, Zur Optimierung von Systemen mit Transportvorgängen. Regelungstechnik und Prozeßdatenverarbeitung **20** (1972) 312–313.
4. *Lunderstädt, R.*, Die Optimierung linearer Übertragungssysteme mit Transportvorgängen und quadratischen Gütekriterien. Regelungstechnik und Prozeßdatenverarbeitung **21** (1973) 38–43.
5. *Pontryagin, L. S.*, Mathematische Theorie optimaler Prozesse. R. Oldenbourg Verlag, München 1967.
6. *Hofer, E., Sagirow, P.*, Über die Wahl optimaler Parameter. Regelungstechnik und Prozeßdatenverarbeitung **12** (1967) 537–540.
7. *Lunderstädt, R.*, The optimization of linear systems with transportation lags for quadratic criteria. Preprints of the "Joint Automatic Control Conference 1973 (JACC)" Columbus Ohio (USA), 677–685.
8. *Lunderstädt, R.*, Optimization of systems with transportation lags. Preprints of the "7th Hungarian Conference in Automation", Budapest, Hungary, 1973, 49–66.

Транспортные системы автоматического управления с оптимальными параметрами

Р. ЛУНДЕРШТЕДТ

(Гамбург)

Рассматривается проблематика оптимизации динамических систем автоматического управления, описываемых с помощью обыкновенных дифференциальных уравнений и так называемых транспортных систем (описываемых дифференциальными уравнеиями в частных производных). Уравнения системы зависят как от функций, описывающих управляющие сигналы, так и от параметров систем. Необходимые условия оптимизации параметров выводятся на основе принципа Гамильтона, с помощью классического вариационного исчисления. Теоретические результаты работы освещаются на практическом примере.

R. Lunderstädt
2000 Hamburg 70
Holstenhofweg 85
Federal Republic of Germany

Problems of Control and Information Theory, Vol. 5 (2), pp. 117—133 (1976)

О ПРИБЛИЖЕНИИ ПРОЦЕССОВ СТОХАСТИЧЕСКОЙ АППРОКСИМАЦИИ СУММАМИ НЕЗАВИСИМЫХ СЛУЧАЙНЫХ ВЕЛИЧИН

М. Б. НЕВЕЛЬСОН

(Москва)

(Поступила в редакцию 19 февраля 1975 г.)

Рассматривается задача оценивания нуля неизвестной функции по измерениям, определенным образом зависящим от значений этой функции и случайных ошибок, относительно которых имеется некоторая априорная информация. При широких предположениях строятся оценки типа стохастической аппроксимации, являющиеся асимптотически нормальными с минимальной в обычном смысле предельной дисперсией. Показывается, что эти оценки могут быть аппроксимированы по вероятности и с вероятностью 1 суммами независимых случайных величин. Отсюда выводятся некоторые свойства конечномерных и бесконечномерных распределений данных оценок. Частным случаем рассматриваемой задачи является задача оценивания сигнала, передаваемого по аддитивному каналу связи, «шумовые» характеристики которого полностью не известны.

1. Введение

Пусть $f(x)$ — некоторая неизвестная функция, определенная на прямой E, такая, что уравнение $f(x) = g$, $g = \text{const}$ имеет единственное решение $x = x_0$, причем $(f(x)-g)(x-x_0) > 0$ при $x \neq x_0$. Рассмотрим задачу оценивания величины x_0 для случая, когда статистик может проводить в любой момент времени $t = 1, 2, \ldots$ «независимые» измерения величины $f(x)$ в произвольной точке $x = x(t)$ с некоторой аддитивной случайной ошибкой, так что результат измерения $Y_0(t, x)$ имеет вид $Y_0(t, x) = f(x) + \xi(t, x)$. (Очевидно, эту задачу можно интерпретировать как задачу управления измерениями значений неизвестной функции). Здесь $\xi(t, x)$, $t = 1, 2, \ldots$, $x \in E$ — совокупность случайных величин, причем

$$M\,\xi(t, x) \equiv 0 \quad \text{при всех } t, x. \tag{1.1}$$

Как известно, в такой ситуации для оценивания x_0 может быть использована процедура стохастической аппроксимации Роббинса—Монро

$$X(t+1) - X(t) = -\frac{a^{-1}}{1+t}\left[Y_0\big(t+1, X(t)\big) - g\right], \quad X(0) = x,\ a > 0, \tag{1.2}$$

которая при определенных условиях относительно $f(x)$, $\xi(t, x)$ оказывается сходящейся к x_0 почти наверное (п. н.), если $t \to \infty$ и при обычной нормировке

асимптотически нормальной. При этом предельный нормальный закон будет обладать оптимальными свойствами (в классе процедур вида (1.2)), если $a = f'(x_0) > 0$.

Кроме того, известны процедуры, асимптотически эквивалентные процедуре (1.2) с $a = f'(x_0)$. Впервые процедура такого рода, основанная на идее усечения была построена Вентером [1] при дополнительном предположении $0 < r_1 < f'(x_0) < r_2$, где r_1, r_2 — известные статистику величины. Позднее Фабиан [2] установил, что это предположение может быть опущено, если рассматривать более сложный способ усечения, зависящий от времени (см. также [3, гл. 6], [4]).

Однако можно представить себе ситуации, когда, во-первых, случайные ошибки не являются обязательно аддитивными, и, во-вторых, когда статистик обладает более общей априорной информацией, чем (1.1).

В связи с этим всюду ниже будем предполагать, что в каждый момент времени t и в любой точке $x \in E$ могут быть измерены величины

$$Y_0(t, x) = F\big(f(x), \xi(t, x)\big) . \tag{1.3}$$

Здесь функция $y = F(x, z)$ определена при $x \in \mathfrak{D}_1$, $z \in E$, где $\mathfrak{D}_1$ — какое-либо борелевское множество, содержащее множество значений функции $f(x)$.

Обозначим через $\mathfrak{D}_2$ такое борелевское множество, что $P\{\xi(t, x) \in \mathfrak{D}_2$ при всех $t = 1, 2, \ldots, x \in E\} = 1$.

Всюду в данной работе будем предполагать, не оговаривая этого особо, что относительно функции $F(x, z)$ и случайных величин $\xi(t, x)$ имеется следующая априорная информация:

(а) $F(g, z)$ — известная монотонно-возрастающая функция,

(б) множество значений, которое принимает функция $F(x, z)$ при $x \in \mathfrak{D}_1$, $z \in \mathfrak{D}_2$, содержится в множестве значений, которое принимает функция $F(g, z)$ при $z \in E$,

(в) известно математическое ожидание $S_0(t, x) = MS\big(\xi(t, x)\big)$ для некоторой неубывающей функции $S(z)$.

Кроме того, если функция $S(z)$ не ограничена, будем считать, что для некоторого $\delta > 0$

$$M \mid S\big(\xi(t, x)\big) \mid^{1+\delta} \leq k(t)\big(1 + r(x)\big) , \tag{1.4}$$

где $k(t)$ — числовая последовательность, а функция $r(x)$ известна статистику. Обычно же в работах, посвященных изучению условий асимптотической нормальности процедур Роббинса—Монро (или ее модификаций), когда $F(x, z) = x + z$, $S(z) \equiv z$, $S_0(t, x) \equiv 0$, предполагается, что условие (1.4) выполнено при $\delta = 1$ и $r(x) = x^2$.

Покажем, что при наличии указанной выше априорной информации и при некоторых весьма общих условиях по наблюдениям (1.3) также можно построить простые рекуррентные оценки $X(t)$ такие, что величины $V(t) = \sqrt{t}(X(t) - x_0)$, во-первых, асимптотически нормальны с оптимальной (в том же смысле, что и процедуры из [1]—[4]) предельной дисперсией, а, во-вторых, могут быть аппроксимированы при больших значениях t суммами независимых случайных величин. Отсюда будут выведены некоторые свойства конечномерных и бесконечномерных распределений процесса $X(t)$. Впервые результаты подобного рода получены в [5] для случая непрерывных процедур стохастической аппроксимации (см. также [3], [6]).

Следует отметить, что рассматриваемые процедуры не требуют наличия никакой априорной информации относительно поведения функции $f(x)$ при $|x| \to \infty$. Аналогичным свойством обладают и модификации процедуры Роббинса—Монро, построенные в [7], [8], [9].

Данная работа выполнена по плану сотрудничества между Академией наук СССР и Венгерской Академией наук.

2. Основные понятия и предположения

Будем считать, что все случайные величины определены на некотором вероятностном пространстве $(\Omega, \mathfrak{A}, P)$. Соотношения между случайными величинами, включая сходимость, будем понимать, если не оговорено противное, с вероятностью 1.

Пусть $S_0(z_1, z_2)$, $-\infty \leq z_1 \leq z_2 \leq \infty$ — множество неубывающих функций $S(z)$ таких, что

(а) если $z_1 \neq z_2$, то функция $S(z)$ строго возрастает на интервале (z_1, z_2),

(б) если $z_1 \neq -\infty$, то $S(z_1 + 0) > S(z)$ при $z_1 > z$,
если $z_2 \neq \infty$, то $S(z_2 - 0) < S(z)$ при $z_2 < z$.

Очевидно, $S_0(-\infty, \infty)$ — совокупность строго возрастающих на E функций, а $S_0(z_1, z_1)$ — совокупность неубывающих функций, разрывных в точке z_1 $(-\infty < z_1 < \infty)$.

Обозначим также через $Q(S)$ множество измеримых (по Борелю) функций $q(z)$ таких, что

(1) $q(z)$ положительна, ограничена и для любых q_1, q_2

$$\inf_{q_1 < z < q_2} q(z) > 0,$$

(б) функция $S(z)\,q(z)$ ограничена и не убывает.

Введем следующее обозначение. Если $a(t)$, $b(t)$ — две последовательности, причем $a(t) < b(t)$, то для любого $x \in E$ положим

$$[x]_{a(t)}^{b(t)} = \begin{cases} x, & \text{если} \quad a(t) \le x \le b(t) \\ a(t), & \text{если} \quad x < a(t) \\ b(t), & \text{если} \quad x > b(t). \end{cases}$$

Обозначим через $\mathfrak{B}$ — σ-алгебру борелевских множеств прямой E, а через $\mathfrak{B}_1$ ($\mathfrak{B}_2$) — σ-алгебру борелевских множеств $\mathfrak{D}_1$ ($\mathfrak{D}_2$). Пусть еще $\mathfrak{A}_t$ — σ-алгебра, порожденная семейством случайных величин $\{\xi(t, x), x \in E\}$, а $\mathfrak{F}_t$ — σ-алгебра, порожденная семейством $\{\xi(u, x), u \le t, x \in E\}$.

Всюду ниже мы постоянно предполагаем, не оговаривая этого особо, что $\xi(t, x)$ — $\mathfrak{A}_t \times \mathfrak{B}$ — измерима и семейство случайных величин $\{\xi(t, x), x \in E\}$ не зависит от $\mathfrak{F}_{t-1}$.

Кроме того, неоднократно будут использоваться следующие предположения, носящие общий характер.

Предположение 2.1. Функция $f(x)$ измерима (по Борелю), а функция $F(x, z)$ — $\mathfrak{B}_1 \times \mathfrak{B}$ измерима, причем

(а) $(f(x) - g)(x - x_0) \ge 0$, $x \in E$ $(F(x, z) - F(g, z))(x - g) \ge 0$, $x \in \mathfrak{D}_1$, $z \in \mathfrak{D}_2$,

(б) для каждого $\varepsilon \in (0, 1)$ справедливы соотношения

$$\inf_{\varepsilon < |x - x_0| < 1/\varepsilon} |f(x) - g| > 0, \quad \sup_{|x| < 1/\varepsilon} |f(x)| < \infty, \quad \sup_{x \in \mathfrak{D}_1,\, Z \in \mathfrak{D}_2} |F(x, z)| < \infty.$$

Обозначим через $z = \psi(y)$ функцию, обратную к функции $y = F(g, z)$ и положим

$$Z(t, x) = \psi(Y_0(t, x)), \quad Y(t, x) = S(Z(t, x)) - S_0(t, x).$$

Предположение 2.2. Функция $R(t, x) = MY(t, x)$ конечна при всех $t = 1, 2, \ldots$ и x из ν — окрестности точки x_0 (ν — некоторое положительное число), имеет при этих значениях t, x производную $\dfrac{\partial R}{\partial x}(t, x) = R'(t, x)$, непрерывную в точке $t = \infty$, $x = x_0$.

Предположение 2.3. Выполнено предположение 2.2, функция $R'(t, x)$ удовлетворяет в ν — окрестности точки x_0 равномерно по t условию Гельдера с показателем $\mu_1 > 0$, причем $[R'(t, x_0) - \lambda]^{\mu_2} \to 0$ при $t \to \infty$, где $\lambda = \lim_{t \to \infty} R'(t, x_0)$, μ_2 — некоторое положительное число.

Предположение 2.4. Справедливо неравенство

$$\sup_{t \ge 0,\, |x - x_0| < \nu_1} M \Phi^2(t, x) < \infty, \tag{2.1}$$

где $\nu_1 > 0$, $\Phi(t, x) = Y(t, x) - R(t, x)$, а функция $A(t, x) = M\Phi^2(t, x)$ непрерывна в точке $t = \infty$, $x = x_0$.

Предположение 2.5. Функция $\Phi(t, x)$ непрерывна в среднем квадратичном в точке x_0 равномерно по t.

Предположение 2.6. Функция $\Phi(t, x)$ удовлетворяет в ν-окрестности точки x_0 равномерно по t условию Гельдера в среднем квадратичном с показателем $\mu > 0$.

Предположение 2.7. Семейство случайных величин $\Phi(t, x)$, $t = 1, 2, \ldots, \ldots, |x - x_0| < \nu_1$ равномерно интегрируемо.

Предположение 2.8. Для некоторого $\varkappa > 0$

$$\sup_{t \geq 0} M \mid S(\xi(t, x)) \mid^{2+\varkappa} < \infty .$$

Рекуррентные процедуры

Как обычно, нам понадобятся оценки величины

$$\lambda = \lim_{t \to \infty} R'(t, x_0).$$

Для построения таких оценок, следуя [1], [2], введем следующую конструкцию. Пусть $X(t)$ — произвольный $\mathfrak{F}_t$-измеримый случайный процесс, а t_i, $i = 1, 2, \ldots$ — последовательность положительных чисел, удовлетворяющая следующим условиям:

$$t_{j+1} > t_j, \quad \lim_{i \to \infty} t_i = \infty, \quad j = 1, 2, \ldots . \tag{3.1}$$

Положим

$$W(t) = W\big(t\,; t_j, X(t_j - 1), c(j), j = 1, \ldots, q(t)\big) =$$

$$= \frac{1}{q(t)} \sum_{j=1}^{q(t)} Y_{0j}\big(t_j, X(t_j - 1)\big), \tag{3.2}$$

где

$$Y_{0j}(t, x) = \frac{Y\big(t, x + c(j)\big) - Y\big(t, x - c(j)\big)}{2c(j)},$$

$q(t)$ — число тех t_j, которые не превосходят t, а

$c(j)$ — некоторая последовательность положительных чисел.

Легко видеть, что значения процесса $W(t)$ в момент времени t можно вычислять по следующим рекуррентным формулам

$$\begin{cases} W(t) = \begin{cases} 0 & \text{при } 0 \leq t < t_i \\ W(t_i) & \text{при } t_i \leq t < t_{i+1}, \ i = 1, 2, \ldots \end{cases} \\ W(t_{i+1}) - W(t_i) = \dfrac{1}{1 + i}\big[-W(t_i) + Y_{0i+1}\big(t_{i+1}, X(t_{i+1} - 1)\big)\big], \ t_0 = 0. \end{cases} \tag{3.3}$$

По аналогии с [2] всюду ниже будем проводить усечение процесса $W(t)$ с помощью некоторых последовательностей положительных чисел.

Рассмотрим сначала процедуру

$$\begin{cases} X(t+1) - X(t) = -\dfrac{Y(t+1, X(t))}{(1+t)\,\widetilde{W}(t)}, \quad X(0) = x \\ \widetilde{W}(t) = [W(t)]_{a(t)}^{b(t)}. \end{cases} \tag{3.4}$$

Здесь $W(t) = W(t;\, t_j,\, X(t_j-1),\, c(j),\, j = 1, \ldots, q(t))$ — процесс, определяемый формулами (3.2) или (3.3), $a(t)$, $b(t)$, $c(t)$ — некоторые последовательности положительных чисел, x — произвольная точка прямой, а t_j удовлетворяют соотношениям (3.1).

Приводимые ниже теоремы показывают, что для ограниченной функции $S(z)$ при широких предположениях можно так выбрать последовательности $a(t), b(t), c(t)$, чтобы величины $\sqrt{t}\,(X(t) - x_0)$ были асимптотически нормальны и приближались с ростом t к суммам независимых случайных переменных. Для того, чтобы можно было гарантировать справедливость аналогичных свойств процедуры (3.4) и в случае неограниченной функции $S(z)$, статистику необходимо иметь дополнительную априорную информацию. В частности, он должен знать, что

$$M\,|S(\varphi(F(f(x),\, \xi(t,\, x))) - S(\xi(t,\, x))| \leq K(1 + |x|), \tag{3.5}$$

$$M\,[S(\varphi(F(f(x),\, \xi(t,\, x))) - MS(\varphi(F(x),\, \xi(t,\, x)))]^2 < K(1 + x^2), \tag{3.6}$$

где K — некоторая постоянная. Так как функция $f(x)$ неизвестна, то условия (3.5), (3.6) могут оказаться трудно проверяемыми. Поэтому в общем случае рассмотрим другую процедуру. Для этого построим сначала процесс $X_0(t)$, сходящийся к x_0 п. н. при $t \to \infty$, а затем с его помощью проведем по аналогии с [10] усечение процесса (3.4).

Именно, положим (см. [11], [12])

$$X_0(t) = \begin{cases} \tilde{x},\ 0 \leq t < t_1 \\ X_0(t_i),\ t_i \leq t < t_{i+1}, \end{cases} \tag{3.7}$$

$$X_0(t_{i+1}) - X_0(t_i) = -\frac{a}{(1+i)\,r(X_0(t_i))}\,[A_{i+1}(Z(i+1, X_0(t_i))) - S_0(i+1, X_0(t_i)))], \tag{3.8}$$

где $\tilde{x}$ — произвольная точка прямой, $a > 0$ и

$$A_t(z) = \frac{S(z)\,q(z)}{q(z) + d(t)}, \quad q(z) \in Q(S), \quad a > 0.$$

Пусть теперь

$$\begin{cases} X(t+1) = \left[X(t) - \dfrac{1}{(1+t)\,\widetilde{W}(t)}\, Y\big(t+1, X(t)\big)\right]_{X_0(t)-\alpha}^{X_0(t)+\alpha}, \quad X(0) = x \\ \widetilde{W}(t) = [W(t)]_{a(t)}^{b(t)} \end{cases} \tag{3.9}$$

Здесь, как и выше, $W(t) = W\big(t; t_i, X(t_j - 1), c(j),\ j = 1, \ldots, q(t)\big)$, $0 < a(t) < b(t)$, $c(t) > 0$, $\alpha > 0$.

Оказывается, процедура (3.9) позволяет опустить условие (3.5) и заменить условие (3.6) на условие $MS^2\big(\xi(t,x)\big) < K(1+x^2)$ (или даже более общее условие), в которое не входит неизвестная функция $f(x)$).

4. Асимптотические свойства распределений

При изучении свойств распределений процедур (3.4), (3.9) будем предполагать, что выполнены следующие условия (А) (относительно смысла этих условий см. [11]). Положим

$$U_\varepsilon(a) = \{x : \varepsilon < |x - a| < 1/\varepsilon\},\ U_\varepsilon^+(a) = \{x : \varepsilon < x - a < 1/\varepsilon\},$$

$$U_\varepsilon^-(a) = \{x : -1/\varepsilon < x - a < -\varepsilon\},$$

$$\bar{f}_\varepsilon^{(1)} = \sup_{\varepsilon < x - x_0 < 1/\varepsilon} [f(x) - g],\ f^{(1)} = \inf_{\varepsilon < x - x_0 < 1/\varepsilon} [f(x) - g],$$

$$\bar{f}_\varepsilon^{(2)} = \sup_{-1/\varepsilon < x - x_0 < -\varepsilon} [g - f(x)],\ f^{(2)} = \inf_{-1/\varepsilon < x - x_0 < -\varepsilon} [g - f(x)].$$

Условия (А). Для любого $\varepsilon \in (0, 1)$ существуют такие множества $C_\varepsilon \in \mathfrak{B}_2$, $\mathfrak{D}_\varepsilon \in \mathfrak{B}_2$, что

(А.1)

$$\bar{\delta}_\varepsilon = \inf_{z \in C_\varepsilon} [S(\alpha_\varepsilon + z) - S(z)] > 0,$$

$$\delta_\varepsilon = \inf_{z \in \mathfrak{D}_\varepsilon} [S(z) - S(z - \beta_\varepsilon)] > 0,$$

где

$$\alpha_\varepsilon = \inf_{x \in \{x : \bar{f}^{(1)} > x - g > f_\varepsilon^{(1)}\} \cap \mathfrak{D}_1,\, z \in C_\varepsilon} \big[\varphi\big(F(x,z)\big) - z\big],$$

$$\beta_\varepsilon = \inf_{x \in \{x : -\bar{f}_\varepsilon^{(2)} < x - g < -f_\varepsilon^{(2)}\} \cap \mathfrak{D}_1,\, z \in \mathfrak{D}_2} \big[z - \varphi\big(F(x,z)\big)\big],$$

(А.2)

$$\sum_{t=0}^{\infty} \frac{1}{(1+t)\,b(t)} \inf_{\varepsilon < x - x_0 < 1/\varepsilon} P\{\xi(t,x) \in C_\varepsilon\} = \infty,$$

$$\sum_{t=0}^{\infty} \frac{1}{(1+t)\,b(t)} \inf_{-1/\varepsilon < x - x_0 < -\varepsilon} P\{\xi(t,x) \in \mathfrak{D}\varepsilon\} = \infty.$$

Нетрудно проверить, что условия (А) выполнены, если справедливо одно из следующих условий:

(а) $S(z) \in S_0(-\infty, \infty)$ и для любого $\varepsilon \in (0,1)$ существует такое ограниченное множество $E_\varepsilon \in \mathfrak{B}_2$, что

$$\inf_{x \in U_\varepsilon(g) \cap \mathfrak{B}_1 z \in E_\varepsilon} |F(x,z) - F(g,z)| > 0, \quad \sum_{t=0}^{\infty} \frac{1}{(1+t)\,b(t)} \inf_{x \in U_\varepsilon(x_0)} P\{\xi(t,x) \in E_\varepsilon\} = \infty$$

или

(б) $S(z) \in S_0(z_1, z_2)$, $-\infty < z_1 \leq z_2 < \infty$ и для любых $\varepsilon \in (0,1)$ и $\nu > 0$

$$\inf_{x \in U_\varepsilon^+(g) \cap \mathfrak{B}_1, z \in [z_1 - \nu, z_2] \cap \mathfrak{B}_2} [F(x,z) - F(g,z)] > 0,$$

$$\inf_{x \in U_\varepsilon^-(g) \cap \mathfrak{B}_1, z \in [z_1, z_2 + \nu] \cap \mathfrak{B}_2} [F(g,z) - F(x,z)] > 0,$$

$$\sum_{t=0}^{\infty} \frac{1}{(1+t)\,b(t)} \inf_{x \in U_\varepsilon^+(x_0)} P\{z_1 - \nu < \xi(t,x) < z_2\} = \infty,$$

$$\sum_{t=0}^{\infty} \frac{1}{(1+t)\,b(t)} \inf_{x \in U_\varepsilon^-(x_0)} P\{z_1 < \xi(t,x) < z_2 + \nu\} = \infty.$$

Справедливы следующие утверждения, доказательство которых даны в приложении.

Теорема 4.1. Пусть

1°. Функция $S(z)$ ограничена.

2°. Выполнены предположения 2.1, 2.2 и условия (А).

3°. Последовательности $a(t)$, $b(t)$, $c(t)$ выбраны так, что

$$a(t) \to 0, \quad b(t) \to \infty, \quad c(t) \to 0, \tag{4.1}$$

$$\sum_{t=0}^{\infty} (1+t)^{-2} [a^{-2}(t) + c^{-2}(t)] < \infty. \tag{4.2}$$

4°. Выполнено предположение 2.4.

Тогда для процесса (3.4) справедливо соотношение $\sqrt{t}\,(X(t) - x_0) \sim N\left(0, \frac{\sigma^2}{\lambda^2}\right)$, где $\sigma^2 = \lim\limits_{t \to \infty, x \to x_0} A(t,x)$, $\lambda = \lim\limits_{t \to \infty} R'(t, x_0)$.

Теорема 4.2. Пусть выполнены условия 1°–3° теоремы 4.1 и предположение 2.5. Тогда для процесса (3.4) имеем: $\sqrt{t}\,(X(t) - x_0) - Z_0(t) \to 0$ по вероятности при $t \to \infty$, где

$$Z_0(t) = -\frac{1}{\lambda \sqrt{t}} \sum_{k=1}^{t} \Phi(k, x_0).$$

Теорема 4.3. Пусть выполнены условия 1°–3° теоремы 4.1, предположения 2.3, 2.6 и , кроме того,

$$c(t) \geq Ct^{-c}, \quad q(t) \geq Qt^{q}, \quad C > 0, \; Q > 0, q > 0. \tag{4.3}$$

Тогда с вероятностью 1 для процесса (3.9) справедливо соотношение*

$$\sup_{t \geq 0} \{t^m [\sqrt{t}\,(X(t) - x_0) - Z_0(t)] < \infty$$

для любого $m < \min\,(\mu/2,\ 1/2,\ qc\mu_1,\ q\mu_2)$.

Теорема 4.4. Пусть

1°. Справедливо неравенство (1.4), где $\delta > 0$, а функция $r(x)$ локально ограничена.

2°. Последовательности $a(t)$, $b(t)$, $c(t)$, $d(t)$, $k(t)$ удовлетворяют соотношениям (4.1), (4.2) и неравенству

$$\sum_{t=0}^{\infty} \left[\frac{d^{\delta}(t)\, k(t)}{1+t} + \frac{1}{(1+t)^2\, d^2(t)} \right] < \infty\,,$$

3°. Выполнены предположения 2.1, 2.2, условия (А) и $\alpha < \nu_1/2$.

4°. Выполнены предположения 2.4, 2.7.

Тогда для процесса (3.9) справедливо заключение теоремы 4.1.

Теорема 4.5. Пусть выполнены условия 1°—3° теоремы 4.4 и предположения 2.5, 2.8. Тогда для процесса (3.9) справедливо заключение теоремы 4.2.

Теорема 4.6. Пусть выполнены условия 1°–3° теоремы 4.4, предположения 2.3, 2.6, 2.8, а последовательности $a(t)$, $c(t)$, $q(t)$ выбраны согласно (4.3). Тогда для процесса (3.9) справедливо заключение теоремы 4.3.

Замечание 4.1. Общее число наблюдений $n(t)$, использованных при построении процесса $X(t)$, описываемого формулами (3.4) или (3.9), равно соответственно $t + 2q(t)$ или $t + 3q(t)$. Если выбрать последовательность t_j так, что $\lim\limits_{j \to \infty} t_j/j = \infty$, то тогда $\lim\limits_{t \to \infty} q(t)/t = 0$. Поэтому при таком выборе t_j в условиях теорем 4.1, 4.4 будем иметь и соотношение $\sqrt{n(t)}\ (X(t) - x_0) \sim N(0,\ \sigma^2/\lambda^2)$.

Обозначим через $V(t)$ стационарный гауссовский процесс, удовлетворяющий стохастическому дифференциальному уравнению Ито $dV(t) = -1/2\ V(t)\,dt + \sigma/\lambda\,d\eta(t)$, где $\eta(t)$ — стандартный винеровский процесс. Пусть, далее, $V_t(s) = \sqrt{te^s}\,(X(te^s) - x_0)$, где t, te^s — целые положительные числа, а $\tilde{V}_t(s)$, $s \geq 0$ — случайная ломаная, которая совпадает с процессом $V_t(s)$ в точках $s_i = \ln\,(1 + i/t)$, $i = 0, 1, 2, \ldots$

* Аналогичное утверждение для оценок максимального правдоподобия установлено в [6].

Из теорем 4.2, 4.3, 4.5, 4.6, теоремы 1 гл. 9§ 4 из [13] (см. также теорему 6.3 гл. 6 [2] и лемму из [6]) вытекают.

Следствие 4.1. Если выполнены условия теоремы 4.2 или 4.5, то конечномерные распределения процесса $\widetilde{V}_i(s)$, $s \geq 0$ сходятся при $t \to \infty$ к конечномерным распределениям процесса $V(s)$.*

Следствие 4.2. Если выполнены условия теорем 4.3 или 4.6, то распределение, порожденное в $C[0, L]$, $L > 0$, процессом $\widetilde{V}_i(s)$, $s \geq 0$ слабо сходится при $t \to \infty$ к распределению, порожденному в $C[0, L]$ процессом $V(s)$.

Приложение

Лемма П.1. Пусть выполнено предположение 2.2, неравенство (2.1) и $X(t_j - 1) \to x_0$ п. н. при $j \to \infty$. Тогда $W(t) \to \lambda$ п. н. при $t \to \infty$ для любой последовательности $c(j) > 0$, удовлетворяющей условиям: $c(j) \to 0$ при $j \to \infty$ и $\Sigma[jc(j)]^{-2} < \infty$. Если же вместо предположения 2.2 выполнено предположение 2.3 и $\sup\limits_{t \geq 1} [t^c c(t) < \infty$, $\inf\limits_{t \geq 1} [q(t)/t^q] > 0$, где $c > 0$, $q > 0$, то $[W(t) - \lambda]t^m \to 0$ п. н. при $t \to \infty$ для любого $m < q \min(c\mu_1, \mu_2)$.

Доказательство этой леммы можно провести аналогично доказательству лемм из [1], [3, гл. 6].

Доказательство теорем 4.1–4.3

1. Пусть выполнены условия 1°–3° теоремы 4.1, а процесс $X(t)$ определяется формулами (3.4). Простые вычисления показывают, что

$$M\{[X(t+1-x_0]^2/X(1), \ldots, X(t)\} \leq (X(t) - x_0)^2 - \\ - \frac{2R(t+1, X(t))(X(t) - x_0)}{(1+t)b(t)} + \frac{c_1}{(1+t)^2 a^2(t)}. \qquad (\text{П.1})$$

(Здесь и ниже c_i — некоторые положительные постоянные).

Очевидно,

$$R(t, x) = M[S(\varphi(F(f(x), \xi(t, x)))) - S(\varphi(F(g, \xi(t, x))))].$$

Так как функция $S(\varphi(y))$ не убывает и $f(x)(x - x_0) \geq 0$, $[F(x, z) - F(g, z)](x - g) \geq 0$, то $R(t, x)(x - x_0) \geq 0$. Отсюда и (5.1) как обычно вытекает существование конечного предела последовательности $|X(t) - x_0|$ при $t \to \infty$ и неравенство

$$\sum_{t=0}^{\infty} \frac{R(t+1, X(t))}{(1+t)b(t)} < \infty. \qquad (\text{П.2})$$

* Для случая, когда $F(x, z) = x + z$ это утверждение при несколько более жестких условиях установлено в [12].

Из условий (А) следует соотношение

$$\sum_{t=0}^{\infty} \frac{r_\varepsilon(t+1)}{(1+t)\,b(t)} = \infty\,, \tag{П.3}$$

где $r_\varepsilon(t) = \inf\limits_{x \in U\varepsilon(x_0)} |R(t, x)|$, а ε — произвольное число из интервала (0, 1).

Из (П.2), (П.3) и того, что $\lim\limits_{t\to\infty} |X(t) - x_0| < \infty$ получаем

$$\lim_{t\to\infty} X(t) = x_0\,. \tag{П.4}$$

В силу леммы (П.1) и (П.4)

$$\lim_{t\to\infty} \widetilde{W}(t) = \lambda\,. \tag{П.5}$$

Согласно предположению 2.2 и (П.4) $R(t+1, X(t)) = \gamma(t)\,[X(t) - x_0]$, где

$$\lim_{t\to\infty} \gamma(t) = \lambda\,. \tag{П.6}$$

Следовательно,

$$X(t+1) - x_0 = \left[1 - \frac{\tilde{\gamma}(t)}{1+t}\right](X(t) - x_0) - \frac{\Phi(t+1, X(t))}{(1+t)\,\widetilde{W}(t)}\,.$$

Здесь $\tilde{\gamma}(t) = \gamma(t)\,\widetilde{W}^{-1}(t)$, причем в силу (П.5), (П.6)

$$\lim_{t\to\infty} \tilde{\gamma}(t) = 1\,. \tag{П.7}$$

Положим

$$\hat{\gamma}(t) = \tilde{\gamma}(t), \quad \text{если } 2 > \tilde{\gamma}(t) > 1/2,\ \hat{\gamma}(t) = 1, \text{ если } \tilde{\gamma}(t) \leqq 1/2 \text{ или } \tilde{\gamma}(t) \geq 2 \tag{П.8}$$

$$\widehat{W}(t) = \widetilde{W}(t), \text{ если } \widetilde{W}(t) > \lambda/2,\ \widehat{W}(t) = 1 \quad \text{если } \widetilde{W}(t) \leq \frac{\lambda}{2} \tag{П.9}$$

и рассмотрим новый процесс $\widehat{X}(t) = \widehat{X}^{s,}\zeta(t)$, который определяется формулами

$$\widehat{X}(t+1) - x_0 = \left[1 - \frac{\hat{\gamma}(t)}{1+t}\right](\widehat{X}(t) - x_0) - \frac{\Phi(t+1, \widehat{X}(t))}{(1+t)\,\widehat{W}(t)}\,,\ \widehat{X}(s) = \zeta, \tag{П.10}$$

где $s \geq 0$, $\zeta - \mathfrak{F}_s$ — измеримая случайная величина с конечной дисперсией.

Как обычно, можно показать, что

$$M\,|\widehat{X}(t) - x_0| = O(1/\sqrt{t}),\ t \to \infty \tag{П.11}$$

Полагая $V(t) = \sqrt{t}\,(\widehat{X}(t) - x_0)$, получим из (П.10) (П.12)

$$V(t+1) = A_{s-1t}\sqrt{s}\,(\zeta - x_0) + \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}}\,\psi(k)\,(\widehat{X}(t) - x_0) -$$

$$- \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \Phi\big(k+1, \widehat{X}(k)\big)\, \widehat{W}^{-1}(k)\,. \tag{П.13}$$

Здесь $\psi(t) = \hat{\gamma}(t) - 1$,

$$A_{kt} = \begin{cases} \prod_{m=k+1}^{t} \big(1 - 1/[2(m+1)]\big), & 1 \le k < t \\ 1, & k = t\,. \end{cases}$$

Очевидно,

$$\lim_{t\to\infty} \psi(t) = 0, \quad \sup_{t\ge 0} |\psi(t)| < c_2\,. \tag{П.14}$$

Нетрудно проверить, что

$$A_{kt} = (1 + \Theta_{kt}) \sqrt{\frac{k+1}{t+1}}, \tag{П.15}$$

где

$$\sup_{t\ge k} |\Theta_{kt}| = O(1/k), \quad k \to \infty\,. \tag{П.16}$$

Первое слагаемое в правой части (П.13) стремится к нулю п. н. при $t \to \infty$. Из (П.11), (П.14) вытекает, что и второе слагемое стремится к нулю по вероятности с ростом t. Следовательно,

$$V(t+1) + \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{\Phi\big(k+1, \widehat{X}(k)\big)}{\widehat{W}(k)} \to 0 \tag{П.17}$$

по вероятности при $t \to \infty$.

2. Пусть выполнены условия п. 1 и предположение 2.4, Тогда, принимая во внимание соотношения $\widehat{X}(t) \to x_0$, $\widehat{W}(t) \to \lambda$ при $t \to \infty$, получим $M\{\Phi^2(k+1, \widehat{X}(k))\, \widehat{W}^{-2}(k) / \widehat{X}(s), \ldots, \widehat{X}(k)\} \to \sigma^2/\lambda_2$ при $k \to \infty$. Отсюда и (П.17) вытекает (см. [14]), что

$$V(t) \sim N(0, \sigma^3/\lambda^2)\,. \tag{П.18}$$

Из определения процесса $\widehat{X}(t)$ и (11.5), (11.7) имеем:

$$P\{X(t) \equiv \widehat{X}^{T, X(T)}(t) \text{ при } t \ge T\} > 1 - \delta \tag{П.19}$$

для любого $\delta > 0$ и некоторого $T = T(\delta)$. Это вместе с (П.18) доказывает теорему 4.1.

3. Пусть выполнены условия п. 1 и предположение 2.5. Положим $\psi(k, x) = M[\Phi(k+1, x_0)] - \Phi(k+1, x_0)]^2$.

Используя ограниченность функции $\Phi(k, x)$, находим, что $\psi\big(k, x(k)\big) \to 0$ для любой последовательности $x(k)$ такой, что $x(k) \to x_0$ при $k \to \infty$. Следовательно,

$$\Psi(k) = M\left[\Phi\left(k+1, \hat{X}(k)\right) - \Phi(k+1, x_0)\right]^2 = M\,\Psi\left(k, \hat{X}(k)\right) \to 0\,. \quad (\text{П}.20)$$

Рассмотрим процесс

$$\zeta(t) = \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{\left[\Phi\left(k+1, \hat{X}(k)\right) - \Phi(k+1\, x_0)\right]}{\hat{W}(k)}\,.$$

Из (П.20) имеем $M\zeta^2(t) \to 0$ и, значит, $\zeta(t) \to 0$ по вероятности. Поэтому, в силу (П.17),

$$V(t+1) + \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{\Phi(k+1, x_0)}{\hat{W}(k)} \to 0\,. \quad (\text{П}.21)$$

по вероятности при $t \to \infty$. Точно также, используя (П.5), неравенство $\hat{W}^{-1}(k) \leq \max(1, \lambda/2)$ и независимость случайных величин $\Phi(k+1, x_0)$, $\hat{W}(k)$ установим, что

$$\sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \left[\hat{W}^{-1}(k) - \lambda^{-1}\right] \Phi(k+1, x_0) \to 0 \quad (\text{П}.22)$$

по вероятности при $t \to \infty$. Из (П.21), (П.22), учитывая соотношения (П.16), получаем утверждение теоремы 4.2.

4. Пусть выполнены условия п. 3, предположения 2.3, 2.6 и соотношения (4.3), (4.4).

Из (П.12), (П.19) вытекает, что $(X(t) - x_0)\, t^{1/2-\varepsilon} \to 0$ при $t \to \infty$ для любого $\varepsilon > 0$. Отсюда и предположения 2.3 получаем: $R(t+1, X(t)) = \gamma(t)\,(X(t) - x_0)$, где

$$\lim_{t\to\infty} \left[\gamma(t) - \lambda\right] t^{m_1} \to 0 \quad \text{для любого} \quad m_1 < \min(1/2\,\mu_1, \mu_2). \quad (\text{П}.23)$$

Кроме того, в силу леммы П.1

$$\lim_{t\to\infty} \left[\hat{W}(t) - \lambda\right] t^{m_3} \to 0 \quad \text{для любого} \quad m_2 < \min(qc\,\mu_1, q\mu_2) \quad (\text{П}.24)$$

Из (П.23), (П.24) вытекает следующее соотношение $\left[\tilde{\gamma}(t) - 1\right] t^{m_3} \to 0$ при $t \to \infty$ для любого $m_3 < \min(qc\mu_1, q\mu_2)$ и, значит,

$$\psi(t)\, t^{m_3} \to 0 \quad (\text{П}.25)$$

для тех же значений m_3.

Используя (П.12), (П.15), (П.25), получаем

$$t^{m_4} \left| \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \psi(k) \left(\hat{X}(k) - x_0\right) \right| \leq k_1\, t^{m_4 - 1/2} \sum_{k=s}^{t} k^{-1/2-m_3+\varepsilon} \to 0 \quad (\text{П}.26)$$

с вероятностью 1, если $m_4 < m_3$ и $\varepsilon < m_3 - m_4$. (Здесь k_i — некоторая, быть может, случайная постоянная).

Далее, из доказательства леммы 4.2 гл. 6 [3], находим, что с вероятностью 1

$$t^{m_5} \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{[\Phi(k+1, \hat{X}(k)) - \Phi(k+1, x_0)]}{\hat{W}(k)} \to 0, \quad m_5 < \mu/2, \tag{П.27}$$

$$t^{m_6} \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} [\hat{W}^{-1}(k) - \lambda^{-1}] \to 0, \quad m_6 < \min(qc\mu_1, q\mu_2). \tag{П.28}$$

Из (П.13), (П.16), (П.26)–(П.28) вытекает утверждение теоремы.

Доказательство теорем 4.4—4.6

Пусть выполнены предположения 1°–3° теоремы 4.4. Повторяя с очевидными изменениями рассуждения, проведенные при доказательстве теоремы 2.3 из [11], получим, что $X_0(t) \to x_0$ п. н. при $t \to \infty$. Следовательно, $X_0(t) = x_0 + \psi(t)$, где $\psi(t) \to 0$ п. н.

Пусть α_0 — произвольное положительное число, а $T_1 = T_1(\alpha_0)$ таково, что

$$P\{|\psi(t)| < \alpha \text{ при всех } t \geq T_1\} > 1 - \frac{\alpha_0}{3}. \tag{П.29}$$

Рассмотрим новый процесс $X_1(t)$, определяемый формулами

$$\begin{cases} X_1(t+1) = \left[X_1(t) - \dfrac{Y(t+1, X_1(t))}{(1+t)\tilde{W}(t)} \right]_{x_0 + [\psi(t)]_{-\alpha}^{\alpha} - \alpha}^{x_0 + [\psi(t)]_{-\alpha}^{\alpha} + \alpha} \\ X_1(T_1) = X(T_1). \end{cases} \tag{П.30}$$

Тогда

$$M\{[X_1(t+1) - x_0]^2 / X_1(T_1), \ldots, X_1(t)\} \leq$$
$$\leq [X_1(t) - x_0]^2 - \frac{2R(t+1, X_1(t))(X_1(t) - x_0)}{(1+t)b(t)} +$$
$$+ [(1+t)a(t)]^{-2} M\{\Phi^2(t+1, X_1(t)) / X_1(T_1), \ldots, X_1(t)\}. \tag{П.31}$$

Неравенство $|X_1(t) - x_0| \leq 2\alpha < \nu_1$ при $t \geq T_1$ вместе с предположением 2.4 дают $M\{\Phi^2(t+1, X_1(t)/X_1(T_1)), \ldots, X_1(t)\} < c_4$ при $t \geq T_1$ Отсюда и из (п. 31) точно так же, как и при доказательстве теоремы 4.1 получаем равенство

$$\lim_{t \to \infty} X_1(t) = x_0. \tag{П.32}$$

Так как, в силу (п.29),

$$P\{X(t) \equiv X_1(t) \text{ при всех } t \geq T_1\} > 1 - \frac{\alpha_0}{3}, \tag{П.33}$$

то из (п.32) и произвольности α_0 следует соотношение

$$\lim_{t\to\infty} X(t) = x_0 . \tag{П.34}$$

Из (П.30) и (п.34) вытекает, что процесс $X_1(t) - [(1+t)\,\widetilde{W}(t)]^{-1}\,Y\big(t+1, X(t)\big)$ с вероятностью 1 выходит из отрезка $\big[x_0 + [\psi]^{\alpha}_{-\alpha} + \alpha,\, x_0 + [\psi]^{\alpha}_{-\alpha} - \alpha\big]$[1] лишь конечное число раз. Следовательно, для некоторого $T_2 = T_2(\alpha_0)$

$$P\{X_1(t) \equiv X_2(t) \text{ при всех } t \geq T_2\} > 1 - \frac{\alpha_0}{3}, \tag{П.35}$$

где процесс $X_2(t)$ определяется при $t \geq T_2$ формулами

$$X_2(t+1) - X_2(t) = -\frac{\hat{Y}\big(t+1, \hat{X}_2(t)\big)}{(1+t)\,\widehat{W}(t)}$$

$$X_2(T_2) = X_1(T_2), \quad \hat{Y}(t,x) = \hat{R}(t,x) + \hat{\Phi}(t,x),$$

$$\hat{R}(t,x) = \begin{cases} R(t,x) & \text{при } |x| < 2\alpha \\ \operatorname{sign} x & \text{при } |x| \geq 2\alpha, \end{cases} \qquad \hat{\Phi}(t,x) = \begin{cases} \Phi(t,x) & \text{при } |x| < 2\alpha \\ \Phi(t,x_0) & \text{при } |x| \geq 2\alpha. \end{cases}$$

Легко проверить, что $X_2(t) \to x_0$. Поэтому

$$X_2(t+1) - x_0 = \left(1 - \frac{\tilde{\gamma}(t)}{1+t}\right)\big(X_2(t) - x_0\big) - \frac{\hat{\Phi}\big(t+1, X_2(t)\big)}{(1+t)\,\widehat{W}(t)},$$

причем $\tilde{\gamma}(t) \to 1$ при $t \to \infty$.

Рассмотрим, наконец, при $t \geq T_3$ процесс $X_3(t)$, определяемый следующим образом:

$$X_3(t+1) - x_0 = \left(1 - \frac{\hat{\gamma}(t)}{1+t}\right)\big(X_3(t) - x_0\big) - \frac{\hat{\Phi}\big(t+1, X_3(t)\big)}{(1+t)\,\widehat{W}(t)}.$$

Здесь постоянная T_3 выбрана из условия

$$P\{X_3(t) \equiv X_2(t) \text{ при } t \geq T_3\} > 1 - \frac{\alpha_0}{3}, \tag{П.36}$$

а величины $\hat{\gamma}(t)$ и $\widehat{W}(t)$ определены формулами (П.8), (п.9).

Как и при доказательстве теоремы 4.1, нетрудно показать, что

$$\sqrt{t}\,\big(X_3(t) - x_0\big) + \sum_{k=T_3}^{t} \frac{A_{kt}}{\sqrt{k+1}}\,\frac{\hat{\Phi}\big(K+1, X_3(k)\big)}{\widehat{W}(k)} \to 0 \tag{П.37}$$

по вероятности при $t \to \infty$.

Далее, точно так же, как и выше, используя формулу (П.37), устанавливаем справедливость утверждений теоремы 4.4–4.6 для процесса $X_3(t)$. Из (П.33), п.35), (п.36) и произвольности α_0 вытекают и требуемые утверждения.

*

Выражаю глубокую благодарность Р. З. Хасьминскому за ряд полезных замечаний.

Литература

1. *Venter, J. N.*, An extension of the Robbins–Monro procedure. Ann. Math. Statist., 1967, **38**, 1, 181–190.
2. *Fabian, V.*, On asymptotic normality in stochastic approximation. Ann. Math. Statist., 1968, **39**, 4, 1327–1332.
3. *Невельсон М. Б., Хасьминский Р. З.* Стохастическая аппроксимация и рекуррентное оценивание. М., Наука, 1972.
4. *Невельсон М. Б., Хасьминский Р. З.* Адаптивная процедура Роббинса—Монро. Автоматика и телемеханика, 1973, *10*, 71—83.
5. *Хасьминский Р. З.* О поведении процесса стохастической аппроксимации при больших значениях времени. Проблемы передачи информации, 1972, **8**, *1*, 81—91.
6 *Ибрагимов И. А., Хасьминский Р. З.* Об аппроксимации статистических оценок суммами иезависимых случайных величин. Доклады АН СССР, 1973, **210**, *6*, 883—887.
7. *Robbins, H., Siegmund, D.*, A Convergence Theorem for Non Negative Almost Supermartingales and Some Application Optimizing Methods in Statistics. Academic Press, N. Y., London, 1971.
8. *Komlós, J., Révész, P.*, A modification of the Robbins–Monro process. Stud. Sci.. Math. hung., 1973, **8**, 3, 329—340.
9. *Fabian, V.*, Asymptotically Efficient Stochastic Approximation, the Case RM. Ann. Math. Statist, 1973, **1**, 3, 486–495.
10. *Hasminskii, R. Z.*, Sequential Estimation and Recursive Asymptotically Optimal Procedure of Estimation and Observation Control. Proceedings of the Prague Symposium on Asymptotic Statistics Charles University Prague 1974, **1**, 157—178
11. *Невельсон М. Б.* О сходимости рекуррентных оценок нуля неизвестной функции. Проблемы передачи информации, 1975, XI, 2, 68—83.
12. *Невельсон М. Б.* On Some Asymptotical Properties of Recursive Estimates. Proceedings of the Colloquia Mathematical Society János Bolyai: Limit Theorems Probability Theory, (Keszthely, 1974), **11**, 193—225, Budapest, 1975.
13. *Гихман И. И., Скороход А. В.* Введение в теорию случайных процессов. М., Наука, 1965.
14. *Sacks, J.*, Asymptotic distribution of stochastic approximation. Ann. Math. Statist, 1958, **29**, 2, 373–405.

On stochastic approximation processes by sums of independent random variables

M. B. NEVELSON

(Moscow)

Let $f(x)$, $-\infty < x < \infty$, be an unknown function, which is negative on interval $(-\infty, x_0)$ and positive on (x_0, ∞). Suppose that we can measure the value $f(x)$ only together with some random error $\xi(t, x)$, so that the value $Y_0(t, x) = F(f(x), \xi(t, x))$ for every $t = 1, 2, \ldots$ and x can be obtained by an experiment. Our aim is to find the value x_0, f we know the expectation $S_0(t, x) = MS(\xi(t, x))$, where $S(z)$ is a nondecreasing function.

We construct under broad conditions on $f(x)$, $\xi(t, x)$, $F(x, z)$ asymptotically normal estimators $X(t)$ of the value x_0, which have optimal limit variance. We will show that this estimators can be approximations in probability or with probability one by sums of independent random variables. Moreover, we consider the properties of finite-dimensional and infinite-dimensional distributions of estimators $X(t)$. In particular we show that distribution generated by the estimators $X(t)$ in the space of continuous function is weakly convergent to the distribution generated by the stationary Gaussian–Markov process, satisfying Ito's stochastic equation

$$\dot{V}(t) = -\frac{1}{2} V(t) + \sigma\dot{\eta}(t)$$

where σ is some constant, $\eta(t)$ is Wiener process.

М. Б. Невельсон
Институт проблем передачи информации АН СССР
СССР, Москва Е-24,
Авиамоторная ул., 8а, корп. 2

Problems of Control and Information Theory, Vol. 5 (2), pp. 135—147 (1976)

PATTERN RECOGNITION IN A MARKOVIAN CASE

L. REJTŐ

(Budapest)

(Received April 10, 1975)

Introduction

We consider the following model of pattern recognition: two sequences $\{\xi_i\}_{i=1}^{\infty}$ and $\{\eta_i\}_{i=1}^{\infty}$ of independent identically distributed random variables (i. i. d. r. v.) are given, with common distributions F and G, respectively. In each step we observe either ξ_i or η_i, i.e. we observe $X_i = (1 - \varrho_i)\,\xi_i + \varrho_i\,\eta_i$ where $\{\varrho_i\}_{i=1}^{\infty}$ is a sequence of $\{0, 1\}$-valued random variables, independent of $\{\xi_i\}_{i=1}^{\infty}$ and $\{\eta_i\}_{i=1}^{\infty}$. The problem is to decide whether we have observed a ξ or an η, i.e., the value of ϱ_i. The most familiar assumption on the ϱ_i's is that they form a sequence of i. i. d. r. v.'s. In this paper we deal with the generalization that the ϱ_i's are not independent, rather they are assumed to form a stationary Markov chain.

In the first part we give a sufficient condition for the error probability to be a decreasing function of the number of sample elements (see Theorem 1.2). The second part of the paper deals with the case when F, G and the transition probabilities of the Markov chain are unknown, rather they are estimated on the basis of a learning process. Theorem 2.1 gives an indication how the "goodness" of the estimation of ϱ_k depends on the length of our learning process.

1. Estimation problem

Let $(\Omega, \mathfrak{A}, P)$ be a probability space, $\xi_1, \xi_2, \ldots, \xi_n$ be independent identically distributed random variables (i. i. d. r. v.) with distribution function $F(x)$ and density function $f(x)$, $\eta_1, \eta_2, \ldots, \eta_n, \ldots$ be i. i. d. r. v. with distribution $G(x)$ and density function $g(x)$. Let $\varrho_1, \varrho_2, \ldots, \varrho_n, \ldots$ be a Markov chain, independent of $\{\xi_i\}_{i=1}^{\infty}$ and $\{\eta_i\}_{i=1}^{\infty}$ with the two states 0 and 1, having stationary transition probabilities. Let $\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$ be the transition probability matrix and

$$P(\varrho_k = 0) = \frac{\beta}{\alpha + \beta}, \qquad P(\varrho_k = 1) = \frac{\alpha}{\alpha + \beta}$$

which are the stationary absolute probabilites.

Let us suppose that in the course of an experiment we observe the random variables $X_1, X_2, \ldots, X_n, \ldots$ where $X_n = (1 - \varrho_n)\,\xi_n + \varrho_n\,\eta_n$, and after every observation we should like to decide about the distribution (or density) of X_n with minimal probability of error.

It is easy to see that our model has the next two properties:

(i) $P(X_n \in A_n \mid X_1, \varrho_1, \ldots, X_{n-1}, \varrho_{n-1}, \varrho_n) = P(X_n \in A_n \mid \varrho_n)$,

(ii) $P(\varrho_n = i \mid X_1, \varrho_1, \ldots, X_{n-1}, \varrho_{n-1}) = P(\varrho_n = i \mid \varrho_{n-1})$

for all $A_n \in \mathfrak{A}$ and $i \in \{0; 1\}$ $(n = 1, 2, \ldots)$.

An easy consequence of (i) and (ii):

Corollary 1.1

$$P(X_{i_1} \in A_{i_1}, X_{i_2} \in A_{i_2}, \ldots, X_{i_n} \in A_{i_n} \mid \varrho_{i_1}, \varrho_{i_2}, \ldots, \varrho_{i_n}) =$$

$$= P(X_{i_1} \in A_{i_1} \mid \varrho_{i_1})\, P(X_{i_2} \in A_{i_2} \mid \varrho_{i_2}) \ldots P(X_{i_n} \in A_{i_n} \mid \varrho_{i_n})$$

where $i_1 < i_2 < \ldots < i_n$ and n is a positive integer.

It is well known that the Bayesian decision function

$$\mathfrak{F}_n(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if} \quad P(\varrho_n = 0 \mid X_1 = x_1, \ldots, X_n = x_n) \geq 1/2 \\ 1 & \text{if} \quad P(\varrho_n = 0 \mid X_1 = x_1, \ldots, X_n = x_n) < 1/2 \end{cases} \tag{1.1}$$

has the minimal probability of error among all decision functions based on the values of $X_1, \ldots, X_n$, i.e., the decision

$$\Delta_n = \mathfrak{F}_n(X_1, \ldots, X_n)$$

minimizes $P\{\Delta_n \neq \varrho_n\}$ (see e.g. [1]).

Let us introduce the random variable $\Pi_n = P(\varrho_n = 0 \mid X_1, \ldots, X_n)$. Clearly

$$\Pi_1 = \frac{\beta f(X_1)}{\beta f(X_1) + \alpha g(X_1)}\,. \tag{1.2}$$

Using (i) and (ii) it is easy to see that

$$\Pi_n = \frac{f(X_n)[\Pi_{n-1}(1 - \alpha - \beta) + \beta]}{f(X_n)[\Pi_{n-1}(1 - \alpha - \beta) + \beta] + g(X_n)[\Pi_{n-1}(\alpha + \beta - 1) + (1 - \beta)]}\,. \tag{1.3}$$

Π_n is a function of X_n and Π_{n-1}, not depending on n. An advantage of the decision function $\mathfrak{F}_n$ is that after every observation $X_n = x_n$ we decide about ϱ_n using only $X_n = x_n$ and the value of Π_{n-1}.

Π_n can also be written as

$$\Pi_n = \frac{f(X_n)\, P(\varrho_n = 0 \mid X_1, \ldots, X_{n-1})}{f(X_n)\, P(\varrho_n = 0 \mid X_1, \ldots, X_{n-1}) + g(X_n)\, P(\varrho_n = 1 \mid X_1, \ldots, X_{n-1})}\,. \tag{1.3}$$

This is analogous to formula (1.2), only the prior probabilities $P(\varrho_1 = 0) = \frac{\beta}{\alpha + \beta}$ and $P(\varrho_1 = 1) = \frac{\alpha}{\alpha + \beta}$ are replaced by the conditional ones

$$P(\varrho_n = 0 \mid X_1, \ldots, X_{n-1}) \text{ and } P(\varrho_n = 1 \mid X_1 \ldots X_{n-1}).$$

From practical point of view a useful equivalent definition of $\mathfrak{F}_n$ is:

$$\mathfrak{F}_n(X_1, \ldots, X_n) = \begin{cases} 0 & \text{if } \mathfrak{D}_n(X_1, \ldots, X_n) \geq 0\,, \\ 1 & \text{if } \mathfrak{D}_n(X_1, \ldots, X_n) < 0 \end{cases}$$

where

$$\mathfrak{D}_n(X_1, \ldots, X_n) = P_n - R_n \tag{1.4}$$

and P_n and R_n are defined recursively by

$$P_1 = \beta f(X_1),\ R_1 = \alpha g(X_1),$$
$$P_n = f(X_n)\,[(1 - \alpha)\,P_{n-1} + \beta R_{n-1}],$$
$$R_n = g(X_n)\,[\alpha P_{n-1} + (1 - \beta)\,R_{n-1}].$$

The error probability e_n of the decision $\Delta_n = \mathfrak{F}_n(X_1, \ldots, X_n)$ is

$$e_n = P\{\Delta_n \neq \varrho_n\} = E(P(\varrho_n = 0, \Pi_n < 1/2 \mid X_1, \ldots, X_n) +$$
$$+ P(\varrho_n = 1, \Pi_n \geq 1/2 \mid X_1, \ldots, X_n)) =$$
$$= E\big(P(\varrho_n = 0(X_1, \ldots, X_n)\,P(\Pi_n < 1/2 \mid X_1, \ldots, X_n, \varrho_n = 0) +$$
$$+ P(\varrho_n = 1 \mid X_1, \ldots, X_n)\,P(\Pi_n \geq 1/2 \mid X_1, \ldots, X_n, \varrho_n = 1)\big)\,.$$

Let us denote by χ_A the indicator function of the set $A \subset \Omega$. Then

$$e_n = E(\Pi_n \chi_{\Pi_n < 1/2} + (1 - \Pi_n)\,\chi_{\Pi_n \geq 1/2}) = E(\min(\Pi_n, 1 - \Pi_n)).$$

Theorem 1.1. The sequence of errors $\{e_n\}_{n=1}^{\infty}$ of the Bayesian decisions Δ_n is nonincreasing.

Proof. We shall prove that $e_k \geq e_{k+1}$ $(k = 1, 2, \ldots)$. In view of the stationarity of our model the error of Δ_k is equal to the error of Δ_{k+1}^*, where

$$\Delta_{k+1}^* = \mathfrak{F}_k(X_2, \ldots, X_{k+1}) = \begin{cases} 0 & \text{if } P(\varrho_{k+1} = 0 \mid X_2, \ldots, X_{k+1}) \geq 1/2, \\ 1 & \text{if } P(\varrho_{k+1} = 0 \mid X_2, \ldots, X_{k+1}) < 1/2. \end{cases}$$

Δ_{k+1}^* is a decision based on $\{X_1, \ldots, X_{k+1}\}$. Thus, the error of the Bayesian decision Δ_{k+1} is not greater than the error of Δ_{k+1}^*, so $e_k \geq e_{k+1}$.

It is possible to decide about the distribution of X_n not considering the dependence between $\varrho_1, \ldots, \varrho_n$ or $X_1, \ldots, X_n$. This means that, as in the independent case, after every sample element X_k we use the decision $\Delta^* = \mathfrak{F}_1(X_k)$. Clearly

$$\mathfrak{F}_1(X) = \begin{cases} 0 \text{ if } & \dfrac{\beta f(X)}{\beta f(X) + \alpha g(X)} \geq \dfrac{1}{2}. \\ 1 \text{ if } & \dfrac{\beta f(X)}{\beta f(X) + \alpha g(X)} < \dfrac{1}{2}. \end{cases} \tag{1.6}$$

The error e of Δ^* is

$$e = \frac{\alpha}{\alpha + \beta} \int_D g(x)\, dx + \frac{\beta}{\alpha + \beta} \int_{\bar{D}} f(x)\, dx \tag{1.7}$$

where $D = \left\{x : \dfrac{\beta f(x)}{\beta f(x) + \alpha g(x)} \geq \dfrac{1}{2}\right\}$ and $\bar{D}$ denotes the complement of D. It is trivial that $e = e_1$. The decision function (1.1) is really useful if $e_k < e$. By virtue of Theorem 1.1 we have that $e_k \leq e$ for all k.

We shall give a condition when $e_k < e$, and after this theorem we shall give an example when $e_k = e$ for all k.

Theorem 1.2. Let us suppose that the decision function $\mathfrak{F}_1(x)$ has the property $1 > \left(\int_D f(x)\, dx\right)\left(\int_{\bar{D}} g(x)\, dx\right) > 0$ and let the range of the function $\dfrac{f(x)}{g(x)}$ be $(0; +\infty)$ on the set where $g(x) \neq 0$; then $e_k < e$ for $k = 2, 3, \ldots$

Proof. We shall prove that in this case $e_1 > e_2$. Clearly, e_1 and e_2 are the errors of $\Delta_1^* = \mathfrak{F}_1(X_2)$ and $\Delta_2 = \mathfrak{F}_2(X_1, X_2)$, respectively. Set

$$D_1 = \{(x; y):\ -\infty < x < +\infty \quad \beta f(y) - \alpha g(y) \geq 0\},$$

$$D_2 = \{(x; y):\ \beta f(y)\,[f(x)\,(1 - \alpha) + g(x)\,\alpha] - \alpha g(y)\,[\beta f(x) + (1 - \beta)\,g(x)] \geq 0\}$$

and let H_1 (H_2) be the projection of D_1 (D_2) on the axis y. Then $H_1 \subseteq H_2$. Let $D = H_2 - H_1$, and $D_{2y} = \{x:\ (x; y) \in D_2$ for the fixed $y \in H_2\}$. Then

$$e_1(\alpha + \beta) = \alpha \int_{H_1} g(y)\, dy + \beta \int_{\bar{H}_1} f(y)\, dy,$$

$$e_2(\alpha + \beta) = \alpha\beta \iint_{D_2} g(y)\, f(x)\, dx dy + \alpha(1 - \beta) \iint_{D_2} g(y)\, g(x)\, dx\, dy +$$

$$+ \beta(1 - \alpha) \iint_{\bar{D}_2} f(y)\, f(x)\, dx dy + \alpha\beta \iint_{\bar{D}_2} f(y)\, g(x)\, dx dy.$$

Using the sets D, H_1, H_2, D_{2y} we have

$$(\alpha+\beta)e_1 = \alpha\beta \int_{H_1}\int_{-\infty}^{+\infty} g(y)\,f(x)\,dxdy + \alpha(1-\beta)\int_{H_1}\int_{-\infty}^{+\infty} g(y)\,g(x)\,dxdy +$$

$$+ \beta(1-\alpha)\int_{\overline{H}_1}\int_{-\infty}^{+\infty} f(y)\,f(x)\,dxdy + \beta\alpha\int_{\overline{H}_1}\int_{-\infty}^{+\infty} f(y)\,g(x)\,dxdy =$$

$$= \alpha\beta\int_{H_1}\int_{D_{2y}} g(y)\,f(x)\,dxdy + \alpha\beta\int_{H_1}\int_{\overline{D}_{2y}} g(y)f(x)\,dxdy +$$

$$+ \alpha(1-\beta)\int_{H_1}\int_{D_{2y}} g(y)g(x)\,dxdy + \alpha(1-\beta)\int_{H_1}\int_{\overline{D}_{2y}} g(y)\,g(x)\,dxdy +$$

$$+ \beta(1-\alpha)\int_{D}\int_{D_{2y}} f(y)\,f(x)\,dxdy + \beta(1-\alpha)\int_{D}\int_{\overline{D}_{2y}} f(y)\,f(x)\,dxdy +$$

$$+ \beta(1-\alpha)\int_{\overline{H}_2}\int_{-\infty}^{+\infty} f(y)\,f(x)\,dxdy + \beta\alpha\int_{D}\int_{D_{2y}} f(y)\,g(x)\,dxdy +$$

$$+ \beta\alpha\int_{D}\int_{\overline{D}_{2y}} f(y)\,g(x)\,dxdy + \beta\alpha\int_{\overline{H}_2}\int_{-\infty}^{+\infty} f(y)\,g(x)\,dxdy$$

and

$$(\alpha+\beta)e_2 = \alpha\beta\int_{H_1}\int_{D_{2y}} g(y)\,f(x)\,dxdy + \alpha\beta\int_{D}\int_{D_{2y}} g(y)\,f(x)\,dxdy +$$

$$+ \alpha(1-\beta)\int_{H_1}\int_{D_{2y}} g(y)\,g(x)\,dxdy + \alpha(1-\beta)\int_{D}\int_{D_{2y}} g(y)\,g(x)\,dxdy +$$

$$+ \beta(1-\alpha)\int_{H_1}\int_{\overline{D}_{2y}} f(y)\,f(x)\,dxdy + \beta(1-\alpha)\int_{D}\int_{\overline{D}_{2y}} f(y)\,f(x)\,dxdy +$$

$$+ \beta(1-\alpha)\int_{\overline{H}_2}\int_{-\infty}^{+\infty} f(y)\,f(x)\,dxdy + \alpha\beta\int_{H_1}\int_{\overline{D}_{2y}} f(y)\,g(x)\,dxdy +$$

$$+ \alpha\beta\int_{D}\int_{\overline{D}_{2y}} f(y)\,g(x)\,dxdy + \alpha\beta\int_{\overline{H}_2}\int_{-\infty}^{+\infty} f(y)\,g(x)\,dxdy.$$

So we have to prove that

$$\int_{H_1}\int_{\overline{D}_{2y}} \alpha g(y)\,[\beta f(x)+(1-\beta)\,g(x)]\,dxdy + \int_{D}\int_{D_{2y}} \beta f(y)\,[(1-\alpha)\,f(x)+\alpha g(x)]\,dxdy >$$

$$> \int_{D}\int_{D_{2y}} \alpha g(y)\,[\beta f(x)+(1-\beta)\,g(x)]\,dxdy + \int_{H_1}\int_{\overline{D}_{2y}} \beta f(y)\,[(1-\alpha)\,f(x)+\alpha g(x)]\,dxdy\,.$$

If the range of $\dfrac{f(x)}{g(x)}$ is $(0;+\infty)$, then the integrals on $D_1 \cap \overline{D}_{2y}$ are not zeros.

Knowing that

$$\beta f(y)\,[(1-\alpha)\,f(x)+\alpha g(x)]<\alpha g(y)\,[\beta f(x)+(1-\beta)g\,(x)]$$

on $D_1 \cap \bar{D}_2$, and on $\{D \otimes (-\infty; +\infty)\} \cap D_2$

$$\beta f(y)\,[(1-\alpha)\,f(x)+\alpha g(x)]\geq \alpha g(y)\,[\beta f(x)+(1-\beta)\,g(x)]$$

we have that $e_2 < e_1$.

Example. Let us consider the density functions:

$$f(x)=\begin{cases}1 & \text{if} \quad x\in(0;\,1],\\ 0 & \text{elsewhere}\end{cases}$$

$$g(x)=\begin{cases}\dfrac{1}{2} & \text{if} \quad x\in(0;\,2],\\ 0 & \text{elsewhere,}\end{cases}$$

and let α, β fulfil the conditions $\beta > \frac{1}{3}1$, $\alpha < \frac{2}{3}$.

In every step the Bayesian decision on ϱ_n is based on Π_n, which will be used in form (1.3)′. Using $\Delta^* = \mathfrak{F}_1(X_n)$, we would adopt $f(x)$ if $X_n \in (0;\,1]$ and $g(x)$ if $X_n \in (1;\,2]$. This decision does not depend on the actual values of α and β as long as $\beta > \dfrac{\alpha}{2}$. It is easy to see that the Bayesian decision based on Π_n in the nth step is independent on $X_1, \ldots, X_{n-1}$ and coincides with the above one because the conditional probabilities also fulfil

$$P(\varrho_n = 0 \mid X_1, \ldots, X_{n-1}) > \frac{1}{3} \quad \text{and} \quad P(\varrho_n = 1 \mid X_1, \ldots, X_{n-1}) < \frac{2}{3}.$$

This means that the probability of error is $e_n = e = e_1$ for every n. Naturally the quotient $\dfrac{f(x)}{g(x)}$ does not satisfy the condition of Theorem 1.2.

2. Pattern recognition

Let us consider the case when f, g, α, β are unknown, f and g are the density functions of ξ and η having values in the r-dimensional Euclidean space $\mathbf{E}^r$.

In this case we suppose that a sample $(X_1, \varrho_1), \ldots, (X_N, \varrho_N)$ is given as a "learning process" and f, g, α, β are estimated from it. These estimates are then used for a decision problem where the observations are $Y_1, \ldots, Y_k$, with $Y_k = (1-\varrho'_k)\,\xi'_k + \varrho'_k\,\eta'_k$, the sequences $\{\varrho'_k\}$, $\{\xi'_k\}$ and $\{\eta'_k\}$ being governed by the same joint distributions as $\{\varrho_k\}$, $\{\xi_k\}$, $\{\eta_k\}$.

Set $\mu_j = \varrho_1 + \varrho_2 + \ldots + \varrho_j$, $j = 1, 2, \ldots, N$. Let $1 \geq a_1 > a_2 > \ldots$ be a monotonically decreasing sequence of positive numbers and let $h(x)$ $(x \in \mathbf{E}^r)$ be a density function. Then, using Corollary 1.1, the functions

$$f_{N-\mu_N}(x) = \frac{1}{N-\mu_N} \sum_{j=1}^{N} \frac{1-\varrho_j}{a^r_{j-\mu_j}} h\left(\frac{x-X_j}{a_{j-\mu_i}}\right)$$

$$g_{\mu_N}(x) = \frac{1}{\mu_N} \sum_{j=1}^{N} \frac{\varrho_j}{a^r_{\mu_j}} h\left(\frac{x-X_j}{a_{\mu_j}}\right)$$

may be used as estimations of f and g (similarly as in [2]) and

$$\alpha_N = \frac{\sum_{j=1}^{N-1} \varrho_{j+1}(1-\varrho_j)}{N-\mu_N}, \qquad \beta_N = \frac{\sum_{j=1}^{N-1} \varrho_j(1-\varrho_{j+1})}{\mu_N}, \qquad \frac{N-\mu_N}{N}, \quad \frac{\mu_N}{N}$$

as the estimations of the probabilities α, β, $P(\varrho_k = 0)$ and $P(\varrho_k = 1)$.

We shall use the estimated decision function

$$\mathfrak{D}_k^{(N)} = P_k^{(N)} - R_k^{(N)} \tag{2.1}$$

to decide about ϱ'_k, where

$$P_1^{(N)} = P_1^{(N)}(Y_1) = \beta_N f_{N-\mu_N}(Y_1),$$
$$R_1^{(N)} = R_1^{(N)}(Y_1) = \alpha_N g_{\mu_N}(Y_1)$$

and

$$P_k^{(N)} = P_k^{(N)}(Y_1, \ldots, Y_k) = f_{N-\mu_N}(Y_k)\,[(1-\alpha_N)\,P_{k-1}^{(N)} + \beta_N\,R_{k-1}^{(N)}]$$
$$R_k^{(N)} = R_k^{(N)}(Y_1, \ldots, Y_k) = g_{\mu_N}(Y_k)\,[\alpha_N\,P_{k-1}^{(N)} + (1-\beta_N)\,R_{k-1}^{(N)}]\,.$$

In the kth step we would adopt $f(x)$ if $\mathfrak{D}_k^{(N)} \geq 0$ and $g(x)$ if $\mathfrak{D}_k^{(N)} < 0$.

Theorem 2.1. Suppose that

a) the density functions $f(x)$ and $g(x)$ $(x \in \mathbf{E}^r)$ are obeying the following conditions:

(i) $f(x)$ and $g(x)$ are vanishing outside an open convex (not certainly bounded) domain T;

(ii) there exist the second order derivatives of $f(x)$ and $g(x)$ in T and

$$\sup_{x \in T} \max \left\{ |f(x)|, \left|\frac{\partial f}{\partial x_1}\right|, \ldots, \left|\frac{\partial f}{\partial x_r}\right|, \left|\frac{\partial^2 f}{\partial x_1^2}\right|, \left|\frac{\partial^2 f}{\partial x_1\,\partial x_2}\right|, \ldots, \left|\frac{\partial^2 f}{\partial x_r^2}\right|, \right.$$
$$\left. |g(x)|, \left|\frac{\partial g}{\partial x_1}\right|, \ldots, \left|\frac{\partial^2 g}{\partial x_r^2}\right| \right\} \leq L$$

for some $0 < L < +\infty$;

(iii) there exists a $\delta > 0$ such that

$$\int_{\mathbf{E}^r} ||x||^\delta f(x)dx < +\infty \quad \text{and} \quad \int_{\mathbf{E}^r} ||x||^\delta g(x)\,dx < +\infty;$$

b) the density function $h(x)$ $(x \in \mathbf{E}^r)$ is obeying the following conditions

(i) $\int_{\mathbf{E}^r} x_i h(x)\,dx = 0$ for $i = 1, \ldots, r$,

$\int_{\mathbf{E}^r} x_i x_j h(x)\,dx$ exist for $i = 1, \ldots, r$

$j = 1, \ldots, r$ and $\lim_{||x|| \to \infty} ||x||^{r+2} h(x) = 0$;

(ii) $|h(x)| < H_0$ for all x in $\mathbf{E}^r$;

(iii) $|h(x) - h(y)| \leq D\,||x - y||^\alpha$ for all $x, y \in \mathbf{E}^r$ where D is a constant and $0 < \alpha \leq 1$;

c) the sequence $\{a_n\}_{n=1}^\infty$ is obeying the following conditions:

(i) $\frac{1}{2} \geq a_1 > a_2 > \ldots$, and $a_k \geq 0$ for $k = 1, 2, \ldots$ and $\lim_{k \to \infty} a_k = 0$;

(ii) there exists an ω $(0 < \omega < 1)$ such that $n^{1-\omega} a_n^r \to \infty$;

d) the transition probabilities are obeying the following conditions:

(i) $\alpha + \beta \neq 1$;

(ii) $0 < \alpha < 1$ and $0 < \beta < 1$.

Then

$$P\Big(\sup_{y_1 \ldots y_k \in T - T_\varrho} |\mathfrak{D}_k^{(N)}(y_1 \ldots y_k) - \mathfrak{D}_k(y_1 \ldots y_k)| > \varepsilon\Big) \leq e^{-c_k N a_N^r + b_1 a_N^r (k-1)} \quad (2.2)$$

where $T_\varrho = \{x: ||x - \mathfrak{F}_r T|| < \varrho\}$ and $\mathfrak{F}_r T$ denotes the frontier of T and

$$P\Big(\int_{\mathbf{E}^r} \int_{\mathbf{E}^r} |\mathfrak{D}_k^{(N)}(y_1 \ldots y_k) - \mathfrak{D}_k(y_1 \ldots y_k)|^p\,dy_1 \ldots dy_k \geq \varepsilon^p\Big) \leq e^{-d_k N a_N^r + b_2 a_N^r (k-1)}. \quad (2.3)$$

Here b_1 and b_2 do not depend on f, g, α, β while c_k and d_k depend on f and g through L only.

The *proof* of this theorem is similar to that of Theorem 2 in [2].

We need only the next lemma, which is a consequence of a result of Formanov [3].

Lemma 2.1. Let $\varrho_1, \varrho_2, \ldots, \varrho_n, \ldots$ be a stationary Markov chain with two states 0 and 1, with transition probability matrix $\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$ where

$\alpha + \beta \neq 1$, $0 < \alpha < 1$, $0 < \beta < 1$; then to any $\varepsilon > 0$ there exist $\lambda_1 > 0$ and $\lambda_2 > 0$ such that

$$P(|\alpha_n - \alpha| > \varepsilon) \leq e^{-\lambda_1 n}$$

$$P(|\beta_n - \beta| > \varepsilon) \leq e^{-\lambda_2 n}.$$

Proof. We have

$$P\left(\left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| > \varepsilon\right) < e^{-q_1 n}$$

and

$$P\left(\left|\frac{\sum_{j=1}^{n-1} \varrho_{j+1}(1-\varrho_j)}{n} - \frac{\alpha\beta}{\alpha+\beta}\right| > \varepsilon\right) < e^{-q_2 n}.$$

The first inequality is a simple application of Lemma 3 of [3]. The second one is the consequence of the fact that the sequence $\{(\varrho_{j+1}(1-\varrho_j);\ \varrho_{j+2}(1-\varrho_{j+1}))\}_{j=1}^{\infty}$ is a Markov chain and to it we can apply the mentioned lemma of Formanov. Clearly

$$P(|\alpha_n - \alpha| > \varepsilon) = P\left(\left|\frac{\frac{\sum_{j=1}^{n-1} \varrho_{j+1}(1-\varrho_j)}{n}}{\frac{n-\mu_n}{n}} - \frac{\frac{\alpha\beta}{\alpha+\beta}}{\frac{\beta}{\alpha+\beta}}\right| > \varepsilon\right) \leq$$

$$\leq P\left(\left|\frac{\frac{\sum_{j=1}^{n-1} \varrho_{j+1}(1-\varrho_j)}{n}}{\frac{n-\mu_n}{n}} - \frac{\frac{\alpha\beta}{\alpha+\beta}}{\frac{\beta}{\alpha+\beta}}\right| > \varepsilon_1 \left| \left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| \leq \varepsilon\right.\right) +$$

$$+ P\left(\left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| > \varepsilon\right).$$

If ε is small enough then

$$0 < C\frac{\beta}{\alpha+\beta} - \varepsilon < \frac{n-\mu_n}{n}.$$

Thus

$$P(|\alpha_n - \alpha| > \varepsilon) \leq P\left(\left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| > \varepsilon\right) +$$

$$+ P\left(\left|\frac{\sum_{j=1}^{n-1} \varrho_{j+1}(1-\varrho_j)}{n} - \frac{n-\mu_n}{n}\alpha\right| > \varepsilon C\right) <$$

$$< P\left(\left|\frac{\sum_{j=1}^{n-1} \varrho_{j+1}(1-\varrho_j)}{n} - \frac{\alpha\beta}{\alpha+\beta}\right| + \alpha\left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| > \varepsilon C\right) +$$

$$+ P\left(\left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| > \varepsilon\right) \leq P\left(\left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| > \varepsilon\right) +$$

$$+ P\left(\left|\frac{\sum_{j=1}^{n-1} \varrho_{j+1}(1-\varrho_j)}{n} - \frac{\alpha\beta}{\alpha+\beta}\right| > \frac{\varepsilon C}{2}\right) + P\left(\left|\frac{n-\mu_n}{n} - \frac{\beta}{\alpha+\beta}\right| > \frac{C\varepsilon}{2\alpha}\right) \leq e^{-\lambda_1 n}.$$

In similar manner it can be proved that

$$P(|\beta_n - \beta| > \varepsilon) \leq e^{-\lambda_2 n}.$$

Proof of Theorem 2.1. First we prove (2.2). We shall use induction on k to the probabilities $P(\sup |P_k^{(N)} - P_k| > \varepsilon)$ and $P(\sup |R_k^{(N)} - R_k| > \varepsilon)$ to prove that

$$P(\sup |P_k^{(N)} - P_k| > \varepsilon) < e^{-c_{k1} N a_N^r + b_{11} a_N^r (k-1)}$$

and

$$P(\sup |R_k^{(N)} - R_n| > \varepsilon) < e^{-c_{k2} N a_N^r + b_{12} a_N^r (k-1)}.$$

If $k = 1$, on account of the above lemma the proof is the same as in [2]. Let us suppose that the inequality is true if $k = l - 1$. It is obvious that

$$|\mathfrak{D}_l^{(N)} - \mathfrak{D}_l| \leq |P_l^{(N)} - P_l| + |R_l^{(N)} - R_l| \tag{2.4}$$

and for example

$$\begin{aligned} |P_l^{(N)} - P_l| \leq{}& (1-\alpha_N) f_{N-\mu_N}(Y_l) |P_{l-1}^{(N)} - P_{l-1}| + P_{l-1} f_{N-\mu_N}(Y_l) |\alpha - \alpha_N| + \\ &+ \alpha P_{l-1} |f_{N-\mu_N}(Y_l) - f(Y_l)| + f_{N-\mu_N}(Y_l) \beta_N |R_{l-1}^{(N)} - R_{l-1}| + \\ &+ f_{N-\mu_N}(Y_l) R_{l-1} |\beta_N - \beta| + R_{l-1} \beta |f_{N-\mu_N}(Y_l) - f(Y_l)|. \end{aligned} \tag{2.5}$$

Thus we shall have to estimate the next type probabilities:

$$P\Big(\sup_{y_1 \cdots y_l} (1-\alpha_N) f_{N-\mu_N}(y_e) |P_{l-1}^{(N)} - P_{l-1}| > \varepsilon\Big) \tag{2.6}$$

$$P\Big(\sup_{y_1 \cdots y_l} P_{l-1} f_{N-\mu_N}(y_l) |\alpha - \alpha_N| > \varepsilon\Big) \tag{2.7}$$

$$P\Big(\sup_{y_1 \cdots y_l} \alpha P_{n-1} |f_{N-\mu_N}(y_l) - f(y_l)| > \varepsilon\Big). \tag{2.8}$$

Clearly:

$$P\Big(\sup_{y_1 \cdots y_l} (1-\alpha_N) f_{N-\mu_N}(y_l) |P_{l-1}^{(N)} - P_{l-1}| > \varepsilon\Big) \leq$$

$$\leq P\big((L+\varepsilon) \sup_{y_1 \cdots y_{l-1}} |P_{l-1}^{(N)} - P_{l-1}| > \varepsilon\big) + P\big(\sup f_{N-\mu_N}(y_l) > L + \varepsilon\big).$$

The first member can be estimated using the induction hypothesis, the second one by the help of Lemma 2.2 and Lemmas 6, 7 of paper [2]. Thus

$$P\left(\sup\,(1-\alpha_N)\,f_{N-\mu_N}(y_l)\,|P_{l-1}^{(N)}-P_{l-1}|>\varepsilon\right)\leq e^{-c_{l-1,1}Na_N^r+b_{1,1}a_N^r(l-2)}+$$

$$+P\left(\sup f_{N-\mu_N}(y_l)>K+\varepsilon,\ N\frac{\alpha}{\alpha+\beta}-N^{\frac{3}{4}}<\mu_N\leq N\frac{\alpha}{\alpha+\beta}+N^{\frac{3}{4}}\right)+$$

$$+P\left(\mu_N\notin\left(N\frac{\alpha}{\alpha+\beta}-N^{\frac{3}{4}};\,N\frac{\alpha}{\alpha+\beta}+N^{\frac{3}{4}}\right)\right)\leq e^{-s_1(l)Na_N^r+b_1a_N^r(l-2)}+$$

$$+\sum_{k=N\frac{\alpha}{\alpha+\beta}-N^{\frac{3}{4}}}^{N\frac{\alpha}{\alpha+\beta}+N^{\frac{3}{4}}}P\left(\sup f_{N-k}>K+\varepsilon\,|\,\mu_N=k\right)P(\mu_N=k)\leq$$

$$\leq e^{-s_2(l)Na_N^r+b_1a_N^r(l-2)}.$$

In the estimation of (2.7) we use that $P_k+R_k\leq 2L^k$, yielding

$$P\left(\sup P_{l-1}f_{N-\mu_N}(y_l)\,|\alpha-\alpha_N|>\varepsilon\right)\leq P\left(\sup f_{N-\mu_N}(y_l)>K+\varepsilon\right)+$$

$$+P\left(2L^{l-1}(K+\varepsilon)\,|\alpha-\alpha_N|>\varepsilon\right)\leq e^{-s_3(l)Na_N^r}.$$

Finally

$$P\left(\sup\alpha P_{l-1}\,|f_{N-\mu_N}(y_l)-f(y_l)|>\varepsilon\right)\leq$$

$$\leq P\left(\mu_N\notin\left(N\frac{\alpha}{\alpha+\beta}-N^{\frac{3}{4}};\,N\frac{\alpha}{\alpha+\beta}+N^{\frac{3}{4}}\right)\right)+$$

$$+\sum_{k=N\frac{\alpha}{\alpha+\beta}-N^{\frac{3}{4}}}^{N\frac{\alpha}{\alpha+\beta}+N^{\frac{3}{4}}}P\left(2L^{l-1}\sup\,|f_{N-k}(y_l)-f(y_l)|>\varepsilon\,|\,\mu_N=k\right)P(\mu_N=k)\leq$$

$$\leq e^{-s_4(l)Na_N^r}.$$

Combining the above results we obtain

$$P\left(\sup_{y_1\ldots y_l\in T-T_\varrho}|\mathfrak{D}_l^{(N)}-\mathfrak{D}_l|\geq\varepsilon\right)\leq P\left(\sup|P_l^{(N)}-P_l|\geq\frac{\varepsilon}{2}\right)+$$

$$+P\left(\sup|R_l^{(N)}-R_l|\geq\frac{\varepsilon}{2}\right)\leq$$

$$\leq P\left(\sup|1-\alpha_N|\,f_{N-\mu_N}|P_{l-1}^{(N)}-P_{l-1}|\geq\frac{\varepsilon}{12}\right)+\ldots+$$

$$+ \ldots + P\left(\sup R_{l-1}\beta\,|g_{N-\mu_N}(y_l) - g(y_l)| \geq \frac{\varepsilon}{12}\right) \leq e^{-c_l N a_N^r + b_1 a_N(l-1)},$$

where the constant c_l depends on L.

To prove (2.3), we shall use induction on k, as well. If $k = 1$ then the proof is the same as in paper [2]. Let us suppose that the theorem is true if $k = l - 1$. By virtue of (2.4) and (2.5) it is enough the deal with the following probabilities:

$$P\Big(\int \ldots \int (1 - \alpha_N)^p f^p_{N-\mu_N}(y_l)\,|P^{(N)}_{l-1} - P_{l-1}|^p\,dy_1 \ldots dy_l > \varepsilon^p\Big),$$

$$P\Big(\int \ldots \int P^p_{l-1}(y_l)\,f_{N-\mu_N}(y_l)\,|\alpha - \alpha_N|^p\,dy_1 \ldots dy_l > \varepsilon^p\Big)$$

and

$$P\Big(\int \ldots \int \alpha^p P^p_{l-1}\,|f_{N-\mu_N}(y_l) - f(y_l)|^p\,dy_1 \ldots dy_l > \varepsilon^p\Big).$$

Clearly

$$P\Big(\int \ldots \int (1 - \alpha_N)^p f^p_{N-\mu_N}(y_l)\,|P^N_{l-1} - P_{l-1}|^p\,dy_1 \ldots dy_l > \varepsilon^p\Big) \leq$$

$$\leq P\big(\sup f_{N-\mu_N}(y_l) \geq K\big) + P\left(\int \ldots \int |P^{(N)}_{l-1} - P_{l-1}|^p dy_1 \ldots dy_{l-1} > \frac{\varepsilon^p}{D_1}\right) \leq$$

$$\leq e^{-t_1 N a_N^r + d_2 a_N^r(l-2)}.$$

$$P\Big(\int \ldots \int P^p_{l-1} f^p_{N-\mu_N}(y_l)\,|\alpha - \alpha_N|^p\,dy_1 \ldots dy_l > \varepsilon^p\Big) \leq$$

$$\leq P\left(\int |f_{N-\mu_N}(y_l) - f(y_l)|^p\,dy_l > \frac{\varepsilon^p}{D_2};\ \mu_N \in \left(N\frac{\alpha}{\alpha+\beta} - N^{\frac{3}{4}};\ N\frac{\alpha}{\alpha+\beta} + N^{\frac{3}{4}}\right)\right) +$$

$$+ P\left(\mu_N \notin \left(N\frac{\alpha}{\alpha+\beta} - N^{\frac{3}{4}};\ N\frac{\alpha}{\alpha+\beta} + N^{\frac{3}{4}}\right)\right) \leq e^{-t_3 N a_N^r} +$$

$$+ \sum_{k = N\frac{\alpha}{\alpha+\beta} - N^{\frac{3}{4}}}^{N\frac{\alpha}{\alpha+\beta} + N^{\frac{3}{4}}} P\left(\int |f_{N-k} - f|^p\,dy_l > \frac{\varepsilon^p}{D_2}\,\middle|\,\mu_N = k\right) P(\mu_N = k) \leq e^{-t_4 N a_N^r}.$$

In the last step we have used result (2) of Theorem 1 of paper [2]. Thus we have

$$P\Big(\int \ldots \int |\mathfrak{D}^{(N)}_l(y_1 \ldots y_l) - \mathfrak{D}_l(y_1 \ldots y_l)|^p\,dy_1 \ldots dy_l \geq \varepsilon\Big) \leq e^{-d_l N a_N^r + b_2 a_N^r(l-1)},$$

where the constant d_l depends on L and p.

Conclusion. Theorem 2.1 shows that it is worthwhile to use the decision function $\mathfrak{D}^{(N)}_k$ if $-c_k N a^r_N + b_1 a^r_N(k-1) < 0$ and $-d_k N a^r_N + b_2 a^r_N(k-1) < 0$. Thus, for given learning process length N, it may not be good to use $\mathfrak{D}^{(N)}_k$ if k is large. Let us consider that the conditions of Theorem 1.2 are fulfilled,

then $e_k < e$ for $k = 2, 3, \ldots$ Then we may fix a k for which

$$-c_k N a_N^r + b_1 a_N^r (k-1) < 0$$

$$-d_k N a_N^r + b_2 a_N^r (k-1) < 0$$

and if $l \leq k$ we use the decision function $\mathfrak{D}_l^N$ and if $l > k$ we use in any step $\mathfrak{D}_k^{(N)}(Y_{l-k+1}, \ldots, Y_{l-1}, Y_l)$.

*

Acknowledgement. The author is indepted to I. Csiszár for his valuable remarks.

References

1. *Rényi, A.*, Statistics and information theory. Studia Sci. Math. Hung. **2** (1967) 249–256
2. *Rejtő, L., Révész, P.*, Density estimation and pattern classification. Problems of Control and Information Theory. **1** (1973) 68–80.
3. *Форманов М. К.* Некоторые предельные теоремы о больших уклонениях для однородных цепей Маркова. Случайные процессы и статистические выводы (выпуск III). Издательство «ФАН» Узбекской ССР, Ташкент 1973 (173—185).

Распознавание образов при марковских процессах

Л. РЕЙТО

(Будапешт)

Рассматривается следующая задача распознавания образов: $\{\xi_i\}$ и $\{\eta_i\}$ представляют собой последовательности независимых случайных переменных, где значения $\xi_i(\eta_i)$ являются независимыми случайными величинами с одинаковыми функциями распределения. Задается некоторая последовательность $\{\varrho_i\}$, значения ϱ_i представляют собой независимые от величин $\{\xi_i\}$ и $\{\eta_i\}$ случайные величины, такие, что они могут принимать значения (0,1). В каждом шаге наблюдается переменная $x_i = (\varrho_i)\ \xi_i + \varrho_i \eta_i$, т. е. если $\varrho_i = 0$, то наблюдается величина ξ_i, а если $\varrho_i = 1$, то — η_i Задача заключается в том, чтобы принять решение с минимальной ошибкой о том, что именно наблюдалось: значение ξ_i или значение η_i, т. е. какое из условий $\varrho_i = 0$ или $\varrho_i = 1$ выполняется. Обычно такого характера задачи распознавания образов рассматриваются при предположении, что случайные величины $\{\varrho_i\}$ являются назависимыми и имеют одинаковое распределение. Это предположение здесь заменяется таким условием, что последовательности $\{\varrho_i\}$ являются стационарными марковскими процессами с однородными вероятностями переходов. В первой части работы предполагается, что предварительно известны плотность распределения величин ξ_i и η_i, и матрица вероятностей переходов величина $\{\varrho_i\}$. В таком случае решение о величине ϱ_i принимается на основе бейесовой функции с применением элементов пробы $x_1, \ldots, x_i$. Во второй части рассматривается такой случай, когда не известны ни плотности вероятностей, ни матрица вероятностей переходов, но задается процесс обучения $(x_1, \varrho_1), \ldots (x_N, \varrho_N)$, с помощью которого можно оценить плотности вероятности и матрицу вероятностей переходов. В основной теореме работы (в теореме 2.1) указывается зависимость „качества" оценки величин ϱ_k от продолжительности процесса обучения.

L. Rejtő
Mathematical Research Institute
Hungarian Academy of Sciences
H-1053 Budapest V., Reáltanoda u. 13—15
Hungary

Problems of Control and Information Theory, Vol. 5 (2), pp. 149—165 (1976)

THE METHOD OF GENERALIZED CHARACTERISTIC EQUATION IN THE THEORY OF LINEAR TIME-VARYING SYSTEMS

F. A. MIKHAILOV

(Moscow)

(Received January 1, 1974)

This paper deals with the linear time-varying systems that are characterized by a single n-order differential equation with respect to a dependent variable x and the time t as an independent one. A method for the analysis of these systems is considered. This method is based on the use of a generalized characteristic equation, i.e. of an equation which is obtained from the original differential equation by the substitution $x = \exp \int \zeta \, dt$. Some new problems stimulating the further development of the method are formulated.

1. Introduction

Consider the system characterized by the equation

$$\frac{d^n x}{dt^n} + b_1(t) \frac{d^{n-1} x}{dt^{n-1}} + \ldots + b_n(t)\, x = 0, \tag{1.1}$$

where $b_1(t), \ldots, b_n(t)$ are real continuous functions. Problems of analysis of processes in this system are simplified when a fundamental set of solutions is known. A classical approach to the analysis problems hinges on the determination of such a set. Since no general procedure exists for its exact determination, the following two problems arise:

(a) constructing methods for obtaining approximations of fundamental sets of solutions,

(b) constructing methods for exact analysis of processes based on the knowledge of approximation of a fundamental set.

The author [1] suggested to solve these problems by use of the equation

$$(\zeta + p)^{n-1}\zeta + b_1(t)\,(\zeta + p)^{n-2}\zeta + \ldots + b_n(t) = 0, \quad p = d/dt \tag{1.2}$$

which follows from Eq. (1.1) due to substitution

$$x = \exp \int \zeta dt. \tag{1.3}$$

The basic idea of the transformation of Eq. (1.1) into Eq. (1.2) is the following: if $b_1(t), \ldots, b_n(t)$ are real functions that belong to some class, it can be searched for solutions of Eq. (1.2) in the complex-valued extension of that class. For example, if the above coefficients are continuous periodic functions and T is their common period, then, as is well known, there exists a complex-

valued T-periodic solution of Eq. (1.2), and hence the problem of its determination can be posed. In [1–4], algorithms employing this idea were introduced for finding sequences of approximate solutions.

Evidently, the success of using Eq. (1.2) depends on the existence on the interval I under consideration at least one of its solutions defined everywhere. The existence problem was considered by the author in [3, 6]. The obtained result is as follows: a solution as described exists. Moreover, it has been shown that there exists such a solution whose $n - 1$ first derivatives are also defined everywhere on I.

In [5], the term *generalized characteristic equation* (GCE) were introduced for denoting Eq. (1.2). In [4], the solutions defined everywhere on I with their above derivatives have been termed the *roots*.

It is evident that if $n = 2$ then there exist solutions of Eq. (1.1) that cannot be represented in the form

$$C_i \exp \int \zeta_i(t)\, dt,$$

where $\zeta_i(t)$ is a root and C_i is a constant. Hence the set of solutions represented in this form is only a part of the full set of solutions. If this part did not include a fundamental set, then the expedience of using GCE could be doubtful. However, these fears are not justified. As shown in [5, 6], a fundamental set $x_1(t), \ldots, x_n(t)$ is included. By relations $\zeta_i(t) = \dot{x}_i(t)/x_i(t)$, $i = 1, \ldots, n$, this corresponds to some sets of roots. The latter has been termed the *fundamental set of roots*.

Since the existence of a fundamental set of roots is proved, it is clear that one can pose a problem for its approximate determination. The advantage of approximate determination of roots in comparison with approximate determination of solutions to Eq. (1.1) is the following: in the first case, the resultant solutions to Eq. (1.1) have a certain structure and, consequently, the field of search of solutions to Eq. (1.1) is narrowed.

Under some restrictions on the coefficients $b_1, \ldots, b_n$ simple methods for approximate determination of roots and approximate formulas for the general solution to Eq. (1.1) may be constructed. Moreover, on using solutions of these problems, efficient methods for exact analysis of linear time-varying systems may be proposed.

2. Methods for approximate determination of roots of GCE

If $n = 1$, then the GCE has the form

$$\zeta + b_1(t) = 0. \tag{2.1}$$

In this case, there exists a unique root which is known: namely, $\zeta_1 = -b_1(t)$.

Omitting this case, we will suppose here and further $n \geq 2$.

For $n \geq 2$ the GCE is a nonlinear differential equation. In order to integrate approximately such an equation, well-known methods, generally, may be applied. However, this approach is not successful due to the necessity of fixing the initial conditions. It is not known what kind of condition corresponds to a root.

Methods for approximate integration of GCE, which are also methods for approximate determination of roots, were suggested in [1–4]. They differ by areas in which they are utilized. Some methods do not require restrictions on the coefficients of Eq. (1.1), except multiple differentiability. Other methods are utilized when the coefficients belong to special classes of functions (periodic functions, functions represented by power series with decreasing degrees of t, etc.).

These methods are iterative: each term of any sequence of approximate roots is determined with the use of information of the preceding one. Approximate roots are easily determined in common cases; moreover, the procedure of calculation gives all n roots of a fundamental system. If sequences converge, then such a method is a method of successive approximation.

To illustrate the character of these methods we consider the following example described in [1, 3, 4] and particularly in [8].

Let the equation

$$\ddot{x} + ct^{\sigma} x = 0, \quad c \neq 0, \tag{2.2}$$

σ being a real number, be given. Then for calculating approximations of GCE roots, we can use the method suggested in [1] and applicable to the case of equations, whose coefficients have power-series representation with decreasing degrees of t. The method in question may be regarded as a transplantation from the theory of differential equations in the complex domain and generalization of a method of Cayley [9]. On using this method, we deal with Eq. (2.2) as follows.

Write the GCE corresponding to Eq. (2.2):

$$\zeta^3 + 3\zeta\dot{\zeta} + \dot{\zeta} + ct^{\sigma} = 0. \tag{2.3}$$

Approximate roots of this one are finding as partial sums of the series'

$$H_{i1}t^{\eta_{i1}} + H_{i2}t^{\eta_{i2}} + \ldots, \qquad i = 1, 2, 3, \tag{2.4}$$

where H_{ij} are complex numbers, and η_{ij} are real numbers satisfying the inequalities $\eta_{i1} > \eta_{i2} > \ldots$

To calculate the first terms $\zeta_i^{(0)} = H_{i1}t^{\eta_{i1}}$ we use the equality

$$p^k \zeta_i^{(0)} = H_{i1}\eta_{i1}(\eta_{i1} - 1) \ldots (\eta_{i1} - k + 1)\, t^{\eta_{i1}-k}. \tag{2.5}$$

Substituting $\zeta = \zeta_i^{(0)}$ into the left-hand side of Eq. (2.3) gives

$$H_{i1}^3 t^{3\eta_{i1}} + 3H_{i1}^2 \eta_{i1} t^{2\eta_{i1}-1} + H_{i1}\eta_{i1}(\eta_{i1} - 1)\, t^{\eta_{i1}-2} + ct^{\sigma}. \tag{2.6}$$

Next, we require that the degrees of t in at least two items of this sum are equal and higher than in other terms. It gives

(i) $\eta_{i1} = \begin{cases} -1 \\ \sigma + 2 \end{cases}$ for $\sigma < -3$,

(ii) $\eta_{i1} = \sigma/3$ for $\sigma \geq -3$.

To determine the coefficients H_{i1} we require that the sum of items with the highest power of t be zero. It yields

(a) $H_{11} = 1,\ H_{21} = 2$ for $\sigma < -3,\ \eta_{11} = \eta_{21} = -1$,

(b) $H_{31} = -\dfrac{c}{(\sigma + 1)(\sigma + 2)}$ for $\sigma < -3,\ \eta_{31} = \sigma + 2$,

(c) H_{i1}, $i = 1, 2, 3$, are the roots of the equation

$$H_{i1}^3 - 3H_{i1}^2 + 2H_{i1} + c = 0$$

for $\sigma = -3$,

(d) $H_{11} = -1,\ H_{21} = \dfrac{1}{2} + \dfrac{\sqrt{-3}}{2},\ H_{31} = \dfrac{1}{2} - \dfrac{\sqrt{-3}}{2}$ for $\sigma > -3$.

It may be easily seen that for every value of σ we find three different pairs (η_{i1}, H_{i1}). It is also easy to verify that the sum of items of the highest degree in (2.6) is zero in each case.

The coefficients η_{i2}, H_{i2}, then η_{i3}, H_{i3} etc. may be determined similarly. Particularly, in the case $\sigma > -3$, according to [4, pp. 108–109], we have

$$\eta_{i2} = -1,\ H_{i2} = \sigma/3, \quad i = 1, 2, 3.$$

In order to appraise the efficacy of the method exploited, suppose $c = 1$ and $\sigma = 1$, and, under these conditions, determine roots approximately as

$$\tilde{\zeta}_i = H_{i1} t^{\eta_{i1}} + H_{i2} t^{\eta_{i2}}, \quad i = 1, 2, 3.$$

We obtain

$$\tilde{\zeta}_1 = -\sqrt[3]{t} + \frac{1}{3t},$$

$$\tilde{\zeta}_{2,3} = \left(\frac{1}{2} \pm \frac{\sqrt{-3}}{2}\right)\sqrt[3]{t} + \frac{1}{3t},$$

Substituting $\zeta = \tilde{\zeta}_1$ in the left-hand side of Eq. (2.3) gives

$$\tilde{\zeta}_1^3 + 3\tilde{\zeta}_1\dot{\tilde{\zeta}}_1 + \ddot{\tilde{\zeta}}_1 + t = 5/9\, t^{-5/3} + 1/9\, t^{-2} + 1/3\, t^{-3} = R(t).$$

One may easily see that for $t \to \infty$ $R(t) = 0(t^{-5/3})$. Moreover, the approximating function $\tilde{\zeta}_1(t)$ for a root of GCE for the equation

$$\dddot{x} + tx = 0$$

is exactly a root for an equation of the form

$$\dddot{x} + [t + 0(t^{-5/3})]x = 0.$$

The analogous results can be obtained by considering the approximate roots $\tilde{\zeta}_2$ and $\tilde{\zeta}_3$.

This example shows that the described method can have wide application including situations where the coefficients of Eq. (1.1) are rapidly time-varying. Generally, this is a common property of many methods for determining the roots of GCE (see [4]) and, particularly, of a new method which we now present.

Define the *root equation* as an n-degree algebraic equation having the roots which are elements of a fundamental set of roots to GCE. It is evident that the coefficients of this equation depend on t. Let the root equation be written in the form

$$\zeta^n + c_1(t)\,\zeta^{n-1} + \ldots + c_n(t) = 0. \tag{2.7}$$

Since there are several (in general, an infinite number) roots of GCE, the coefficients of Eq. (2.7) are not uniquely defined by the coefficients of Eq. (1.1). Evidently, the problem of approximate determination of roots is equivalent to the problem of approximate determination of coefficients $c_1(t), \ldots, c_n(t)$. In order to solve the latter problem, it is necessary to obtain differential equations with respect to these coefficients. Let us find the equations.

Fulfilling the operations inside the parentheses in Eq. (1.2) rewrite this equation in the form

$$\zeta^n + b_1(t)\,\zeta^{n-1} + \ldots + b_n(t) = -\vartheta, \tag{2.8}$$

where ϑ represents the sum of items necessary to add to the left side of Eq. (2.8) in order to obtain the left side of Eq. (1.2).

By substituting $\zeta = \zeta_i$ into Eq. (2.8) and by varying i from 1 to n we obtain a set of n equations

$$\zeta_i^n + b_1, \zeta_i^{n-1} + \ldots + b_n = -\vartheta_i, \; i = 1, \ldots, n,$$

where the ϑ_i's are some functions of $p\zeta_i, \ldots, p^{n-1}\zeta_i, t$. It is evident that for each $\zeta_i(t)$ there exist functions $d_1(t), \ldots, d_n(t)$ satisfying the equality

$$\vartheta_i = d_1(t)\,\zeta_i^{n-1} + d_2(t)\,\zeta_i^{n-2} + \ldots + d_n(t). \tag{2.9}$$

Let us require the identities of these sets of functions for all equalities. It leads to the following system of linear algebraic equations

$$d_1(t)\,\zeta_i^{n-1} + \ldots + d_n(t) = \vartheta_i, \quad i = 1, \ldots, n. \tag{2.10}$$

The determinant of the coefficients matrix of this system is the well-known Vandermonde's determinant. It does not vanish if all roots are distinct. Since the roots form jointly a fundamental set of roots to GCE, there do not exist subintervals of I on which some roots are repeated. Therefore this determinant is nonzero-valued almost everywhere on I and system of Eqs. (1.10) is solvable on any subintervals that do not involve points in which some roots are repeated.

Solving Eqs. (1.10), we obtain representation for the coefficients $d_1, \ldots, d_n$ as rational algebraic functions of the roots and their first $n - 1$ derivatives with coefficients dependent on the coefficients $b_1, \ldots, b_n$. These functions are symmetric with respect to roots and to their derivatives of each order. Using the formulas:

$$\begin{aligned}
&\zeta_1 + \ldots + \zeta_n = s_1,\\
&\zeta_1^2 + \ldots + \zeta_n^2 = s_2,\\
&\ldots\ldots\ldots\ldots\ldots\\
&\zeta_1^n + \ldots + \zeta_n^n = s_n,
\end{aligned}$$

in which $s_1, \ldots, s_n$ are expressed in terms of $c_1, \ldots, c_n$, we obtain, after their differentiation $1, 2, \ldots$ times and solving resultant systems of equations with respect to the highest roots' derivatives, representations for the first $n - 1$ roots' derivatives as rational algebraic functions of roots with coefficients dependent on the coefficients $c_1, \ldots, c_n$ and their first $n - 1$ derivatives. Thus we obtain the coefficients d_i as functions of roots. These functions are also symmetric. It allows one to eliminate the roots and hence to represent the coefficients in question as functions of coefficients $b_1, \ldots, b_n, c_1, \ldots, c_n$ and the first $n - 1$ derivatives of the c_i's. These are rational algebraic functions.

Comparing Eqs. (2.8) and (2.9) with Eq. (2.7) results in

$$c_i = b_i + d_i, \quad i = 1, \ldots, n. \tag{2.11}$$

Therefore the coefficients d_i are terms which complete the coefficients b_i transforming them into the coefficients of the root equation. If these terms are represented by expressions obtained by the above procedure, then the coeffi-

cients $c_1, \ldots, c_n$ are related by differential equations. When $n = 2$ the equations are of the form

$$c_1 = b_1 + \frac{\dot{c}_1 c_1 - 2\dot{c}_2}{c_1^2 - 4c_2}, \qquad c_2 = b_2 + \frac{2\dot{c}_1 c_2 - c_1 \dot{c}_2}{c_1^2 - 4c_2}. \tag{2.12}$$

In this case as well as in the case $n > 2$, it is easy to construct the iterative processes for approximate integration of these equations. Particularly, if all derivatives of the coefficients of Eq. (1.1) exist and $n = 2$, then the iterative process may be determined as follows:

$$c_1^{(k+1)} = b_1 + \frac{\dot{c}_1^{(k)} c_1^{(k)} - 2\dot{c}_2^{(k)}}{(c_1^{(k)})^2 - 4c_2^{(k)}},$$

$$c_2^{(k+1)} = b_2 + \frac{2\dot{c}_1^{(k)} c_2^{(k)} - c_1^{(k)} \dot{c}_2^{(k)}}{(c_1^{(k)})^2 - 4c_2^{(k)}}. \tag{2.13}$$

It is seen that the following additional condition is required: $(c_1^{(k)})^2 - 4c_2^{(k)} \neq 0$ for all k and t.

In the case $n > 2$, the iterative process may be determined similarly, i.e.

$$c_i^{(k+1)} = b_i + d_i^{(k)}, \quad i = 1, \ldots, n,$$

where the $d_i^{(k)}$'s may be expressed in terms of $c_i^{(k)}$'s by analogy with the relations between the d_i's and c_i's (see above). Of course the additional conditions such as the existence of all or some number of first derivatives of the b_i's and some inequalities are required.

The original situation may be determined by $c_i^{(0)} = b_i$, $i = 1, \ldots, n$. It gives the exact solutions in the case of time-invariant coefficients b_i for $k = 0$ and of the Euler equation for $k = 1$ and $n = 2$ provided the iterative process is defined by Eqs. (2.13). It is instructive to note that the above choice of original situation and iterative relation is not unique. Of course one can construct other iterative processes and can choose other original situations. The analysis of those possibilities might clear out the correspondence between the method of calculation and the class of equations for which this method would be sufficiently efficient.

Note the following salient characteristics of the iterative process as described with the above original situation:

a) it requires the algebraic operations and the operation of differentiation only,

b) if the coefficients $b_1, \ldots, b_n$ are given in the algebraic form, then the coefficients $c_1^{(k)}, \ldots, c_n^{(k)}$ are also obtained in the algebraic form,

c) for many cases the process gets quite good approximations on its initial stages (for example, $k = 1$).

To illustrate these statements consider the following equation

$$\ddot{x} + tx = 0. \tag{2.14}$$

Applying Eqs. (2.13) and taking into account an interval $[a, \infty)$ with $a > 0$, we conclude that

$$c_1^{(1)} = \frac{1}{2t}, \quad c_2^{(1)} = t.$$

On treating the roots of equation

$$\tilde{\zeta}^2 + \frac{1}{2t}\tilde{\zeta} + t = 0$$

as an approximation of some roots of GCE, we obtain

$$\zeta_{1,2} \cong \tilde{\zeta}_{1,2} \equiv -\frac{1}{4t}\left(1 \pm i\sqrt{16t^3 - 1}\right),$$

$$i = \sqrt{-1}.$$

In Section 3, we will apply this result to the calculation of a particular solution of Eq. (2.14), and letting $a = 2$ show that this approximation is quite good.

3. The use of the above results in the analysis of processes

The use of the above results in the analysis of processes may be successful in two ways:

a) Using the formula for the general solution of Eq. (1.1)

$$x(t) = C_1 x_1(t) + \ldots + C_n x_n(t),$$

where $C_1, \ldots, C_n$ are constants and $x_1(t), \ldots, x_n(t)$ form jointly a fundamental set of solutions, and choosing these latter ones in the form

$$x(t) = \exp \int_{t_0}^{t} \zeta_i(t)\, dt,$$

a new approach to the approximate integration of Eq. (1.1) may be proposed. This is as follows: the roots are found approximately, and then the solution is calculated by the formula

$$x(t) \cong C_1 \exp \int_{t_0}^{t} \tilde{\zeta}_1(t)\, dt + \ldots + C_n \exp \int_{t_0}^{t} \tilde{\zeta}_n(t)\, dt, \tag{3.1}$$

where $\tilde{\zeta}_1, \ldots, \tilde{\zeta}_n$ are approximate roots, and $C_1, \ldots, C_n$ are found in accordance with initial conditions.

If the coefficients of Eq. (1.1) are given in the analytical form and one of the above-mentioned methods for the approximate searching of roots is used, then either the approximate roots or the coefficients of the root equation are represented similarly. In this case, the described method for approximate integration of Eq. (1.1) may also be recommended as a method for its numerical integration. The recommendation is based on the fact that the problem of the numerical integration of a linear system of first-order differential equations reduces here to the problem of integrals

$$\int_{t_0}^{t} \tilde{\zeta}_1(t)\,dt, \ldots, \int_{t_0}^{t} \tilde{\zeta}_n(t)\,dt$$

calculation and of calculation the roots of algebraic equations.

Using the impulsive response formed as a particular solution of Eq. (1.1) and the convolution integral, a method for approximate integration of non-homogeneous differential equation

$$\frac{d^n x}{dt^n} + b_1(t)\frac{d^{n-1}x}{dt^{n-1}} + \ldots + b_n(t)\,x = f(t)$$

may be formed similarly. As in the homogeneous case, this method can be exploited as a numerical one provided the coefficients $b_1(t), \ldots, b_n(t)$ are of the analytical form.

b) When the coefficients of Eq. (1.1) are represented in analytical form, new possibilities of free and forced motions analysis arise.

To illustrate the application of the results of Section 2 to the approximate integration of Eq. (1.1) consider Eq. (2.14) during interval of time [2,10]. Let

$$x(2) = -0.015,\ \dot{x}(2) = -1.097.$$

Applying Eq. (3.1) to calculate the corresponding particular solution and using the approximation of GCE roots which has been obtained in Section 2 yields the result given in Table 1 in the row marked by II.

For comparison, the exact solution is also given in this table (see row I).

One can find the details of calculation in [1, 3, 4].

It is seen that the error of approximation does not suceed the value 0.008.

Consider now the application of the results of Section 2 to the problems of processes' analysis. In order to simplify the question, suppose the case of free motion.

If the approximate formulas for roots, which form jointly a fundamental set, are obtained, then the following is made possible:

Table 1

A particular solution of Eq. (2.14) and its approximation

t	2.0	2.25	2.5	2.75	3.0	3.25	3.5
I	−0.015	−0.282	−0.510	−0.658	−0.695	−0.603	−0.392
II	−0.015	−0.281	−0.506	−0.652	−0.687	−0.595	−0.385
t	3.75	4.0	4.25	4.5	4.75	5.0	5.25
I	−0.096	0.220	0.482	0.618	0.584	0.382	0.063
II	−0.093	0.219	0.477	0.610	0.576	0.375	0.061
t	5.5	5.75	6.0	6.25	6.5	6.75	7.0
I	−0.274	−0.519	−0.582	−0.435	−0.123	0.236	0.498
II	−0.272	−0.513	−0.575	−0.428	−0.119	0.235	0.492
t	7.25	7.5	7.75	8.0	8.25	8.5	8.75
I	0.550	0.362	0.011	−0.344	−0.533	−0.459	−0.152
II	0.542	0.355	0.009	−0.340	−0.526	−0.452	−0.148
t	9.00	9.25	9.5	9.75	10.0		
I	0.233	0.492	0.480	0.197	−0.199		
II	0.231	0.486	0.473	0.193	−0.198		

1) to obtain in analytical form the approximate formula for the general solution of Eq. (1.1);

2) to estimate particular solutions of Eq. (1.1);

3) to determine the stability conditions.

This list is not complete.

Consider the above questions sequentially.

The first possibility follows from Eq. (3.1).

To estimate particular solutions of Eq. (1.1) we transform this equation into a system of differential equations with respect to new dependent variables $z_1, \ldots, z_n$ by

$$x = A(t)z, \tag{3.2}$$

where $x = \operatorname{col}\left(x, \frac{dx}{dt}, \ldots, \frac{d^{n-1}x}{dt^{n-1}}\right)$, $z = \operatorname{col}(z_1 \ldots, z_n)$,

$$A(t) = \begin{Vmatrix} 1 & \ldots & 1 \\ \xi_1(t) & & \xi_n(t) \\ [\xi_1(t) + p]\xi_1(t) & \ldots & [\xi_n(t) + p]\xi_n(t), \\ \cdot\ \cdot\ \cdot

while $\zeta_1(t), \ldots, \zeta_n(t)$ are $n-1$-times differentiable complex-valued functions satisfying jointly the following condition:

$$\det A(t) \not\equiv 0 \text{ for } t \in I.$$

To complete the determination of the transformation, let $\xi_1, \ldots, \xi_n$ be $\tilde{\zeta}_1, \ldots, \tilde{\zeta}_n$. Then the resulting system will have the form

$$\dot{z}_i = \tilde{\zeta}_i z_i + \sum_{j=1}^{n} h_{ij} z_j, \quad i = 1, \ldots, n, \tag{3.3}$$

where

$$h_{ij} = -[(\tilde{\xi}_j + p)^{n-1} \tilde{\xi}_j + b_1(\tilde{\xi}_j + p)^{n-2} \tilde{\xi}_j + \ldots b_{n-1} \tilde{\xi}_j + b_n] \frac{(-1)^{n+i} v_{ni}}{W},$$

$W = \det A$, while v_{ni} is the determinant of the matrix that is obtained from matrix A by excluding the nth row and the ith column.

In the case $\tilde{\zeta}_i = \zeta_i$, $i = 1, \ldots, n$, the coefficients matrix of Eqs. (3.3) is diagonal. Under this condition Eqs. (3.3) have the form

$$\dot{z}_i = \zeta_i z_i, \quad i = 1, \ldots, n.$$

In view of this property of the transformation represented by Eq. (3.2), one of Wazewski's [10] theorems may be efficiently applied. This yields [3, p. 63]

$$|x(t)| \leq r_0 \sqrt{n} \exp \int_{t_0}^{t} \mu_n(t)\, dt, \tag{3.4}$$

where μ_n is the greatest eigenvalue of the Hermitian matrix $1/2\,(B + B^*)$, in which expression B is the coefficients matrix of Eqs. (3.3), and $r_0 = \sqrt{|z_1(t_0)|^2 + \ldots + |z_n(t_0)|^2}$.

These estimations may be replaced by more precise ones. For example, using the scheme suggested by the author [1] yields the sequences of upper and lower estimations which converge to the exact solution.

Now consider some results related to the stability problem and based on the use of approximate roots.

Preliminary to the discussion of stability, we point out the exploited definition of stability. Letting $u_1, \ldots, u_s$ be continuous functions of x, $\frac{dx}{dt}, \ldots, \frac{dt^{n-1}x}{dt^{n-1}}$, t, in spirit of [1–4], this takes the form:

Definition. The processes are stable with respect to $u_1, \ldots, u_s$ *if for all solutions to* Eq. (1.1) *the resultant functions* $u_1 = f_1(t), \ldots, u_s = f_s(t)$ *are bounded*

as $t \to \infty$. *They are stable asymptotically if in addition the equalities*

$$\overline{\lim_{t\to\infty}}\ u_i = 0, \quad i = 1, \ldots, s,$$

hold.

Note that the interval I in the investigation of stability is regarded as unlimited on the right.

Supposing $x = x_1, \dfrac{dx}{dt} = x_2 \ldots, \dfrac{d^{n-1}x}{dt^{n-1}} = x_n$, we obtain a set of differential equation

$$\begin{aligned}
\dot{x}_1 &= x_2, \\
\dot{x}_2 &= x_3, \\
&\ldots \\
\dot{x}_n &= -b_{n-1}x_1 - b_{n-2}x_2 - \ldots - b_1 x_n.
\end{aligned}$$

This allows one to compare our definition with well-known ones, particularly, of Lyapunov [11, pp. 13–14]. The result of comparison would be as follows. If $u_1, \ldots, u_s$ are independent of t, these definitions are equivalent. The details were given in [3, 4].

Referring to [4], we give here without proving the following two results.

Theorem 1. Let s linear forms,

$$u_i = e_{i1}x + e_{i2}px + \ldots + e_{in}p^{n-1}x, \quad i = 1, \ldots, s,$$

be chosen so that $e_{ij} = e_{ij}(t)$ *are continuous functions, let the system be characterized by* Eq. (1.1), *and let* $A(t) \neq 0$ *for* $\xi_i = \zeta_i$, $i = 1, \ldots, n$, $t \geq t_0$, *and*

$$g_1(t) = \sqrt{n}\, \exp \int_{t_0}^{t} \mu_n(t)\, dt\,,$$

$$g_2(t) = \sqrt{\sum_{i=1}^{n} |\tilde{\zeta}_i(t)|^2}\, \exp \int_{t_0}^{t} \mu_n(t)\, dt\,,$$

$$\ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots$$

$$g_n(t) = \sqrt{\sum_{i=1}^{n} |[\tilde{\zeta}_i(t) + p]^{n-2}\, \tilde{\zeta}_i(t)|^2}\, \exp \int_{t_0}^{t} \mu_n(t)\, dt\,.$$

Then the processes are stable with respect to $u_1, \ldots, u_n$ *if the inequalities*

$$\overline{\lim_{t\to\infty}} \max_j |\, e_{ij}(t)\, g_j(t)\,| < \infty, \quad i = 1, \ldots, n,$$

hold.

Theorem 2. If the processes are stable and in addition the equalities

$$\overline{\lim_{t\to\infty}} \max_j | e_{ij}(t)\, g_j(t) | = 0, \quad i = 1, \ldots, n,$$

hold, then they are asymptotically stable.

It is easily shown that the sufficient condition for stability coincides with a necessary one whenever $\tilde{\zeta}_i = \zeta_i$, $i = 1, \ldots, n$, and there exist a number T and a value m of the index i such that $\operatorname{Re} \zeta_m \geq \operatorname{Re} \zeta_i$, $i = 1, \ldots, n$, for $t > T$.

4. The problem of uniqueness for fundamental set of roots

Methods for the analysis based on approximate integration of GCE are specified by dealing with a set of such approximate solutions of Eq. (1.1) that are represented in the form

$$\exp \int_{t_0}^{t} \tilde{\zeta}_i(t)\, dt,$$

where $\tilde{\zeta}_i(t)$ is an approximate root of GCE.

It is clear from the foregoing sections that the concept of root of GCE depends on the interval I under consideration; i.e., a solution of GCE that is a root may lose its root property if the interval I is replaced by an interval $I' \subset I$ and the solution is continued on it. The question arises: can one obtain a unique fundamental set of roots by extension of I? If a positive solution of this problem exists, then the theory of free motion for time-varying systems may be elaborated by the analogy with the time-invariant case.

However, unlike the situation for characteristic roots in the case of time-invariant systems, the solution of this problem is negative even if I is $(-\infty, \infty)$. It is easily shown by the following example.

Example 1. Consider the case $n = 2$ of Eq. (1.1). Let the coefficient b_1 and b_2 be constants (this case may be regarded as a particular case of time-varying systems) and the characteristic roots (in the classical sense) λ_1 and λ_2 be real and distinct. Then in addition to the fundamental set of roots of GCE

$$\zeta_1(t) = \lambda_1, \quad \zeta_2(t) = \lambda_2 \tag{4.1}$$

one may construct an infinite number of other fundamental sets by using the following procedure.

Pairs of particular solutions to Eq. (1.1) are constructed as follows:

$$x_{1,2}(t) = C_1 \exp \lambda_1 t \pm \sqrt{-1}\, C_2 \exp \lambda_2 t, \tag{4.2}$$

where C_1 and C_2 are real and nonzero-valued. The corresponding roots are determined as their logarithmic derivatives.

The possibility of a positive solution of the problem in question appears at least for some classes of equations if we extend the definition of the equation on the full complex plane. Particularly, if it is done in the above example, then all solutions of GCE that are constructed as logarithmic derivatives of $x_{1,2}(t)$ given by Eq. (4.2) cease to be roots, since if k is integral and

$$t = a_{1,2} \equiv \frac{\ln C_1 \pm \pi/2\sqrt{-1} - \ln C_2 + 2k\pi\sqrt{-1}}{\lambda_2 - \lambda_1},$$

then in these points $x_{1,2} = 0$ and $\zeta_{1,2}$ are undefined. Meanwhile the root property of functions defined by Eqs. (4.1) preserves. Thus in this case, the fundamental set of roots of GCE is unique.

In view of that particular result, the extension of the domain of definition of Eq. (1.1) on the full complex plane seems to be hopeful. However, due to new situation, the existence problem again arises. For solving this problem, a new approach based on the theory of differential equations in the complex domain is required

It is easily shown that in the more general case than of Example 1, namely, in the case of an arbitrarily finite-order differential equation with constant coefficients, a fundamental set of roots of GCE is unique whenever the domain of definition of Eq. (1.1) is the full complex plane and the ordinary characteristic roots are all distinct. Moreover, the roots of GCE and characteristic roots coincide here.

The modification of Example 1 assuming multiple roots, however, leads to a negative solution of the existence problem. To verify this consider the next example.

Example 2. Let $n = 2$, the coefficients b_1 and b_2 be constants and the characteristic roots be repeated, i.e. $\lambda_1 = \lambda_2$. Then it is clear that any element $x_i(t)$ of any fundamental set may be represented in the form

$$x_i(t) = (C_1 t + C_2) \exp \lambda_1 t, \tag{4.3}$$

where C_1 and C_2 are constants. Obviously, any fundamental set contains an element for which $C_1 \neq 0$. For such an element, there exists a complex number c, namely $c = -C_2/C_1$, such that $x_i(c) = 0$ and $| \dot{x}_i/x_i | \to \infty$ as $t \to c$. Thus, if Eq. (1.1) is regarded on the full complex plane, the fundamental set of roots does not exist.

Since the existence problem is of great importance, it is interesting to regard this example from the point of view of seeking such its modification that ensures the positive solution of the problem. Clearly the existence of the fundamental set of GCE roots may be ensured if we eliminate the point $t = 0$ from the domain of definition of Eq. (1.1). In this case, the two GCE roots

forming jointly a fundamental set may be determined as logarithmic derivatives of the following solutions to Eq. (1.1):

$$x_1(t) = \exp \lambda_1 t, \; x_2(t) = t \exp \lambda_1 t. \tag{4.4}$$

These roots are of the form

$$\zeta_1(t) = \lambda_1, \; \zeta_2(t) = \lambda_1 + 1/t. \tag{4.5}$$

Note that this fundamental set is unique because

(a) for Eq. (1.1), the logarithmic derivative of any solution of the form $C_1 \exp \lambda_1 t$ or $C_2 \text{texp} \lambda_1 t$, where C_1 and C_2 are nonzero constants, is one of (4.5);

(b) independently of the values of nonzero constants C_1 and C_2, there exists a point $t = c \neq 0$ such that

$$x_3(t) = C_1 x_1(t) + C_2 x_2(t) = 0$$

as $t = c$ and hence $\dot{x}_3/x_3$ is not a root.

Reverting to the case of general time-varying system and taking into account the above consideration, we may formulate the following two problems:

(a) picking out classes of equations for which a fundamental set of roots exists and is unique if their domains of definition are

(i) the full complex plane,

(ii) the complex plane without one point,

(iii) the complex plane without several points;

(b) constructing methods for obtaining the succesive approximation of a fundamental set of roots for such equations.

If the first problem would be solved, the theory of time-varying systems might be essentially simplified. This is based on the possibility to introduce the concept of *main fundamental set of roots* of GCE as a set containing such elements which do not lose the root property when the initial domain of definition of Eq. (1.1) is extending on the corresponding complex domain.

Methods for determining approximate roots, which were mentioned in Section 2, in some cases, may represent a solution of the second problem.

Thus by transferring the free motion equation into the complex domain a new basis for methods of time-varying systems analysis appears.

5. Conclusion

The generalized characteristic equation (GCE) method described in this paper provides a general approach to a few problems of time-varying system analysis. It is applicable to the systems characterized by the n-order differential equation.

The purpose of this paper has been to delineate some methods for solving basic problems of time-varying systems analysis by using the GCE conception. There are methods for obtaining an approximate formula for general solution of original differential equation, for estimation a particular solution of this equation, and for determining the stability conditions.

In the case of appropriate coefficients of original differential equation, there are some connections between the theory of GCE and the theory of linear differential equation in the complex domain. The investigation of this connections may be useful.

References

1. *Mikhailov, F. A.*, "The free motion of time-varying linear systems", Proceedings of the Moscow Institute of Aircraft Engineering, **135** (1961), pp. 1–270 (in Russian).
2. *Mikhailov, F. A.*, "A method for the analysis of the free motion of time-varying linear systems using canonic transforms of the free motion equation", Proceedings of the Moscow Institute of Aircraft Engineering, **189** (1970), pp. 5–32 (in Russian).
3. *Mikhailov, F. A., Teryaev, E. D., Bulekov, V. P., Dankov, G. J., Salikov, L. M., Stepanyancs, G. A.*, "Dynamics of time-varying linear systems", Izdat. "Nauka", Moscow, 1967 (in Russian).
4. *Mikhailov, F. A., Teryaev, E. D., Bulekov, V. P., Salikov, L. M., Dikanova, L. S.*, "Dynamics of continuous linear systems with deterministic and random parameters", Izdat. "Nauka", Moscow, 1971 (in Russian).
5. *Mikhailov, F. A.*, "Two theorems to the theory of linear ordinary differential equation", Isvestia Vysceshikh Uchebnikh Zavedenii, Mathematica, **2** (39), 1964, pp. 120–126 (in Russian).
6. *Mikhailov, F. A.*, "On a method for stability analysis of linear time-varying systems", Isvestia AN SSSR, Tekhnicheskaya Kibernetika, **4**, 1963, pp. 171–179 (in Russian).
7. *Mikhailov, F. A.*, "Random processes in the time-varying linear systems", ser. Automatika, Isdat. "Energiya", Moscow, 1969 (in Russian).
8. *Mikhailov, F. A.*, "The methods of time-variable systems analysis based on new trends in theory of these systems", Proceedings of the Fifth IFAC Congress, June 12–17, 1972, Paris; Instrument Society of America, Pittsburgh, 1972, pp. 35.3, 1–9.
9. *Cayley, A.*, "Note on the theory of linear differential equations", Journ. reine und angewandte Mathematik **100** (1887), pp. 286–295.
10. *Wazewski, T.*, "Sur la limitation des intégrales des systémes d'équations différentielles linéaires ordinaires", Studia Mathematica, **10** (1948), pp. 48–59.
11. *Lyapunov, A. M.*, "The general problem of the stability of the motion", Complete proceedings, vol. **2**, Isdat. AN SSSR, Moscow–Leningrad, 1956, pp. 5–263 (in Russian).

Метод обобщенного характеристического уравнения в теории нестационарных линейных систем

Ф. А. МИХАЙЛОВ

(Москва)

Рассматривается метод анализа процессов в непрерывных нестационарных линейных системах, использующий обобщенное характеристическое уравнение (1.2), получаемое из уравнения свободных колебаний (1.1) в результате замены зависимой переменной по формуле (1.3).

Приводятся заимствованные из [4] понятия *корня* и *фундаментальной системы корней* обобщенного характеристического уравнения (ОХУ) и следующей доказанный в [5, 6] результат: если коэффициенты уравнения (1.1) непрерывны, то фундаментальная система корней *существует*.

Показываются возможности расчета и анализа процессов в системах рассматриваемого класса, связанные с использованием ОХУ. При этом непрерывность коэффициентов уравнения (1.1) предполагается. Приводимые методы базируются на предварительном приближенном вычислении фундаментальной системы корней ОХУ. Указывается на ряд способов решения последней задачи, предложенных в [1-4]. Характер способов иллюстрируется на примере уравнения (2.2).

Приводится новый способ, в котором используется «корневое уравнение» (2.7). Это уравнение строится как алгебраическое уравнение, корнями которого являются элементы фундаментальной системы корней ОХУ. Предлагается приближенно вычислять коэффициенты корневого уравнения, а затем вычислять корни получаемой аппроксимации этого уравнения. Эти корни могут рассматриваться как аппроксимации корней ОХУ.

Для приближенного вычисления коэффициентов корневого уравнения предлагается использовать дифференциальные уравнения, связывающие их с коэффициентами уравнения (1.1) Указывается путь получения таких уравнений и определяется итеративный процесс их приближенного решения.

Предлагается приближенная формула общего решения уравнения (1.1) в виде (3.1). Указывается, как организовать процесс приближенного интегрирования уравнения (1.1) на основе этой формулы.

Возможности анализа процесса свободных колебаний иллюстрируются результатами решения двух вопросов:

а) оценки частного решения уравнения (1.1),

б) получения достаточных условий устойчивости процессов.

Результаты базируются на преобразовании уравнения (1.1) в систему (3.3) согласно подстановке (3.1). Преобразование такого вида в более ранних работах автора названо *каноническим.*

Каноническое преобразование уравнения (1.1) позволяет эффективно применить теорему Важевского [10] и получить оценку (3.4). Указывается на путь уточнения оценки, предложенный в [1].

Приводятся достаточные условия устойчивости процессов (теорема I) и асимптотической устойчивости (теорема II).

Рассматривается проблема единственности фундаментальной системы корней ОХУ. Показывается, что в общем случае положительного решения этой проблема не существует. Указывается, что возможность положительного решения появляется при распространении области определения уравнения (1.1) на комплексную плоскость. Однако проблема существования фундаментальной системы корней при этом должна рассматриваться заново. Указывается на возможность для некоторых классов уравнений найти такую область определения (на комплексной плоскости), при которой фундаментальная система корней ОХУ существует и является единственной. Такая фундаментальная система корней при рассмотрении уравнения (1.1) в вещественной области может рассматриваться как *главная.* Отмечается перспективность этого направления.

Ф. А. Михайлов
Московский авиационный институт им. С. Орджоникидзе
СССР, Москва А-80,
Волоколамское ш., д. 4

Problems of Control and Information Theory, Vol. 5 (2), pp. 167—180 (1976)

A METHOD OF ANALYSIS OF A LIMIT CYCLING RELAY SYSTEM

M. A. SULTAN

(Cairo)

(Received January 6, 1975)

The aim of the present work is to investigate the stability of self-periodic solution in a relay feedback limit cycling system under sinusoidal and random disturbances. The work introduces a uniquely simplified approach to solve these problems. This approach is based on a method of linearization, which has been published by the author, and its application has been extended to solve these problems.

1. Introduction

The relay feedback system under consideration is shown in Fig. 1, in which the linear part $W_L(s)$ is given by

$$W_L(s) = \frac{k}{s(T_1 s + 1)(T_2 s + 1)} \tag{1}$$

where k is a gain, T_1, T_2 are time constants, and s is the Laplace variable.

It is known that the given relay feedback system is limit cycling, or in other words it is a system in which a self-periodic solution exists. The problems of interest to the control engineer for analysis of such a system are:

(a) Analysis to determine the conditions of existence of forced periodic solutions, namely, when $f(t)$ in Fig. 1 is sinusoidal with amplitude A_f and frequency Ω.

(b) Analysis to determine the conditions of existence of self-periodic solution when $f(t)$ is a stationary random signal having standard deviation σ_f.

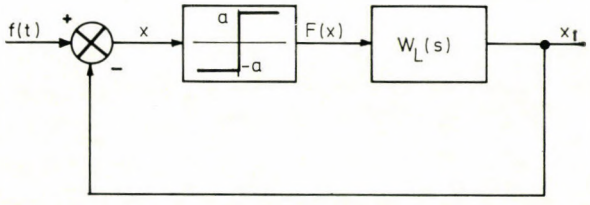


Fig. 1. Limit cycling relay feedback system:

$$W_L(s) = \frac{k}{s(T_1 s + 1)(T_2 s + 1)}$$

The solution of these problems can be found in literature. However, the techniques used are rather tedious, and also lack rigorousness, particularly with regard to the second problem. It is the object of this work to introduce a simplified approach to solve the above-mentioned problems. The present approach to solve the second problem postulated above is based on a method of linearization which has already been published [1], and is given again here in Appendix I.

2. Application of the method of linearization to the given relay system

Referring to the relay system of Fig. 1, we have

$$x = f(t) - x_1, \tag{2}$$

$$kF(x) = [(T_1 s + 1)(T_2 s + 1)] s x_1 \tag{3}$$

where

$$F(x) = a \operatorname{sign} x. \tag{4}$$

From Eqs (2) and (3) we have

$$\Phi(t) - kF(x) = [(T_1 s + 1)(T_2 s + 1)]\, sx \tag{5}$$

where

$$\Phi(t) = [(T_1 s + 1)(T_2 s + 1)]\, sf(t). \tag{6}$$

Therefore, the given system of Fig. 1 can be represented by the block diagram of Fig. 2. In Figs 2 and 3 $W'_L(s) = \dfrac{1}{k} W_L(s)$.

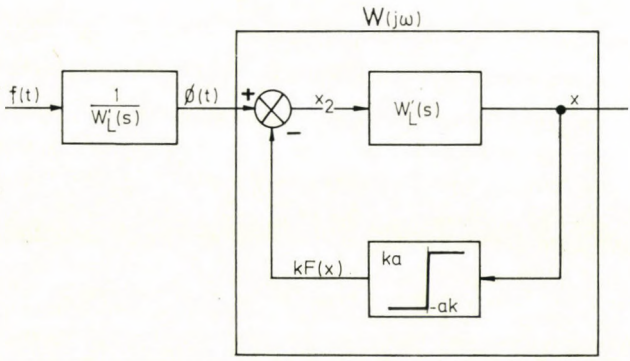


Fig. 2. Block diagram of relay system equivalent to Fig. 1:

$$W'_L(s) = \frac{1}{s(T_1 s + 1)\,(T_2 s + 1)}$$

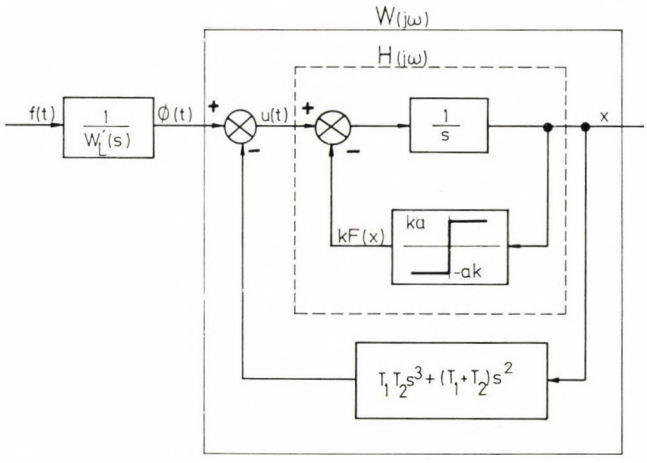


Fig. 3. Reoriented block diagram of Fig. 2

Now referring to the method of linearization in Appendix I, the part of the system of Fig. 3 in dashed square is described by

$$\frac{dx}{dt} = u(t) - ak \operatorname{sign} \dot{x}. \tag{7}$$

Equation (7) is nonlinear. It is of the type described by Eqs (51) in Appendix I, where $n = 1$, $F(x) = ak \operatorname{sign} x$. Also the unit step response from rest is monotonic as it follows from

$$x(t) = Kt \tag{8}$$

where

$$K = 1 - ak. \tag{9}$$

The process of linearization can be applied now with respect to this part of system by applying Eqs (59) and (60). Therefore, we have the following expression for the amplitude frequency characteristic $|H(\omega)|$ of the nonlinear part of Fig. 3

$$|H(\omega)| = \frac{K}{\omega}. \tag{10}$$

From (10) it is evident that the phase frequency characteristic associated with the amplitude frequency characteristic $|H(\omega)|$ is constant and equals $-\pi/2$. Therefore,

$$H(j\omega) = \frac{K}{\omega} \underline{/ -\pi/2} = \frac{K}{j\omega}. \tag{11}$$

The transfer locus $W(j\omega)$ of the part of the system indicated in Fig. 2 is therefore given by

$$W(j\omega) = \frac{H(j\omega)}{1 - [(T_1 + T_2)\,\omega^2 + T_1 T_2\,\omega^3 j]\,H(j\omega)}\,. \tag{12}$$

By substituting for $H(j\omega)$ as given by (11) we have

$$W(j\omega) = \frac{K}{-K(T_1 + T_2)\,\omega^2 + j\omega(1 - T_1 T_2 K\omega^2)}\,. \tag{13}$$

3. Analysis of realy system for forced periodic solutions

In this section the first problem postulated in Section 1 is solved. By the analysis of forced periodic solutions it is meant the determination of a threshold amplitude $A_{f_{th}}$ of the forced harmonic oscillations $f(t) = A_f \sin(\Omega t + \theta)$, which will put off the self-periodic oscillation and let the system oscillate with the forced frequency Ω.

Referring to Fig. 2, let a forced periodic solution exist in the system, therefore, we have the following fundamental harmonic oscillations in the system

$$x_2(t) = A_2 \sin \Omega t, \tag{14}$$

$$x(t) = A \sin(\Omega t + \alpha) \tag{15}$$

$$\Phi(t) = A_\Phi \sin(\Omega t + \gamma) \tag{16}$$

where α is the phase angle between the output and input of the linear element $W'_L(s)$.

The response of the relay element to the sinusoidal function at its input has a fundamental given by

$$kF(x) = \frac{ak\pi}{2} \sin(\Omega t + \alpha)\,. \tag{17}$$

It can be seen from Fig. 2 that

$$\Phi(t) = x_2 + kF(x) \tag{18}$$

or

$$A_\Phi \sin(\Omega t + \lambda) = A_2 \sin \Omega t + \frac{ak\pi}{2} \sin(\Omega t + \alpha)\,. \tag{19}$$

It can also be seen from the vector diagram of Fig. 4 that A_Φ follows from:

$$A_\Phi = \sqrt{A_2^2 + \left(\frac{ak\pi}{2}\right)^2 - A_2\,ak\pi \cos\alpha}\,. \tag{20}$$

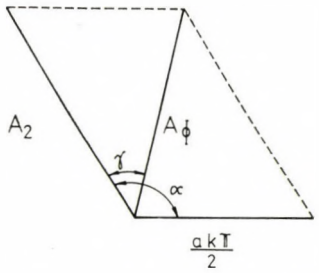


Fig. 4. Vector diagram for relation (19)

Minimizing A_Φ with respect to A_2, we have for $\frac{dA_\Phi}{dA_2} = 0$,

$$A_2 = \frac{ak\pi}{2} \cos \alpha \,. \tag{21}$$

Therefore, the minimum value of A_Φ follows from

$$A_\Phi|_{\min} = \frac{ak\pi}{2} \sin \alpha \,. \tag{22}$$

Now referring to Fig. 2, it can be easily shown that

$$\sin \alpha = \frac{(1 - T_1 T_2 \Omega^2)}{\sqrt{(1 + T_1^2 \Omega^2)(1 + T_2^2 \Omega^2)}} \,. \tag{23}$$

Also from Fig. 2 the amplitudes A_f and A_Φ are related by

$$A_\Phi = [\sqrt{(1 + T_1^2 \Omega^2)(1 + T_2^2 \Omega_2^2)}]\, A_f \Omega \,. \tag{24}$$

Therefore, from (22), (23) and (24) we have

$$A_f|_{\min} = \frac{ak\pi}{2\Omega} \frac{(1 - T_1 T_2 \Omega^2)}{(1 + T_1^2 \Omega^2)(1 + T_2^2 \Omega^2)} \,. \tag{25}$$

It can therefore be concluded that the condition of existence of forced periodic solution in the relay system of Fig. 1 is that the forced harmonic oscillations $f(t)$ have a threshold amplitude $A_{f_{th}}$ given by

$$A_{f_{th}} \geq A_f|_{\min} \,. \tag{26}$$

From expression (25), it can be seen that

(a) The self-periodic solution, which corresponds to $A_f|_{\min} = 0$, exists at a frequency Ω_0 given by

$$\Omega_0 = \frac{1}{\sqrt{T_1 T_2}} \,. \tag{27}$$

(b) For $T_1\Omega < 1$, $T_2\Omega < 1$, and hence $T_1T_2\Omega^2 \ll 1$, we have

$$A_f|_{\min} \cong \frac{ak\pi}{2\Omega}. \tag{28}$$

4. Analysis of relay system for random input signals

In this section the second problem postulated in Section 1 is solved. In the limit cycling relay system of Fig. 1 if $f(t)$ is a stationary random signal having a level designated by its standard deviation σ_f, then it is known that if the level of the signal exceeds a certain threshold level $\sigma_{f_{th}}$, the self-periodic solution in the system will not exist. In order to determine $\sigma_{f_{th}}$ by applying the present linearization technique to the relay system, we proceed as follows:

Consider Fig. 2 and a stationary random signal $\Phi_r(t)$ having power density spectrum $S_\Phi(\omega)$. The power density spectrum $S_x(\omega)$ of the signal x would then be given by

$$S_x(\omega) = |W(j\omega)|^2 S_\Phi(\omega) \tag{29}$$

where $|W(j\omega)|$ is the amplitude frequency characteristic of $W(j\omega)$ given by (13). Therefore, we have

$$|W(j\omega|^2 = \frac{K^2}{\omega^2[2K^2(T_1+T)^2\omega^2 + (1-KT_1\omega^2)^2]}. \tag{30}$$

Now if self-periodic solution exists in the system, the power density spectrum $S_x(\omega)$ by referring to Fig. 2 can also be given by

$$S_x(\omega) = |W'_L(j\omega)|^2 S_\Phi(\omega) - |W'_L(j\omega)|^2 S_r(\omega) \tag{31}$$

where

$$|W'_L(j\omega|^2 = \frac{1}{\omega^2[(1-T_1T_2\omega^2)^2 + (T_1+T_2)^2\omega^2]}; \tag{32}$$

and $S_r(\omega)$ is the power density spectrum of the output signal of the relay element, shown in Fig. 5, when self-periodic solution exists in the system.

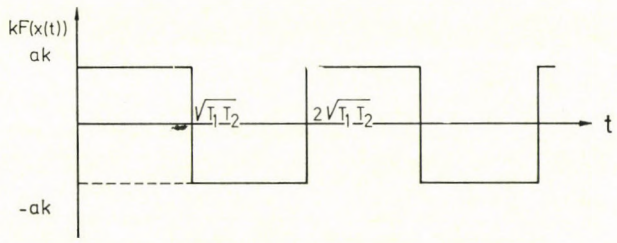


Fig. 5. Square wave signal

Therefore, from (29) and (31) we have the following relation in case of existence of self-periodic solution in the relay system

$$S_{\Phi}(\omega) = \frac{|W'_L(j\omega)|^2}{|W'_L(j\omega)|^2 - |W(j\omega)|^2} \cdot S_r(\omega) . \tag{33}$$

In case that self-periodic solution does not exist in the system, the above equality (33) would obviously be disturbed, namely,

$$S_{\Phi}(\omega) \gtrless \frac{|W'_L(j\omega)|^2}{|W'_L(j\omega)|^2 - |W(j\omega)|^2} \cdot S_r(\omega) . \tag{34}$$

However, it is physically evident that the self-periodic oscillations will be put off if the random input signal exceeds a certain level. The self-periodic solution will therefore still exist in the system if $S_{\Phi}(\omega)$ is less than the right-hand side of (33). Therefore, self-periodic solutions under random disturbances will exist in the system if

$$S_{\Phi}(\omega) \leq \frac{|W'_L(j\omega)|^2}{|W'_L(j\omega)|^2 - |W(j\omega)|^2} \cdot S_r(\omega) . \tag{35}$$

From Fig. 2 it can be seen that the power density spectrum $S_f(\omega)$ of the input random signal $f_{(t)}$ is related to $S_{\Phi}(\omega)$ by

$$S_f(\omega) = |W'_L(j\omega)|^2 S_{\Phi}(\omega) \tag{36}$$

Therefore, from (35) and (36) the condition of existence of self-periodic solution with random disturbances is

$$S_f(\omega) \leq \psi(\omega) S_r(\omega) \tag{37}$$

where:

$$\psi(\omega) = \frac{|W'_L(j\omega|^4}{|W'_L(j\omega)|^2 - |W(j\omega)|^2} . \tag{38}$$

Substituting for $|W'_L(j\omega)|$, $|W(j\omega)|$ in (38), we have

$$\psi(\omega) = \frac{1 + 2K\omega^2[K(T_1^2 + T_2^2) - T_1T_2] + K^2\omega^4 T_1^2 T_2^2}{\omega^2[(1-K^2) + \omega^2 K(T_1^2 K + T_2^2 K - 2T_1T_2[(1 - T_1T_2\omega^2)^2 + (T_1 + T_2)^2\omega^2]} . \tag{39}$$

The power density spectrum $S_r(\omega)$ of the square wave signal shown in Fig. 5 is given by

$$S_r(\omega) = \left(\frac{2ak}{\pi}\right)^2 \left[\delta(\omega - \Omega_0) + \frac{\delta(\omega - 3\Omega_0)}{9} + \frac{\delta(\omega - 5\Omega_0)}{25} + \dots\right] \tag{40}$$

where Ω_0 is the frequency of self-periodic solution given by (27), and $\delta(\omega - \Omega_0)$, $\delta(\omega - 3\Omega_0)$, $\delta(\omega - 5\Omega_0)$ . . . are delta functions.

Now the mean square value σ_f^2 of the input random sianal $f(t)$ is related to $S_f(\omega)$ by

$$\sigma_f^2 = \frac{1}{\pi} \int_0^\infty S_f(\omega)\, d\omega \qquad (\omega > 0)\,. \tag{41}$$

From (37) and (41) it follows that the condition for existence of self-periodic solution with random input signals is

$$\sigma_f^2 \leq \frac{1}{\pi} \int_0^\infty \psi(\omega)\; S_r(\omega)\, d\omega\,. \tag{42}$$

Substituting for $S_r(\omega)$ as given by (40), and making use of the known relation

$$\int_0^\infty \chi(\tau)\delta(\tau - \tau_0) = \chi(\tau_0) \tag{43}$$

expression (42) will be

$$\sigma_f^2 \leq \frac{4a^2 k^2}{\pi^3}\left[\psi(\Omega_0) + \frac{\psi(3\Omega_0)}{9} + \frac{\psi(5\Omega_0)}{25} + \ldots\right]. \tag{44}$$

Now from (39), it can be seen that the second and higher terms in expression (44) can be neglected. Therefore, expression (44) reduces to

$$\sigma_f^2 \leq \frac{4a^2 k^2}{\pi^3}\,\psi(\Omega_0) \tag{45}$$

and also from (39) we have

$$\psi(\Omega_0) = \frac{T_1 T_2\,[1 + K(K-2) + 2\beta K^2)}{(\beta + 2)\,[1 - K(K+2) + \beta K^2]} \tag{46}$$

where

$$\beta = \frac{T_1^2 + T_2^2}{T_1 T_2}\,. \tag{47}$$

Therefore, expression (45) will be

$$\sigma_f^2 \leq \frac{4a^2 k^2}{\pi^3}\;\frac{T_1 T_2\,[1 + K(K-2) + 2\beta K^2)}{(\beta + 2)\,[1 - K(K+2) + \beta K^2]}\,. \tag{48}$$

In practice $ka \gg 1$. Therefore, $K \cong -ak$ and expression (48) reduces then to

$$\sigma_f^2 \leq \left(\frac{4a^2 k^2}{\pi^3}\right)\frac{T_1 T_2(2\beta + 1)}{(\beta - 1)\,(\beta + 2)} \tag{49}$$

giving the condition for existence of self-periodic solution in the relay feedback system with random input signal having mean square value σ_f^2.

For the special case when $T_1 = T_2 = T$, we have $\beta = 2$ and expression (49) will be

$$\sigma_f^2|_{T_1=T_2} \leq \frac{5a^2 k^2 T^2}{\pi^3}. \tag{50}$$

Appendix I. A method of linearization for a class of nonlinear systems

The objectives of the method of linearization proposed here arose in consideration of the fact that the known methods of analysis and synthesis of nonlinear systems are far lagging behind the corresponding linear techniques as far as rigorousness and simplicity are concerned. For a class of nonlinear systems the pertinent linear spectral techniques can be applied for the analysis of the nonlinear system by applying the present method of linearization.

In essence, that main aspect of this method is that approximate transfer loci can be obtained for the part containing the nonlinear effects in a given system (Fig. 6). This nonlinear part or element which will be completely defined in the next section of this Appendix is described by a set of canonical equations of the form

$$\frac{dX_1}{dt} = X_2, \frac{dX_2}{dt} = X_2, \ldots,$$

$$\frac{dX_n}{dt} = Bu(t) - F(X_1, X_2, \ldots X_n) \tag{51}$$

where B is a constant, $u(t)$ is a steering variable, and $F(X_1, X_2, \ldots, X_n)$ is a nonlinear function with respect to the phase coordinates $X_1, X_2, \ldots, X_n$.

Hence, it is possible that the methods known in the linear theory be applied to a nonlinear system incorporating nonlinear elements of the type given by (51).

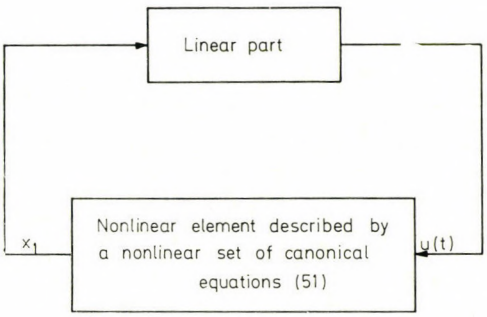


Fig. 6. Nonlinear system comprising nonlinear element of the type described by (51)

(i) *Defining a class of nonlinear elements of automatic control*

The transfer function $G(s)$ of linear elements of automatic control with noncomplex negative poles and no zeros is generally given by

$$G(s) = \mathfrak{L}\left[\frac{X_1(t)}{u(t)}\right] = \frac{B}{s^{n-1}\prod_{i=1}^{l}(sT_i + 1)}; \; l \leq n, \tag{52}$$

where X_1 is the regulated variable, T_i is the time constant, real positive, and s is the Laplace variable.

This type of linear elements can also be described by the following set of canonical equations

$$\frac{dX_1}{dt} = X_2, \frac{dX_2}{dt} = X_3, \ldots,$$

$$\frac{dX_n}{dt} = Bu(t) - F_L(X_2, X_2, \ldots X_n) \tag{53}$$

in which $F_L(X_1, X_2, \ldots, X_n)$ is a linear function with respect to the phase coordinates such that

$$F_L(X_1, X_2, \ldots, X_n) = -F_L(-X_1, -X_2, \ldots, -X_n) \text{ and } F_L(0) = 0. \tag{54}$$

The linear element given by (52) or (53) has its poles on the negative real axis including the origin in the complex plane. Therefore, an oscillatory mode for the given linear element does not exist if a step disturbance is applied to it. An evidence therefore that the linear canonical equations (53) belong to the transfer function (52) would be a monotonous step response from rest of the given element, namely

$$| X_1(t + \Delta t) | \geq | X_1(t) | \qquad 0 < t < \infty. \tag{55}$$

In actual cases the derivation of the expression of the transfer function starts by putting down the equations describing an element which are of the type given by (53) and then transforming these equations by Laplace. In most cases in practice the function $F_L(X_1, X_2, \ldots, X_n)$ is nonlinear with respect to the phase coordinates as a result of the inherent nonlinearities appearing in the characteristics of a given element. It is only through incremental or if possible piecewise linearization of these characteristics that the function $F_L(X_1, X_2, \ldots, X_n)$ can be made linear so that the equations describing the given element may be transformed. Therefore, it can be said that the set of canonical equations (51) which as Eqs (53) have the properties given by (54) and (55) but in which the function $F(X_1, X_2, \ldots, X_n)$ is nonlinear with respect

to the phase coordinates represent in fact a more general case of a wide class of elements of automatic control met in practice. It is this class of nonlinear elements that a method for obtaining approximate transfer loci corresponding to them is proposed for. This implies in fact a linearization process carried out with regard to such nonlinear elements as long as transfer loci can be defined only for linear elements. The linearization process is postulated in the next section and may be carried out with regard to any nonlinear element of the above-mentioned type which might be incorporated in a system of automatic control.

(ii) *Postulation of a method of linearization for a class of nonlinear systems*

In the system shown in Fig. 7, $1(t)$ is a unit step function applied to a linear element having a transfer function $H(s)$ of the following form

$$H(s) = B/s^n. \tag{56}$$

The time response denoted by $X_1L(t)$ for the linear element (56) when starting from rest is given by

$$X_{1L}(t) = \frac{Bt^n}{n!}. \tag{57}$$

Also for such a linear element the amplitude of the transfer locus $| H(\omega) |$ follows from

$$| H(\omega) | = B/\omega^n, \tag{58}$$

where ω is the frequency.

From expressions (57) and (58) it can be seen that the following relations exist between the amplitude of the transfer locus $| H(\omega) |$ and the time response $X_1L(t)$ of the linear element (56), namely

$$X_{1L}(t) = \varepsilon \, | H(\omega) |, \tag{59}$$

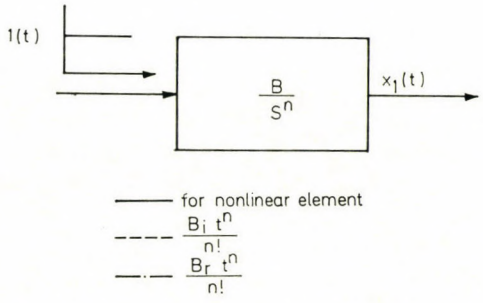


Fig. 7. Linear element B/s^n with unit step input

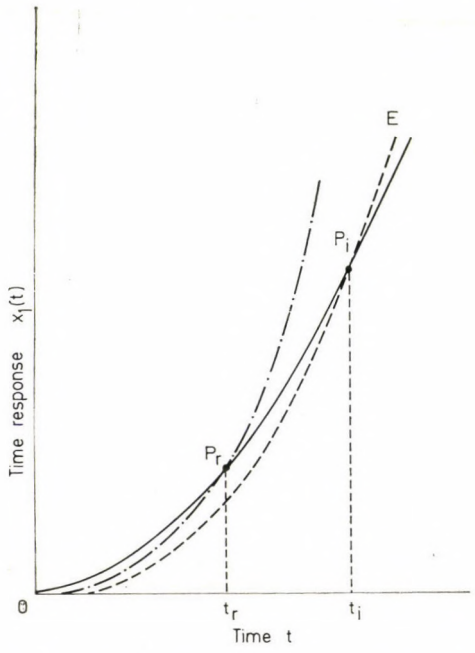


Fig. 8. Illustration of the method of linearization

$$\omega t = (\varepsilon n!)^{1/n}, \tag{60}$$

where ($\varepsilon \neq 0$) is a proportionality factor.

Expressions (59) and (60) deduced in the above represent in fact the fundamental relations upon which the proposed method of linearization is based. An explanation of the proposed process of linearizaion can be now pursued as follows:

It is evident that in the absence of the function $F(X_1, X_2, \ldots, X_n)$ Eqs (51) will reduce to

$$\frac{dX_1}{dt} = X_2, \frac{dX_2}{dt} = X_3, \ldots, \frac{dX_n}{dt} = Bu(t) \tag{61}$$

which in fact describe the linear element (56) for which the above relations (59) and (60) have been derived. Consider now the diagram of Fig. 8 in which the solid line represents the step response $X_1(t)$ from rest of the nonlinear element (51) specified in the last section. At a certain point P_i on this line, there is a linear model of the form B_i/s^n having a unit step response from rest which passes through point P_i as shown by the dotted line OE in Fig. 8. Referring to expression (57) it can be seen that B_i follows from

$$B_i = \frac{X_1(t)}{t_i^n/n!}. \tag{62}$$

From expressions (59) and (60) we see that for this linear element at point P there is an amplitude $| H(\omega_i) |$ of the transfer locus given by:

$$X_1(t_i) = \varepsilon \, | H(\omega_i) |, \tag{63}$$

where

$$\omega_i = \frac{(\varepsilon_n)^{1/n}}{t_i}. \tag{64}$$

For the entirety of points on the solid line of Fig. 8 there will be a continuously varying linear model of the form B_r/s^n in which B_r represents the variability and which has at any particular moment t_r the same value $X_1(t_r)$ and hence $\varepsilon \, | H(\omega_r) |$ is the generally nonlinear element (51). Therefore, an amplitude of the transfer locus corresponding to the nonlinear element can be defined first by obtaining the unit step response from rest of the nonlinear element (51) and then applying relations (59) and (60). The factor of proportionality ε appearing then in the expression of the transfer locus amplitude may be cancalled out in some problems, and may be determined by considering the physical conditions existing in a system in other problems. A phase frequency characteristic can then be associated with the amplitude of the transfer locus by applying the known graph-analytical methods of the linear theory. Once the transfer locus corresponding to the nonlinear element is defined, the known linear spectral methods can be applied for the analysis or synthesis of a control system containing such a nonlinear element.

References

1. *Sultan, M. A.*, A method of linearisation for a class of nonlinear systems with random disturbances. Regelungstechnik **2** (16. Jahrgang 1968).
2. *Graham, D., McRuer, D.*, Analysis of nonlinear control systems. John Wiley & Sons, New York, 1961.
3. *Popov, E. P., Paltov, U. P.*, Approximate methods for investigating nonlinear automatic control systems. Fizmatgiz, Moscow, 1960.

Метод анализа предельного цикла релейной системы автоматического регулирования

М. А. СУЛТАН
(Каир)

Рассматривается исследование устойчивости периодического решения предельного цикла релейной системы автоматического регулирования с обратной связью при синусоидальной и случайной помехах. Предлагается упрощенное приближенное решение данной задачи. Метод основан на применении метода линеаризации, разработанного и опубликованного автором ранее в [1].

M. A. Sultan
International Atomic Energy Agency
A-1011 Vienna, P.O.B. 645
Kärtner Ring 11
Austria

Problems of Control and Information Theory, Vol. 5 (2), pp. 181–188 (1976)

NEW CLASS OF THE CONTROLLED SHIFT-REGISTERS. I

Z. GRODZKI

(Lublin)

(Received March 19, 1975)

A new class $C_{k,m,p}$ of the controlled shift-registers will be introduced. It is a very general class and it covers all cases of the registers considered in the literature before.

The fundamental properties of the sets of sequences generated by the registers of the class $C_{k,m,p}$ will be given.

Three aspects of the complexity problem for such registers will be investigated too.

1. Introduction

A linear k-shift-register (briefly k-register) is a physical object which generates only infinite cyclic sequences with 2^k as the maximal length of the cycle (see Golomb's monograph [3]).*

Many technical applications of the linear k-registers have been discussed in literature (see [1, 2, 3, 7]). For lack of space we shall list only some fields: automatic regulation, criptology, radar, simulation of the deterministic or nondeterministic processes.

The more general class is Boolean k-register class. But this class is very weakly elaborated from the mathematical point of view. There are only a few papers on this subject. A partial bibliography with a short abstract of the results can be found in Yoeli's paper [8].

The aim of this paper is to introduce a very general class $C_{k,m,p}$ of the controlled shift-registers.

Every register of the class $C_{k,m,p}$ consists of the memory with k cells, where symbols of a finite alphabet may be put in and also of m transformations $\varphi_1, \ldots, \varphi_m$ of k arguments every. The selection of the transformation is done by a register control ψ in every moment $np + 1$ for $n = 0, 1, 2, \ldots$ Let us assume that only one transformation operates at the moment.*

To every initial state located in the memory the control ψ assigns unique transformation φ_i. During p successive moments the controlled shift-register

* A technical realization as well as the fundamental properties of the sets of all sequences generated by the k-registers have been given in [4].

* A case $p = 1$ has been considered in [5].

"works" as the k-register (with a single transformation φ_i) and it generates a sequence $t_1, \ldots, t_{p+k}$. Then to the sequence $t_{p+1}, \ldots, t_{p+k}$ (in the moment $p+1$) the control ψ assigns a new transformation φ_j which generates a sequence $t_{p+1}, \ldots, t_{2p+k}$ and so on.

When $M = \{0, 1\}$ and the transformations $\varphi_1, \ldots, \varphi_m$ are linear (Boolean) then such registers are called linear (Boolean) controlled shift-registers. The technical construction of such registers is rather simple.

The formal definition as well as the fundamental properties of the sets of sequences generated by the registers of the class $C_{k,m,p}$ will be given. The great part of this paper is devoted to the complexity problem of such registers. Three aspects of this problem will be investigated here.

2. Preliminaries

The set of all nonnegative integers (briefly numbers) is denoted by N and the empty set by $\emptyset$. The numbers are denoted by lower case Latin letters i, j, k, m, n, q, r, s.

Let M be a finite set consisting at least of two elements and let M^* (M^∞) denote the set of all finite (infinite) sequences, over M. Let us assume thatthe empty sequence Λ belongs to M^*.

The nonempty subsets of M^∞ (or of M^*) are denoted by upper case Latin letters E, F, G, H (perhaps with subscripts). The cardinality of the set E is denoted by card (E); in particular χ_0, the cardinality of N.

The elements of M are denoted by lower case Latin letters t, u, v, w, x, y, z with subscripts and nonempty sequences of M^∞ (or of M^*) by upper case Latin letters T, U, V, W, X, Y, Z (possibly with subscripts).

$1(T)$ denotes the number of elements (length) in the sequence T, if T is finite, or ∞— otherwise.

A restricted sequence $t_i, \ldots, t_j$ for every i, j, $1 \leq i \leq j \leq 1(T)$ is denoted by $T|_{i,j}$ · $T|_i$ denotes a sequence $t_i, \ldots, t_m$, if $1(T) = m$ or an infinite sequence $t_i, t_{i+1}, \ldots$, if T is infinite.

The concatenation of the sequences T and U $(1(T) < \infty)$ is denoted by TU.

The functions are denoted by lower case Greek letters μ, π, φ, ψ (with or without subscripts).

The collection of the functions $\langle \varphi_1, \ldots, \varphi_m \rangle$ is denoted by upper case Greek letter Φ.

* We shall distinguish physical objects (the controlled (k, m, p)-shift-registers) from formal notions for these objects (the controlled (k, m, p)-machines). We shall use only formal notions.

The symbols $\neg$, $\vee$, $\&$, $\Rightarrow$, $\Leftrightarrow$ denote the usual logical connectives and the symbols $\exists$, $\forall$ are the abbreviations for "there exists" and "for every" respectively.

For brevity we shall write $\exists_{p \leq q} R(\ldots p \ldots)$ and $\forall_{p \leq q} R(\ldots p \ldots)$ instead of $\exists_p[1 \leq p \leq q \,\&\, R(\ldots p \ldots)]$ and $\forall_p[1 \leq p \leq q \Rightarrow R(\ldots p \ldots)]$.

3. Basic definitions

In this section we shall give several basic definitions which are necessary to understand the following discussion.

For the further consideration let k, m, p be the arbitrary (fixed) numbers and let M be a finite set such that car $(M) = n$ for $n > 1$.

Definition 1. By a controlled (k, m, p)-machine $A_{k,m,p}$ (briefly (k, m, p)-machine) we mean an $(m+2)$tuple $\langle M, \varphi_1, \ldots, \varphi_m, \psi\rangle$ where every φ_i for $1 \leq i \leq m$ is a total function of M^k into M and ψ is total function of $M^k \times H$ into $\Phi = \{\varphi_1, \ldots, \varphi_m\}$ where $H = \{i \in N : i = jp + 1; j = 0, 1, 2, \ldots\}$ and M^k denotes the kth Cartesian product of M.

Every φ_i, $1 \leq i \leq m$ will be called a transition function and ψ — a control of $A_{k,m,p}$.

We shall assume that $\varphi_i \neq \varphi_j$ for $1 \leq i < j \leq m$.

Definition 2. Let $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi\rangle$ be an arbitrary controlled (k, m, p)-machine and let $\Phi = \{\varphi_1, \ldots, \varphi_m\}$.

An infinite sequence $T = t_1, t_2, \ldots \in M^\infty$ is said to be a computation of $A_{k,m,p}$ iff the following condition is satisfied:

$$\forall_{n \geq 0} \forall (t_{np+1}, \ldots, t_{(n+1)p+k}) \in M^* \forall_{\varphi i} \in \Phi [\psi(t_{np+1}, \ldots, t_{np+k}, np+1) =$$
$$= \varphi_i \Rightarrow \forall_q (np+1) \leq q \leq (n+1)p \Rightarrow t_{k+q} = \varphi_i(t_q, \ldots, t_{k+q-1}))].$$

The set of all computations of $A_{k,m,p}$ will be denoted by $C_{A_{k,m,p}}$.

Definition 3. A sequence $T \in M^\infty$ is said to be a (k, m, p)-computation iff there is a (k, m, p)-machine $A_{k,m,p}$ such that T is its computation.

Definition 4. A set $E \subseteq M^\infty$ is said to be a (k, m, p)-computation set iff there is a (k, m, p)-machine $A_{k,m,p}$ such that $E = C_{A_{k,m,p}}$.

Definition 5. Let k, q be two numbers such that $q > k \geq 1$ and let $E \subseteq M^*$ be a set consisting of only sequences of the length q.

E is said to be k-homogeneous iff the following condition is satisfied:

$$\forall_{T \in E} \forall_{U \in E} \forall_{i \leq l(T)-k} \forall_{j \leq l(U)-k} T|_{i, i+k-1} =$$
$$= U|_{j, j+k-1} \Rightarrow T|_{i+k, i+k} = U|_{j+k, j+k}].$$

Definition 6. Let $E \subseteq M^{\infty}$ be a nonempty set and let k, p be two numbers.

By a (k, p)-division of E (Notation: $\mathrm{div}_{k,p}(E)$) (if such exists) we mean a set $F = \{F_1, F_2, \ldots\}$ (finite or infinite) such that:

(1) every set $F_i \in F$ consists of only sequences of the length $k + p$ and it is a k-homogenous set;

(2) for every $T \in E$ and $i \geq 0$ there is $F_j \in F$ such that $T|_{ip+1.(i+1)p+k} \in F_j$;

(3) $F_i \cap F_j = 0$ for $i \neq j$.

Definition 7. Two controlled machines $A_{i,j,p}$ and $A_{k,m,p}$ are said to be equivalent (Notation: $A_{i,j,p} \equiv A_{k,m,p}$) iff $C_{A_{i,j,p}} = C_{A_{k,m,p}}$.

Definition 8. Let $E \subseteq M^{\infty}$ be a set of all computations of all equivalent (k, m, p)-machines (k, p — the fixed numbers and m — an arbitrary one).

A triple (k, q, p) is said to be a degree of complexity of E (Notation: $\deg(E) = (k, q, p)$) iff q is the smallest number of all numbers m for which there is a (k, m, p)-machine $A_{k,m,p}$ such that $E = C_{A_{k,m,p}}$.

Definition 9. Let $E \subseteq M^{\infty}$ be a set of all computations of all equivalent (k, m, p)-machines (k, m — the fixed numbers and p — an arbitrary one).

A triple (k, m, q) is said to be a measure of complexity of E (Notation: $\mathrm{Meas}(E) = (k, m, q)$ iff q is the largest number of all numbers p for which there is a (k, m, p)-machine $A_{k,m,p}$ such that $E = C_{A_{k,m,p}}$.

The last two definitions have been introduced because from the economical point of view for solving the synthesis problem it is convenient to use only such (k, m, p)-machines which consist of a few transition functions, every of which "works" in a long time interval.

The number of cells of the (k, m, p)-machine $A_{k,m,p}$ such that $E = C_{A_{k,m,p}}$ is determined by the cardinality of E.

4. The fundamental properties of the (k, m, p)-computation sets

A necessary and sufficient condition for a set $E \subseteq M^{\infty}$ to be a (k, m, p)-computation one will be proved.

Theorem 1. A nonempty set $E \subseteq M^{\infty}$ is a (k, m, p)-computation set iff the following conditions are satisfied:

(1) $\forall_{T \in M^k} \exists_{U \in E} (T = U|_{1,k})$;

(2) $\forall_{T \in E} \forall_{i \geq 0} (T|_{ip+1} \in E)$;

(3) there is a set $F = \mathrm{div}_{k,p}(E)$ such that $\mathrm{card}(F) = m$.

Proof. Necessity. Let us consider a (k, m, p)-machine $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$ such that $E = C_{A_{k,m,p}}$.

Since every transition functions φ_i of $A_{k,m,p}$ is a total function of M^k into M then condition (1) holds. Because $E = C_{A_{k,m,p}}$ then condition (2) holds. As every transition function φ_i of $A_{k,m,p}$ "works" during p successive moments then there is a set $F = \mathrm{div}_{k,p}(E)$. From the assumption that there exists m transition functions of $A_{k,m,p}$ we obtain card $(F) = m$.

Sufficiency. Let us consider a set $E \subseteq M^\infty$ for which conditions (1)–(3) of Theorem 1 are satisfied. Let be $F = \{F_1, \ldots, F_m\} = \mathrm{div}_{k,p}(E)$.

For every i, $1 \leq i \leq m$ let us define a total function $\varphi_i : M^k \to M$ as follows:

$\varphi_i(u_1, \ldots, u_k) = u_{k+1}$ iff there is $T \in F_i$ and a number $j \leq 1(T)$ such that $u_1, \ldots, u_{k+1} = T\,|_{j,j+k}$, otherwise we put $\varphi_i(u_1, \ldots, u_k) = 1$. Since every F_i, $1 \leq i \leq m$ is a k-homogeneous set then φ_i is a function.

Let us define now a function $\psi : M^k \times H \to \{\varphi_1, \ldots, \varphi_m\}$, where $H = \{i \in N : i = jp + 1; j = 0, 1, 2 \ldots\}$ as follows: $\psi(T, np + 1) = \varphi_q$ for every $n \geq 0$ and some $q \leq m$ iff there is $U \in F_q$ such that $U\,|_{1,k} = T$.

Let us denote $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$. It follows from the construction that $E \subseteq C_{A_{k,m,p}}$ and from conditions (1)–(2) that $E = C_{A_{k,m,p}}$. Q. E. D.

Corollary 1. The cardinality of an arbitrary (k, m, p)-computation set E is equal to n^k, where card $(M) = n$ for $n > 1$.

Corollary 2. Let $E \subseteq M^\infty$ be an arbitrary set of all computations of all equivalent (k, m, p)-machines (k, p — the fixed numbers and m — an arbitrary one).

For every $q \geq 1$, deg $(E) = (k, q, p)$ iff q is the smallest number of all numbers m for which there is a set $F = \mathrm{div}_{k,p}(E)$ such that card $(F) = m$.

Corollary 3. Let $E \subseteq M^\infty$ be a set of all computations of all equivalent (k, m, p)-machines (k, m — the fixed numbers and p — an arbitrary one).

For every $q \geq 1$, Meas $(E) = (k, m, q)$ iff q is the largest number of all numbers p such that for the set E the conditions (2) and (3) of Theorem 1 are satisfied.

Corollary 4. Let $M = \{0, 1\}$ and let $E \subseteq M^\infty$ be an arbitrary set. Let us denote by $\bar{E}$ a set obtained from E by replacing 0 into 1 and conversely in all its sequences.

For all numbers k, m, p, $\bar{E}$ is a (k, m, p)-computation set iff E is a (k, m, p)-computation set.

Theorem 2. Let $E, F \subseteq M^{\infty}$ be the (i, j, n)- and (k, m, p)-computation sets. Then we have:

(1) there are numbers q, r, s such that $E \cup F$ is a (q, r, s)-computation subset;

(2) if $E \cap F$ and $E - F$ are the nonempty sets then they are (i, j, n)-computation subsets.

Proof. It follows from Theorem 1 and Corollary 2 that card $(E \cup F) \leq$ $\leq n^i + n^k$ and there is a set $F = \mathrm{div}_{q,s}(E \cup F)$ such that card $(F) = r$ for some numbers q, r, s. Then $E \cup F$ is a (q, r, s)-computation subset.

Proof of (2) is obvious. Q. E. D.

5. The complexity rank of the (k, m, p)-computation sets

In this section a case when M is two element alphabet $\{0, 1\}$ will be considered. All Boolean functions which will be used as the transition functions of the (k, m, p)-machines are in the disjunctive normal form. The number of the occurrence of the disjunction symbol $\vee$ in the function φ_i is called a range of φ_i (Rg (φ_i)) and $\max_{1 \leq i \leq m}$ Rg (φ_i) — a range of $\langle \varphi_1, \ldots, \varphi_m \rangle$. With respect to this notion a complexity rank of the (k, m, p)-computation set will be introduced. Then a necessary and sufficient condition that a (k, m, p)-computation set E is of the complexity rank n will be proved.

At the beginning we shall define several notions. Let $P = \{x_1, x_2, \ldots\}$ be an infinite set of variables and let $Q = \{\vee, \wedge, \neg, f_1, \ldots, f_m, g, h\}$ be finite set of functions.

Let us define inductively the classes $A^k_{\vee}$ and $A^k_{\wedge}$ of formulas.

The classes $A^1_{\vee} = A^1_{\wedge}$ consist of only formulas $x_1, x_2, \ldots, \neg x_1, \neg x_2, \ldots$.

The classes $A^k_{\vee}$ and $A^k_{\wedge}$ for $k > 1$ are defined as follows:

(1) if $y_1 \in A^1_{\vee}, \ldots, y_k \in A^1_{\vee}$ then $y_1 \wedge \ldots \wedge y_k \in A^k_{\wedge}$ $(y_1 \vee \ldots \vee y_k \in A^k_{\vee})$;

(2) if $f, g \in A^k_{\vee}(A^k_{\wedge})$ then $f \vee g \in A^k_{\vee}$ $(f \wedge g \in A^k_{\wedge})$;

(3) nothing else does not belong to the classes $A^k_{\vee}$ and $A^k_{\wedge}$ unless it follows from (1)–(2).

The elements of the class $A^k_{\vee}(A^k_{\wedge})$ are called disjunctive (conjunctive) normal formulas.

We shall consider only disjunctive normal formulas which we shall call briefly formulas.

Let us define inductively the range of the formula h (Rg (h)) as follows:

(1) if $h \in A^1_{\wedge}$ then Rg $(h) = 0$;

(2) if $h = f \vee g$ then Rg (f) = Rg (f) + Rg (g) + 1.

Let us consider an m-tuple $\langle f_1, \ldots, f_m \rangle$ where $f_i \in A_\vee^k$ for $1 \leq i \leq m$. A number j is said to be a range of this m-tuple (Notation: Rg $\langle f_1, \ldots, f_m \rangle$) iff $j = \max_{1 \leq p \leq m} \text{Rg}\,(f_p)$.

Let us assume that the variables $x_1, \ldots, x_n$, ranging over the $\{0, 1\}$. Let us assign to every formula of $A_\vee^k$ the identity or the negation functions and to the formulas of $A_\vee^k$ $(k > 1)$ — the disjunction function of k arguments.

These functions will be denoted (as previously) by μ, π, φ, ψ or sometimes by $I, \neg, \vee$.

By the range of the Boolean function φ (Rg (φ)) we mean the range of the respective formula, analogically for the collection $\langle \varphi_1, \ldots, \varphi_m \rangle$ of functions.

Let $E \subseteq M^\infty$ be an arbitrary (k, m, p)-computation set. A number $q \geq 1$ is said to be a complexity rank of E (Notation: rk $(E) = q$) iff the following condition is satisfied:

there exists a (k, m, p)-machine $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$ such that $E = C_{A_{k,m,p}}$, Rg $\langle \varphi_1, \ldots, \varphi_m \rangle = q$ and the number q is minimal as above.

Let $E \subseteq M^\infty$ be an arbitrary (k, m, p)-computation set for $k \geq 1$, $m > 1$, $p \geq 1$ and let $F = \{F_1, \ldots, F_m\} = \text{div}_{k,p}(E)$.

Let us assign to every set F_i, $1 \leq i \leq m$ a set $G_i = \{t_1, \ldots, t_{k+1} : t_{k+1} = 1 \;\&\; \exists_{U \in E}\, \exists_{j \leq 1(U)-k} (t_1, \ldots, t_{k+1} = U\,|_{j,j+k})\}$.

Theorem 3. For every $n \geq 0$, rk $(E) = n$ iff the following condition is satisfied:

(1) $\max_{1 \leq i \leq m} (\text{card}\,(G_i)) = n + 1$.

Proof of necessity is obvious and will be omitted.

Sufficiency. Let us consider a (k, m, p)-computation set $E \subseteq M^\infty$ and a set $\{G_1, \ldots, G_m\}$ defined as above. Let us assume that condition (1) holds.

We have to construct a (k, m, p)-machine $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$ such that $E = C_{A_{k,m,p}}$ and Rg $\langle \varphi_1, \ldots, \varphi_m \rangle = n$.

Let us consider an arbitrary set $G_i = \{U_1, \ldots, U_q\}$ for $1 \leq i \leq m$. Let us define an auxiliary function π_j, $1 \leq j \leq q$ as follows: if $U_j = u_1, \ldots, u_{k+1}$ then $\pi_j(u_1, \ldots, u_k) = v_1 \wedge \ldots \wedge v_k$ where $v_i = u_i$ when $u_i = 1$ or $v_i = \neg u_i$ when $u_i = 0$. It is obvious that $\pi_j(u_1, \ldots, u_k) = 1 = u_{k+1}$. Let be $\varphi_i = \pi_1 \vee \ldots \ldots \vee \pi_q$.

Now let us define a total function $\psi : M^k \times H \to \{\varphi_1, \ldots, \varphi_m\}$ where $H = \{i \in N : i = jp + 1;\ j = 0, 1, 2, \ldots\}$ as follows: $\psi(T, j) = \varphi_i$ iff there is $U \in E$ such that $U\,|_{j,j+k+p} \in F_i$ and $T = U\,|_{j,j+k-1}$. Let be $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$. Assumption (1) implies that rk $(E) = n$. Q. E. D.

References

1. *Birkhoff, G., Bartee, T.*, Modern applied algebra, McGraw-Hill Book Company (1970), 367—397.
2. *Geffe, P. R.*, How to protect data with ciphers that are really to break, Electronics **46** (1973) 99—111.
3. *Golomb, S. W.*, Shift-register sequences, San Francisco Holden Day (1967), 24—89.
4. *Grodzki, Z.*, The theory of shift-registers, Information and Control **21** (1972) 196—205.
5. *Grodzki, Z.*, The controlled shift-registers, Elektronische Informationsverarbeitung und Kybernetik **3** (1975), 143—150.
6. *Peterson, W. W.*, Error-correcting codes, MIT Press, Massachusetts Institute of Technology, 1961.
7. *Stone, H. S.*, Discrete ;mathematical structures and their applications, Science Research Associates I. N. C. Chicago Palo Alto, Toronto, Henley — on Thames, Sydney (1973) 296—361.
8. *Yoeli, M.*, Counting with nonlinear feedback shift-registers, IEEE Transactions on Electronic Computers (1963), 357—361.

Новый класс управляющих регистров сдвига

З. ГРОДЗКИ

(Люблин)

В настоящей статье вводится новый класс $C_{k,m,p}$ управляемых регистров сдвига. Регистры этого класса охватывают все случаи регистров, рассматриваемых в литературе.

Приводятся основные свойства множеств последовательностей, генерируемых этими регистрами.

Рассматриваются три аспекта сложности этих регистров.

Z. Grodzki
Higher Engineering School
Pl-20–950 Lublin
Dabrowskiego 13
Poland

THE 4TH INERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE

The Fourth International Conference on Artificial Intelligence held in Tbilissi (USSR) between September 3 and 8, 1975 was organized by the Academy of Sciences of the USSR, the Georgian Republican Academy of Sciences and the Joint International Council on Artificial Intelligence.

The conference attended by more than 100 researchers from 35 countries had the following sessions: mathematical and theoretical aspects of artificial intelligence; representation of problems and solutions; problem planning and solution; learning, debugging and automatic programming; artificial intelligence software; search methods; understanding of natural language; understanding of speech; processing of visual information; robots and methods for productivity enhancement; and phsychological aspects of artificial intelligence.

The most interesting papers of the session "Learning, debugging and automatic programming" (by D. E. Show, U. R. Swartaut and K. K. Green "Examples of program synthesis in LISP", and by D. L. Waterman "Some rules for automatic program synthesis", USA) were devoted to automatic synthesis of programs, i.e. to generation of computer programs proceeding from a given original problem formulation. The automatic program synthesis system combines ability to program with a great amount of knowledge in the related field. The major feature of such programs is the necessity to introduce into the computer means related to the programming technique itself which formalize the basic stages of this process. It is also important to select such a representation of the "world of problems" for which programs are to be synthesized that would enable fairly effective synthesis. In the papers above the "world of problems" is rather poor, and they illustrate possible ways to automatic program synthesis rather than solution of this problem.

Interesting papers by V. M. Bryabrin, V. A. Serebryakov and V. M. Yuffa (USSR) "LORD: a LISP-oriented solver and data base" and Sandwell (Sweden) "Some considerations about data base control in LISP" were presented at the session on "Artificial intelligence software".

The interest to LISP is not occasional. Its list structure enables effective data representation, and its set of functions provides convenient realization of procedures characteristic of artificial intelligence problems.

The paper by Bryabrin et al. describes a programming system for artificial intelligence problems providing a solver based on a language of LISP type. Interaction with the user is ensured by a metaprocessor constructing syntactic analyser for a particular version of the input language on the basis of a given grammar.

The paper by A. P. Ershov, I. A. Mel'chuk and A. S. Narin'yani, USSR (the session "Understanding of natural language") describes and experimental system RITA (picture-information-text-author) intended for texts in a natural language and simple geometric compositions. The general philosophy and architecture of the system are briefly discussed, and the major targets of the project are considered.

Development of large program systems more and more requires a practical solution to the problem of man-computer dialogue in a natural language.

One of the major goals of the RITA project is the study and experimental verification of linguistic and program components of the dialogue system in a natural language. Development of the system should also confirm or correct the assumption that the functional parts of the linguistic processor may be separated as a standard subprocessor for a rather wide class of machine systems.

Researchers from the University of Toronto (Canada) J. Milopulos, A. Bordgida, P. Kohen, N. Russopulos, J. Zozos and H. Wang developed and realized a prototype of data-handling system using information about graduating students as data base. Their system TORUS understands a natural language and is an interface between an unprepared user and the data base control system. The system uses a semantic network for understanding each input sentence and determining response. This network gives general information on certain field (students and the learning process in the University of Toronto, in particular) and particular information extracted from dialogues with the user. Procedures are provided for updating the network by new sentences and for separation of that part of the network which contains information required for the answer.

The paper by W. Scragg (Switzerland) describes a program answering questions about processes. It is assumed that answers may be generated through analysis of the data involved into a procedure of action execution (simulation). The same simulation procedure of sandwich preparation may be presented as a list of actions explaining sandwich making in terms of the anticipated result and side-effects such as knife becoming dirty or cheese con-

sumption. Answers may be obtained via procedures differing only in minor details. An example is presented illustrating operation of the system.

Analysis of the theoretical foundations of the artificial intelligence as a field of scientific research is of principal importance. One of the major approaches to the analysis is comparison of human intellect and machine operation, identification of likeliness and difference between them. Such an analysis is impossible without recourse to philosophy and psychology which are traditionally involved into the study of the human intellect. These problems were raised by O. K. Tikhomirov, Dr. of Psychology in his paper "Philosophical and psychological problems of artificial intelligence" discussing relations between human intellect and computer functioning. Differences are demonstrated between heuristic search in the artificial intelligence and functioning of the human intellect. The "Machine centrism" of direct comparison of human and machine functions is criticized. Possible ways to improvement of the artificial intelligence are suggested, and it is demonstrated that the aim of approaching the human intellect is posed without (or only with limited) regard to the psychological data. Psychological problems involved into the development and use of the artificial intelligence are formulated.

In the session on "Mathematical and theoretical aspects of artificial intelligence" papers by T. Gergely, G. Andreka, I. Némety (Hungary) "Definition theory as the basis for construction of a problem solver having creativity" and D. A. Pospelov (USSR) "Semiotic models in the problems of artificial intelligence" should be mentioned. The latter paper states that at the today's level of studies in the field of artificial intelligence a theory of semiotic systems is required which should play now the role of formal systems used there at early stages. Basic goals of the theory of semiotic models and systems are discussed.

Among papers presented at the session "Representation of problems and solutions" the study by V. L. Stefanyuk (USSR) should be mentioned devoted to representation of "hard-nut-to-crack" problems. The author employs local heuristics enabling classification of situations for the problem of covering a set of squares by dominoes. Representations in the form of classes are shown to be deeply related to the problem itself because they give the necessary and sufficient conditions of perfect covering. Solution of a problem often depends on the selection of an appropriate covering. A brilliant example presented in the paper is the "hard-nut-to-crack" problem requiring to demonstrate that an 8×8 field of squares where two diagonally opposite squares are eliminated cannot be covered by dominoes.

Among papers presented at the session "Search methods", those by C. Morgan ("Automated generation of hypotheses by means of an enlarged

inductive resolvence"), L. A. Rasstrigin, R. Kh. Erenstein ("Collective of algorithms"), and V. M. Ozernoi, M. G. Gaft ("A method for selection of the most preferable solutions in multi-objective problems") should be mentioned. The decision theory is one of the most rapidly developing areas in the field of artificial intelligence. Although not identical, decision making and problem solution are closely related. Problem solution implies selection of means for attaining a goal which is visible, but directly unaccessible. Decision making is generally understood as identification of all possible ways leading to the goal, comparison of their advantages and selection of the most preferable one. The most important element of decision making is the decision maker who has incomplete information about environment, possible results and values of various results. Complicated practical problems and interest raised by them to the involved scientific difficulties have stimulated intensive development of the decision making theory.

The paper by D. E. Okhotsimskii and A. K. Platonov "Preceptive robot in the three-dimensional space" presented at the session devoted to "Robots and methods for productivity enhancement" caused manifest interest. It has, for the first time, solved the problem of the six-legged robot moving in an environment close to real. At the same session F. Hees proposed a representation enabling robot to work out a strategy in terms of its initial goal. The paper by V. S. Gurfinkel et al. "Some aspects of sensibilization of robots and manipulators" shows that principles underlying human sensor system are applicable to the development of sensibilized manipulators and robot lower extremities.

The field of artificial intelligence application becomes wider and wider. All these programs are characterized by attempts to formalize the human experience gained over a long period. This formalization involves both means of representation of initial data and "environment" model structure, and selection of some heuristic procedures for data handling.

N. I. Rodionova

BOOK REVIEW

E. JESSEN: Architecture of digital computers. Springer-Verlag, Berlin–Heidelberg–New York, 1975, 246 pages, 97 figures (in German)

This book is one of the volumes of the series "Informatik" published by the Springer Publishing House. It deals with the general principles of digital computers and digital computer systems. It tries to give quantitative expressions which can serve as a quide for engineers designing and using computers and computer systems. The reading and application of this book assume the knowledge of electronics, design principles of logic circuits and programming techniques in machine code.

After a short summary of the essential notations and definitions used in the field of the computers, the general structure of data processing systems is treated. Readers get acquainted with the von Neumann's principle, the different types of computers and the different modes of operation such as local batch processing, multiprogramming, time-shared modes etc. Various kinds of formats, codes are here also discussed. This part of the book is completed with the discussion of the hierarchy.

Next the volume deals with the central processors. Here a detailed discussion is given on the machine instructions, the instruction set, the adressing techniques, the program interrupts, the binary arithmetic and the basic parts of the central processor, that is, control unit, arithmetic unit. This part of the book is completed with the treatise of microprogramming, emulation, reliability and maintenance.

The next part is devoted to the main memory and the auxiliary storages. Definitions of the most important parameters, such as access time, capacity etc. are also given here.

The remaining part deals with the I/O (input/output) devices, the I/O units, channels and processors, then, with the computer centres, computer networks and some types of unorthodox machines.

The last chapter summarizes the brief history of the computers and predicts the future on this field.

Author's and Subject Index complete the book.

This book is recommended to engineers and students who want to get acquainted with the basic principles of computers and computer systems.

T. Gál

C. STRIEBEL: Optimal control of discrete time stochastic systems. Springer–Verlag, Berlin–Heidelberg–New York, 1975, 20 references, 208 pages

The theory of discrete time stochastic control systems has received impetus from two directions. The area, which is followed in the book concerning terminology and notation, is first that of engineering applications; and second the area of sequential statistical decision theory.

This monograph is intended as a self-contained, but not a comprehensive treatment of discrete time stochastic control systems. Relationships with other works are pointed out (primarily in the introductory sections), but all essential results are proved directly without appeal to the literature of stochastic control theory. The principal aim of the author is to develop workable algorithms for the construction of optimal control laws. Hence the emphasis is on sufficient conditions for optimality.

Chapter 1 contains the Introduction, formulation of the model and makes comparison with other works.

Chapter 2 is devoted to the distribution theory required to deal with incompleteness of information, furthermore a general estimation theory is considered.

The properties of the sufficient statistics defined for the linear Gaussian model are treated in Chapter 3. The definition of a sufficient statistic given here is slightly different from others. The emphasis in Chapter 3 is on methods for determining sufficient statistics for a particular loss structure.

Chapters 2 and 3 are devoted to distribution theory which is motivated by, but is theoretically independent of, the control problems. The theory of Chapters 2 and 3 depends heavily on appendix A.1 which is a collection of standard results on stochastic kernels not conveniently available elsewhere.

Perhaps the most innovative results are in Chapter 4 where a new definition for the minimal conditional loss functional is proposed.

In Chapter 5 a selection theorem is exploited to obtain constructive existence theorems. Most of the Chapter 5 is devoted to the problem that in the presence of a sufficient statistic the question of existence of an optimal Markov law and the existence of an optimal law in the statistics are essentially equivalent.

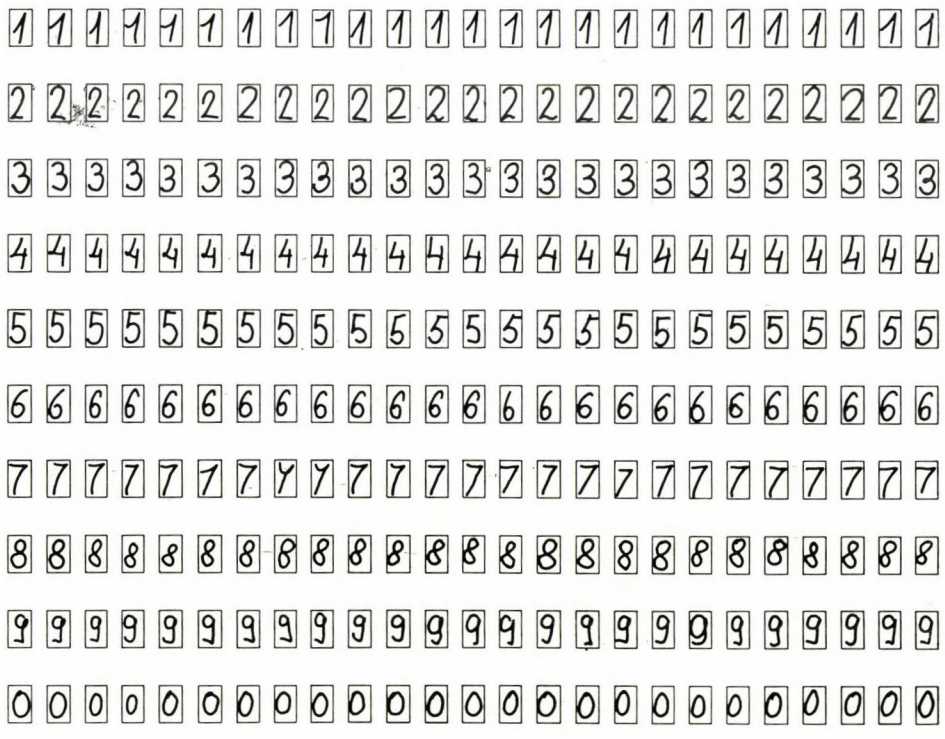
Chapters 6 and 7 provide examples of the application of the optimal control theory of Chapter 5 to the linear Gaussian model.

L. Keviczky

CORRIGENDA

Correction to the paper "Structural Character Recognition by Forming Projections" by P. Brauer, M. Vajta, Jr. (Problems of Control and Information Theory, vol. 4, No. 4, pp. 339–352).

A sorryful misunderstanding happened during the edition in the Press. Figure 4 (p. 347) is wrong because it shows 25 printed numerals instead of 25 handwritten ones. There is an essential difference between them from the point of view of character recognition. The correct figure is as follows:



Printed in Hungary
AKADÉMIAI KIADÓ, BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

PROBLEMS OF CONTROL AND INFORMATION THEORY VOL. 5(2)
ENGLISH TRANSLATION OF THE COMMUNICATION IN RUSSIAN

ON STOCHASTIC APPROXIMATION PROCESSES BY SUMS OF INDEPENDENT RANDOM VARIABLES

M. B. NEVELSON

(Moscow)

Estimation of the zero of an unknown function is considered done by means of measurements depending, in a certain way, on the values of this function and on random errors about which there is some a priori information. Under broad assumptions, estimates of the type of stochastic approximation are constructed that are asymptotically normal to the minimal, in the conventional sense, limit variance. It is shown that these estimates may be approximated in probability or with probability 1 by sums of independent random variables. Hence some properties of finite- and infinite-dimensional distributions of these estimates are deduced.

1. Introduction

Let $f(x)$ be an unknown function defined over straight line E such that equation $f(x) = g$, $g = \text{const.}$ has a unique solution $x = x_0$, and also $(f(x) - g) - (x - x_0) > 0$ at $x \neq x_0$. Consider estimation of x_0 for a case where at any moment $t = 1, 2, \ldots$ the statistician can carry out "independent" measurements of $f(x)$ value in an arbitrary point $x = x(t)$ with some additive random error, so that the measurement result $Y_0(t, x)$ has the form of $Y_0(t, x) = f(x) + \xi(t, x)$. Here $\xi(t, x)$, $t = 1, 2, \ldots, x \in E$ is the totality of random variables and

$$M\xi(t, x) \equiv 0 \text{ under all } t, x. \tag{1.1}$$

It is known that in such a situation x_0 may be estimated by the Robbins–Monro stochastic approximation procedure

$$X(t+1) - X(t) = -\frac{a^{-1}}{1+t}\left[Y_0\big(t+1, X(t)\big) - g\right], \quad X(0) = x, \quad a > 0 \tag{1.2}$$

which converges to x_0 almost surely under certain assumptions about $f(x)$, $\xi(t, x)$ if $t \to \infty$. Moreover, this procedure is asymptotically normal under conventional normalization. In this case, the limiting normal law will have optimal properties (in the class of procedures of type (1.2)) if $a = f'(x_0) > 0$.

Moreover, procedures are known asymptotically equivalent to (1.2) with $a = f'(x_0)$. For the first time, a procedure of this kind based on the idea

of truncation was constructed by Venter [1] under an additional assumption that $0 < r_1 < f'(x_0) < r_2$ where r_1, r_2 are values known by the statistician. Later, Fabian established [2] that this assumption might be omitted if one considers a more complicated time-dependent truncation (see also [3, Ch. 6; 4]).

One can, however, conceive a situation where, first, random errors are not necessarily additive, and, second, the statistician has more general a priori information than (1.1).

In this connection, we shall assume below that at every time t and in any point $x \in E$, values

$$Y_0(t, x) = F(f(x), \xi(t, x)) \tag{1.3}$$

may be measured. Here function $y = F(x, z)$ is defined for $x \in \mathfrak{D}_1$, $z \in E$ where $\mathfrak{D}_1$ is some Borel set containing the set of function $f(x)$ values.

Denote by $\mathfrak{D}_2$ a Borel set such that $P\{\xi(t, x) \in \mathfrak{D}_2$ for all $t = 1, 2, \ldots, \ldots, x \in E\} = 1$. It will be assumed throughout this paper, that there is the following a priori information about function $F(x, z)$ and random variables $\xi(t, x)$:

(a) $F(g, z)$ is a known monotone increasing function;

(b) the set of values taken by function $F(x, z)$ over $x \in \mathfrak{D}_1$, $z \in \mathfrak{D}_2$ is contained in the set of values taken by function $F(g, z)$ at $z \in E$; and

(c) for some nondecreasing function $S(z)$, the expectation $S_0(t, x) = MS(\xi(t, x))$ is known.

In addition, if function $S(z)$ is not bounded, it will be assumed that for some $\delta > 0$ the following holds:

$$M \mid S(\xi(t, x)) \mid^{1+\delta} \leq k(t)\,(1 + r(x))\,, \tag{1.4}$$

where $k(t)$ is a numerical sequence, and function $r(x)$ is known. Usually, papers dealing with the asymptotical normality conditions for Robbins–Monro procedures (or their modifications) where $F(x, z) = x + z$, $S(z) \equiv z$, $S_0(t, x) \equiv 0$, assume that condition (1.4) is met at $\delta = 1$ and $r(x) = x^2$.

We are going to show that, having the above a priori information and under some very general conditions, one also can construct by observations (1.3) simple recurrent estimates $X(t)$ such that values $V(t) = \sqrt{t}\,(X(t) - x_0)$, first, are asymptotically normal to the optimal (in the same sense, as in [1–4]) limit variance, and, second, may be approximated for large t by sums of independent random variables. Hence some properties of finite- and infinite-dimensional distributions of process $V(t)$ will be deduced. For the first time results of this kind were obtained in [5] for continuous procedures of stochastic approximation (see also [3, 6]).

It should be noted that the procedures under consideration do not require any a priori information about the behaviour of function $f(x)$ under $x \to \infty$. Modifications of the Robbins–Monro procedures constructed in [7–9] have a similar property.

This work has been carried out in continuance of the Cooperation agreement between the Academy of Sciences of the USSR and the Hungarian Academy of Sciences.

2. Basic notions and assumptions

All random variables are assumed to be defined over some random space $(\Omega, \mathfrak{A}, P)$. If not otherwise stated, relations between random variables, including convergence, will be understood with probability 1.

Let $S_0(z_1, z_2)$, $-\infty \leq z_1 \leq z_2 \leq \infty$ be a set of nondecreasing functions $S(z)$ such that:

(a) if $z_1 \neq z_2$, function $S(z)$ strictly increases over interval (z_1, z_2); and

(b) if $z_1 \neq -\infty$ then $S(z_1 + 0) > S(z)$ under $z_1 > z$, if $z_2 \neq \infty$ then $S(z_2 - 0) < S(z)$ under $z_2 < z$.

It is evident that $S_0(-\infty, \infty)$ is a class of functions strictly increasing over E, and $S_0(z_1\, z_1)$ is a class of nondecreasing functions discontinuous in point $z_1(-\infty < z_1 < \infty)$.

Denote also by $Q(S)$ a set of measureable (in the sense of Borel) functions $q(S)$ such that

(a) $q(S)$ is positive, bounded and for any $q_1 q_2$

$$\inf_{q_1 < z < q_2} q(z) > 0$$

(b) function $S(z)\ q(z)$ is bounded and nondecreasing.

Introduce the following notation. If $a(t)$, $b(t)$ are two sequences and also $a(t) < b(t)$, let us denote

$$[x]_{a(t)}^{b(t)} = \begin{cases} x & \text{if} \quad a(t) \leq x \leq b(t) \\ a(t) & \text{if} \quad x < a(t) \\ b(t) & \text{if} \quad x > b(t). \end{cases}$$

Denote by $\mathfrak{B}$ a σ-algebra of Borel sets of straight line E, and by $\mathfrak{B}_1(\mathfrak{B}_2)$ a σ-algebra of Borel sets $\mathfrak{D}_1(\mathfrak{D}_2)$. Let, in addition, $\mathfrak{A}_t$ be a σ-algebra generated by a family of random variables $\{\xi(t, x), x \in E\}$, and $\mathfrak{F}_t$ be a σ-algebra generated by family $\{\xi(u, x), u \leq t, x \in E\}$.

If not otherwise stated, it is always assumed below that $\xi(t, x)$ is $\mathfrak{A}_t \times \mathfrak{B}$-measurable, and random variable family $\{\xi(t, x), x \in E\}$ is independent of $\mathfrak{F}_{t-1}$.

Moreover, the following assumptions of general nature will be repeatedly used.

Assumption 2.1. Function $f(x)$ is measurable (in the sense of Borel), and function $F(x, z)$ is $\mathfrak{B}_1 \times \mathfrak{B}$-measurable, and also

(a) $(f(x) - g)(x - x_0) \geq 0, x \in E$

$$(F(x, z) - F(g, z))(x - g) \geq 0, \quad x \in \mathfrak{D}_1, z \in \mathfrak{D}_2,$$

(b) for each $\varepsilon \in (0, 1)$ the following relations hold

$$\inf_{\varepsilon < |x - x_0| < 1/\varepsilon} |f(x) - g| > 0, \quad \sup_{|x| < 1/\varepsilon} |f(x)| < \infty, \quad \sup_{x \in \mathfrak{D}_1, z \in \mathfrak{D}_2} |F(x, z)| < \infty .$$

Denote by $z = \varphi(y)$ a function inverse to function $y = F(g, z)$ and let

$$Z(t, x) = \varphi(Y_0(t, x)), \quad Y(t, x) = S(Z(t, x)) - S_0(t, x) .$$

Assumption 2.2. Function $R(t, x) = MY(t, x)$ is finite under all $t = 1, 2, \ldots$ and x from ν-vicinity of point x (where ν is some positive number), and has under these values of t and x derivative $\dfrac{\partial R}{\partial x}(t, x) = R'(t, x)$ continuous in point $t = \infty$, $x = x_0$.

Assumption 2.3. Assumption 2.2 is true, function $R'(t, x)$ satisfies Hölder's condition with index $\mu_1 > 0$ in ν-vicinity of point x_0 uniformly with respect to t, and also $[R'(t, x_0) - \lambda] t^{\mu_2} \to 0$ under $t \to \infty$ where $\lambda = \lim_{t \to \infty} R'(t, x_0)$, and μ_1 is some positive number.

Assumtpion 2.4. The following inequality holds

$$\sup_{t \geq 0,\ |x - x_0| < \nu_1} M \Phi^2(t, x) < \infty \tag{2.1}$$

where $\nu_1 > 0$, $\Phi(t, x) = Y(t, x) - R(t, x)$ and function $A(t, x) = M\Phi^2(t, x)$ is continuous in point $t = \infty$, $x = x_0$.

Assumption 2.5. Function $\Phi(t, x)$ is continuous in mean-square in point x_0 uniformly with respect to t.

Assumption 2.6. Function $\Phi(t, x)$ satisfies Hölder's condition in mean-square with index $\mu > 0$ in ν-vicinity of point x_0 uniformly with respect to t.

Assumption 2.7. Family of random variables $\Phi(t, x)$, $t = 1, 2, \ldots$ $\ldots | x - x_0 | < \nu_1$, is uniformly integrable.

Assumption 2.8. For some $\varkappa > 0$

$$\sup_{t \geq 0} M |S(\xi(t, x))|^{2+\varkappa} < \infty .$$

3. Recurrent procedures

As usually, estimates of

$$\lambda = \lim_{t \to \infty} R'(t, x_0)$$

will be required. To construct them, introduce, as in [1, 2], the following. Let $X(t)$ be an arbitrary $\mathfrak{F}_t$-measurable stochastic process, and t_i, $i = 1, 2, \ldots$, be a sequence of positive numbers satisfying the following conditions:

$$t_{j+1} > t_j, \quad \lim_{i \leftarrow \infty} t_i = \infty, \quad j = 1, 2, \ldots \tag{3.1}$$

Let

$$W(t) = W(t; t_j, X(t_j - 1), c(j), \ j = 1, \ldots, q(t)) =$$

$$= \frac{1}{q(t)} \sum_{j=1}^{q(t)} Y_{0j}(t_j, X(t_j - 1)),$$

where

$$Y_{0j}(t, x) = \frac{Y(t, x + c(j)) - Y(t, x - c(j))}{2c(j)}, \tag{3.2}$$

$q(t)$ is the number of those t_j which do not exceed t, and $c(j)$ is some sequence of positive numbers.

It is easy to see that values of process $W(t)$ at time t may be computed through the following recurrent formulae:

$$\begin{cases} W(t) = \begin{cases} 0 & \text{under} \quad 0 \leq t < t_i \\ W(t_i) & \text{under} \quad t_i \leq t < t_{i+1}, \quad i = 1, 2, \ldots \end{cases} \\ W(t_{i+1}) - W(t_i) = \dfrac{1}{1+i}\left[- W(t_i) + Y_{0i+1}(t_{i+1}, X(t_{i+1} - 1))\right], \ t_0 = 0. \end{cases} \tag{3.3}$$

By analogy with [2], we will truncate below process $W(t)$ with some sequences of positive numbers.

Consider first procedure

$$\begin{cases} X(t+1) - X(t) = - \dfrac{Y(t+1, X(t))}{(1+t)\,\widetilde{W}(t)}, \ X(0) = x \\ \widetilde{W}(t) = [W(t)]_{a(t)}^{b(t)}. \end{cases} \tag{3.4}$$

Here $W(t) = W(t; t_j, X(t_j - 1), c(j), j = 1, \ldots, q(t))$ is a process defined by formulae (3.2) or (3.3); $a(t)$, $b(t)$, $c(t)$ are some sequences of positive numbers, x is an arbitrary point in the straight line; and t_j satisfies relations (3.1).

Theorems below show that under broad assumptions one can select for bounded function $S(z)$ sequences $a(t)$, $b(t)$, $c(t)$ so that values of $\sqrt{t}\,(X(t) - x_0)$

are asymptotically normal and approach, as t increases, sums of independent random variables. In order to be sure of validity of similar properties of procedure (3.4) also in the case when function $S(z)$, is not bounded the statistician ought to have additional a priori information. In particular, he should know that

$$M|S(\varphi(F(f(x),\ \xi(t,\ x)))\ -\ S(\xi(t,\ x))|\ \leq\ K(1\ +\ |x|)\,, \tag{3.5}$$

$$M\left[S(\varphi(F(f(x),\ \xi(t,\ x)))\ -\ MS(\varphi(F(f(x),\ \xi(t,\ x)))\right]^2\ \leq\ K(1\ +\ x^2)\,, \tag{3.6}$$

where K is some constant. Since function $f(x)$ is unknown, it might be difficult to verify conditions (3.5) and (3.6). Therefore, in the general case, we shall discuss another procedure. To this end, construct first process $X_0(t)$ almost surely converging to x_0 under $t \to \infty$, and then truncate with its aid, like in [10], process (3.4).

Namely, assume (see [11, 12]) that

$$X_0(t) = \begin{cases} \tilde{x},\ 0 \leq t < t_1 \\ X_0(t_i),\ \ t \leq t < t_{i+1} \end{cases} \tag{3.7}$$

$$X_0(t_{i+1}) - X_0(t_i) =$$

$$= -\ \frac{a}{(1\ +\ i)\ r(X_0(t_i))}\left[A_{i+1}(Z(i\ +\ 1,\ X_0(t_i))) - S_0(i\ +\ 1,\ X_0(t_i))\right], \tag{3.8}$$

where $\tilde{x}$ is an arbitrary point in the straight line; $a > 0$; and

$$A_t(z) = \frac{S(z)\,q(z)}{q(z)\ +\ d(t)}\ ,\ \ q(z) \in Q(S),\ \ a > 0\,.\ \ a > 0\,.$$

Let now

$$\begin{cases} X(t\ +\ 1) = \left[X(t)\ -\ \dfrac{1}{(1\ +\ t)\ \widetilde{W}(t)}\ Y(t\ +\ 1,\ X(t))\right]_{X_0(t)-\alpha}^{X_0(t)+\alpha}\ ,\ \ X(0) = x \\ \widetilde{W}(t) = [W(t)]_{b(t)}^{b(t)} \end{cases} \tag{3.9}$$

As above, here $W(t) = W(t;\ t_j,\ X(t_{j-1}),\ c(j),\ j = 1,\ \ldots,\ q(t))$, $0 < a(t) < b(t)$, $c(t) > 0$, $d > 0$.

It turns out that procedure (3.9) enables omission of condition (3.5) and use of condition $MS^2(\xi(t,\ x)) \leq K(1\ +\ x^2)$ (or even of a more general condition which does not involve unknown function $f(x)$, instead of condition (3.6),

4. Asymptotic properties of distributions

In studying properties of distributions of procedures (3.4) and (3.9), it will be assumed that the following conditions (A) are satisfied (about these conditions see [11]).

Take

$$U_\varepsilon(a) = \{x : \varepsilon < |\, x - a\,| < 1/\varepsilon\},\ U_\varepsilon^+(a) = \{x : \varepsilon < x - a < 1/\varepsilon\},$$

$$U_\varepsilon^-(a) = \{x : -1/\varepsilon < x - a < -\varepsilon\},$$

$$\bar{f}_\varepsilon^{(1)} = \sup_{\varepsilon < x - x_0 < 1/\varepsilon} [f(x) - g], \qquad f_\varepsilon^{(1)} = \inf_{\varepsilon < x - x_0 < 1/\varepsilon} [f(x) - g],$$

$$\bar{f}_\varepsilon^{(2)} = \sup_{-1/\varepsilon < x - x_0 < -\varepsilon} [g - f(x)], \quad f_\varepsilon^{(2)} = \inf_{-1/\varepsilon < x - x_0 < -\varepsilon} [g - f(x)]$$

Condition (A). For any $\varepsilon \in (0, 1)$ there are sets $C_\varepsilon \in \mathfrak{B}_2$, $\mathfrak{D}_\varepsilon \in \mathfrak{B}_2$ such that

$$\bar{\delta}_\varepsilon = \inf_{z \in C_\varepsilon} [S(\alpha_\varepsilon + z) - S(z)] > 0\,,$$

$$\delta_\varepsilon = \inf_{z \in \mathfrak{D}_\varepsilon} [S(z) - S(z - \beta_\varepsilon)] > 0\,. \tag{A.1}$$

where

$$\alpha_\varepsilon = \inf_{x \in \{x : \bar{f}_\varepsilon^{(2)} > x - g > \underline{f}_\varepsilon^{(1)}\} \cap \mathfrak{D}_1,\, z \in C_\varepsilon} \big[\varphi\big(F(x, z)\big) - z\big],$$

$$\beta_\varepsilon = \inf_{x \in \{x : -\bar{f}_\varepsilon^{(1)} < x - g < -\underline{f}_\varepsilon^{(2)}\} \cap \mathfrak{D}_1,\, z \in \mathfrak{D}_\varepsilon} \big[z - \varphi\big(F(x, z)\big)\big] \tag{A.2}$$

$$\sum_{t=0}^{\infty} \frac{1}{(1 + t)\, b(t)} \inf_{\varepsilon < x - x_0 < 1/\varepsilon} P\{\xi(t, x) \in C_\varepsilon\} = \infty\,.$$

It is easy to verify that conditions (A) are satisfied if one of the following conditions is valid:

(a) $S(z) \in S_0(-\infty, \infty)$ and for any $\varepsilon \in (0, 1)$ there is bounded set $E_\varepsilon \in \mathfrak{B}_2$ sucht that

$$\inf_{x \in U_\varepsilon(g) \cap \mathfrak{D}_1,\, z \in E_\varepsilon} |\, F(x, z) - F(g, z)\,| > 0\,, \sum_{t=0}^{\infty} \frac{1}{(1 + t)\, b(t)} \inf_{x \in U_\varepsilon(x_0)} P\{\xi(t, x) \in E_\varepsilon\} = \infty$$

or

(b) $S(z) \in S_0(z_1, z_2)$, $-\infty < z_1 \leq z_2 < \infty$ and for any $\varepsilon \in (0, 1)$ and $\nu > 0$:

$$\inf_{x \in U_\varepsilon^+(g) \cap \mathfrak{D}_1,\, z \in [z_1 - \nu,\, z_2] \cap \mathfrak{D}_2} [F(x, z) - F(g, z) > 0\,,$$

$$\inf_{x \in U_\varepsilon^-(g) \cap \mathfrak{D}_1,\, z \in [z_1,\, z_2 + \nu] \cap \mathfrak{D}_2} [F(g, z) - F(x, z)] > 0.$$

$$\sum_{t=0}^{\infty} \frac{1}{(1 + t)\, b(t)} \inf_{x \in U_\varepsilon^+(x_0)} P\{z_1 - \nu < \xi(t, x) < z_2\} = \infty$$

$$\sum_{t=0}^{\infty} \frac{1}{(1+t)\,b(t)} \inf_{x \in U_{\varepsilon}^{-}(x_0)} P\{z_1 < \xi(t,x) < z_2 + \nu\} = \infty .$$

The following statements are valid (see for proofs the Appendix).

Theorem 4.1. Let:

1. Function $S(z)$ be bounded.
2. Assumptions 2.1, 2.2 and Condition (A) be satisfied.
3. Sequences $a(t)$, $b(t)$, $c(t)$ be chosen so that

$$a(t) \to 0,\ b(t) \to \infty,\ c(t) \to 0 \tag{4.1}$$

$$\sum_{t=0}^{\infty} \frac{1}{(1+t)^2} \left[\frac{1}{a^2(t)} + \frac{1}{c^2(t)}\right] < \infty . \tag{4.2}$$

4. Assumption 2.4 be satisfied.

Then for process (3.4) relation

$$\sqrt{t}\ (X(t) \sim x_0) \sim N(0, \sigma^2/\lambda^2)$$

holds where

$$\sigma^2 = \lim_{t \to \infty,\, x \to x_0} A(t,x), \quad \lambda = \lim_{t \to \infty} R'(t, x_0) .$$

Theorem 4.2. Let Conditions 1, 2 and 3 of Theorem 4.1 and Assumption 2.5 be satisfied. Then the following statement is valid for process (3.4):

$$\sqrt{t}\ (X(t) - x_0) - Z_0(t) \to 0$$

in probability under $t \to \infty$ where

$$Z_0(t) = -\frac{1}{\lambda \sqrt{t}} \sum_{k=1}^{\infty} \Phi(k, x_0) .$$

Theorem 4.3. Let Conditions 1, 2 and 3 of Theorem 4.1, Assumptions 2.3 and 2.6 be satisfied, and, moreover,

$$c(t) \leq Ct^{-c},\ q(t) \geq Qt^{q},\ C > 0,\ Q > 0,\ q > 0 . \tag{4.3}$$

Then for any $m < \min(\mu/2,\ 1/2,\ qc\mu_1,\ q\mu_2)$, the following relation* holds with probability 1 for process (3.9)

$$\sup_{t \geq 0} \{t^m [\sqrt{t}\,(X(t) - x_0) - Z_0(t)]\} < \infty .$$

Function $S(z)$ is not assumed to be bounded in the theorems below. However, according to Section 1, we take that Condition (1.4) is met.

* A similar statement for maximal likelihood estimates is established in [6].

Theorem 4.4. Let:

1. Inequality (1.4) hold where $\delta > 0$ and function $r(x)$ is locally bounded,
2. Sequences $a(t), b(t), c(t), d(t), k(t)$ satisfy relations (4.1), (4.2) and inequality

$$\sum_{t=0}^{\infty} \left[\frac{d^{\delta}(t)\, b(t)}{1+t} + \frac{1}{(1+t)^2\, d^2(t)} \right] < \infty .$$

3. Assumptions 2.1, 2.2, Conditions (A) and $\alpha < \nu_1/2$ be satisfied, and
4. Assumptions 2.4 and 2.7 be satisfied. Then Theorem 4.1 holds for process (3.9).

Theorem 4.5. Let Conditions 1, 2 and 3 of Theorem 4.4 and Assumptions 2.5, 2.8 be satisfied. Then Theorem 4.2 holds for process (3.9).

Theorem 4.6. Let Conditions 1, 2 and 3 of Theorem 4.4, and Assumptions 2.3, 2.6, 2.8 be satisfied; and let sequences $a(t), b(t), c(t)$ be chosen according to (4.3). Then Theorem 4.3 holds for process (3.9).

Remark 4.1. The total number of observations $n(t)$ used in constructing process $X(t)$ described by (3.4) or (3.9) is, respectively, $t + 2q(t)$ or $t + 3q(t)$. If sequence t_j is chosen so that $\lim_{j\to\infty} t_j/j = \infty$, then $\lim_{t\to\infty} q(t)/t = 0$. Therefore, under such a choice of t_j in conditions of Theorems 4.1 and 4.2, the following will hold:

$$\sqrt{n(t)}\,(X(t) - x_0) \sim N(0,\ \sigma^2/\lambda^2) .$$

Denote by $V(t)$ a stationary Gaussian process satisfying Ito's stochastic differential equation

$$dV(t) = -\,1/2\, V(t)\, dt + \frac{\sigma}{\lambda}\, d\eta(t)$$

where $\eta(t)$ is a standard Wiener process. Further, let $V_t(s) = \sqrt{te^s}\,(X(te^s) - x_0)$ where t, te^s are positive integers, and $\widetilde{V}_t(s)$, $s \geq 0$ is a random broken line coinciding with process $V_t(s)$ in points $s_i = \ln(1 + i/t)$, $i = 0, 1, 2, \ldots$

It follows from Theorems 4.2, 4.3, 4.5 and 4.6 (see also Theorem 1 in [13, Ch. 9, Section 4] and Theorem 6.3 [2, Ch. 6] and Lemma [6])

Corollary 4.1. If conditions of Theorems 4.2 or 4.5 are satisfied, then finite-dimensional distributions of process $\widetilde{V}_t(s)$, $s \geq 0$ converge under $t \to \infty$ to finite-dimensional distributions of process $V(s)$.*

* For a case where $F(x, z) = x + z$ this statement was established in [12] under somewhat more strict conditions.

Corollary 4.2. If conditions of Theorems 4.3 or 4.6 are satisfied, distribution generated in $C[0, L]$, $L > 0$ by process $\widetilde{V}_t(s)$, $s \geq 0$ weakly converges under $t \to \infty$ to a distribution generated in $C[0, L]$ by process $V(s)$.

Appendix

Lemma A.1. Let Assumption 2.2 be satisfied, inequality (2.1) hold, and $X(t_j - 1) \to x_0$ almost sure under $j \to \infty$. Then $W(t) \to \lambda$ almost certainly under $t \to \infty$ for any sequence $c(j) > 0$ satisfying conditions

$$\sum_{j=1}^{\infty} \frac{1}{j^2 c^2(j)} < \infty, \quad \lim_{j \to \infty} c(j) = 0.$$

If, instead of 2.2, Assumption 2.3 is satisfied, and if $\sup_{t \geq 1} [t^c c(t)] < \infty$, $\inf_{t \geq 1} [q(t)/t^q] > 0$ where $c > 0$, $q > 0$, then almost certainly $[W(t) - \lambda] t^m \to 0$ under $t \to \infty$ for any $m < q \min(c\mu_1, \mu_2)$.

Proof of this lemma may follow that of corresponding lemmas from [1; Ch. 6 in 3].

Proof of Theorems 4.1, 4.2 and 4.3

1. Let Conditions 1, 2 and 3 of Theorem 4.1 be satisfied, and process $X(t)$ be defined by formulae (3.4).

Simple calculations result in

$$M\{[X(t+1) - x_0]^2 / X(1), \ldots, X(t)\} \leq (X(t) - x_0)^2 - \\ + \frac{2R(t+1, X(t))(X(t) - x_0)}{(1+t)\, b(t)} + \frac{c_i}{(1+t)^2 a^2(t)}. \qquad \text{(A. 1)}$$

(Here and below, c_i are some positive constants). Moreover

$$R(t, x) = M\left[S(\varphi(F(f(x), \xi(t, x)))) - S(\varphi(F(g, \xi(t, x))))\right].$$

$R(t, x)(x - x_0) \geq 0$ because function $S(\varphi(y))$ does not decrease and $f(x)(x - x_0) \geq 0$, $[F(x, z) - F(g, z)](x - g) \geq 0$. From this and (A.1) follow, as usually, both existence of almost certainly finite limit of sequence $|X(t) - x_0|$ und $t \to \infty$, and inequality

$$\sum_{t=0}^{\infty} \frac{R(t+1, X(t))}{(1+t)\, b(t)} < \infty. \qquad \text{(A. 2)}$$

From conditions (A) an inequality

$$\sum_{t=0}^{\infty} \frac{r_\varepsilon(t+1)}{(1+t)\, b(t)} = \infty \qquad \text{(A. 3)}$$

follows where

$$r_\varepsilon(t) = \inf_{x \in U_\varepsilon(x_0)} |R(t, x)|$$

and ε is an arbitrary number from interval (0, 1).

From (A.2), (A.3) and the fact that $\lim_{t \to \infty} | x(t) - x_0 | < \infty$ obtain

$$\lim_{t \to \infty} X(t) = x_0 . \tag{A. 4}$$

It follows from Lemma A.1 and (A.4) that

$$\lim_{t \to \infty} \widetilde{W}(t) = \lambda . \tag{A. 5}$$

According to Assumption 2.2 and (A.4),

$$R(t + 1, X(t)) = \gamma(t) [X(t) - x_0]$$

where

$$\lim_{t \to \infty} \gamma(t) = \lambda . \tag{A. 6}$$

Consequently,

$$X(t + 1) - x_0 = \left[1 - \frac{\tilde{\gamma}(t)}{1 + t}\right] \big(X(t) - x_0\big) - \frac{\Phi\big(t + 1, X(t)\big)}{(1 + t)\, \widetilde{W}(t)} .$$

Here $\tilde{\gamma}(t) = \gamma(t) \widetilde{W}^{-1}(t)$, and by virtue of (A.5) and (A.6)

$$\lim_{t \to \infty} \tilde{\gamma}(t) = 1 . \tag{A. 7}$$

Take

$$\hat{\gamma}(t) = \begin{cases} \tilde{\gamma}(t) & \text{if} \quad 2 < \tilde{\gamma}(t) < 1/2 \\ 1 & \text{if} \quad \tilde{\gamma}(t) \leq 1/2 \text{ or } \tilde{\gamma}(t) \geq 2 \end{cases} \tag{A. 8}$$

$$\widehat{W}(t) = \begin{cases} \widetilde{W}(t) & \text{if} \quad \widetilde{W}(t) > \lambda/2 \\ 1 & \text{if} \quad \widetilde{W}(t) \leq \lambda/2 \end{cases} \tag{A. 9}$$

and consider new process $\hat{X}(t) = \hat{X}^{s,\zeta}(t)$ defined by formulae

$$\begin{aligned} &\hat{X}(t + 1) - x_0 = \left[1 - \frac{\hat{\gamma}(t)}{1 + t}\right] \big(\hat{X}(t) - x_0\big) - \frac{\Phi\big(t + 1, \hat{X}(t)\big)}{(1 + t)\, \widehat{W}(t)} \\ &X(s) = \zeta \end{aligned} \tag{A. 10}$$

where $s \geq 0$, and ζ is an $\mathfrak{F}_s$-measurable random variable with finite variance.

It may be shown in the usual fashion that

$$M | \hat{X}(t) - x_0 | = 0\left(\frac{1}{\sqrt{t}}\right), \quad t \to \infty \tag{A. 11}$$

and

$$\lim_{t\to\infty} |X(t) - x_0|\, t^{1/2-\varepsilon} = 0 \quad \text{for any} \quad \varepsilon > 0\,. \tag{A. 12}$$

By taking $V(t) = \sqrt{t}\,(\widehat{X}(t) - x_0)$ obtain from (A.10)

$$V(t+1) = A_{s-1\,t}\sqrt{S}\,(\zeta - x_0) + \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}}\,\psi(k)\,\big(\widehat{X}(t) - x_0\big) -$$

$$- \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}}\,\Phi\big(k+1, \widehat{X}(k)\big)\,\widehat{W}^{-1}(k)\,. \tag{A. 13}$$

Here

$$A_{kt} = \begin{cases} \prod\limits_{m=k+1}^{t} \left(1 - \dfrac{1}{2(m+1)}\right), & 1 \le k < t \\ 1 & k = t \end{cases}$$

$$\psi(t) = \gamma(t) - 1.$$

It is evident that

$$\lim_{t\to\infty} \psi(t) = 0\,, \quad \sup_{t\le 0} |\psi(t)| < C_2\,. \tag{A. 14}$$

It is easy to verify that

$$A_{k,\,t} = (1 + \Theta_{kt})\sqrt{\frac{k+1}{t+1}}\,, \tag{A. 15}$$

where

$$\sup_{t\ge k} |\Theta_{kt}| = O\left(\frac{1}{k}\right), \quad k \to \infty\,. \tag{A. 16}$$

The first summand in the right side of (A.13) almost surely tends to zero under $t \to \infty$. It follows from (A.11) and (A.14) that the second summand also tends to zero in probability as t grows. Consequently,

$$V(t+1) + \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}}\,\frac{\Phi\big(k+1, \widehat{X}(k)\big)}{\widehat{W}(k)} \to 0 \tag{A. 17}$$

in probability under $t \to \infty$.

2. Let conditions of Section 1 and Assumption 2.4 be met. Then, allowing for relations $\widehat{X}(t) \to x_0$ and $\widehat{W}(t) \to \lambda$ under $t \to \infty$, obtain

$$M\left\{\left[\frac{\Phi\big(t+1, \widehat{X}(k)\big)}{\widehat{W}(k)}\right]^2 \Big/ \widehat{X}(s), \ldots, \widehat{X}(k)\right\} = \frac{A\big(k+1, \widehat{X}(k)\big)}{\widehat{W}^2(k)} \to \frac{\sigma^2}{\lambda^2}$$

under $k \to \infty$. It follows from this and (A.17) (see [14]) that

$$V(t) \sim N(0,\, \sigma^2/\lambda^2). \tag{A.18}$$

From the definition of process $\widehat{X}(t)$, (A.5) and (A.7) obtain

$$P\{X(t) \equiv \widehat{X}^{T, X(T)}(t) \text{ under } t \geq T\} > 1 - \delta \tag{A.19}$$

for any $\delta > 0$ and some $T = T(\delta)$. Together with (A.18) this proves Theorem 4.1.

3. Let conditions of Section 1 and Assumption 2.5 be met. Take

$$\psi(k, x) = M[\Phi(k+1, x) - \Phi(k+1, x_0)]^2.$$

By using the fact $\sup_{k>1,\, x} |\Phi(k, x)| < \infty$ find that $\psi(k, x(k)) \to 0$ for any sequence $x(k)$ such that $x(k) \to x_0$ under $k \to \infty$. Consequently,

$$\tilde{\psi}(k) = M[\Phi(k+1, \widehat{X}(k)) - \Phi(k+1, x_0)]^2 = M\tilde{\psi}(k, \widehat{X}(k)) \to 0. \tag{A.20}$$

Consider process

$$\zeta(t) = \sum_{k=5}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{[\Phi(k+1, \widehat{X}(k)) - \Phi(k+1, x_0)]}{\widehat{W}(k)}.$$

From (A.20) obtain $M\zeta^2(t) \to \infty$ as $t \to \infty$ and, hence, $\zeta(t) \to 0$ in probability. Therefore, by virtue of (A.17),

$$V(t+1) + \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{\Phi(k+1, x_0)}{\widehat{W}(k)} \to 0 \tag{A. 21}$$

in probability under $t \to \infty$. In a similar way, using (A.5) inequality $\widehat{W}^{-1}(k) \leq$ $\leq \max(1, \lambda/2)$ and independence of random variables $\Phi(k+1, x_0)$, $\widehat{W}(k)$ establish that

$$\sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \left[\frac{1}{\widehat{W}(k)} - \frac{1}{\lambda}\right] \Phi(k+1, x_0) \to 0 \tag{A. 22}$$

in probability under $t \to \infty$. From (A.21) and (A.22), (A.16) obtain the statement of Theorem 4.2.

4. Let conditions of Section 3, Assumptions 2.3, 2.6 and relations (4.3), (4.4) be satisfied.

It follows from (A.12) and (A.19) that $(X(t) - x_0) t^{1/2-\varepsilon} \to 0$ under $t \to \infty$ for any $\varepsilon < 0$. This and Assumption 2.3 lead to $R(t+1, X(t)) = \gamma(t)(X(t) - x_0)$ where

$$\lim_{t \to \infty} [\gamma(t) - \lambda] t^{m_1} = 0 \tag{A. 23}$$

or any $m_1 < \left(\frac{1}{2}\mu_1, \mu_2\right)$.

Moreover, in virtue of Lemma 3.1,

$$\lim_{t\to\infty} [\hat{W}(t) - \lambda]\, t^{m_2} = 0 \tag{A. 24}$$

under $m_2 < \min(qc\mu_1, q\mu_2)$.

The following relation results from (A.23) and (A.24)

$$\lim_{t\to\infty} [\tilde{\gamma}(t) - 1]\, t^{m_3} = 0$$

for any $m_3 < \min(qc\mu_2, q\mu_2)$, and, consequently,

$$\lim_{t\to\infty} \psi(t)\, t^{m_3} = 0 \tag{A. 25}$$

for the same values of m_3.

Using (A.12), (A.15) and (A.25), obtain

$$t^{m_4} \left| \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \psi(k) \left(\hat{x}(k) - x_0\right) \right| \leq \frac{k_1 t^{m_4}}{\sqrt{t+1}} \sum_{k=s}^{t} k^{-1/2-m_3+\varepsilon} \to 0 \tag{A. 26}$$

with probability 1 if $m_4 < m_3$ and $\varepsilon < m_3 - m_4$. (Here k_i is some constant, may be random.)

Further, from the proof of Lemma 4.2 [3, Ch. 6] the following holds with probability 1:

$$t^{m_5} \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{\left[\Phi\left(k+1, X(k)\right) - \Phi(k+1, x_0)\right]}{\hat{W}(k)} \to 0, \quad m_5 < \mu/2 \tag{A. 27}$$

$$t^{m_6} \sum_{k=s}^{t} \frac{A_{kt}}{\sqrt{k+1}} \left[\frac{1}{\hat{W}(k)} - \frac{1}{\lambda}\right] \to 0, \quad m_6 < \min(qc\mu_1, q\mu_2)\,. \tag{A. 28}$$

From (A.13), (A.16), (A.26), (A.27) and (A.28) obtain the statement of the theorem.

Proofs of Theorems 4.4, 4.5 and 4.6

Let Assumptions 1, 2 and 3 of Theorem 4.4 be satisfied. By repeating, with evident changes, the reasoning used in the proof of Theorem 2.3 of [11], obtain that $X_0(t) \to 0$ under $t \to \infty$. Hence, $X_0(t) = x_0 + \psi(t)$ where $\psi(t) \to 0$ almost surely.

Let α_0 be an arbitrary positive number, and $T_1 = T_1(\alpha_0)$ be such that

$$P\{|\psi(t)| < \alpha \text{ under all } t \geq T_1\} > 1 - \alpha_0/3\,. \tag{A.29}$$

Consider a new process, $X_1(t)$ defined by formulae

$$\begin{cases} X_1(t+1) = \left[X_1(t) - \dfrac{Y\left(t+1, X_1(t)\right)}{(1+t)\,\hat{W}(t)}\right]_{x_0+[\psi(t)]_{-\alpha}^{\alpha}}^{x_0+[\psi(t)]_{-\alpha}^{\alpha}+\alpha} \\ X_1(T_1) = X(T_1)\,. \end{cases} \tag{A. 30}$$

Then

$$M\{[X_1(t+1)-x_0]^2/X_1(T_1),\ldots,X_1(t)\}\leq$$
$$\leq[X_1(t)-x_0]^2-\frac{2R\big(t+1,X_1(t)\big)\big(X_1(t)-x_0\big)}{(1+t)\,b(t)}+$$
$$+\frac{M\{\Phi^2\big(t+1,X_1(t)\big)/X_1(T_1),\ldots,X_1(t)\}}{(1+t)^2\,a^2(t)}. \tag{A. 31}$$

Together with Assumption 2.4, inequality $|X_1(t)-x_0|\leq 2\alpha<\nu$ under $t\geq T_1$ gives

$$M\{\Phi^2(t+1,X_1(t))/X_1(T_1),\ldots,X_1(t)\}=A(t+1,X_1(t))<c_4,\ t\geq T_1.$$

In a like manner as in the proof of Theorem 4.1, obtain from this and (A.31) the following equality:

$$\lim_{t\to\infty}X_1(t)=x_0\,. \tag{A. 32}$$

Since, in virtue of (A.29),

$$P\{X(t)\equiv X_1(t)\ \text{under all}\ t\geq T_1\}>1-\frac{\alpha_0}{3} \tag{A. 33}$$

it follows from (A.32) and the arbitrariness of α_0 that

$$\lim_{t\to\infty}X(t)=x_0\,. \tag{A. 34}$$

From (A.30) and (A.34), it follows that process

$$X_1(t)-\frac{Y\big(t+1,X_1(t)\big)}{(1+t)\,\widetilde{W}(t)}$$

leaves with probability 1 segment $[x_0+[\psi]_{\alpha}^{-\alpha}+\alpha,\ x_0+[\psi]_{-\alpha}^{\alpha}-\alpha]$ only finite number of times. Hence, for some $T_2=T_2(\alpha_0)$

$$P\{X_1(t)\equiv X_2(t)\ \text{for all}\ t\geq T_2\}>1-\frac{\alpha_0}{3} \tag{A. 35}$$

where process $X_2(t)$ is defined under $t\geq T$ by formulae

$$X_2(t+1)-X_2(t)=-\frac{\widehat{Y}\big((t+1,X_2(t)\big)}{(1+t)\,\widetilde{W}(t)},$$

$$X_2(T_2)=X_1(T_2),\quad \widehat{Y}(t,x)=\widehat{R}(t,x)+\widehat{\Phi}(t,x),$$

$$\widehat{R}(t,x)=\begin{cases}R(t,x) & \text{under}\quad |x|<2\alpha\\ \operatorname{sign}x & \text{under}\quad |x|\geq 2\alpha\end{cases},$$

$$\widehat{\Phi}(t, x) = \begin{cases} \Phi(t, x) & \text{under} \quad |x| < 2\alpha \\ \Phi(t, x_0) & \text{under} \quad |x| \geq 3\alpha \end{cases}.$$

It is easy to verify that $X_2(t) \to x_0$. Therefore

$$X_2(t+1) - x_0 = \left(1 - \frac{\hat{\gamma}(t)}{1+t}\right)\left(X_2(t) - x_0\right) - \frac{\widehat{\Phi}\left(t+1, X_2(t)\right)}{(1+t)\,\widehat{W}(t)}.$$

and also $\tilde{\gamma}(t) \to 1$ under $t \to \infty$.

Finally, consider under $t \geq T_0$ process $X_3(t)$ defined as follows:

$$X_3(t+1) - x_0 = \left(1 - \frac{\tilde{\gamma}(t)}{1+t}\right)\left(X_2(t) - x_0\right) - \frac{\widehat{\Phi}\left(t+1, X_3(t)\right)}{(1+t)\,\widetilde{W}(t)}.$$

Constant T_3 is chosen here with respect to condition

$$P\{X_3(t) \equiv X_2(t) \text{ under } t \geq T_3\} > 1 - \alpha_0/3 \tag{A.36}$$

and variables $\hat{\gamma}(t)$ and $\widehat{W}(t)$ are defined by formulae (A.8) and (A.9).

It may be readily shown, as in the proof of Theorem 4.1, that

$$\sqrt{t}\left(X_3(t) - x_0\right) + \sum_{k=T_3}^{t} \frac{A_{kt}}{\sqrt{k+1}} \frac{\widehat{\Phi}\left(k+1, X_3(k)\right)}{\widehat{W}(k)} \to 0 \tag{A.37}$$

in probability under $t \to \infty$.

Further, by using formula (5.37) like above, establish validity of the statements of Theorems 4.4, 4.5 and 4.6 for process $X_3(t)$. The required statements follow from (A.33), (A.35), (A.36) and the arbitrariness of α_0.

*

My thanks are due to R. Z. Khasminskii for his useful criticism.

References

1. *Venter, J. N.*, An extension of the Robbins-Monro procedure. Ann. Math. Statist. 1967, **38**, 1, pp. 181—190.
2. *Fabian, V.*, On asymptotic normality in stochastic approximation. Ann. Math. Statist. 1968, **39**, 4, pp. 1327—1332.
3. *Nevelson, M. B., Khasminskii, R. Z.*, Stochastic approximation and recurrent estimation, Nauka Publ., Moscow, 1972 (in Russian).
4. *Nevelson, M. B., Khasminskii, R. Z.*, An adaptive Robbins—Monro procedure. Avtomatika and Telemechanica, 1973, **10**, pp. 71—83.
5. *Khasminskii, R. Z.*, On behaviour of the stochastic approximation process under large values of time. Problemi peredachi informatsii, 1972, **8**, 1, pp. 81—91 (in Russian).
6. *Ibragimov, I. A., Khasminskii, R. Z.*, On approximation of statistical estimates by sums of independent random variables. Dokladi AN SSSR, 1973, **210**, 6, pp. 883—887 (in Russian).

7. *Robbins, H., Siegmund, D.*, A convergence theorem for non-negative almost supermartingales and some applications optimizing methods in statistics. Academic Press, N. Y., London, 1971.
8. *Komlós, J., Révész, P.*, A modification of the Robbins-Monro process. Stud. Sci. Math. Hung., 1973, **8**, 3, pp. 329—340.
9. *Fabian, V.*, Asymptotically efficient stochastic approximation, the case RM. Ann. Math. Statist., 1973, **1**, 3, pp. 486—495.
10. *Khasminskii, R. Z.*, Sequential estimation and recursive asymptotically optimal procedure of estimation and observation control. Proceeding of the Prague symposium on asymptotic statistics. Charles University Prague 1974, **1**, 157—178.
11. *Nevelson, M. B.*, On convergence of recurrent estimates of zero of an unknown function. Problemi peredachi informatsii, 1975, XI, 2, 68—83.
12. *Nevelson, M. B.*, On some asymptotical properties of recursive estimates. Proceeding of the colloquia mathematical societaty János Bolyai. Limit theorems of probability theory (Keszthely, 1974), **11**, 193—225. Budapest, 1975.
13. *Gikhman, I. I., Skorokhod, A. V.*, Introduction into the theory of stochastic processes, Nauka Publ., Moscow, 1965 (in Russian).
14. *Sacks, J.*, Asymptotic distribution of stochastic approximation Ann. Math. Statist., 1968, **29**, 2, pp. 373—405.

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

H–1111 Budapest XI., Goldmann György tér 3, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4–5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которых должен содержать рисунки, таблицы и литературу) направляются

Е. Д. Теряеву – Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский Проспект, 14, Москва В-71, СССР

или

Я. Кочишу – Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета
Н–1111 Будапешт XI, площадь Голдманн Д., 3,
Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре.

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4–5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру .

Статье должна предшествовать аннотация объемом 50–100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме – на русском языке, если статья написана на английском, и наоборот – должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем резюме не должен превышать 10–15% объема статьи.

Авторам высылается корректура статьи, которую они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Lunderstädt, R.: Transportation systems with optimal parameters (*Лундерштедт Р.* Транспортные системы автоматического управления с оптимальными параметрами) 109

Невельсон М. Б. О приближении процессов стохастической аппроксимации суммами независимых случайных величин (*Nevelson, M. B.:* On stochastic approximation processes by sums of independent random variables) 117

Rejtő, L.: Pattern recognition in a Markovian case (*Рейто Л.* Распознавание образов при марковских процессах) 135

Mikhailov, F. A.: The method of generalized characteristic equation in the theory of linear time-varying systems (*Михайлов Ф. А.* Метод обобщенного характеристического уравнения в теории нестационарных линейных систем) 149

Sultan, M. A.: A method of analysis of a limit cycling relay system (*Султан М. А.* Метод анализа предельного цикла релейной системы автоматического регулирования) 167

Grodzki, Z.: New class of the controlled shift-registers. I (*Гродзки З.* Новый класс управляющих регистров сдвига) 181

Rodionova, N. I.: The 4th International Conference on Artificial Intelligence 189

Book Review 193

316.920

VOL. 5 * NUMBER 3
TOM ТОМ НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1976

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

is an international quarterly sponsored jointly by the Presidium of the Academy of Sciences of the USSR and of the Hungarian Academy of Sciences. It offers publicity for original papers and short communications of the following topics:

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

— theory and methods of adaptation, learning, identification and pattern recognition

— methods of information processing; application of digital computers in control and communication systems

— new physical principles in constructing technical devices for automation, control and information processing

The six issues published per year make up a volume of some 480 pp.

While this quarterly is mainly a publication forum of the research results achieved within the USSR and Hungary, also papers of international interest from other countries are welcome.

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

H-1389 Budapest 62, P.O.B. 149, Hungary

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

H-1363 Budapest P.O.B. 24, Hungary

Академия наук СССР и Академия наук ВНР совместно издают журнал

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал публикует материалы о новых результатах научных исследований, разработок и краткие сообщения о достижениях в следующих областях науки:

— общая теория процессов управления и теория систем
— теория автоматов
— теория информации
— исследование операций; теория сложных систем
— теория управления экономическими системами; модели систем
— теория и методы адаптации, обучения, идентификации и распознавания образов
— методы обработки информации; применение вычислительных машин в системах управления и передачи информации
— новые физические принципы создания технических средств автоматики, управления и передачи информации

Журнал выходит шесть раз в год общим объёмом 480 печатных страниц.

Журнал публикует статьи о научных достижениях обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

H-1389 Будапешт 62, п. о. 149, Венгрия

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии наук Венгрии

H-1363 Будапешт п. о. 24, Венгрия

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)
F. CSÁKI (Budapest)

DEPUTY EDITORS

V. S. PUGACHEV (Moscow)
V. I. SIFOROV (Moscow)
S. CSIBI (Budapest)

COORDINATING EDITORS

M. A. GAVRILOV (Moscow)
L. KALMÁR (Szeged)

S. V. EMELYANOV (Moscow)
I. CSISZÁR (Budapest)

N. N. KRASOVSKII (Sverdlovsk)
A. PRÉKOPA (Budapest)

E. P. POPOV (Moscow)
L. VARGA (Budapest)

E. D. TERYAEV (Moscow)
J. KOCSIS (Budapest)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)
Ф. ЧАКИ (Будапешт)

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Ш. ЧИБИ (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

М. А. ГАВРИЛОВ (Москва)
Л. КАЛМАР (Сегед)

С. В. ЕМЕЛЬЯНОВ (Москва)
И. ЧИСАР (Будапешт)

Н. Н. КРАСОВСКИЙ (Свердловск)
А. ПРЕКОПА (Будапешт)

Е. П. ПОПОВ (Москва)
Л. ВАРГА (Будапешт)

Е. Д. ТЕРЯЕВ (Москва)
Я. КОЧИШ (Будапешт)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 5 (3), pp. 195—206 (1976)

ДЕКОДИРОВАНИЕ И ДИОФАНТОВЫ ПРИБЛИЖЕНИЯ

В. Д. ГОППА

(Поступила в редакцию 20 декабря 1974 г.)

Развивается новая точка зрения на декодирование (L, g)-кодов как на задачу наилучшего рационального приближения в g-адических топологиях. Предлагается алгоритм декодирования в пределах конструктивной границы для (L, g)-кодов, являющийся аналогом алгоритма Малера рационального приближения g-адических чисел.

1. Нормирование и псевдонормирование полей

Пусть K — произвольное поле. Любая вещественнозначная функция на поле K называется нормированием поля K, если она обладает следующими свойствами [1]:

(1): $w(0) = 0$, $w(a) > 0$ при $a \neq 0$;

(2): $w(a + b) \leq w(a) + w(b)$;

(3): $w(a \cdot b) = w(a) \cdot w(b)$

для любых $a, b, \in K$.

Если вместо условия (3) выполняется более слабое условие (3′):

$$w(a \cdot b) \leq w(a)\, w(b)\,,$$

то $w(a)$ называется псевдонормированием поля K.

Из этих определений легко вывести следующие свойства $w(a)$:

$$w(-a) = w(a),$$

$$w(a) - w(b) \leq w\,(a - b) \leq w\,(a) + (b),$$

$$|\, w(a) - w(b)\,| \leq w(a - b),$$

$$w(\pm\, 1) = 1 \qquad \text{(если } w \text{ — норма)}$$

$$w(\pm\, 1) \geq 1 \qquad \text{(если } w \text{ — псевдонорма).}$$

Каждое поле имеет по крайней мере одно нормирование, а именно — тривиальное:

$$w_0(a) = \begin{cases} 0 & \text{если} \quad a = 0 \\ 1 & \text{если} \quad a \neq 0. \end{cases}$$

Для тривиального нормирования неравенство треугольника принимает более сильную форму:

(2′): $$w_0(a+b) \leq \max\left(w_0(a),\ w_0(b)\right)$$

(ультраметрическое неравенство).

Любое нормирование или псевдонормирование, для которого выполняется неравенство (2′), называется неархимедовым, в противном случае нормирование называется архимедовым.

Таким образом, нормирование $|\ a\ |$ поля рациональных чисел — архимедово.

В дополнение к $|\ a\ |$ и $w_0(a)$ поле Q имеет также бесконечное множество p-адических нормирований $|\ a\ |_p$, которые определяются следующим образом:

(*i*): $|\ 0\ |_p = 0$;

(*ii*): если $a \neq 0$, то a может быть представлено единственным способом в виде $a = p^n \frac{r}{s}$, где r и s не делятся на p, а n — целое число. Положим

$$|\ a\ |_p = p^{-n}.$$

Легко проверить, что $|\ a\ |_p$ является неархимедовым нормированием.

Если q — простое число, то

$$|q|_p = \begin{cases} \dfrac{1}{p} & \text{если} \quad q = p \\ 1 & \text{если} \quad q \neq p\,. \end{cases}$$

Это означает, что p-адические нормирования, соответствующие различным простым p, различны.

Таким образом, Q обладает бесконечным множеством различных нормирований $|\ a\ |$, $w_0(a)$,

$$\text{и } |\ a\ |_p \text{ для различных } p.$$

Эти нормирования связаны друг с другом фундаментальным тождеством $|\ a\ | \cdot \prod\limits_p |\ a\ |_p = w_0(a)$ для всех $a \in Q$, которое эквивалентно основной

теореме арифметики об однозначном разложении целых чисел на простые множители.

Пусть теперь K — произвольное поле, а $K(x)$ — его простое трансцендентное расширение (т. е. множество всех дробнорациональных функций от x). $K(x)$ является полем частных кольца $K[x]$.

Обозначим через p неприводимый полином в $K[x]$ и выберем вещественную константу $0 < \theta < 1$. Если $r \neq 0$ — произвольный элемент $K[x]$, будем обозначать через $\deg r$ степень r и $\operatorname{ord}_p r$ — порядок r по p, т. е. наибольшее целое m такое, что p^m делит r.

Каждый элемент $a \neq 0$ поля $K(x)$ может быть записан единственным способом в виде $a = \dfrac{r}{s}$, где r и s — взаимно простые полиномы, а s — ненулевой полином со старшим коэффициентом, равным 1. Обозначим

$$\| a \| = \begin{cases} 0 & \text{если} \quad a = 0 \\ \theta^{\deg s - \deg r} & \text{если} \quad a \neq 0 \end{cases}$$

$$\| a \|_p = \begin{cases} 0 & \text{если} \quad a = 0 \\ \theta^{\operatorname{ord}_p r - \operatorname{ord}_p s} & \text{если} \quad a \neq 0. \end{cases}$$

Нетрудно проверить, что эти две функции являются неархимедовыми нормированиями поля $K(x)$ и что $w_0(a)$, $\| a \|$, $\| a \|_p$ для всех неприводимых многочленов p образуют бесконечную систему различных нормирований, связанных фундаментальным тождеством:

$$\| a \| \cdot \prod_p \| a \|_p^{\deg p} = w_0(a), \qquad a \in K(x).$$

Каждое нормирование порождает в поле топологию, в которой базу окрестностей нуля составляют шары

$$V_w^{\varepsilon} = \{ a \in K : w(a) \leq \varepsilon \}, \quad \varepsilon > 0.$$

С помощью известного метода Коши нормированное поле можно вложить в полное поле, в котором исходное поле будет всюду плотным множеством. Если две нормы определяют одну и ту же топологическую структуру, то эти нормы приводят к одним и тем же пополнениям. Такие нормы называются эквивалентными. Так, из любой нормы $w(a)$ можно получить новую норму $w^*(a) = (w(a))^{\lambda}$, где $0 < \lambda < 1$, которая, очевидно, эквивалентна старой.

Все нормирования поля Q исчерпываются с точностью до эквивалентности нормами $w_0(a)$, $| a |$ и $| a |_p$ (теорема Островского [2]) и они приводят к различным пополнениям — Q, R (поле вещественных чисел) и Q_p (поле p-адических чисел) соответственно.

2. g-адическое псевдонормирование поля $K(x)$

Пусть $g \in K[x]$, где K — произвольное поле. Допустим, что многочлен g имеет следующее разложение на неприводимые множители:

$$g = p_1^{e_1} p_2^{e_2} \dots p_r^{e_r}.$$

Фиксируем числа $\lambda_1, \lambda_2, \dots, \lambda_r$ так, чтобы

$$|| g ||_{p_1}^{\lambda_1} = || g ||_{p_2}^{\lambda_2} = \dots = || g ||_{p_r}^{\lambda_r} = \theta < 1,$$

т. е. положим $\lambda_i = \dfrac{1}{e_i}$, $i = 1, 2, \dots, r$.

Определим теперь величину

$$|| a ||_g = \max (|| a ||_{p_1}^{\lambda_1}, \dots, || a ||_{p_r}^{\lambda_r}).$$

Легко проверить, что $|| a ||_g$ является неархимедовой псевдонормой на $K(x)$.

Шар $V_g^1 = \{a \in K(x): || a ||_g \leq 1\}$ является кольцом. Действительно, если $a, b \in V_g^1$, то

$$|| a \pm b ||_g \leq \max (|| a ||_g, || b ||_g) \leq 1,$$
$$|| ab ||_g \leq || a ||_g || b ||_g \leq 1.$$

Пусть $a \in K(x)$ представлен в несократимой форме $a = \dfrac{r}{s}$. Если s имеет общий делитель с g, например, p_i^l, то $|| a ||_{p_i} = \theta^{-l} > 1$, поскольку $\theta < 1$, поэтому $|| a ||_g > 1$. Обратно, если s и g взаимно просты, то

$$|| a ||_g = \max (|| r ||_g, 1) \leq 1 .$$

Таким образом, кольцо V_g^1 состоит из всех дробно-рациональных функций, знаменатели которых взаимно просты с g:

$$V_g^1 = \left\{ \frac{r}{s} \in K(x) : (s, g) = 1 \right\}.$$

Если A — некоторое кольцо (коммутативное с единицей), то мультипликативным подмножеством в A называется всякое подмножество $S \subset A$, содержащее I и замкнутое относительно умножения. Бинарное отношение $\equiv$ на $A \times S$: $(a, s) \equiv (b, t) \Leftrightarrow (at - bs)u = 0$ для некоторого $u \in S$, является отношением эквивалентности. Если a/s обозначает класс эквивалентности пары (a, s), а $S^{-1}A$ — множество классов эквивалентности, то на $S^{-1}A$ можно

ввести структуру кольца, определив сложение и умножение «дробей» a/s по формуле

$$(a/s)+(b/t)=(at+bs)/st,$$
$$(a/s)\,(b/t)=(ab/st).$$

Кольцо $S^{-1}A$ называется кольцом частных A относительно S [3] и является обобщением известной конструкции поля частных.

Если выбрать $A=K[x]$, $S_g=\{f\in A\colon (f,g)=1\}$, то, очевидно,

$$V_g^1=S_g^{-1}K[x].$$

Рассмотрим далее шар

$$V_g^\theta=\{a\in K(x):\|a\|_g\leq\theta\}.$$

Очевидно, $V_g^\theta\subset V_g^1$. Если $a,b\in V_g^\theta$, $c\in V_g^1$,

то
$$\|a\pm b\|_g\leq\max(\|a\|_g,\|b\|_g)\leq\theta,$$
$$\|ac\|_g\leq\|a\|_g\cdot\|c\|_g\leq\theta\cdot 1=\theta.$$

Таким образом, V_g^θ является идеалом кольца V_g^1. Поскольку $\|g\|_g=\theta$, то $g\in V_g^\theta$.

Далее,

$$\|a/g\|_g=\max(\|a/g\|_{p_1}^{\lambda_1},\ldots,\|a/g\|_{p_r}^{\lambda_r})=\max\left(\frac{\|a\|_{p_1}^{\lambda_1}}{\theta},\ldots,\frac{\|a\|_{p_r}^{\lambda_r}}{\theta}\right)=\frac{1}{\theta}\|a\|_g,$$

отсюда следует:

$$a\in V_g^\theta\Leftrightarrow a/g\in V_g^1.$$

Таким образом, V_g^θ состоит из всех произведений $a=a'g$, где $a'\in V_g^1$, т. е. V_g^θ совпадает с главным идеалом (g) кольца V_g^1.

В случае, когда g является неприводимым многочленом, идеал V_g^θ является максимальным, причем единственным максимальным идеалом кольца V_g^1. Он совпадает в этом случае со всеми необратимыми элементами кольца V_g^1. Таким образом, в этом случае V_g^1 — локальное кольцо.

3. Алгебраические коды. Ультраметрическое мажорирование метрики Хэмминга

Пусть E^n — векторное пространство размерности n над полем Галуа $GF(q)$, где q — степень простого числа. Предположим, что фиксирован некоторый базис этого пространства. Число ненулевых координат вектора $x\in E^n$ по этому базису называется весом Хэмминга $\|x\|_H$ вектора x. Вес Хэмминга является архимедовой нормой на E^n.

Подпространство $C \subset E^n$ называется алгебраическим (или линейным) кодом [4]. Отношение k/n, где k — размерность C, называется скоростью передачи, а $d = \min\limits_{x \in C-\{0\}} || x ||_H$ — кодовым расстоянием, или весом кода.

Декодирование для кода C заключается в нахождении вектора с наименьшим весом Хэмминга в смежном классе $x + C$.

Выберем

$$L = GF(q^m),\ m > 0,\ n = q^m.$$

Отображение

$$x = (a_1, a_2, \ldots, a_n) \xrightarrow{\varphi} \sum_{i=1}^{n} \frac{a_i}{z - \alpha_i} = \xi_x(z) = \frac{\psi(z)}{f(z)},$$

где $\{\alpha_i\}$ — все элементы поля L, линейно, инъективно и переводит вес Хэмминга $|| x ||_H$ в степень знаменателя дроби $\xi_x(z)$.

Это отображение, названное рациональным представлением кодов [5], позволяет мажорировать метрику Хэмминга неархимедовыми метриками поля рациональных функций $L(z)$. Действительно, пусть $g \in L[z]$ не имеет корней в L, а R обозначает образ E^n при отображении φ. Тогда $R \subset V_g^1$, $|| \xi_x(z) ||_g = || \psi(z) ||_g$, $|| f(z) || = \theta^{\deg f(z)} = \theta^{-||x||_H}$, и если значение $||\xi_x(z)||_g$ мало, то значение $|| f(z) ||$, а, следовательно, и значение $|| x ||_H$ велико.

Выберем в качестве кодов пересечения идеалов V_g^θ с пространством R. Это линейные коды, состоящие из дробей, числители которых делятся на многочлен g ((L, g)-коды [5]).

Фактор-пространство R/V_g^θ состоит из многочленов степени $< r = \deg g$. Каждый элемент $S(z)$ этого пространства называется синдромом. Задача декодирования для (L, g)-кодов принимает следующую форму: для каждого $S(z) \in R/V_g^\theta$ найти дробь $\dfrac{\psi(z)}{f(z)} \in R$, такую, что

$$\left\| \frac{\psi(z)}{f(z)} - S(z) \right\|_g \leq \theta$$
$$|| f(z) || \quad \text{минимальна.} \tag{1}$$

В такой формулировке эта задача декодирования очень похожа на задачи теории диофантовых приближений [1].

Классическая задача такого рода, сформулированная еще Гюйгенсом, заключается в нахождении рационального числа p/q с возможно малым значением $| q |$, для которого $| p/q - \alpha | \leq \varepsilon$, где α — произвольное вещественное число.

Как известно, существует эффективный алгоритм решения этой задачи с помощью непрерывных дробей, который позволяет найти приближение p/q с $|q| < 1/\sqrt{\varepsilon}$.

Малер сформулировал g-адический аналог этой задачи и обобщил алгоритм непрерывных дробей на этот случай [1]. Его алгоритм позволяет для любого g-адического числа α найти рациональное число p/q такое что

$$\begin{aligned} &|p/q - \alpha|_g \leq \varepsilon, \\ &\max(|p|, |q|) < 1/\sqrt{\varepsilon}. \end{aligned} \tag{2}$$

Алгоритм Малера применим и для нахождения дробно-рациональной функции, удовлетворяющей условию (1).

Заметим, что для любого многочлена S существует единственная дробная функция P/Q такая, что

$$P/Q - S \equiv 0 \pmod{g},$$

$$\deg Q \leq 1/2 \deg g, \quad \deg P < 1/2 \deg g.$$

Такую пару многочленов и позволяет найти алгоритм Малера при помощи разложения в цепную дробь.

4. Полиномиальные цепные дроби

Пусть u, v — два многочлена над произвольным полем K, $\deg u \geq \deg v$. Тогда существует единственное представление

$$u = q_0 v + r_1, q_0, r_1 \in K[z], \deg r_1 < \deg v.$$

Разделим на v и обозначим $\alpha_1 = \dfrac{v}{r_1}$:

$$\frac{u}{v} = \alpha_0 = q_0 + \frac{1}{\alpha_1}.$$

Очевидно, $$\|\alpha_1\| > 1, \ \|\alpha_0\| = \|q_0\|.$$

Применив такое же разложение для α_1 и т. д., получаем разложение α_0 в цепную дробь

$$\alpha_0 = q_0 + \frac{1}{q_1}\,\frac{1}{q_2}\cdots\frac{1}{\alpha_n}. \tag{3}$$

Дробь α_n называется n-остатком дроби α_0. Это разложение можно продолжить до тех пор, пока очередной шаг алгоритма Эвклида не приведет к нулевому остатку.

Для любой цепной дроби

$$\alpha_0 = [q_0, q_1, \ldots, q_m]$$

подходящие дроби P_k/Q_k определяются по формулам

$$\left\{\begin{matrix} P_{-1} = 1 \\ Q_{-1} = 0 \end{matrix}\right\}, \quad \left\{\begin{matrix} P_0 = q_0 \\ Q_0 = 1 \end{matrix}\right\}, \quad \left\{\begin{matrix} P_k = q_k P_{k-1} + P_{k-2} \\ Q_k = q_k Q_{k-1} + Q_{k-2} \end{matrix}\right\} \quad k = 1, 2, \ldots, m. \tag{4}$$

Тождество

$$P_{k-1} Q_k - P_k Q_{k-1} = (-1)^k \tag{5}$$

показывает, что все подходящие дроби несократимы.

Имеет место следующее выражение для α_0:

$$\alpha_0 = \frac{P_{k-1}\alpha_k + P_{k-2}}{Q_{k-1}\alpha_k + Q_{k-2}}, \tag{6}$$

которое легко доказывается по индукции.

Из (5) и (6) следует:

$$\alpha_0 - \frac{P_{k-1}}{Q_{k-1}} = \frac{(-1)^{k-1}}{Q_{k-1}(Q_{k-1}\alpha_k + Q_{k-2})}. \tag{7}$$

5. Алгоритм Малера

Пусть g, S — произвольные многочлены, $\deg S < \deg g$. Требуется найти пару многочленов P, Q такую, что

$$|| QS - P ||_g \leq \theta \tag{8}$$

с малыми значениями $|| Q ||$, $|| P ||$. Заметим, что неравенство (8) эквивалентно сравнению $QS - P \equiv 0 \pmod g$. Здесь P и Q не обязательно взаимно простые, но мы предположим, что их общие делители являются обязательно делителями g. Разложим $\alpha_0 = \dfrac{S}{g}$ в цепную дробь:

$$\alpha_0 = [q_1, q_2, \ldots, q_m].$$

Пусть R_k/Q_k — подходящие дроби для α_0. Положим

$$P_k = S \cdot Q_k - g R_k. \tag{9}$$

Так как Q_k и R_k взаимно просты, то общий делитель P_k и Q_k является обязательно делителем g. Кроме того, пара (P_k, Q_k) удовлетворяет условию (8).

Из равенства (7), далее, следует, что

$$\frac{S}{g} - \frac{R_k}{Q_k} = \frac{(-1)^k}{Q_k(Q_k \alpha_{k+1} + Q_{k-1})},$$

так что

$$P_k = \frac{(-1)^k g}{Q_k \alpha_{k+1} + Q_{k-1}}, \quad k = 0, 1, \ldots, m-1. \qquad (10)$$

Далее, $\alpha_{k+1} = q_{k+1} + \beta$, $|| \beta || < 1$. Поэтому $|| Q_k \alpha_{k+1} + Q_{k-1} || = || (Q_k q_{k+1} + Q_{k-1}) + Q_k \beta || = || Q_{k+1} ||$, поскольку $\deg Q_{k+1} \geq \deg Q_k$, а β — правильная дробь.

Таким образом,

$$|| P_k || \cdot || Q_{k+1} || = || g ||, \quad k = 0, 1, \ldots, m-1,$$

т. е.

$$\deg P_k + \deg Q_{k+1} = \deg g, \quad k = 0, 1, \ldots, m-1. \qquad (11)$$

Рассмотрим теперь 2 случая. Если $\deg Q_m > \frac{1}{2} \deg g$, то найдется такой номер j, $0 \leq j \leq m-1$, что $\deg Q_j \leq \frac{1}{2} \deg g < \deg Q_{j+1}$, поскольку

$$0 = \deg Q_0 < \deg Q_1 < \ldots < \deg Q_m.$$

Для такого номера j будет справедливо и неравенство $\deg P_j < \frac{1}{2} \deg g$ согласно (11).

Если же $\deg Q_m \leq \frac{1}{2} \deg g$, то это означает, что $\deg (S, g) \geq \frac{1}{2} \deg g$, поскольку $\frac{R_m}{Q_m} = \frac{S}{g}$.

Но $Q_m \cdot S = R_m \cdot g \equiv 0 (\mathrm{mod}\ g)$ и можно выбрать $P_m = 0$:

$$|| Q_m \cdot S - 0 ||_g \leq 0, \quad \deg Q_m \leq \frac{1}{2} \deg g.$$

Таким образом, построен эффективный алгоритм, позволяющий найти единственную пару многочленов P и Q таких, что

$$QS - P \equiv 0 \ (\mathrm{mod}\ g),$$
$$\deg Q \leq 1/2 \ \deg g,$$
$$\deg P < 1/2 \deg g.$$

для любых многочленов S и g.

6. Декодирование (L, g) - кодов

Сформулируем алгоритм декодирования (L, g)-кодов, основанный на алгоритме Малера, и позволяющий исправлять t ошибок, если $\deg g = 2t$.

1. Вычислить синдром $S(z)$:

$$S(z) = \sum_{i=1}^{n} a_i \frac{g^{-1}(\alpha_i)\, g(z) - 1}{z - \alpha_i}.$$

Здесь $\{a_i\}$ — координаты вектора на выходе, $a_i \in GF(q)$, $\{\alpha_i\}$ — элементы поля $GF(q^n)$, служащие для нумерации координат.

2. Разложить $\dfrac{S(z)}{g(z)}$ в цепную дробь, одновременно вычисляя подходящие дроби P_k/Q_k. Вычисления прекратить, как только $\deg Q_k$ станет больше t.

Многочлен Q_{k-1} является локатором ошибок.

3. Вычислить $P_{k-1} = Q_{k-1}\, S(z) - g(z)\, R_{k-1}$.

Дробь P_{k-1}/Q_{k-1} является искомой.

Для иллюстрации алгоритма рассмотрим два примера.

1. Двоичный *БЧХ*-код длины 31, исправляющий 3 ошибки, является (L, g)-кодом с $g(z) = z^6$, $L = GF(2^5) - \{0\}$.

Занумеруем все позиции слова с помощью степеней α^i корня примитивного многочлена $x^5 + x^2 + 1$.

Допустим, что вектор на выходе имеет единицы на местах, соответствующих α^0, α^1 и α^4. Тогда

$$S(z) = \frac{z^6 - 1}{z - 1} + \frac{\alpha^{-6} z^6 - 1}{z - 2} + \frac{\alpha^7 z^6 - 1}{z - \alpha} = \alpha^{20} z^5 + \alpha^{10} z^4 + \alpha^{22} z^3 + \alpha^{10} z^2 +$$

$$+ \alpha^{11} z + \alpha^{21} \frac{S(z)}{z^6} = [0, \alpha^{11} z + \alpha, \alpha^2 z + \alpha^{-7}, \alpha^8 z + \alpha, \alpha^{20} z + \alpha^6, \alpha^{20} z^2 + \alpha^{15}].$$

q_k		0	$\alpha^{11} z + \alpha$	$\alpha^2 z + \alpha^{-7}$	$\alpha^8 z + \alpha$
R_k	1	0	1	$\alpha^2 z + \alpha^{-7}$	$\alpha^{10} z^2 + \alpha^6 z + \alpha^{21}$
Q_k	0	1	$\alpha^{11} z + \alpha$	$\alpha^{13} z^2 + \alpha^{21} z + \alpha^{21}$	$\alpha^{21} z^3 + \alpha^7 z^2 + \alpha^{16} z + \alpha^{26}$

Легко проверить, что

$$Q_3 = \alpha^{21} z^3 + \alpha^7 z^2 + \alpha^{16} z + \alpha^{26} = \alpha^{21}(z - 1)(z - \alpha)(z - \alpha^4),$$

$$P_3 = Q_3 S - R_3 z^6 = \alpha^{21} z^2 + \alpha^{16}.$$

(заметим, что для двоичных кодов не нужно вычислять R_k).

2. Пусть $L = GF(2^5)$, $g = (z^3 + z + 1)^2$ (неприводимый код), ошибки произошли на местах, соответствующих 0 и 1, т. е. локатор ошибок имеет вид $(z + 1)z$.

Вычисляем синдром:

$$S(z) = (z^5 + z) + (z^5 + z^4 + z^3 + z^2) = z^4 + z^3 + z^2 + z,$$

$$\frac{S(z)}{g(z)}\,[0, z^2 + z, z^4 + z^3 + z^2 + z]\,.$$

q_k		0	$z^2 + z$	$z^4 + z^3 + z^2 + z$
R_k	1	0	1	$z^4 + z^3 + z^2 + z$
Q_k	0	1	$z^2 + z$	$z^6 + z^2 + 1$

$Q_2 = z^2 + z = (z + 1)z$

$P_2 = Q_2 S - R_2 g = (z^2 + z)(z^4 + z^3 + z^2 + z) - (z^6 + z^2 + 1) = 1.$

7. Заключение

Предлагаемый алгоритм декодирования (L, g)-кодов по сложности примерно такой же, что и известный итеративный алгоритм Берлекэмпа [4]. Одна итерация алгоритма Берлекэмпа заключается в умножении двух многочленов. Одна итерация алгоритма цепных дробей состоит в делении с остатком одного многочлена на другой и нахождении подходящей дроби (умножение двух многочленов). Хотя одна итерация алгоритма Берлекэмпа выполняется проще, но число итераций в нем фиксировано $(= 2t)$, где t — число исправляемых ошибок, а в алгоритме цепных дробей — переменно, всегда $\leq t$, а для многих синдромов оказывается строго меньше t.

Кроме того, деление в конечном поле выполняется обычно также с помощью разложения в цепную дробь. Применение одного и того же алгоритма (с небольшими различиями) и для деления, и для декодирования приводит к некоторому единству аппаратуры, что облегчает проектирование систем связи.

Таким образом, в настоящее время известно несколько эффективных алгоритмов декодирования (L, g)-кодов. Все эти алгоритмы позволяют исправлять ошибки только в пределах конструктивной границы. В то же время известно, что можно выбрать порождающий многочлен g кода так, что соответствующий код будет лежать на границе Гилберта, т. е. конструктивная граница далека от фактической для большинства (L, g)-кодов. Декодирование

полного числа ошибок для таких кодов является трудной задачей диофантовых приближений с ограничением: нужно аппроксимировать данный многочлен по g-адической норме дробями специального вида — знаменатель каждой такой дроби должен распадаться на линейные множители в фиксированном поле $GF(q''')$. Существует надежда, что какой-либо из разработанных методов декодирования в пределах конструктивной границы (или идея, связанная с этим методом) окажется исходным этапом на пути решения одной из важнейших проблем теории кодирования — декодирования на границе Гилберта.

Литература

1. *Mahler K.*, Lectures on diophantine approximations. Part I: g-adic numbers and Roth's theorem. University of Notre Dame, 1961.
2. *Боревич З. И., Шафаревич И. Р.* Теория чисел. »Наука«, М., 1972.
3. *Атья М., Макдональд И.* Введение в коммутативную алгебру. »Мир«, М., 1972.
4. *Берлекэмп Е.* Алгебраическая теория кодирования. »Мир«, М., 1971.
5. *Гоппа В. Д.* Рациональное представление кодов и (L, g)-кодов. Проблемы передачи информации, 1971, **7**, 3, 41—49.

Decoding and diophantine approximations

V. D. GOPPA

(Moscow)

Let $g(z)$ be a generating polynomial of (L, g)-code and $S(z)$ be a syndrome. Decoding algorithm for these codes is a method of solving the key equation:

$$Q(z)\ S(z) - P(z) \equiv 0 \pmod{g(z)}$$

with respect to $Q(z)$ and $P(z)$.

In this paper a new method is presented for solving such equations:

1. split $S(z)/g(z)$ into continuous fraction;
2. find the first convergent R_{k+1}/Q_{k+1} with $\deg Q_{k+1} > t$, where $\deg g(z) = 2t$; then Q_k is an error-locator polynomial;
3. compute $P_k = Q_k S - gR_k$.

This algorithm is the analogue of Mahler's algorithm of rational approximation of g-adic numbers.

В. Д. Гоппа
ЦЕМИ АН СССР
СССР, Москва В—333,
ул. Вавилова 44, корп. 2

Problems of Control and Information Theory, Vol. 5 (3), pp. 207—225 (1976)

A GLOBAL OPTIMIZATION PROCEDURE FOR DYNAMIC SYSTEMS

S. RACZYŃSKI

(Kraków)

(Received October 24, 1974)

In [1–24] some theoretical problems connected with control systems were considered. These works have been based on the theory of contingent equations of Marchaud and Zaremba [25–27] and on the orientor field theory. The approach of orientor field theory appears to lead to some practical algorithms for solving control problems.

One of the basic definitions introduced indepedently by Marchaud and Zaremba in 1935 is the definition of the contingent equation emission zone corresponding to the system reachable set defined in the classical control theory. In this paper a practical way of the reachable set (RS) shape and a suboptimal control estimation is described. The searching procedure which we discuss is of a global type. It can be implemented as the first (global) stage of optimization and combined with some local optimization algorithm.

1. Some properties of the reachable set

Let us consider the control system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \tag{1}$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ is the state vector, $\mathbf{u} = (u_1, \ldots, u_p)$ is the control vector (or simply the control), t is the time and $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ is a vector function with components $f_k(\mathbf{x}, \mathbf{u}, t)$, $k = 1, 2, \ldots, n$. The dot on top denotes time differentiation.

The n- and p-dimensional Euclidean spaces are denoted as R^n and R^p, respectively. We introduce the notation

$$\Omega = J \cap R^n \cap A,$$

where J is a time interval $0 \leq t \leq T$, $0 < T < \infty$, and A is a close limited set belonging to R^p space.

We assume that the function $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ is defined, continuous and limited in the region Ω. The control $\mathbf{u}$ is a function of time. Its components are assumed to be measurable and limited on J. The control $\mathbf{u}$ fulfils the condition

$$\mathbf{u} \in C(\mathbf{x}, t)\,, \tag{2}$$

where $C(\mathbf{x}, t)$ denotes the set called the control domain. It is assumed that this set is contained in the p-dimensional sphere with constant radius, and changes continuously with respect to $\mathbf{x}$ and t.

Finally, we assume that the function $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ fulfils the generalized Lipschitz condition (as in [11]): for an arbitary $\mathbf{x}'$ and $\mathbf{x}''$ such that $(\mathbf{x}', \mathbf{u}, t) \in \Omega$ and $(\mathbf{x}'', \mathbf{u}, t) \in \Omega$ the inequality

$$|\mathbf{f}(\mathbf{x}' \mathbf{u}, t) - \mathbf{f}(\mathbf{x}'' \mathbf{u}, t)| \leq \omega (|\mathbf{x}' - \mathbf{x}''|, t) \tag{3}$$

holds, where $\omega(v, t)$ is such a function that $\omega(v, t) \geq 0$ and the differential equation $\dot{v} = \omega(v, t)$ has the unique solution $v(t) \equiv 0$ passing through the point $(0, 0)$ on an arbitrary interval $0 \leq t \leq h$, $h < \infty$.

It should be noted that the function $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ is not assumed to be differentiable. Moreover, convexity of the control domain $C(\mathbf{x}, t)$ and of the system orientor $N(\mathbf{x}, t)$ (see the Appendix) is not required. System (1) fulfilling all the assumptions introduced above is called here the system $\mathbb{S}$.

Now we direct our attention to the set of all trajectories $\mathbf{x}(t)$ of the system $\mathbb{S}$ with the initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$. We denote the union of the graphs of all these trajectories by Z and call it the emission zone, or the reachable set of the system $\mathbb{S}$ with respect to the point $(\mathbf{x}_0, t_0)$. We denote

$$S(t_1) = \theta(t_1) \cap Z$$

where $\theta(t_1)$ is the hyperplane $t = t_1$. We call this set the time section of the reachable set.

It is interesting to note that the reachable set properties for a convex-control case were already known and proved in 1935 by Marchaud and Zaremba. The only difference between their approach and that of the control theory developed much later is in terminology. In fact, the trajectories of a contingent (or paratingent as in [25], [26]) equation are identical with those of a control system fulfilling some regularity and convexity assumptions [1]. The generalization of the contingent equation approach to control systems has been done by Wazewski [1–10]. It permitted in the early sixtieth's to find and prove properties of the control system reachable set for a non-convex control case with relatively weak assumptions.

Considering the set of all quasitrajectories of the system $\mathbb{S}$ (see the Appendix) we can define the RS for the system quasitrajectories. It was found [23] that the system quasitrajectories can be treated as identical with trajectories with respect to the point attainability introduced in [23]. On the other hand, it was proved [6] that the set $S(t_1)$ for the system $\mathbb{S}$ quasitrajectories are connected for all $0 < t_1 < \infty$. It means that, for a practical goal, the time section of the RS is a connected set for all $0 < t_1 < \infty$. The other result of

the orientor field theory [11] is that all quasitrajectories of the system $\mathcal{S}$ can be treated as limits of sequences of trajectories. It appears, that the RS for the system quasitrajectories obtained by applying the tendor ("bang-bang" type) control is identical with that obtained by controlling with the full control domain. It is equivalent to say that the trajectories lying on the RS boundary can be approximated by the trajectories corresponding to a "bang-bang" control, even in a non-convex control case.

The method discussed in this paper utilizes the RS properties mentioned above. It can be used for the determination of the RS "lower" estimate. Differently stated, we can determine the set of points in the $(\mathbf{x}, t)$-space contained in the RS. If a point of this space appears to be laid inside the determined set, we can say that it is certainly attainable. Hence, the method provides some information about point attainability. Simultaneously, the trajectory corresponding to the best value of some quality function is determined. We are searching the global optimum of quality function and, hence, the searching process must be extended to whole interior of the reachable set. It seems, that a reasonable course of the action is to construct an algorithm which determines the system reachable set and the global optimal control simultaneously in a unique searching process.

Possibilities of the RS determination

The reachable set can be determined by the following methods.

(a) By solving the partial differential equation of the first order, describing the RS boundary [9], [10].

(b) By solving the system of $2n$ differential equations of Hamilton type, obtained from some classical optimization methods, e.g. the maximum principle.

(c) By the system modelling and searching of the RS interior.

Method (a) leads to method (b). It might appear that the use of methods of this type should provide better results than method (c). However, it should be noted that the use of Hamilton's equations permits the problem to be reduced to that having only one dimension less. Generally speaking, the RS of an n-dimensional system is the region in the $n + 1$-dimensional $(\mathbf{x}, t)$-space and its boundary is an n-dimensional hyperspace. If n is great, the profit obtained by decreasing the problem dimensionality by unity has no practical meaning. On the other hand, some difficulties in the integration of the set of adjoint variable equations could appear. The appearance of the two-point boundary value problem also causes great difficulties in many practical cases. In this paper method (c) is considered. For the search of the RS interior we

apply the density forming (DF) algorithm [24] being a modification of the pure random search method. This algorithm generates the control function for a certain number of trajectories in such a way that the trajectory end points $\mathbf{x}(T)$ have an almost uniform probability density inside the RS time section $S(T)$. Hence, we can estimate the shape of the RS after generating a reasonably great number of trajectories. The trajectory corresponding to the best value of quality function is remembered. This trajectory can be utilized as a starting point of some local optimization procedures.

System model preparation

For a practical reason it is necessary to restrict the class of permitted controls. The restriction of this class to tendor control is of great importance. In view of our previous considerations the system trajectories lying on the RS boundary can be approximated by trajectories with tendor control. The other reason for introducing the tendor control is that it can be generated by a searching algorithm far easier than a measurable control restricted by condition (2) only. To carry out this change of the control function class we have to find the tendor kernel of the set $C(\mathbf{x}, t)$ (see the Appendix). To make the method more feasible, we replace the system $\mathbb{S}$ by the system $\mathbb{S}'$ defined as follows

$$\dot{\mathbf{x}}(t) = \mathbf{f}\big(\mathbf{x}(t), \psi(\mathbf{u}'(t), \mathbf{x}(t), t), t\big) \tag{4}$$

$$u'_k(t) = 0 \text{ or } 1 \text{ for } k = 1, 2, \ldots, p' \tag{5}$$

where $\mathbf{u}' = (u_1, \ldots, u_{p'})$ and the following conditions are fulfilled:

(a) the unique transformation exists

$$\mathbf{u} = \psi(\mathbf{u}', \mathbf{x}, t)$$

for all $\mathbf{u}'$ fulfilling condition (5);

(b) the orientor $N'(\mathbf{x}, t)$ of the system $\mathbb{S}'$ fulfils the relationship

$$N'(\mathbf{x}, t) = Q(\mathbf{x}, t) = \text{tendor } \big(N(\mathbf{x}, t)\big)$$

for all $(\mathbf{x}, t)$, $\mathbf{x} \in R^n$, $t \in J$.

If the tendor $Q(\mathbf{x}, t)$ contains the finite number of points such a transformation ψ does exist. In this case the dimensionality p' of the new control u' may be less than p. With an infinite number of points in the set $Q(\mathbf{x}, t)$ we have to approximate this set by the finite number of points. However, in this case some difficulties may appear.

It was pointed out that the orientor fields $N(\mathbf{x}, t)$ and $Q(\mathbf{x}, t)$ (see Wazewski [1–4]. For our system $\mathbb{S}$ fulfils condition (3), it can be proved that the reachable sets of both systems, $\mathbb{S}$ and $\mathbb{S}'$, are identical to each other if we neglect their boundaries (the RS of the system $\mathbb{S}'$ is an open set). It should be noted that the relationship $Q(\mathbf{x}, t) \subset N(\mathbf{x}, t)$ holds. Hence, all trajectories of the system $\mathbb{S}'$ are at the same time the trajectories of the system $\mathbb{S}$. As a consequence we may suppose that the procedure proposed below may provide some useful information about point attainability and suboptimal control also for systems which do not fulfil condition (3). However, in such cases the probability of missing some regions of the RS cannot be estimated.

Introducing the "bang-bang" control we have to (also for practical reason) estimate the number of switching points. For many systems the number of switching points for trajectories lying on the RS boundary is limited. We are not going to consider here methods of this number determination. Such methods are known for some (e.g. linear) systems. Here we restrict our attention to systems having a reasonably small number of switching points. This approach allows the control function to be expressed as a function of time and a finite number of parameters (the switching times).

The time discretization is the second important step in a permitted control class restricting. In many cases the control function appears to be discrete in time, though the system equations are continuous. This situation arises in those systems in which the decision-making process is discrete, e.g. in models of economical and management systems and some technological process models. We distinguish between time discretization for the control function (related to switching times) and time discretization resulting from the use of some numerical integration method. Let H be a step for the switching time discretization. We assume that the time-step for the trajectory integration is smaller than H and equal to $h = H/k$, where k is a positive integer number. It should be emphasized that our above considerations concerning the RS properties were related to continuous systems. As long as a discrete system is an approximation of a continuous one, these properties are approximately valid. For systems with relevant time discretization, the RS properties (e.g. connectivity) may be absolutely different. On the other hand, it is important to note that if we integrate a system trajectory we are sure that all points of the $(\mathbf{x}, t)$-space belonging to the graph of this trajectory are reachable. Therefore the algorithm of RS shape determination proposed in this paper can also provide some information in a discrete-time case.

The system model fulfilling the above requirements is to be written as a set of differential (or difference) equations. The right-hand sides of these equations have to be coded as a separate program segment called further

RHS and prepared in such a way as a description of right-hand sides used by standard librarian subroutines of ordinary differential equation integration. The RHS segment must refer to the common block, the components of control vector are stored in. The general block diagram of a part of the computer program using the method discussed is shown in Fig. 1. The main program

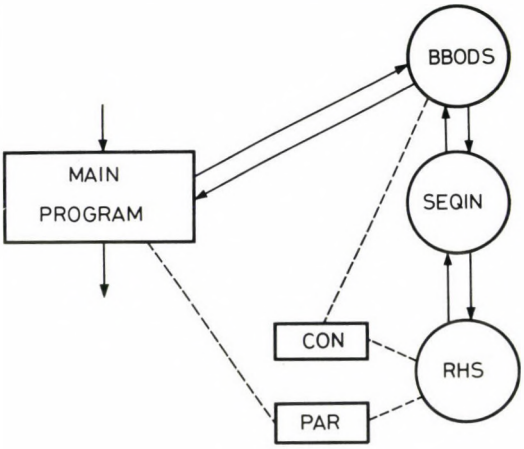


Fig. 1

calls for the subroutine BBODS, to be dealt with in the next section, in which the essential part of the proposed algorithm is coded. This subroutine uses the segment SEQIN which is the procedure of system trajectory integration. It may contain, for example, a subroutine of Runge–Kutta type or, in a discrete case, a subroutine computing the components of the state vector as follows

$$\mathbf{x}(t + H) = \mathbf{x}(t) + H \cdot \mathbf{d}(\mathbf{x}(t),\ \mathbf{u}(t),\ t)$$

where $\mathbf{d}(\mathbf{x}, \mathbf{u}, t)$ is a "velocity vector" corresponding to the function $\mathbf{f}(\mathbf{x}, \mathbf{u}, t)$ (as in (1)) in a continuous case. In a continuous linear case some other algorithms can be used for trajectory integration, e.g. the algorithm based on the state transition method [28].

The components of the vector $\mathbf{d}$ or of the function $\mathbf{f}$ are evaluated in the subroutine RHS. Note, that the subroutine RHS is not called for by the segment BBODS immediately. Therefore the control vector generated in segment BBODS has to be transferred to RHS segment through the common block, called CON in Fig. 1. The common block PAR is optional and may be used to give values of some constant parameters appearing in the system equa-

tions. Such a program structure makes the subprogram BBODS more flexible. The segments SEQIN and RHS may be written for a particular case, including an arbitrary integration procedure and a description of an arbitrary dynamical system.

The algorithm

The computer algorithm for a reachable set and optimal control determination is summarized in Fig. 2.

We introduce the notations corresponding to FORTRAN variables as follows

U(K) — the value of the control u'_k
X(K) — the value of the state variable x_k
I(W) — the number of steps H in the time interval $(0, T)$
L — the time-step consecutive number
LT — the trajectory consecutive number
IP(K) — the number of switching points of the control u'_k
OP — the optimal value of the system quality function. We assume the actual value of this function to be the value of the nth component of the state vector. Hence, OP is equal to the minimal value of $x_n(T)$ reached in the searching process
XMA(K, L) — the maximal value of $x_k(t)$, where $t = L \cdot H$, reached in the searching process. In the array XMA(K, L) for L = 1, 2, . . . , IW the upper envelope curve of the projection of the reachable set on the (x_k, t)-plane is stored
XMI(K, L) — the same as XMA, regarding lower envelope curve
KPY(J, K) — the array in which the control switching times are stored. The values of KPY(J, K) for $J = 1, 2, \ldots,$ IP(K) denote the step numbers in which the control u'_k is switched
N — the state-vector dimensionality
IU — the control-vector dimensionality

After entering the subroutine BBODS, some initial statements are executed and the loop is opened in which IT of system trajectories are to be integrated. The trajectories are integrated in series of M of them. At the beginning of each trajectory the state vector components are given the initial values. The first series of M trajectories is integrated with the switching points generated in pure random way (operation 4), i.e. for each trajectory, components of the array KPY are generated randomly from the sector 1 to IW.

Initial values of the controls (0 or 1) are also random ones. An initial value of the control U(K) is remembered in the Kth element of the appropriated array JU.

Then, the loop is opened in which the trajectory is integrated step by step. At the beginning of each time step operation 7 is executed accordingly to the statements

$$\mathsf{XMA(K, L)} = \mathsf{AMAX1(XMA(K, L), X(K))}$$

$$\mathsf{XMI(K, L)} = \mathsf{AMIN1(XMI(K, L), X(K))}$$

in a loop $K = 1, \ldots, N$. In such a way the envelope curves for the reachable set are corrected immediately with the trajectory integration. Next, the switching of controls follows. The values of controls are switched or not, accordingly to the information contained in the array KPY. The segment SEQIN evaluates the components of the state vector for the next time step. After when the whole trajectory is integrated (L = IW + 1) the trajectory end point $\mathbf{x}(T)$ is recorded in the array element XK(LT, K) where LT is the trajectory consecutive number and K is the vector $\mathbf{x}$ component number. Then, the values of arrays KPY and JU are recorded in the "archival" arrays KPA and JZ. The variables KPY (J, K) are transfered into the variables KPA(J, K, LP) and the variables JU(K) into JZ(K, LP) (for $J = 1, \ldots,$ IP(K) and $K = 1, \ldots,$ IU).

If the component $x_n(T)$, equal to $X(N)$, appeared to be less than the variable OP value the operation 9 is executed in which the variable OP is given the value of $X(N)$. In this operation the array KPY is transferred into the array KPZ and the array JU into the array MU. Hence, in the KPZ and MU arrays the control switching times and initial conditions are stored. Those parameters characterize the trajectory having had a best quality function among the trajectories integrated by the algorithm.

In such a way the first M trajectories are integrated with completely random switching of controls. If the counter LT has reached the value greater than M the density forming (DF) algorithm begins to act. This algorithm [24] consists of the following operations. The MDP (minimal density point) determination is the operation in which the set of end points $\mathbf{x}(T)$ of the last M trajectories, remembered in the array XK, is scanned. The point is found for which the distance to the nearest of other end points is a maximal one. This point is an estimate of the point position at which the probability density of trajectory end points is minimal in the set $S(T)$. The consecutive number, in the last series of M trajectories, of the point found is moved to a variable KK. Next, the parameters of the trajectory corresponding to this minimal density point are moved from the "archival" arrays KPA and JZ to the arrays

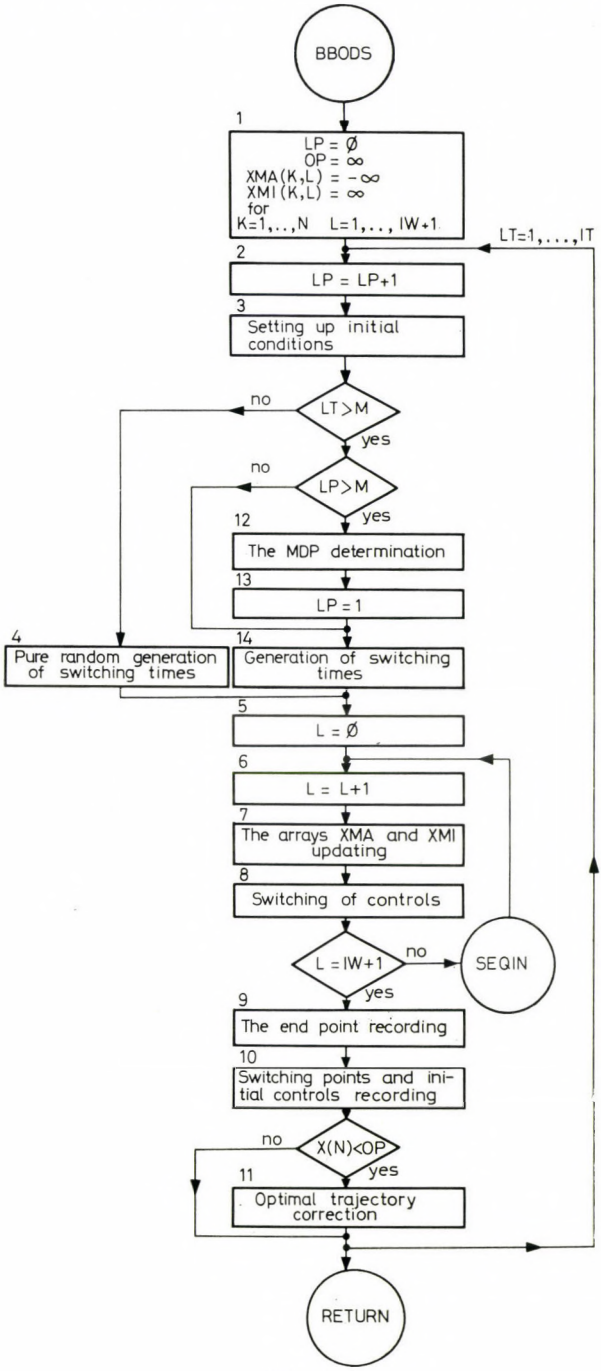

BBODS
1
LP = Ø
OP = ∞
XMA(K,L) = -∞
XMI(K,L) = ∞
for
K=1,..,N L=1,..,IW+1
LT=1,...,IT
2
LP = LP+1
3
Setting up initial conditions
no
LT>M
yes
no
LP>M
yes
12
The MDP determination
13
LP = 1
4
Pure random generation of switching times
14
Generation of switching times
5
L = Ø
6
L = L+1
7
The arrays XMA and XMI updating
8
Switching of controls
L = IW+1
no
SEQIN
yes
9
The end point recording
10
Switching points and initial controls recording
no
X(N)<OP
yes
11
Optimal trajectory correction
RETURN

Fig. 2

KPX and JU, respectively, that is the substitution is executed KPX(J, K) = = KPA(J, K, KK) and JU(K) = JZ(K, KK) for $J = 1, \ldots,$ IP(K), $K = 1, \ldots,$ IU. An auxiliary counter LP is set equal to 1. Hence, operations 12 and 13 are executed for LT = $M + 1$, $2M + 1$, $3M + 1$, ... Operation 14 is executed for all trajectories except the first M ones. It consists in generation of the switching times and moving them into the array KPY. This generation process is a random one but not as random as at operation 4. The numbers generated have their expectation values equal to the "pattern" values recorded in the previously prepared array KPX. The distribution of those numbers is of discrete type, being an approximation of the normal truncated distribution. The standard deviation of the numbers generated is greater than 1 and less than IW. Generally, it have to be chosen in the practical trials way. As a result of operations 12 and 14 the probability density of the trajectory end points for currently integrated trajectory series is made high in the neighbourhood of the point in which the density for earlier generated trajectories is low. This algorithm is a practical realization of the DF algorithm described in [24]. If some additional assumptions are introduced it can be proved that the probability density of the set of points $\mathbf{x}(T)$ becomes nearly uniform in the set $S(T)$.

If the RETURN statement is being executed, the results of subroutine BBODS have been placed in the arrays XMI, XMA, XK, KPZ, MU and in the variable OP. The arrays XMI and XMA contain the envelope curves of the projections of RS on the planes (x_k, t). For example, if the projection on the plane (x_k, t) is required we should print out the sets of numbers XMI(K, L) (the lower envelope curve) for $L = 1, \ldots,$ IW. Those numbers can be printed out as plots shown in Fig. 3. The reachable set connectivity property may be

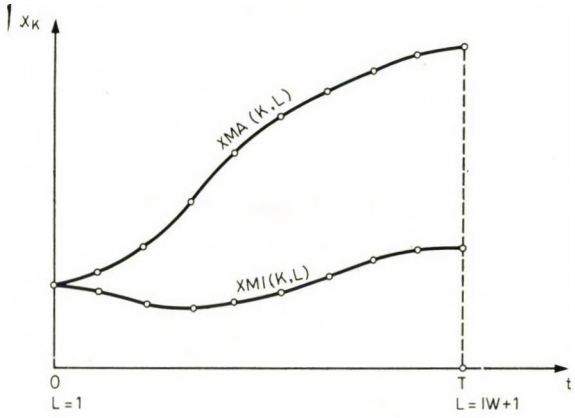


Fig. 3

utilized at the stage of the results printing out. Since the set $S(T)$ of the system $\S$ is connected for $0 < t \leq T$ all points lying between XMA and XMI plots can be treated as attainable ones. It should be emphasized that the RS connectivity property is valid in a continuous case. If a system has a relevant time discontinuity the attainability of points positioned between XMA and XMI curves is only a hypothesis. The set of trajectory end points remembered in the array XK can be used to print out the projections of the set $S(T)$ on the plains to be given (the RS time sections). The suboptimal control is entirely characterized by the values recorded in the arrays KPZ and MU. It can be used as a starting point by some local optimization subroutine at the next optimization stage.

The storage required by BBODS segment depends first of all on the dimensionality of the arrays KPA and XK. The array KPA has dimensions (IPM, IU, M) where IPM is the maximal number of switching points for one control component, IU is control vector dimensionality, M is an amount of trajectories being integrated in one series. The array XK has dimensions (IT, N) where IT is the maximal number of trajectories being integrated in one subprogram run, and N is the system state vector dimensionality. This array storage requirement could be decreased if operation 9 (the end point recording) be performed not for all state vector components but for a certain one only. In such case the RS time section might be printed out as its projections on the planes corresponding to these components only. This restriction provides operation 12 to be performed for a certain components of the vector $\mathbf{x}(T)$, e.g. the distances between trajectory end points are computed in a certain hyperspace of the space R^n. This fact makes the DF algorithm to act in the hyperspace. In other words, the density of $\mathbf{x}(T)$ points converges to the uniform one in this hyperspace. In practice, if such a storage limitation is necessary, operations 9 and 12 are to be performed for some most relevant state vector components only.

The CPU time required is not less than $IT \cdot t_i$ where t_i is a time required for integration of one trajectory. In order to decrease the CPU time of a BBODS segment, one can perform the scanning in operation 12 not for all trajectory end points but for a certain amount of them only, i.e. chosen randomly.

Example

A system is considered

$$\mathbf{V}(t) = \mathbf{M}\dot{\mathbf{V}}(t) + \mathbf{w}(t) + \mathbf{V}_0(t) \tag{6}$$

describing the growth of a multisector economical system. Here $\mathbf{V}$ is the state vector $\mathbf{V} = (V_1, \ldots, V_n)$, $\mathbf{w}$ is the control $\mathbf{w} = (w_1, \ldots, w_n)$ and $\mathbf{V}_0$ is the function of time to be given, $\mathbf{V}_0(t) = (V_{01}(t), \ldots, V_{0n}(t))$, $\mathbf{M}$ is the matrix $n \times n$. The following conditions must be fulfilled

$$V_i(t) \geq 0, \qquad \dot{V}_i(t) \geq 0, \qquad w_i(t) \geq 0 \tag{7}$$

for $i = 1, 2, \ldots, n$.

The economical interpretation of this system can be found in [29].

As an example we consider the system describing the growth of two stock-forming sectors. In such a case the matrix $\mathbf{M}$ has an inverse matrix $\mathbf{H}\{h_{ij}\}$ and the system can be described by the equation

$$\dot{\mathbf{V}}(t) = H(\mathbf{V}(t) - \mathbf{z}(t)) \tag{8}$$

where $\mathbf{z} = \mathbf{w} + \mathbf{V}_0$. Now, conditions (7) become

$$\begin{aligned} &h_{11}(V_1 - z_1) + h_{12}(V - z_2) \geq 0 \\ &h_{21}(V_1 - z_1) + h_{22}(V_2 - z_2) \geq 0 \\ &z_1 > V_{01} \qquad z_2 > V_{02}. \end{aligned} \tag{9}$$

We consider some particular cases in which $h_{11} < 0$, $h_{22} < 0$, $h_{12} > 0$, $h_{21} > 0$ and $\frac{h_{21}}{h_{22}} < \frac{h_{11}}{h_{12}}$. One of the possibilities of the shape of the orientor $N(\mathbf{V}, t)$ of the system under consideration is shown in Fig. 4. Using (6) we can find the region $C(\mathbf{V}, t)$ (the control domain) on the z_1, z_2 plane (see Fig. 5). The

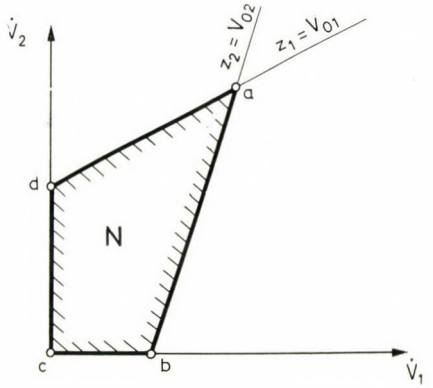


Fig. 4

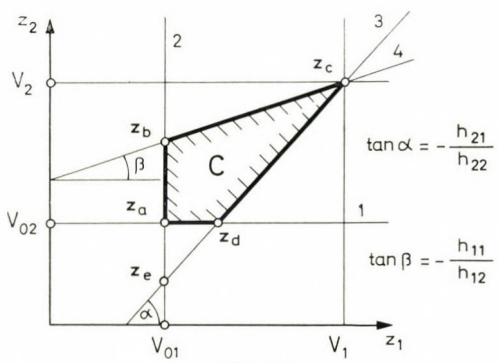


Fig. 5

extremal points a, b, c, d of the orientor $N(\mathbf{V}, t)$ correspond to the extremal points $\mathbf{z}_a$, $\mathbf{z}_b$, $\mathbf{z}_c$, $\mathbf{z}_d$ of the region $C(\mathbf{V}, t)$. Introducing the "bang-bang" control we restrict the control domain to these four points. According to the notation introduced in the Appendix the set of points a, b, c, d is the tendor of the orientor $N(\mathbf{V}, t)$ and the set of points $\mathbf{z}_a$, $\mathbf{z}_b$, $\mathbf{z}_c$, $\mathbf{z}_d$ is the tendor kernel of the control domain. Now, an auxiliary control vector $\mathbf{u}' = (u_1', u_2')$ is introduced, components of which are equal to 0 or 1. The combinations (00), (01), (10), (11) are ascribed the points $\mathbf{z}_a$, $\mathbf{z}_b$, $\mathbf{z}_c$, $\mathbf{z}_d$ as follows

$$\mathbf{z} = \begin{cases} \mathbf{z}_a & \text{if} \quad \mathbf{u}' = (0, 0) \\ \mathbf{z}_b & \text{if} \quad \mathbf{u}' = (0, 1) \\ \mathbf{z}_c & \text{if} \quad \mathbf{u}' = (1, 0) \\ \mathbf{z}_d & \text{if} \quad \mathbf{u}' = (1, 1). \end{cases} \tag{10}$$

The above relationship is the function ψ appearing in (5). The components of the vectors $\mathbf{z}_a$, $\mathbf{z}_b$, $\mathbf{z}_c$, $\mathbf{z}_d$ can be easily found from geometric considerations in terms of the current values of the vectors $\mathbf{V}$ and $\mathbf{V}_0$ and the matrix $\mathbf{H}$ coefficients (see Fig. 5). There are some other possibilities of the point configuration which must be taken into account.

The essential flow diagram of the RHS subroutine for a considered case is shown in Fig. 6. The first operation is the configuration of points $\mathbf{z}_a$, $\mathbf{z}_b$, $\mathbf{z}_c$, $\mathbf{z}_d$ verification, which must be done because the point $\mathbf{z}_a$ may appear to be positioned below the point $\mathbf{z}_a$ or above the point $\mathbf{z}_b$. Next, the evaluation of the coordinates of the set C extremal points follows (operations 2,3 or 4). These coordinates are substituted into the vector $\mathbf{z}$ components in terms of the control $\mathbf{u}'$ coordinates. In Fig. 6 the computation of the extremal points coordinates for the point configuration as in Fig. 5 is exemplified only. The remaining two cases are summarized as operations 2 and 4. The values of the vector $\mathbf{z}$ components are substituted to Eq. (8) to determine the system equation right-hand sides.

An example of the projection of the RS on the plane (V_1, t) is shown in Fig. 7. The asterisks denote reachable points.

It should be noted that the linearity of the system considered has no relevant meaning in this approach. Consider, for example, the system

$$\dot{x}_1 = x_2 \cdot u_1 + x_1$$

$$\dot{x}_2 = (\exp(x_2 \cdot u_1) + x_1) \cdot u_2$$

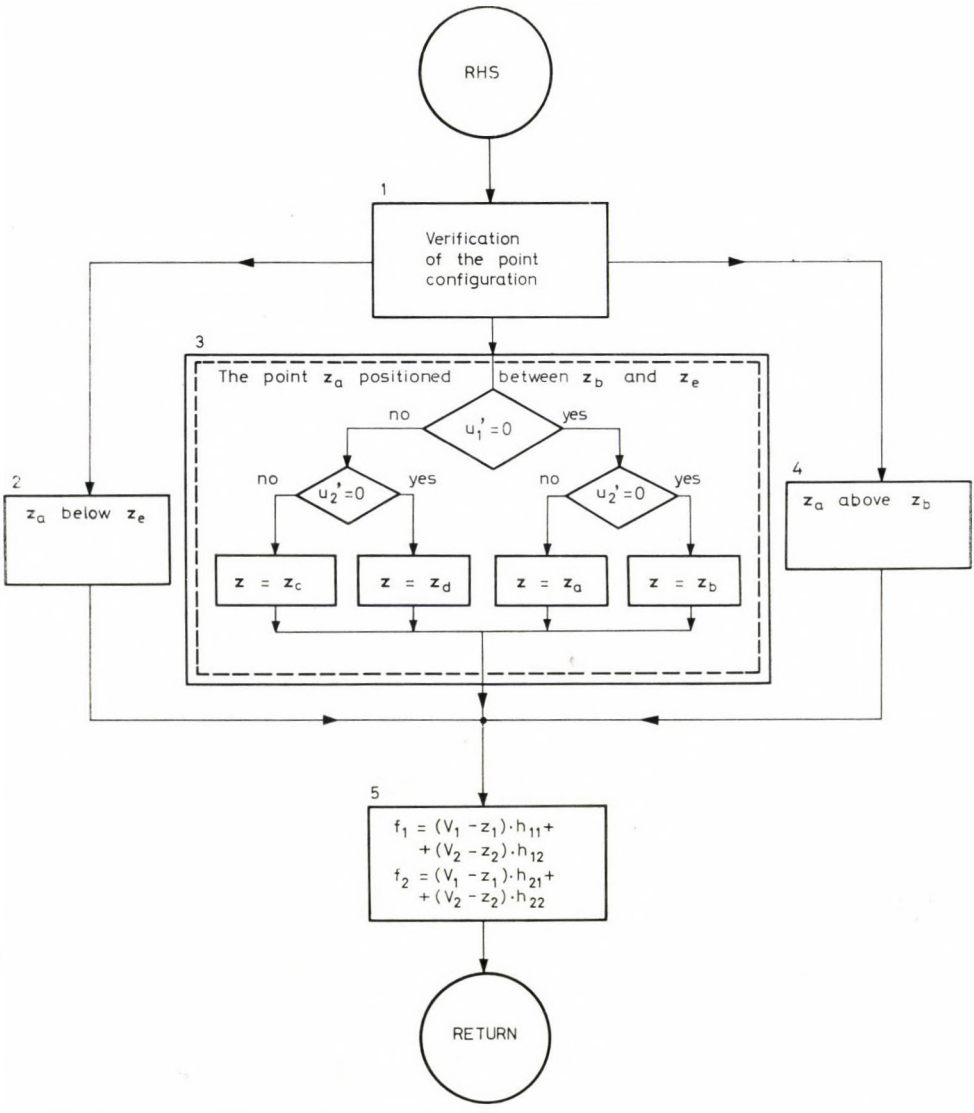


Fig. 6

with $\mathbf{x}(0) = (1, 0)$, $u_1 \in [0, 1]$, $u_2 \in [0, 1]$. It is easy to show, that the tendor set of this system consists of four points: $(x_1, 0)$, $(x_2 + x_1, 0)$, $(x_1, 1 + x_1)$, and $(x_2 + x_1, \exp(x_2) + x_1)$. Note, that for an arbitrary initial condition, this is not in general the case. The tendor set of the same system in the point $\mathbf{x} = (-2, 1)$ consists of an infinite number of points.

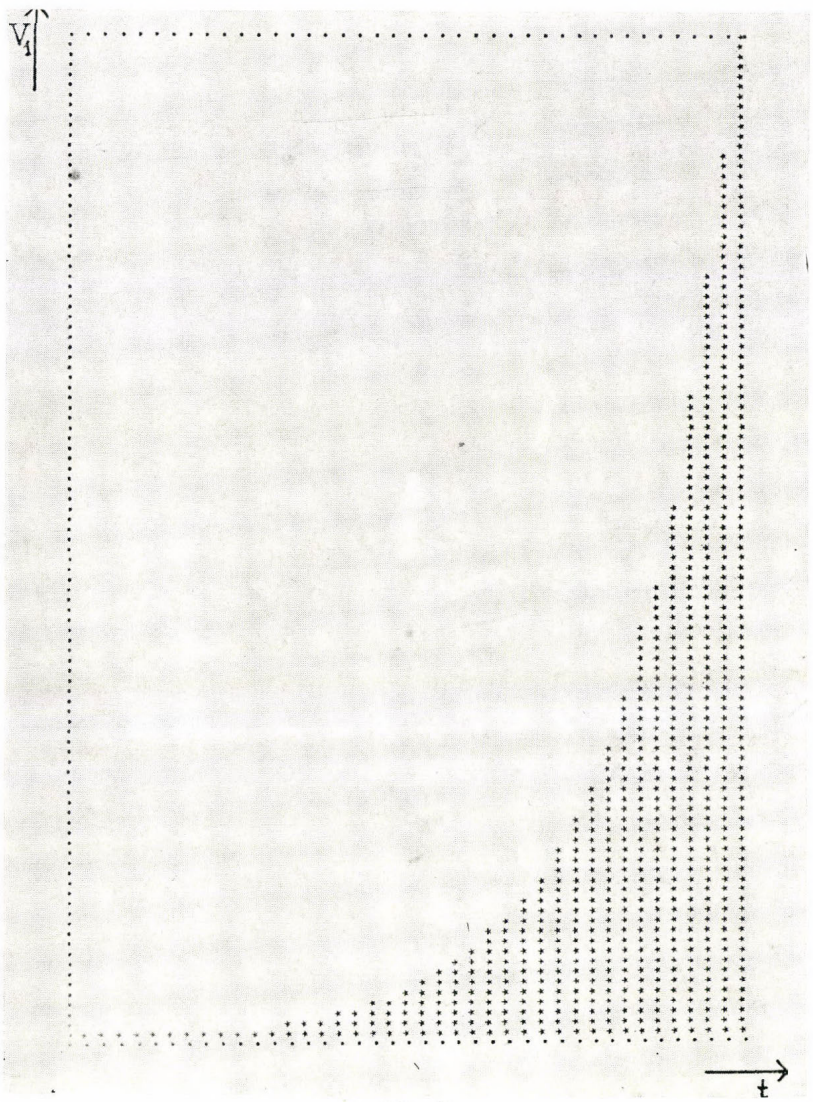


Fig. 7

Conclusions

The discussed method is, in fact, some practical approach to the method of elimination of a control variable suggested by Wazewski ([10], [12]. Instead of passing from a control system $\mathcal{S}$ to an associated orientor field and seeking its RS and optimal trajectory, we determine another control system having an associated orientor field identical to the tendor field of the system under

consideration. Since the control $\mathbf{u}$ of the system $\mathbb{S}$ can be expressed by the relation (4), no inverse problem appears (as in [10]). A system $\mathbb{S}'$ investigation is, in the most cases, far easier than that of the system $\mathbb{S}$, because the former system has a control domain consisting of few points. Besides, a dimensionality of the new control $\mathbf{u}'$ may be less than the control $\mathbf{u}$ dimensionality. In the example discussed, the set $Q(\mathbf{V}, t)$ was determined from geometric considerations. The other approach might be suggested, in which the points of the set Q are determined automatically by some appropriated subroutines.

Interesting results are expected to be provided by this method in the multicriterial system optimization problem. In this case the reachable sets in the space of system performance indexes can be determined. Such sets, printed out in a graphical form, would be helpful in man-computer system analyses and designing.

Appendix

The set of values denoted as $\mathbf{x} = (x_1, \ldots, x_n)$ is treated here both as a vector and as a point in the R^n space. Hence, we can say that the set of vectors is convex when the set of points having coordinates identical with the vector components is a convex one. Let us assume that the set of vectors corresponds to the point of $n + 1$-dimensional $(\mathbf{x}, t)$ space. Such a set is denoted as $N(\mathbf{x}, t)$ and called the orientor. The same notation is used for a set-valued function $N(\mathbf{x}, t)$ ascribing the orientor to a point $(\mathbf{x}, t)$. This function is called the orientor field.

The condition

$$\dot{\mathbf{x}} \in N(\mathbf{x}, t), \tag{A1}$$

where $\dot{\mathbf{x}} = \dfrac{d\mathbf{x}}{dt}$ is called the orientor condition.

The absolutely continuous function $\mathbf{x}(t)$ fulfilling almost everywhere condition (A1) is called the trajectory of the orientor field $N(\mathbf{x}, t)$.

The absolutely continuous function $(\mathbf{x}, t)$ is called the quasitrajectory of the orientor field $N(\mathbf{x}, t)$ if there exists such a sequence of absolutely continuous functions $\mathbf{x}_i(t)$ that $\mathbf{x}_i(0) = \mathbf{x}(0)$ and

$$\dot{\mathbf{x}}_i(t) \to \mathbf{x}(t)$$

and

$$r\big(\dot{\mathbf{x}}_i(t), N(\mathbf{x}_i(t), t)\big) \to 0$$

if $i \to \infty$. The symbol $r(\mathbf{a}, A)$ denotes the smallest distance from the point $\mathbf{a}$ to the set A. It should be noted that the functions $\mathbf{x}_i(t)$ may not be trajectories

of the field $N(\mathbf{x}, t)$ because they may not fulfil the condition

$$\dot{\mathbf{x}}_i(t) \in N(\mathbf{x}_i, t).$$

The following orientor fields are introduced

$$E(\mathbf{x}, t) = \text{env } (N(\mathbf{x}, t))$$
$$Q(\mathbf{x}, t) = \text{tendor } (N(\mathbf{x}, t)),$$

where env (A) is the smallest convex envelope of the set A and tendor (A) is the smallest set such that env (tendor (A)) = env (A).

Let us consider the control system $\mathbb{S}$ introduced in section 1. It is described by the vector differential equation

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \tag{A2}$$

with the condition

$$\mathbf{u} \in C(\mathbf{x}, t). \tag{A3}$$

Now, the transformation

$$\mathbf{v} = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)$$

is introduced. Let $\mathbf{u}$ scan all points of the set $C(\mathbf{x}, t)$ for a certain fixed point $(\mathbf{x}, t)$. Then the point $\mathbf{v}$ scans the points belonging to the set $N(\mathbf{x}, t)$. The set $N(\mathbf{x}, t)$ is called the orientor associated to the control system $\mathbb{S}$.

The absolutely continuous function $\mathbf{x}(t)$ is called the trajectory of the system $\mathbb{S}$ on the interval $0 \leq t \leq T$ if there exists such a measurable function $\mathbf{u}(t)$ fulfilling condition (A3) that Eq. (A2) holds for almost all t on the interval $0 \leq t \leq T$.

The absolutely continuous function $\mathbf{x}(t)$ is called the quasitrajectory of the system $\mathbb{S}$ if there exists such a sequence of absolutely continuous functions $x_i(t)$ with the initial condition $\mathbf{x}_i(0) = \mathbf{x}(0)$ and a sequence of measurable functions $\mathbf{u}_i(t)$ that

$$\mathbf{u}_i(t) \in C(\mathbf{x}_i(t), t)$$
$$\dot{\mathbf{x}}_i(t) - \mathbf{f}(\mathbf{x}_i(t), \mathbf{u}_i(t), t) \to 0$$
$$\mathbf{x}_i(t) \to \mathbf{x}(t)$$

for $i \to \infty$. Functions $\mathbf{x}_i(t)$ may not be trajectories of the system $\mathbb{S}$.

The system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)$$
$$\mathbf{u} \in D(\mathbf{x}, t)$$

is denoted by $\mathbb{S}_1$. The set $D(\mathbf{x}, t)$ is the greatest set such that the orientor $N_1(\mathbf{x}, t)$ associated to this system is identical with the convex envelope of the system $\mathbb{S}$ orientor $N(\mathbf{x}, t)$, that is $N_1(\mathbf{x}, t) = E(\mathbf{x}, t) = \text{env}(N(\mathbf{x}, t))$.

The system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t)$$

$$\mathbf{u} \in R(\mathbf{x}, t)$$

is denoted by $\mathbb{S}_2$. Here the set $R(\mathbf{x}, t)$ is the greatest set such that $N_2(\mathbf{x}, t) = Q(\mathbf{x}, t) =$ tendor $(N(\mathbf{x}, t))$, where the orientor N_2 is associated to the system $\mathbb{S}_2$. The set $R(\mathbf{x}, t)$ corresponds to the "bang-bang" type of control. This set is called the tendor ("bang-bang") kernel of the control domain $C(\mathbf{x}, t)$.

Wazewski [1–4] proved that the control systems $\mathbb{S}$, $\mathbb{S}_1$, $\mathbb{S}_2$, and the orientor fields N, E, Q have the same quasitrajectories. Next, the quasitrajectories of the system $\mathbb{S}_1$ and those of the field E are identical with the trajectories of $\mathbb{S}_1$ and E. It was proved, that if the moving target is attainable by at least one system quasitrajectory there exists the time-optimal quasitrajectory reaching the target in minimal time.

It was pointed out [16] that the quasitrajectory may not be a limit of a sequence of trajectories. The conditions which must be fulfilled in order to make the system quasitrajectories be limits of trajectory sequences are at the same time the conditions of "bang-bang" control applicability [23].

References

1. *Ważewski, T.*, Systèmes de commande et équations au contingent, Bulletin de L'Académie Polonaise des Sciences, Ser. des sci. math., astr. at phys., **IX,** *3*, 1961.
2. *Ważewski, T.*, Sur une condition d'existance des fonction implicites measurable, Ibid., **IX,** *12*, 1961.
3. *Ważewski, T.*, Sur une condition équivalente à l'équation au contingent, Ibid., **IX,** *12*, 1961.
4. *Ważewski, T.*, Sur la sémicontinuité inferieure du "tendeur" d'un ensamble compact variant d'une façon continue, Ibid., **IX,** *12*, 1961.
5. *Ważewski, T.*, Sur une généralisation des solutions d'une équation au contingent, Ibid., **X,** *1*, 1962.
6. *Ważewski, T.*, Sur les systèmes de commande non lineaires dont le contredomaine de commande n'est par forcément convexe, Ibid., **X,** *1*, 1962.
7. *Ważewski, T.*, Sur quelques définitions equivalentes des quasitrajectoires des systèmes de commande, Ibid., **X,** *9*, 1962.
8. *Ważewski, T.*, Sur un système de commande dont les trajectoires coincident avec les quasitrajectoires du système de commande donné, Ibid., **XI,** *3*, 1963.
9. *Ważewski, T.*, O problemie optymalnego sterowania w przypadku nieliniowym, Archiwum Automatyki i Telemechaniki, **VII,** *1–2*, 1962, Warsaw.
10. *Ważewski, T.*, On an optimal control problem (in connection with the theory of orientor fields of A. Marchaud and S. K. Zaremba), Proceedings of the Conference "Differential equations and their applications" held in Prague, September 1962.
11. *Turowicz, A.*, Sur les trajectoires et quasitrajectoires des systèmes de commande nonlineaires, Bulletin de L'Academie Polonaise des Sciences, Ser. des sci. math., astr. at phys., **X,** *10*, 1962.
12. *Turowicz, A.*, Remarque sur un travail de R. V. Gemkralidze relatif aux régimes optimaux glissants, Ibid., **X,** *11*, 1962.
13. *Turowicz, A.*, Sur les zones d'émission des trajectoires et des quasitrajectoires des systèmes de commande nonlinéaires, Ibid., **XI,** *2*, 1963.

14. *Turowicz, A.*, Remarque sur les zones d'emission des trajectoires et des quasitrajectoires des systèmes de commande non-linéaires, Ibid., **XI,** *5*, 1963.
15. *Turowicz, A.*, Remarque sur la definition des quasitrajectoires d'un système de commande nonlinéaire, Ibid., **XI,** *6*, 1963.
16. *Pliś, A.*, Trajectories and quasitrajectories of an orientor field, Ibid., **XI,** *6*, 1963.
17. *Pliś, A.*, Measurable orientor fields, Ibid., **XIII,** *8*, 1965.
18. *Pliś, A.*, On trajectories of orientor fields, Ibid., **XIII,** *8*, 1965.
19. *Sedziwy, S.*, On the trajectories and quasitrajectories of a two-dimensional orientor field, Ibid., **XII,** *3*, 1964.
20. *Lasota, A., Opial, Z.*, L'application du principle de Pontriagin à l'évaluation de l'intervalle d'existance et d'unicité des solutions d'un problème aux limites, Ibid., **XI,** *2*, 1963.
21. *Lasota, A.*, Sur une généralization du premier théorème de Fredholm, Ibid., **XI,** *3*, 1963.
22. *Lasota, A., Opial, Z.*, An application of the Kakutani–Ky Fan theorem in the theory of ordinary differential equations, Ibid., **XIII,** *11–12*, 1965.
23. *Raczyński, S.*, Applicability of the tendor control (of "bang-bang" type) in the non-linear optimum systems, Ibid., Série des sciences techniques, **XVII,** *4*, 1969.
24. *Raczyński, S.*, On the determination of the reachable sets and optimal control by the random method, IFAC Symposium "Optimization methods", Varna, Bulgaria 1974.
25. *Zaremba S. K.*, O równaniach paratyngensowych, Rocznik Polskiego Towarzystwa Matematycznego, **9,** 1935, supplement.
26. *Zaremba, S. K.*, Sur les equations au paratingent, Bulletin des Sciences Mathematiques, **60,** 1936.
27. *Marchaud, A.*, Sur les champs de demi-cônes et les équations differentielles du premier ordre, Bulletin de la Soc. Math. de France, **62,** 1934.
28. *Capehart, B. L., Schneider, D. P.*, A state variable integration routine for IBM system 360/CSMP, Simulation, **18,** *4*, 1972.
29. *Dubovskii, S. V., Dyukalov, A. N., Ivanov, Yu. N., Tokarev, V. V., Uzdemir, A. P., Fatkin, Yu. M.*, Construction of an optimal economic plan, Automation and Remote Control, **33,** *8*, 1972.

Метод глобальной оптимизации динамических систем

С. РАЧИНСКИ

(Краков)

Рассматривается метод глобальной оптимизации и определения области достижимости динамических систем и приводится соответствующая программа для ЭВМ. Предлагаемый метод основан на теории контингентных уравнений Маршо и Зарембы и теории ориенторных полей. Показано, что геометрическая интерпретация проблемы динамической оптимизации, введенная Т. Важевским, приводит к интересным практическим приложениям. Описанный метод состоит в случайном бросании внутри области достижимости системы. Он гарантирует равномерную плотность конечных точек траектории системы внутри области достижимости в пространстве состояний системы (а не в пространстве управлений, как в методе чисто случайного бросания). Предлагаемый алгоритм не использует классических методов таких, как принцип максимума или динамическое программирование. Потому здесь не фигурируют уравнения системы сопряженных переменных, что позволяет избежать вычислительных трудностей.

S. Raczyński
Academy of Mining and Metallurgy
Kraków, al. Mickiewicza 30
Poland

Problems of Control and Information Theory, Vol. 5 (3), pp. 227—238 (1976)

TWO OPTIMIZATION METHODS IN NEAREST NEIGHBOR PATTERN CLASSIFICATION

J. FRITZ

(Budapest)

(Received April 25, 1975)

The well-known nearest neighbor classification rule has very nice asymptotic properties, but the classification error may depend rather strongly on the training points and on the metric defining the nearest neighbor in case of a finite training sequence. Here we propose two stochastic gradient type algorithms to minimize the classification error probability. In both cases the classification rule itself is given by a so-called artificial training sequence of fixed size. The first procedure minimizes the error probability as a function of the points of the artificial training set, while the second one approaches the optimal metric of the sample space. The optimization procedures are based on a natural training sequence, convergence with probability one is proven for them. The second method seems to be applicable for solving feature extraction problems.

1. Introduction

Let us consider a couple (ω, ϱ) of random variables, where ω is a randomly selected point from the Euclidean r-space E_r, while the possible values of its label ϱ are the integers $1, 2, \ldots, s$. We say that there are s classes $C_1, C_2, \ldots, C_s$ and the observed value of ω belongs to C_i if $\varrho = i$. The basic problem is that we have to classify the observed value of ω into one of the classes, that is we have to tell ϱ when only ω is known. Such a decision rule can be defined by a function $\mathbf{Z} = \mathbf{Z}(\mathbf{x})$ taking values in the set $\{1, 2, \ldots, s\}$ for $\mathbf{x} \in E_r$; we classify ω into C_i by θ if $\theta(\omega) = i$, where the probability of misclassification will be $P_e = P[\varrho \neq \theta(\omega)]$. Suppose that, given $\varrho = i$, the conditional density function $f_i(\mathbf{x})$ of ω exists for each i, and denote $p_i = P(\varrho = i)$ the a priori probability of the ith class; the p_i's and the $f_i(\mathbf{x})$'s determine the distribution of (ω, ϱ).

It is well known that P_e is minimized by the Bayesian decision $\theta^*(\mathbf{x}) = i$ if $p_j f_j(\mathbf{x})$ is maximal for $j = i$. Unfortunately, in practical situations the a priori probabilities and the conditional densities are never known. Instead of a complete information on (ω, ϱ) we are given a training sequence (ω_1, ϱ_1), $(\omega_2, \varrho_2), \ldots, (\omega_n, \varrho_n), \ldots$ of preclassified samples, that is the couples (ω_n, ϱ_n) are idependent and identically distributed random variables with the same common distribution as that of (ω, ϱ). In the mathematical theory of statistical pattern recognition several methods have been developed to construct

optimal or quasioptimal classification rules on the basis of the training sequence. In this paper we investigate only the nearest neighbor rule that classifies the given sample into the class of its nearest neighbor from a finite training sequence (see [1], [2], [3]). Not so as in the cited works, the nearest neighbor decision will be defined by means of an artificial training sequence (ATS) and the natural training sequence will be used only to optimize the ATS.

We motivate this indirect approach by the following simple examples. Let $s = 2$, $p_1 = \frac{1}{4}$, $p_2 = \frac{3}{4}$ and $f_1(\mathbf{x}) = f_2(\mathbf{x}) = \frac{1}{V}$ on a domain of volume V. Then the Bayesian error probability is $\frac{1}{4}$, while the nearest neighbor error converges to $\frac{3}{8} > \frac{1}{4}$ although one point belonging to C_2 is enough to realize the Bayesian decision according to the nearest neighbor principle. A less trivial example is $p_1 = p_2 = \ldots = p_s = \frac{1}{s}$, $f_i(\mathbf{x}) = \varphi(|\mathbf{x} - m_i|)$, where φ is a non-negative, decreasing function and $|\mathbf{x} - m_i|$ is the distance of $\mathbf{x}$ and $\mathbf{m}_i \in E_r$ $i = 1, 2, \ldots, s$ (e.g. the conditional densities are normal with the same covariances and expectations $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_s$). Using $(\mathbf{m}_1, 1), (\mathbf{m}_2, 2), \ldots, (\mathbf{m}_s, s)$ as the ATS and $|\,.\,|$ as the metric defining the nearest neighbor we can reach the Bayesian error, while an arbitrarily large natural training sequence will yield a larger error probability. This example shows that besides the ATS, the metric may also have a strong influence on the probability of misclassification; here the Bayesian error is attained only if the nearest neighbor is defined by the same metric $|\,.\,|$ as figuring in the analytical expression of the conditional densities.

2. Formulation of the problem

First we introduce some notations. Let $|\,.\,|_{\mathbf{w}}$ denote an Euclidean norm on E_r of type

$$|\mathbf{x}|_{\mathbf{w}} = \left[\sum_{i=1}^{r} w_i^2 x_i^2\right]^{1/2}$$

where x_i and w_i denote the ith coordinate of $\mathbf{x} \in E_r$ and $\mathbf{w} \in E_r$, respectively. The ATS is defined as an arbitrary sequence of labelled points, $(\mathbf{x}_1, \theta_1), (\mathbf{x}_2, \theta_2), \ldots$ $\ldots, (\mathbf{x}_m, \theta_m)$, $\mathbf{x}_j \in E_j$, $\theta_j \in \{1, 2, \ldots, s\}$ for $j = 1, 2, \ldots, m$, m is fixed. The nearest neighbor $\mathbf{x}'$ of a point $\mathbf{x} \in E_r$ from the ATS and its label θ' are: $\mathbf{x}' = \mathbf{x}'(\mathbf{x}) = \mathbf{x}_j$ and $\theta' = \theta'(\mathbf{x}) = \theta_j$ if j is the first integer for which $|\mathbf{x} - \mathbf{x}_j|$ is minimal. Both $\mathbf{x}'$ and $\mathbf{Z}'$ depend on $\mathbf{w}$ and on the sequence $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$

of training points, too. Since the conditional distributions of ω are absolutely continuous in our case, it has no importance how one breaks ties in the definition of $\mathbf{x}'$ and $\mathbf{Z}'$.

Any ATS induces a classification rule given by $\mathbf{Z}'(\mathbf{x})$ whose error probability will be

$$P_e = P[\theta'(\omega) \neq \varrho] = 1 - \sum_{i=1}^{m} P[\theta'(\omega) = \varrho = i] =$$

$$= 1 - \sum_{j=1}^{m} \sum_{i=1}^{m} p_i \theta_{ij} v \int_{V_j} f_i(\mathbf{x})\, d\mathbf{x}\,, \tag{1}$$

where $\theta_{ij} = 1$ if $\theta_j = i$, $\theta_{ij} = 0$ if $\theta_j \neq i$, and V_j is the set of such values of $\mathbf{x}$ for which $\mathbf{x}$ will have nearest neighbor $\mathbf{x}_j$ (i.e. $\mathbf{x} \in V_j$ if j is the smallest integer minimizing $|\mathbf{x} - \mathbf{x}_j|_{\mathbf{w}}$, see [2], [3]). Since ties are broken in the definition of the nearest neighbor, the V_j's are disjoint convex sets jointly covering E_r; $\mathbf{Z}'(\mathbf{x})$ is constant on each of them. It is plain that any classification rule, including the Bayesian one, can be approximated by means of such nearest neighbor rules, provided that the ATS is large enough and the θ_j's are appropriately choosen. The most interesting cases, however, are those where a relatively small ATS can induce a very good classification. It seems that such situations are not so rare in practice (cf. the second example in Section 1).

We shall consider P_e as a function of $\mathbf{X}$ and $\mathbf{w}$, throughout this paper the θ_j's will be fixed. Our first algorithm minimizes P_e as a function of $\mathbf{X}$ (i.e. $\mathbf{w}$ is fixed), while the second one minimizes P_e as a function of $\mathbf{w}$. In both cases the stochastic gradient method will be used; this recursive procedure approaches local minima by moving in the direction opposite to the estimated value of the gradient vector at each step. The main problem is here to find a good estimate for the gradient of P_e. It is easy to check that P_e is a differentiable function of $\mathbf{X}$ and $\mathbf{w}$, provided that the conditional densities are continuous, but the expression of the derivatives is rather complicated. The gradient vector of P_e will be estimated by means of a sequence of kernels approximating the differential operator in the same sense as singular integrals do the unit operator.

3. The stochastic gradient method

Let $f(\mathbf{y})$ be a nonnegative, differentiable function on the k-dimensional Euclidean space E_k, $\mathbf{U}(\mathbf{y})$ denotes its gradient vector at $\mathbf{y} \in E_k$. Local minima of f are usually approximated by the gradient algorithm

$$\mathbf{y}_1 \in E_k,\ \mathbf{y}_{n+1} = \mathbf{y}_n - a_n \mathbf{U}(\mathbf{y}_n)\,, \qquad n = 1, 2, \ldots,$$

where the step length $a_n > 0$ need not depend on n. In our case neither f nor $\mathbf{U}$ (i.e. P_e and its gradient) are explicitly given, but we have a possibility to estimate them on the basis of the natural training sequence. The following stochastic version of the gradient algorithm will be applied to find the minimum of f:

$$\mathbf{Y}_1 \in E_k, \quad \mathbf{Y}_{n+1} = \mathbf{Y}_n - a_n \mathbf{Z}_{n+1}, \quad n = 1, 2, \ldots, \tag{2}$$

where $\mathbf{Z}_{n+1} \in E_k$ is an estimate of $\mathbf{U}(\mathbf{Y}_n)$ satisfying the conditions below.

Let $\mathfrak{A}_1, \mathfrak{A}_2, \ldots, \mathfrak{A}, \ldots$ be an increasing sequence of σ-algebras in an overall probability space $(\Omega, \mathfrak{A}, P)$ such that $\mathbf{Y}_1$ is measurable on $\mathfrak{A}_1$ and $\mathbf{Z}_{n+1}$ is measurable on $\mathfrak{A}_{n+1}$ for each n, then $\mathbf{Y}_n$ will be a random vector measurable on $\mathfrak{A}_n$. We suppose that $\mathbf{Y}_1$ has finite second moments and the a_n's are given positive numbers (e.g. $a_n = n^{-1}$). $E(\,.\,|\,\mathfrak{A}_n)$ denotes conditional expectation, given $\mathfrak{A}_n$, a Σ without specifications means summation over the set of natural numbers, $|\,.\,|$ and $(.\,,.)$ denote the usual Euclidean norm and —scalar product in E_k ($|\,.\,|$ may be different from $|\,.\,|_{\mathbf{w}}$). For the convergence of process (2) we need assumptions (A)–(F), where M, R, L are positive constants and $\mathbf{y}_0$ is a fixed point in E_n.

(A) $\quad \Sigma\, a_n = +\infty, \ \Sigma\, a_n^2 < +\infty.$

(B) $\quad \Sigma\ a_n \,|\, E(\mathbf{Z}_{n+1} \,|\, \mathfrak{A}_n) - \mathbf{U}(\mathbf{Y}_n) \,| < +\infty$ a.s.

(C) $\quad \Sigma\ a_n^2 E[\,|\mathbf{Z}_{n+1} - E(Z_{n+1} \,|\, \mathfrak{A}_n)|^2 \,|\, \mathfrak{A}_n] < +\infty$ a.s.

(D) $\quad |\,\mathbf{U}(\mathbf{y})\,|^2 \leq M(1 + \mathbf{y} - \mathbf{y}^0\,|^2)$ for each $\mathbf{y} \in E_k$.

(E) $\quad (\mathbf{y} - \mathbf{y}_0,\ \mathbf{U}(\mathbf{y})) \geq 0$ if $|\,\mathbf{y} - \mathbf{y}_0\,| \geq R$.

(F) $\quad |\,\mathbf{U}(\mathbf{y}) - \mathbf{U}(\mathbf{y}')\,| \geq L\,|\,\mathbf{y} - \mathbf{y}'\,|$ for $\mathbf{y},\ \mathbf{y}' \in E_k$.

As a consequence of Theorem 1 in [4] we obtain the following auxiliary result.

Theorem A. Under assumptions (A)–(F) we have $\lim \mathbf{U}(\mathbf{Y}_n) = 0$ for process (2). Moreover, if the stationary points of f are isolated from each other, then $\mathbf{Y}_n$ converges with probability one to one of them.

Let us remark that P_e has at least two stationary points, namely its minimum and maximum; this is why we need a stochastic approximation theorem for regression functions with several roots. This theorem is the corresponding version of Venter's [5] result for the Robbins–Monro process.

The estimates $\mathbf{Z}_{n+1}$ for $\mathbf{U}(\mathbf{Y}_n)$ will be constructed by means of a vector valued kernel $\mathbf{T}(\mathbf{u}) \in E_n$ defined on E_k. The necessary properties of $\mathbf{T}$ are the

following: such functions can be obtained by differentiating a symmetric density function with expectation $\mathbf{0} \in E_n$. For example,

$$\mathbf{T}(\mathbf{u}) = \mathbf{u}(2\pi)^{-k/2} \exp\left(-\frac{1}{2}|\mathbf{u}|^2\right)$$

satisfies (i)–(iv) below.

(i) $$\int_{E_k} \mathbf{T}(\mathbf{u})\ d\mathbf{u} = 0\,.$$

(ii) $$\int_{E_k} \mathbf{T}(\mathbf{u})\ (\mathbf{u}, \mathbf{U})\, d\mathbf{u} = \mathbf{U} \text{ if } \mathbf{U} \in E_k\,.$$

(iii) $$\int_{E_k} |\mathbf{u}|^2\ |\mathbf{T}(\mathbf{u})|\ du < +\infty\,.$$

(iv) $$\int_{E_k} |\mathbf{T}(\mathbf{u})|^2\ d\mathbf{u} < +\infty\,.$$

Lemma A is a special case of Lemmas 5 and 6 in [4].

Lemma A. If function f is bounded and satisfies (F), while (i)–(iv) hold for $\mathbf{T}$, then

$$|\mathbf{U}(\mathbf{y}) - \lambda^{k+1} \int_{E_k} T[\lambda(\mathbf{u} - \mathbf{y})]\ f(\mathbf{u})\ d\mathbf{u}| \leq \lambda^{-1} K_1\,, \tag{3}$$

$$\lambda^{2k+2} \int_{E_k} |T[\lambda(\mathbf{u} - \mathbf{y})]|^2 f(\mathbf{u})\ d\mathbf{u} \leq \lambda^{k+2} K_2\,, \tag{4}$$

where $\mathbf{y} \in E_k$, $\lambda > 0$ are arbitrary, and K_1 and K_2 do not depend on them.

Using this lemma we can construct an estimate for the gradient of P_e having bias $0(\lambda^{-1})$ and variance $0(\lambda^{k+2})$. Therefore, Theorem A can be applied only if λ goes to infinity in such a way that

(B′) $$\Sigma\, a_n\, \lambda_n^{-1} < +\infty,$$

(C′) $$\Sigma\, a_n^2\, \lambda_n^{k+2} < +\infty.$$

E.g. $a_n = \dfrac{1}{n}$ and $\lambda_n = (\log n)^2$ satisfy (A), (B′) and (C′).

4. Algorithm I

In this section we consider P_e as a function of the mr-dimensional vector $\mathbf{X}$ formed by the points $\mathbf{x}_k$ of the ATS, the weight vector and the labels in the ATS will be fixed. Denote $\widetilde{\mathbf{U}}_k(\mathbf{X})$ the vector of partial derivatives of P_e with respect to the coordinates of the kth point $\mathbf{x}_k$, $k = 1, 2, \ldots, m$, and let $\mathbf{T}(\mathbf{u})$

be a mapping of E_r into itself satisfying (i)–(iv) with $k = r$. At this point it has no importance how the Euclidean structure of E_r is choosen.

Let $\tilde{I}_j(\mathbf{x}, \mathbf{X})$ denote the indicator of V_j, that is $\tilde{I}_j = 1$ if $\mathbf{x}_j$ is the nearest neighbor of $\mathbf{x}$, $I_j = 0$ otherwise. Then (1) becomes

$$P_e = \tilde{P}_e(\mathbf{X}) = 1 - \sum_{i=1}^{s} \sum_{j=1}^{m} p_i \, \theta_{ij} \int_{E_r} \tilde{I}_j(\mathbf{x}, \mathbf{X}) \, f_i(\mathbf{x}) \, d\mathbf{x}$$

and in view of Lemma A we obtain the following approximate value $\tilde{\mathbf{U}}_{k\lambda}$ of the kth component of the gradient vector $\tilde{\mathbf{U}}(\mathbf{X}) = (\tilde{\mathbf{U}}_1(\mathbf{X}), \tilde{\mathbf{U}}_2(\mathbf{X}), \ldots, \tilde{\mathbf{U}}_r(\mathbf{X}))$ of $\tilde{P}_e$:

$$\tilde{\mathbf{U}}_{k\lambda}(\mathbf{X}) = - \sum_{i=1}^{s} \sum_{j=1}^{m} p_i \, \theta_{ij} \int_{E_r} \tilde{\mathbf{S}}_{jk}(\lambda, \mathbf{x}, \mathbf{X}) \, f_i(\mathbf{x}) \, d\mathbf{x} , \tag{5}$$

where

$$\tilde{\mathbf{S}}_{jk}(\lambda, \mathbf{x}, \mathbf{X}) = \lambda^{r+1} \int_{E_r} \tilde{I}_{jk}(\mathbf{x}, \mathbf{u}) \, \mathbf{T}(\lambda(\mathbf{u} - \mathbf{x}_k)) \, d\mathbf{u} \tag{6}$$

and $\tilde{I}_{jk}(\mathbf{x}, \mathbf{u}) = \tilde{I}_j(\mathbf{x}, \mathbf{X})$ with $\mathbf{u} = \mathbf{x}_k$ (the other components of $\mathbf{X}$ are fixed). (5) has been obtained from

$$\tilde{\mathbf{U}}_{k\lambda}(\mathbf{X}) = \lambda^{r+1} \int_{E_r} \mathbf{T}(\lambda(\mathbf{u} - \mathbf{x}_k)) \; \tilde{P}_{ek}(\mathbf{u}) \, d\mathbf{u}$$

by (i) and the Fubini theorem; $\tilde{P}_{ek}(\mathbf{u}) = \tilde{P}_e(\mathbf{X})$ with $\mathbf{x}_k = \mathbf{u}$.

We are now in a position to construct a process minimizing $\tilde{P}_e(\mathbf{X})$; the natural training sequence (ω_n, ϱ_n) will be used to estimate $\tilde{\mathbf{U}}(\mathbf{X})$. Define $\varrho_{in} = 1$ if $\varrho_n = i$, $\varrho_{in} = 0$ if $\varrho_n \neq$ i, and observe that $\tilde{U}_{k\lambda_n}(\mathbf{X})$ is just the expectation of the random vector

$$\mathbf{Z}_{k,n+1}(\mathbf{X}) = - \sum_{i=1}^{s} \sum_{j=1}^{m} \theta_{ij} \, \varrho_{im} \, \tilde{\mathbf{S}}_{jk}(\varrho_n, \omega_n, \mathbf{X}) . \tag{7}$$

Since $\tilde{\mathbf{S}}_{jk}$ is an explicitly given function, the following sequence can be computed recursively. Let $\mathbf{X}_1 \in E_{mr}$ be arbitrary and form

(I) $$\mathbf{X}_{n+1} = \mathbf{X}_n - a_n \tilde{\mathbf{Z}}_{n+1} , \qquad n = 1, 2, \ldots ,$$

where a_n and λ_n are given positive numbers satisfying (A), (B′), (C′) with $k = r$ and $\tilde{\mathbf{Z}}_{n+1} \in E_{mr}$ is an m-tuple having $\tilde{\mathbf{Z}}_{k,n+1}(\mathbf{X}_n)$ as its kth component.

Theorem 1. Suppose that the gradient vector $\tilde{\mathbf{U}}(\mathbf{X})$ of $\tilde{P}_e(\mathbf{X})$ satisfies (D), (E) and the uniform Lipschitz condition (F), then $\lim \tilde{U}(\mathbf{X}_n) = 0$ a.s. If the stationary points of $\tilde{P}_e$ are isolated from each other, then $\tilde{\mathbf{X}}_n$ converges to one of them with probability one.

Proof. Denote $\mathfrak{A}_n$ the minimal σ-algebra on which $(\omega_1, \varrho_1), (\omega_2, \varrho_2), \ldots$ $\ldots, (\omega_{n-1}, \varrho_{n-1})$ are measurable ($\mathfrak{A}_1$ is the trivial σ-algebra). Since (ω_n, ϱ_n) is an independent sequence, Lemma A implies that

$$| E[\tilde{\mathbf{Z}}_{k,n+1}(\mathbf{X}_n) \mid \mathfrak{A}_n] - \tilde{\mathbf{U}}_n(\mathbf{X}_n) | = 0(\lambda_n^{-1}),$$

$$E[\tilde{\mathbf{Z}}_{k,n+1}(\mathbf{X}_n)]^2 = 0(\lambda_n^{r+2}),$$

whence (B) and (C) follow directly by (B′) and (C), thus the statement is an immediate consequence of Theorem A.

It is more difficult to explain how this algorithm works. First we investigate the vectors $\tilde{\mathbf{S}}_{jk}(\lambda, \mathbf{x}, \mathbf{X})$ in the most natural case when $\mathbf{T}$ has a rotational symmetry with respect to $| \, . \, |_{\mathbf{w}}$. Suppose that $\mathbf{x}_j$ is the nearest and $\mathbf{x}_p$ is the second nearest neighbor of $\mathbf{x} = \omega_n$ in $\mathbf{X}$. Then $\tilde{I}_{jj} \neq 0$ if $| \mathbf{u} - \mathbf{x} |_{\mathbf{w}} < | \mathbf{x}_j - \mathbf{x} |_{\mathbf{w}}$, $\tilde{I}_{pj} \neq 0$ if $| \mathbf{u} - \mathbf{x} |_{\mathbf{w}} < | \mathbf{x}_j - \mathbf{x} |_{\mathbf{w}}$, and $I_{qj} = 0$ if $q \neq j$ or $q \neq p$; thus $\tilde{\mathbf{S}}_{jj} + \tilde{\mathbf{S}}_{pj} = 0$ in view of (i). Further, since $\mathbf{T}$ is symmetric, $\tilde{\mathbf{S}}_{ji}$ has the same direction as the vector $\mathbf{x} - \mathbf{x}_j$, and the direction of $\tilde{\mathbf{S}}_{pj}$ is the opposite one, while $\tilde{\mathbf{S}}_{qj} = 0$ for $q \neq j$ or $q \neq p$. Similarly, for $k \neq j$ we have that $\tilde{I}_{jk} \neq 0$ if $| \mathbf{u} - \mathbf{x} |_{\mathbf{w}} > | \mathbf{x}_j - \mathbf{x} |_{\mathbf{w}}$, $\tilde{I}_{kk} \neq 0$ if $| \mathbf{u} - \mathbf{x} |_{\mathbf{w}} < | \mathbf{x}_j - \mathbf{x} |_{\mathbf{w}}$, $\tilde{I}_{qk} = 0$ if $q \neq j$ or $q \neq k$, whence $\tilde{\mathbf{S}}_{jk} + \tilde{\mathbf{S}}_{kk} = 0$, $\tilde{\mathbf{S}}_{qk} = 0$ if $q \neq j$ or $q \neq k$, and $\tilde{\mathbf{S}}_{jk}$ is directed as $\mathbf{x}_k - \mathbf{x}$, $\tilde{\mathbf{S}}_{kk}$ as $\mathbf{x} - \mathbf{x}_k$.

Therefore, the algorithm modifies the ATS on the bass of (ω_k, ϱ_k) as follows: we discuss the behavior of the nearest neighbor (the jth point) and of other points, separately. If $\varrho_n = \theta_p = 0$ or $\varrho_n \neq \theta_j$, $\varrho_n \neq \theta_p$, then the nearest neighbor of ω_n remains unchanged, while it moves towards ω_n or in the opposite direction, according to $\varrho_n = \theta_j \neq \theta_p$ or $\varrho_n = \theta_p \neq \theta_j$. Similarly, the kth point of the ATS ($k \neq j$) moves towards ω_j or in the opposite direction if $\varrho_n = \theta_k \neq \theta_j$ or $\varrho_n = \theta_j \neq \theta_k$, respectively, and nothing happens with it otherwise.

Roughly speaking, Algorithm I approaches local minima of $\tilde{P}_e$ in such a way that the points of the ATS get closer to or farther from the considered training point according as their labels are the same or not as given in the natural training sequence. Due to the stochastic gradient principle, the step lengths here have been determined in such a way that this difficult motion results in a convergence to the set of stationary points of $\tilde{P}_e(X)$.

5. Algorithm II

In this section the ATS will be fixed and we consider P_e as a function $P_e = \hat{P}_e(\mathbf{w})$ of the weight vector $\mathbf{w} \in E_r$. Denote $\hat{\mathbf{U}}(\mathbf{w}) = (\hat{U}_1(\mathbf{w}), \hat{U}_2(\mathbf{w}), \ldots, \hat{U}_r(\mathbf{w}))$ the gradient of P_e and let $T(u)$ be a real function satisfying (i)–(iv) with $k = 1$. The usual norm and scalar product in E_r are denoted by $|\ .\ |$ and $(\ .\ ,\ .\ )$, respectively; $|\ .\ |_{\mathbf{w}}$ is the Euclidean norm defined by the weight vector $\mathbf{w}$.

Observe first that $\hat{P}_e$ is a homogeneous function of $\mathbf{w}$; i.e. $\hat{P}_e(\mathbf{w}) = \hat{P}_e(c\mathbf{w})$ for any real number c, whence it follows by differentiation that

$$(\mathbf{w},\ \hat{\mathbf{U}}(\mathbf{w})) = 0 \quad \text{and} \quad c\ \hat{\mathbf{U}}(c\mathbf{w}) = \hat{\mathbf{U}}(\mathbf{w}). \tag{8}$$

The second equality shows that $\hat{P}_e$ is not differentiable at $\mathbf{w} = 0$; nevertheless we can proceed almost in the same way as above.

Let $\hat{I}_j(\mathbf{x}, \mathbf{w})$ denote the indicator of V_j, that is $I_j = 1$ if the jth point of the ATS is the nearest neighbor of $\mathbf{x}$ with respect to $|\ .\ |_{\mathbf{w}}$, $\hat{I}_j = \mathbf{d}$ otherwise. Then

$$\hat{P}_e(\mathbf{w}) = 1 - \sum_{i=1}^{s} \sum_{j=1}^{m} p_i\, \theta_{ij} \int_{E_r} \hat{I}_j(\mathbf{x},\ \mathbf{w})\ f_i(\mathbf{x})\ \mathrm{d}\mathbf{x},$$

whence by a formal application of Lemma A we obtain an approximation $\hat{U}_{k\lambda}$ for the kth coordinate of $\hat{\mathbf{U}}_\lambda$ defined as follows:

$$\hat{U}_{k\lambda}(\mathbf{w}) = -\sum_{i=1}^{s} \sum_{j=1}^{m} p_i\, \theta_{ij} \int_{E_r} \hat{S}_{jk}(\lambda,\ \mathbf{x},\ \mathbf{w})\, f_i(\mathbf{x})\ \mathrm{d}\mathbf{x}, \tag{9}$$

where

$$\hat{S}_{jk}(\lambda,\ \mathbf{x},\ \mathbf{w}) = \lambda^2 \int_{E_1} \hat{I}_{jk}(\mathbf{x},\ u)\, T(\lambda(u - w_k))\ \mathrm{d}u \tag{10}$$

and $\hat{I}_{jk}(\mathbf{x},\ u) = \hat{I}_j(\mathbf{x},\ \mathbf{w})$ with $u = w_k$. $\hat{S}_{jk}$ is an explicitly given function of its variables and $\hat{U}_{k\lambda_n} f(\mathbf{w})$ is just the expectation of

$$\hat{Z}'_{k,n-1} = -\sum_{i=1}^{s} \sum_{j=1}^{m} \theta_{ij}\, \varrho_{in}\, \hat{S}_{jk}(\lambda_n, \omega_n, \mathbf{w}), \tag{11}$$

thus we may hope that the vector $\hat{\mathbf{Z}}'_{n+1}$ formed by these components will be a good estimate of $\hat{\mathbf{U}}(\mathbf{w})$.

Some unnecessary difficulties will be avoided by taking into consideration the first statement of (8). Indeed, as $\hat{\mathbf{U}}(\mathbf{w})$ is orthogonal to $\mathbf{w}$, it is reasonable

to consider an estimate of $\hat{\mathbf{U}}$ satisfying this condition. This is why we define Algorithm II as follows: $\mathbf{W}_1 \in E_r$ is arbitrary with $|W_1| = 1$, and

$$\mathbf{W}_{n+1} = \mathbf{W}_n - a_n \hat{\mathbf{Z}}_{n+1}, \qquad n = 1, 2, \ldots, \tag{II}$$

where $\hat{\mathbf{Z}}_{n+1}$ is obtained from (11) by putting $\mathbf{w} = \mathbf{W}_n$,

$$\hat{\mathbf{Z}}_{n+1} = \hat{\mathbf{Z}}'_{n+1} - |\mathbf{W}_n|^{-2}(\mathbf{W}_n, \hat{\mathbf{Z}}_{n+1})\,\mathbf{W}_n \tag{12}$$

is the projection of $\hat{\mathbf{Z}}'_{n+1}$ to the hyperplane orthogonal to $\mathbf{W}_n$, further a_n and λ_n are given positive numbers satisfying (A), (B′), (C′) with $k = 1$. Since $\hat{\mathbf{Z}}_{n+1}$ is orthogonal to $\mathbf{W}_n$, we obtain from (II) that

$$|\mathbf{W}_{n-1}|^2 = |\mathbf{W}_n|^2 + a_n^2 |\hat{\mathbf{Z}}_{n+1}|^2 = 1 + \sum_{i=1}^{n} a_i^2 |\hat{\mathbf{Z}}_{i+1}|^2 \leq 1 + \sum_{i=1}^{n} a_i^2 |\hat{\mathbf{Z}}'_{i+1}|^2,$$

but from (10) it follows that

$$|\hat{S}_{jk}(\lambda, \mathbf{x}, \mathbf{w},)| \leq \lambda \int E_1 |T(u)|\,\mathrm{d}u = \lambda K_3$$

independently of j, k, $\mathbf{x}$, $\mathbf{w}$; thus $|\hat{\mathbf{Z}}_{k,n+1}| \leq \lambda_n m K_3$, whence

$$1 \leq |\mathbf{W}_{n+1}|^2 \leq 1 + r m^2 K_3^2 \sum_{i=1}^{\infty} a_i^2 \lambda_i^2 < +\infty. \tag{13}$$

This means that the sequence $\mathbf{W}_n$ is bounded away from $\mathbf{0}$ where $\hat{P}_e$ is not even continuous. On the other hand, $\lim \hat{\mathbf{U}}(\mathbf{W}_n) = \mathbf{0}$ shall imply that the limit points of the algorithm will be stationary points of $\hat{P}_e$, although $\lim \hat{\mathbf{U}}(\mathbf{W}_n) = 0$ even if $\lim |\mathbf{W}_n| = +\infty$. Now we are in a position to prove that Algorithm II approaches the set of stationary points of $\hat{P}_e$.

Theorem 2. Suppose that $\hat{P}_e$ is twice differentiable with bounded second derivatives on the surface of the unit sphere of E_r. Then $\lim \hat{\mathbf{U}}(\mathbf{W}_n) = 0$ a.s., and $\mathbf{W}_n$ converges a.s. to one of the stationary points of $\hat{P}_e$ if they are isolated from each other.

Proof. We shall apply Theorem A to the function $\hat{P}'_e(\mathbf{w}) = \varphi(|\mathbf{w}|^2)\,\hat{P}_e(\mathbf{w})$, where φ is a twice differentiable real function with a bounded second derivative, $\varphi(t) = 0$ if $t \leq 0$, $\varphi(t) = 1$ if $t \geq 1/2$. Denote $\sigma(t)$ the first derivative of φ; since $0 = \sigma(0)$ and the second derivative of φ is bounded, $\varphi(|\mathbf{w}|^2) = 0(|\mathbf{w}|^4)$ if $\mathbf{w}$ goes to $\mathbf{0}$, while $0 \leq \hat{P}_e \leq 1$, $\hat{P}_e$ will be differentiable even at $\mathbf{w} = 0$ (cf. (8)) and its gradient vector is

$$\hat{\mathbf{U}}'(\mathbf{w}) = 2\,\varrho(|\mathbf{w}|^2)\,\hat{P}_e(\mathbf{w})\,\mathbf{w} + \varphi(|\mathbf{w}|^2)\,\hat{\mathbf{U}}(\mathbf{w}). \tag{14}$$

Of course, $\widehat{P}'_e(\mathbf{0}) = 0$, $\widehat{\mathbf{U}}'(\mathbf{0}) = \mathbf{0}$ and $\widehat{P}'_e(\mathbf{w}) = \widehat{P}_e(\mathbf{w})$, $\widehat{\mathbf{U}}'(\mathbf{w}) = \widehat{\mathbf{U}}(\mathbf{w})$ if $|\mathbf{w}| \geq$ $\geq 1/2$. Now the validity of (D) and (F) is an immediate consequence of the boundedness of the derivatives of $\widehat{P}_e$ on the unit sphere and of the second equality in (8), while (E) follows from the first statement of (8) with $\mathbf{y}_0 = \mathbf{0}$ and $R = 1/2$. (C) is also obviously true, since $|\widehat{\mathbf{Z}}_n| \leq |\widehat{\mathbf{Z}}_n| = 0(\lambda_{n+1})$ and a_n, λ_n satisfy (C′).

The basic assumption of Theorem A is (B), it will be deduced from Lemma A. Let A_n denote the σ-algebra generated by (ω_1, ϱ_1), (ω_2, ϱ_2), $\ldots$, $(\omega_{n-1}, \varrho_{n-1})$, then

$$E(\widehat{\mathbf{Z}}_{n+1}/A_n) = \widehat{\mathbf{U}}_{\lambda n}(\mathbf{W}_n),$$

where the kth coordinate of $\widehat{\mathbf{U}}_\lambda$ is defined by (9); we have to show that

$$|\widehat{\mathbf{U}}_{\lambda n}(\mathbf{W}_n) - \widehat{\mathbf{U}}'(\mathbf{W}_n)| = 0(\lambda_n^{-1}), \tag{15}$$

since from (8) and (12) it follows that

$$\begin{aligned}
&|E(\widehat{\mathbf{Z}}_{n+1} \mid A_n) - \widehat{\mathbf{U}}'(\mathbf{W}_n)| = \\
&= |\widehat{\mathbf{U}}_{\lambda n}(\mathbf{W}_n) - \widehat{\mathbf{U}}'(\mathbf{W}_n) - |\mathbf{W}_n|^{-2}(\mathbf{W}_n, \widehat{\mathbf{U}}_{\lambda n}(\mathbf{W}_n))\,\mathbf{W}_n| = \\
&= [\,|\widehat{\mathbf{U}}_{\lambda n}(\mathbf{W}_n) - \widehat{\mathbf{U}}'(\mathbf{W}_n)|^2 - |\mathbf{W}_n|^{-2}(\mathbf{W}_n, \widehat{\mathbf{U}}_{\lambda n}(\mathbf{W}_n))^2]^{1/2} \leq \\
&\leq |\widehat{\mathbf{U}}_{\lambda n}(\mathbf{W}_n) - \widehat{\mathbf{U}}'(\mathbf{W}_n)|.
\end{aligned}$$

We know that $|\mathbf{W}_n| \geq 1$, thus $\widehat{\mathbf{U}}(\mathbf{W}_n) = \widehat{\mathbf{U}}'(\mathbf{W}_n)$ for each n. Therefore, it is enough to show that

$$|\widehat{\mathbf{U}}_{k\lambda}(\mathbf{w}) - \widehat{\mathbf{U}}_n(\mathbf{w})| = 0(\lambda^{-1}) \text{ if } |\mathbf{w}| \geq 1. \tag{16}$$

Set $\widehat{P}_{ek}(u) = \widehat{P}_e(\mathbf{w})$ and $\widehat{P}'_{ek}(u) = \widehat{P}'_e(\mathbf{w})$ with $u = w_k$ (the other coordinates of $\mathbf{w}$ are fixed), then P_{ek} satisfies the conditions of Lemma A if $w_i \neq 0$ for some $i \neq k$, but in view of (9),

$$\widehat{U}_{k\lambda}(\mathbf{w}) = \lambda^2 \int_{E^1} \widehat{P}_{ek}(u)\, T\big(\lambda(u - w_k)\big)\, \mathrm{d}u\,,$$

whence (16) follows by Lemma A in this case. On the other hand, if $w_i = 0$ for $i \neq k$, then $|w_n| \geq 1$, thus

$$\begin{aligned}
&\lambda^2 \Big|\int_{E_1} (\widehat{P}_{ek}(u) - \widehat{P}_{ek}(u))\, T(\lambda(u - w_k))\, \mathrm{d}u\Big| \leq \\
&\leq \lambda^2 \int_{E_2} 4\,(u - w_k)^2\, |\,T(\lambda(u - w_k))\,|\,\mathrm{d}u\,| = \\
&= \lambda \mathrm{V}_1 \int_{E^1} u^2\, |\,T(u)\,|\,\mathrm{d}u,
\end{aligned}$$

since $\widehat{P}'_{ek}(u) = \widehat{P}_{ek}(u)$ if $|u| \geq 1/2$ and $4(u - w_k)^2 \geq 1 \geq |\widehat{P}'_{ek}(u) - \widehat{P}_{ek}(u)|$ otherwise. Taking into account that P'_{ek} satisfies (F) and $\widehat{\mathbf{U}}_\lambda(\mathbf{w}) = \widehat{\mathbf{U}}'_\lambda(\mathbf{w})$, we obtain (16) applying Lemma A to $\widehat{P}_{ek}(u)$.

We have shown that $\widehat{P}_e$ and Algorithm II satisfy the conditions of Theorem A, thus $\lim \widehat{\mathbf{U}}'(\mathbf{W}_n) = \mathbf{0}$ a.s. However, $\widehat{\mathbf{U}}(\mathbf{W}_n) = \widehat{\mathbf{U}}'(\mathbf{W}_n)$ for each n, which proves the statement of Theorem 2.

The heuristic background of Algorithm II is much more complicated than that of Algorithm I. We are not in a position to discuss this problem in details. Let us point out only the main effect of the nth step, when $\varrho_n = \theta_j$ and the jth training point of the ATS is the nearest neighbor of ω_n: the components of the weight vector are decreased or increased, according as the nearest neighbor minimizes, or does not do, the distance between the corresponding coordinates of ω_j and of the points in the ATS.

It seems that the coordinates associated with large weights in case of an optimal weight vector are the most important ones from a point of view of classification. Although this statement cannot be justified by theoretical arguments, this method may be a useful tool in solving data reduction (feature selection) problems.

$\widehat{\mathbf{Z}}_{n+1}$ would have been defined by means of an r-dimensional $\mathbf{T}$ in the same way as in Algorithm I. This choice, however, would involve some difficulties in computing the corresponding $\widehat{\mathbf{S}}_{jk}$'s. In our case only integrals of T over intervals should be calculated, and the limits of these intervals can be determined from the ATS directly.

Algorithms I and II can be used simultaneously. Theorem A implies as convergence of this compound algorithm, too. Unfortunately, it is not powerful enough to deduce convergence of such methods where also the labels of the ATS are corrected from step to step.

References

1. *Cover, T., Hart,* P. (1967) Nearest neighbor pattern classification, IEEE **IT–13,** 21–27.
2. *Wagner, T.* (1971) Convergence of the nearest neighbor rule, IEEE **IT–17,** 566–571.
3. *Fritz, J.* (1975) Distribution-free exponential error bound for nearest neighbor pattern classification, IEEE **IT–21** (to appear)
4. *Fritz, J.* (1973) Stochastic approximation for finding local maxima of probability densities, Studia Sci. Math. Hungarica **8,** 309–322.
5. *Venter, J. H.* (1966) On Dvoretzky stochastic approximation theorems, Ann. Math. Stat. **37,** 1534–1544.

Два способа оптимизации при использовании метода классификации по правилам «ближайших соседей»

Й. ФРИЦ

(Будапешт)

Как известно, при использовании метода классификации по правилам »ближайших соседей« выявляются положительные асимптотические свойства, но в то же время ошибка классификации зависит в существенной мере от выбора точек обучения и меры, определяющей ближайшего соседа в случае исследования ограниченного количества проб.

Для минимизации вероятности ошибочной классификации предложены два стохастических градиентных алгоритма. В обоих алгоритмах правило классификации задается в виде »искусственной« последовательности обучения, объем которого зафиксирован. В первом алгоритме вероятность ошибочного решения минимизируется в зависимости от множества искусственных точек обучения, во втором — предлагается методика приближения оптимальной меры пространства проб. Методика оптимизации в обоих алгоритмах основывается на применении «естественной» последовательности обучения; доказана сходимость методов с вероятностью 1. Второй метод применяется и для решения проблем автоматического выбора показателей качества.

J. Fritz
Mathematical Research Institute
Hungarian Academy of Sciences
H–1053 Budapest, V. Reáltanoda u. 13—15.
Hungary

Problems of Control and Information Theory, Vol. 5 (3), pp. 239–246 (1976)

SIMULATION MODELS OF COMPUTING SYSTEMS WITH TWO DIFFERENT MEMORY STRUCTURES

D. DONKOV, L. GLUSHKOV

(Sofia)

(Received 25 April, 1974)

A number of potential bottlenecks comprises the configuration of commonly used on-line and real-time computing systems, where some big queues of messages could arise at peak loads on the system. The corollary of these are congestion of the whole system and drastic detention of the response time. Thus the total throughput capability of the system could be determined by the throughput capabilities of a particular bottleneck. In a well-selected computer configuration, the different bottlenecks should evenly balance and should have adequate throughput capabilities. With the on-line and real-time systems, the secondary memory appears to be one of the general bottlenecks. The importance of this critical point for the systems being discussed, is determined by the fact that in a number of cases the read and write time for information stored in files on the secondary memory, is the basic, longest lasting component of the response time of systems in general. This is valid in particular for business-oriented on-line systems, where each transaction being processed, requires a number of accesses to the files. On the other hand, the accumulation of transaction queues, waiting for servicing by the secondary memory, results in additional loading of the primary memory that is one of the most critical resources in modern computers. The discussed systems show a doubtless tendency for increasing the number and capacity of the direct access units used as secondary memory, and their value tends to comprise a considerable amount from the total prise of the computer.

For a rationally structured secondary memory of a particular computer system attending a determined application, it is necessary to create a memory hierarchy, that is, to ensure the necessary efficiency with the possible least cost in mind.

As it is shown in [1], the analysis of systems with hierarchically structured memory is performed at two stages. Originally models of the studied hierarchic structures are created that permit for quantitative determination of the most important characteristics. Then some value parameters are incorporated, allowing for the selection of the most suitable variant.

The basic characteristics supposed to be quantitatively determined at the first stage of the analysis are:

(a) What is the mean time a request spends in the system, its standard deviation, and what is practically the maximum possible time that a request spends in the system at different arrival rate.

(b) How are well balanced the elements of the tract (channel-control unit-storage) and where are the longest queues of requests formed.

(c) The sensitivity of the systems on the work load changes (for instance, increase of the arrival rate of requests).

The analysis of the hierarchic systems could be performed by analytical models using the queueing theory with certain approximation, as the reliability of the results in a number of cases deviates considerably with the increase of the system utilization. On the other hand, the practical analytical models are inevitably oriented on particular structures and specific modes of work; facts that exclude versatility and make impossible their vast application for the analysis of arbitrary computing systems [2].

Nowadays for the study of such system simulation methods are used; though time-consuming, they offer possibility for wider scope studies and more reliable results. In this paper, the principles and results of two simulation programs, describing the systems' performance with different structure of the secondary memory are discussed. The studies are carried out in an environment typical for on-line and real-time systems.

Description of the simulated systems

It is supposed that the basic information files are indexly recorded on disc units. This is considered to be one of the most popular type of organization with the on-line systems.

Two simulation programs — model "A" and model "B" were created. With model "A", the index tables are recorded on the same disc unit with the information files. They are situated in the midlle of the seek area on each disc pack [3] to reduce the mean access time to the index tables. With model "B" only the information files are recorded on disc pack, while the index tables are situated at a higher level of the secondary memory: a magnetic drum or head-per-track disc unit. In both models all disc units are connected to the channel via one control unit. It is considered several cases — with 2, 3, 4, 6 and 8 disc units incorporated to the system.

In view of versatility, the technical characteristics of units being used are inputs to the programs as input data. The simulations were based on the

ES 5052 disc unit, made in Bulgaria, and ES 5033 magnetic drum made in the USSR, as a secondary memory to the Unified System of Computers.

Model "A" performs as follows:

1. A request for reading information from the file is generated. The input traffic of requests is Poissonian. A pseudorandom number generated determines the particular disc unit where the request arrived newly should be directed. This request joins a queue, waiting for the unit to be released if busy at the moment of arrival.

2. After the release of any disc unit, a check is performed to determine whether more requests are waiting in the queue. If any, the first one in turn is being serviced. Another check is performed to find out whether the channel is free. If busy, the current activity is being awaited to complete. Then follows an order for moving the read-write heads, from their current position, to the tracks where the required index tables are situated.

3. After the time necessary for positioning the heads expires, the request joins a queue where is awaiting for the channel to be released, if busy. Since the request occupies the channel, the rotational latency time, which is generated in the model as a pseudorandom number, starts running. A reading of the index tables follows and the information is sent to the Central Processing Unit. The read in, from the index table, address generates an order for moving the read-write heads to the required track. In this particular version it is supposed that there is even probability for access to the different records, though the model allows for considering the difference of the access activities on particular records on the various cylinders of the disc pack. After the completion of the above activities, the request releases the channel.

4. Since the time necessary for transition of the read-write heads to the track, where the necessary information is recorded, expires, the release of the channel is awaited and the rotational latency time added. In the model this is achieved in a manner that follows the searching in the index tables described in item 3. The required record is read in and information is transferred to the Central Processor.

The record length in the model can be specified as certain probabalistic distribution. In this case the read time is assumed to be constant. Since all these activities are completed, the request releases the channel and queues up for processing by the Central Processor.

5. The CPU processing time in this particular case is considered uniformly distributed. It is also accepted that the CPU is performing no other activities, but servicing the requests. As the time for queueing up in the CPU is considerably short as compared to the total request's servicing time, the

influence of this simplification has no considerable effect on the behaviour of the model.

6. As a result of the processing, it is determined in the CPU whether a particular request will update or not the record being read. In this variant of the model an equal probability for both activities is accepted. All requests which will not perform any modernization, leave the system after being processed.

7. Requests, which will perform updating on the information are queueing up, waiting for the channel to be released. When the channel is occupied by a request, the rotational latency is also considered. When computing this time, it is realized that it is the same record to be updated, from which the information has been read and, therefore, the updating could be performed only after an integer number of revolutions of the disc pack made after reading. This is achieved in the model by considering the CPU processing time and the queueing time of releasing the channel for each request. As a general practice in many systems, in the following revolution a check read procedure is performed.

8. After a preliminary determined number of requests is processed, the input arrival rate is increased. The sequence of the above activities is repeated until a quasi-saturation point is reached. This is followed by an increase of the number of disc units, the model is set to the starting point and the run commences with the new parameter set.

The performance of model "B" differs from the above scheme only in the manner of reading the index tables. In that case they are recorded on a magnetic drum. Hence the request is joining a queue, waiting for the drum to be released, if busy at that moment. Other way the processing of the request commences immediately.

After the drum is released, the processing of the next request that follows in the queue immediately starts. Pseudorandom numbers are used to determine the drum sector where the necessary index table is situated. The rotational latency time is evaluated, timing the period for the required index table to appear under the read-write heads. For this reason, in model "B", the current angle declination of the drum is permanently considered.

As to make correlation on equal bases for the results obtained by the two simulation programs, it is accepted in model "B" that only one drum is connected to the control unit, and only the index tables are recorded on the drum.

Both models are programmed using the simulation language C.S.L. All runs were performed on ICL-1904 A computer.

Results

Some preliminary experiments were made to determine the time necessary for stabilizing the models. To achieve more quickly the desired stability, the models were not set to the initial condition with the change from lower to higher arrival rates, thus the queues of requests formed at each unit were preserved. It was determined the satisfactory stability is reached after processing of some 3000 requests.

The following characteristics of the studied systems were determined: mean time requests spend in the system, waiting and being served $\bar{T}_q$; the standard deviation of the above time σ_{T_q}; mean time for servicing the request; mean awaiting time for releasing of the disc unit; mean channel-servicing time; average number of requests, waiting for the channel to be released; the probability for the channel to be free; the probability for immediate servicing of the request (no waiting); average number of requests queueing up for servicing by the disc unit; the maximum time a request spends in the system $T_{q_{max}}$, etc.

In addition the following characteristics were obtained by model "B": average number of requests waiting for the drum to become free; mean time for waiting and servicing of a request by the drum.

All characteristics were obtained with different arrival rate and with different number of the disc units in the studied systems.

Figure 1 shows the mean time a request spends in the system for both models, depending on the arrival rate. It is evident from the given characteristics that organizing the index tables separately on magnetic drum, results in increased throughput capabilities of the investigated systems when servicing requests for access to the files. The throughput capabilities for a particular application increase by 8—9 requests per second when simulated by model "B", in comparison to the system simulated by model "A". This is an increase of about 45—90% depending on the number of the disc units being used.

A considerable cut down of the mean time requests spend in the system $\bar{T}_q$ is observed also. With one request per second initial arrival rate, the reduction of $\bar{T}_q$ is about 20%. This cut down increases with the growth of the arrival rate and in areas where the curves for $\bar{T}_q$ in model "A" reach saturation, the reduction is approximately doubled in "B".

With both models, the throughput capabilities do not increase in proportion to the number of the disc units K.

By the increase of K from 6 to 8, the general throughput ability of the system increases negligibly, which is especially expressed in model "B".

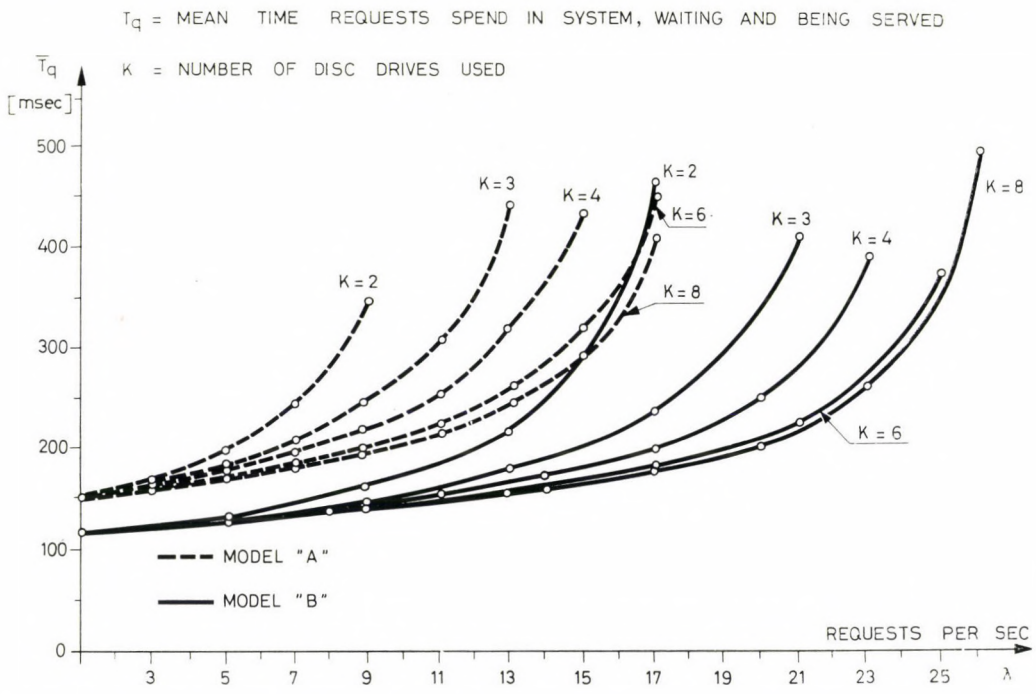


Fig. 1. Mean time requests spend in the system T_q versus the arrival rate λ, with different number of disc units in the system K

The reason is the load on the channel, which at $K > 4-5$ and with the increase of the arrival rate becomes the basic curbing factor on the general throughput capability. On the other hand, the use of a greater number of disc units contributes for the reduction of $\overline{T}_q$ especially at higher arrival rate. The main reason to the above phenomenon is the considerable reduction of the mean awaiting time for a disc unit to become free.

The analysis of the curves shown in Fig. 2 about the standard deviation of the time the requests spend in the system (σ_{T_q}) demonstrates, in this particular case that the recording of index tables on magnetic drum or any other unit of similar type, offers considerable advantages. A significant reduction is noted with model "B" which, for the systems being considered, is of importance. In the areas where the systems are approaching saturation, a drastic rise of σ_{T_q} is observed with both models, but the system simulated with model "B" is reaching the point at considerably higher arrival rate.

The histograms of the time the requests spend in the system of both models show an increase of the maximum time $T_{q_{\max}}$ from 2.5 $\overline{T}_q$ to approxi-

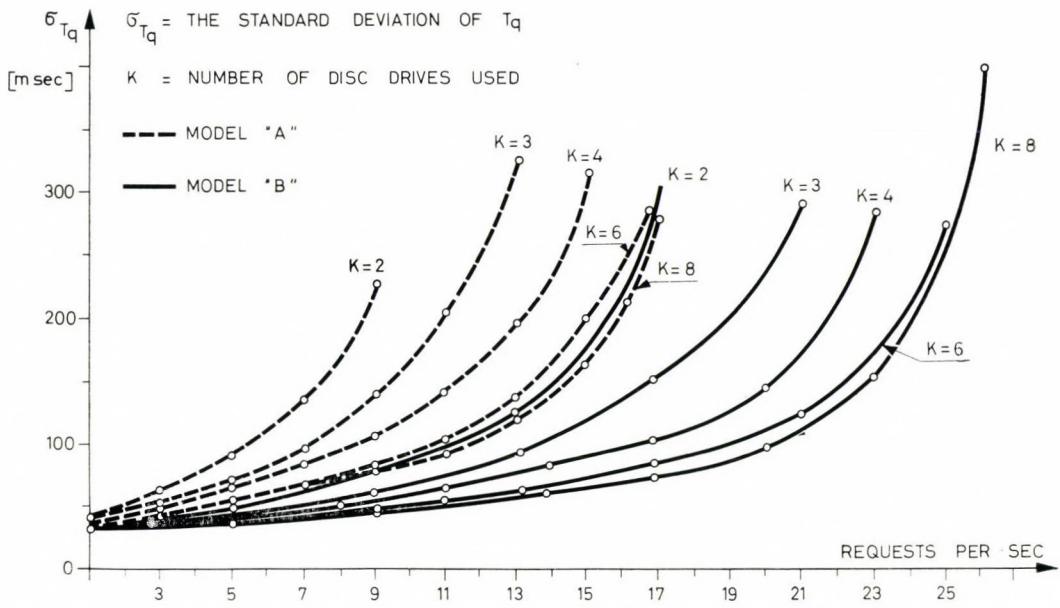


Fig. 2. Standard deviation of the mean time requests spend in the system σ_{Tq} versus arrival rate, with different number of disc units in the system K

mately 6 $\bar{T}_q$ with the increase of the arrival rate and depends relatively slightly on the number of the disc units used.

The corelation between stability of servicing the requests and changes in the work load is of great significance for the on-line and real-time systems — e.g. in what extent changes the time requests spend in the system with the increase of the arrival rate. In this particular case an indicative for stability is the slope of the graphs for $\bar{T}_q$. The advantages of the system, simulated by model "B", are evident; $\bar{T}_q$ increases remarkably more monotonously with the increase of the arrival rate and smoothly goes into saturation. With both models the stability of the suggested servicing increases with the number of disc units used in the system.

The investigation of the characteristics derived from the models shows that the throughput capabilities of the different elements of the tract (channel-controller-storage units) vary considerably with the number of the disc units being used. Using two or three disc units in the system, the basic curbing factor on the total throughput capability in both models are the disc units, where queues are formed and whose average length extends with the arrival rate. On the other hand, having five or more disc units, the total throughput capability is determined by the queues, waiting for service by the channel.

For this particular case, the system is most successfully balanced using four disc units connected to one disc control unit with model "A", and 5—6 disc units connected to one disc control unit with the system simulated by model "B". The queues of requests having such a configuration, are almost evenly distributed in front of the different elements comprising the discussed tract. If an additional number of disc units is required, say for extention of the files, rather than the optimum, better results will be achieved if these units are connected to more disc control units and have separate channels, in the case when the type of the computer used, allows for that.

The quantitative analysis performed, using the two simulation models, allows for selection of a more suitable variant for a particular application by introducing value parameters.

References

1. *Gecsei, J., Lukes, J. A.*, A Model for the Evaluation of Storage Hierarchies, "IBM Syst. J." 1974, **13**, 2, 163—178.
2. *Estrin, G., Muntz, R. R., Uzgalis, R. C.*, Modeling Measurement and Computer Power, "AFIPS Cond. Proc Vol. **40**, Spring Joint Comput. Conf. Atlantic City, N. Y., 1972" Montvale, N. Y., 1972, 725—738
3. *Saleeb, S., Riad, M.*, Use of Simulation for Operating System Optimization, Information Processing 74, **2**, 1974, 281—285.

Модели вычислительных систем с двумя различными структурами памяти

Д. ДОНКОВ, Л. ГЛУШКОВ

(София)

Разработаны две моделирующие программы для вычислительных систем с двумя различными структурами памяти, используемые в реальном масштабе времени, и проведено их исследование. В первой модели рабочие массивы и таблицы индексов расположены на дисках. Во второй модели таблицы индексов вынесены на более высокий уровень иерархии памяти — на магнитный барабан.

При пуссоновском распределении входного потока получены зависимости среднего времени пребывания заявок в системе и его среднеквадратического отклонения как функции входной интенсивности заявок при различном числе подключенных к системе дисков. Установлено повышение пропускной способности системы на 45 — 90W при использовании иерархической структуры памяти с барабаном. Показано лимитирующее влияние канала при увеличении числа дисков.

При моделировании использовались диски типа ЕС-5052, изготовляемые в Болгарии, и магнитные барабаны типа ЕС-5033, созданные в СССР, а также вторичная память унифицированной системы вычислительных машин.

Д. Донков, Л. Глушков
ВЦ АОНСУ; Институт социального управления
ул. Пионерски път 21
София
Болгария

Problems of Control and Information Theory, Vol. 5 (3), pp. 247–261 (1976)

ЛОГИКА ПОВЕДЕНИЯ ЦЕЛЕНАПРАВЛЕННЫХ СИСТЕМ

Е. И. ЕФИМОВ

(Москва)

(Поступила в редакцию 20 декабря 1974 г.)

Создание теории целенаправленных систем требует решения двух проблем: 1) построения семиотической аксиоматико-дедуктивной логики управления и 2) разработки на языке вышеуказанной логики семиотической целенаправленной системы. В работе рассматривается лишь первая проблема, относящаяся к логике поведения целенаправленных систем. Строится языковая система, включающая в себя языки различных порядков, связанных между собой трансляционными языками. Приводится синтаксис, семантика и прагматика такой системы и дается понятие вывода.

Довольно типичным актом поведения целенаправленной системы является формирование ею решения следующей задачи. Задана исходная ситуация и цель деятельности системы. Цель в общем случае отражает только некоторые желаемые свойства конечной ситуации. Системе необходимо сформировать решение задачи достижения указанной цели в виде некоторой последовательности действий и реализовать её. Сформулированная задача является задачей ситуационного управления, получившего в последнее время широкое распространение в исследовании сложных систем [1, 2].

Особенностью данной задачи в отличие от ранее решаемых в ситуационном управлении является то, что требуемое решение, как правило, не выбирается из числа априори заданных, а формируется заново. Таким образом, в зависимости от поставленной перед системой цели и исходной ситуации система должна будет вырабатывать свою последовательность действий, которая в дальнейшем при аналогичных задачах может использоваться как уже готовое решение.

Формирование вышеуказанных решений конкретных задач поведения системы будем представлять как автоматическое доказательство теорем в некоторой логической системе.

Здесь возможны два пути. Первый путь состоит в развитии эффективных методов дедуктивного вывода в рамках некоторых модификаций формальных систем математической логики. Примером первого пути служат многочисленные американские работы, посвященные различным модификациям алгоритмов поиска доказательства, основанного на методах резолю-

ции и парамодуляции [3-9]. Пока эти алгоритмы работают в логиках первого порядка. Однако это обстоятельство по признанию самих американских ученых [9, 10] значительно снижает эффективность поиска доказательства из-за излишней детализации вывода, что делает вывод громоздким и непривычным для человека. К этому необходимо добавить, что и сами методы избирательного поиска доказательства требуют привлечения различных эвристических приемов. Попытки разработать универсальные методы резолюнции и парамодуляции применительно к логикам высшего порядка пока не привели к каким-либо заметным успехам [10].

Второй путь заключается в том, чтобы, изучив специфику управления и принятия решений человеком, попытаться построить специальную логику — логику управления. Эта специфика заставляет отойти от традиционных схем построения *формальных* аксиоматико-дедуктивных систем математической логики и строить логику управления как динамическую семиотическую аксиоматико-дедуктивную систему. Последнее означает, что такая логика существенно связывается с моделью (предметной областью) и её динамикой через аксиоматику и понятие «общезначимая формула» здесь теряет свой универсальный смысл.

Из вышесказанного следует, что проблема непротиворечивости и полноты семиотической аксиоматико-дедуктивной системы должны решаться каждый раз при фиксации её модели.

Одно из возможных направлений второго пути и рассматривается в данной работе. Логику управления будем представлять как некоторую языковую систему L, при построении которой используем два основных момента:

а) основную идею ситуационного управления — описание ситуации через состояния и связи (отношения) элементов внешнего мира и их действий;

б) выбор в качестве описательных средств языка математической логики с тем, чтобы в дальнейшем расширить его применительно к специфике задач поведения целенаправленных систем.

В соответствии со спецификой принятия решений человеком достаточно принять, что язык L включает в себя язык L_i низшего i-го порядка, язык L_j — высшего j-го порядка и трансляционный язык L_{ij}, служащий для внутреннего диалога между языками L_i и L_j.

Синтаксис языка L.

Синтаксис языков L_i и L_j.

Оба языка по синтаксису совершенно одинаковы, поэтому временно индекс j опустим.

Алфавит

1. Знаки пунктуации: , ( ) [ ] { };

2. Имена элементов (с индексами или без них) w, $w_1, \ldots$;

3. Имена действий (с индексами или без них) f, $f_1, \ldots$;

4. Переменные значения по j-му признаку (с индексами или без них) $y^j, y_1^j, \ldots, y \geq 1$:

5. Функциональные переменные (с индексами или без них) $\psi, \psi_1, \ldots$;

6. Константы (с индексами или без них) α, $\alpha_1, \ldots$;

7. Предикатные переменные и константы (с индексами или без них) $g, g_1, \ldots$;

8. Высказывания (с индексами или без них) q, $q_1, \ldots$;

9. Символы логических операций — (отрицание), $\Rightarrow$ (импликация), $\rightarrow$ — (строгая импликация), $\Leftrightarrow$ (эквивалентность).

Выражения

Термы

1. Каждый символ α и y есть терм;

2. Если $t_1, \ldots, t_n$ — количественные термы, то ψ $(t_1, \ldots, t_n$ — терм;

3. Никакие другие выражения не являются термами.

Индивидуальные переменные

1. Каждый переменный терм есть индивидуальная переменная;

2. Каждый символ w и f есть индивидуальная переменная;

3. Никакие другие выражения не являются индивидуальными переменными.

Правильно построенные формулы—ППФ

1. Каждая буква q есть ППФ;

2. Если $t_1, \ldots, t_n$ — термы и t_n — время, то $g(t_1, \ldots, t_n)$ — ППФ;

3. Если $x \in \{w, f\}$, $t_1 \ldots, t_n$ — термы и t_n — время, то $g(x_1, \ldots, t_n, t)$ — ППФ;

4. Если $x_1, \ldots, x_n \in \{w, f\}$ и t — время, то $g(x_1, x_2, \ldots, x_n, t)$ — ППФ;

5. Если A — ППФ, то $\bar{A}$ — ППФ;

6. Никакие другие выражения не являются ППФ.

Предложения

1. Каждая ППФ есть предложение;
2. Если B_1 и B_2 — предложения, то B_1B_2 и B_2B_1 — предложения,
3. Если B — предложение, то $\overline{B}$ — предложение;
4. Никакие другие выражения не являются предложениями.

Аксиомы

Будем различать два вида аксиом: статические и динамические (аксиомы действий). Первые позволяют осуществлять вывод в рамках одновременного среза модели поведения, вторые — в рамках разных временных срезов. Пусть A и B — предложения с определенной семантикой в модели языка L_i. Тогда имеет место:

а) статическая аксиома $A \Rightarrow B$, если и только если A выражает достаточные, но не необходимые посылки для заключения B;

в) статическая аксиома $A \Leftrightarrow B$, если и только если A и B выражают достаточные и необходимые посылки друг для друга;

с) аксиома действия $A \rightarrow B$, если и только если A выражает необходимые и достаточные условия применения действия, соотнесенные с моментом t_1, а B выражает результат этого действия, соотнесенный с моментом t_2 и $t_2 > t_1$.

Никакие другие выражения не являются аксиомами.

На практике удобно в качестве предикатных констант и переменных g использовать русские слова. Например, константа ВПЕРЕДИ в ППФ ВПЕРЕДИ (x_1, x_2) — «x_1 впереди x_2» или переменная МЕСТО в ППФ МЕСТО (x_1, x_2) — «x_1 имеет местом x_2». В последнем случае в качестве значений переменной МЕСТО могут выступать константы НА, В, ВНУТРИ и т. д.

Аксиомы системы L_i отражают пространственные, временные и функциональные отношения между элементами и действиями в ситуации. Примером пространственной аксиомы может служить

МЕСТО (w_1, w_2, t) СУБЪЕКТ $(f, w_1, t) \Leftrightarrow$ МЕСТО (f, w_2, t).

Примером временной аксиомы может служить:

$$\text{БЫЛ } (x, t) \Rightarrow \text{БЫЛ } (x, t + \Delta t), \text{ где } \Delta t \geq 0.$$

Примером функциональной аксиомы может служить (для шахмат):

НА (w, x_1, t) ХОД (w, t) ДОСТУП $(w, x_2, t) \rightarrow$ НА $(w, x_2, t + 1)$.

Грамматика

В качестве метавыражений будем использовать секвенции двух типов: $\Gamma B_1 \vdash B_2$ и $\vdash B_1 \oplus B_2$, где B_1 и B_2 — предложения (не обязательно разные) и Γ — список аксиом (в том числе и ранее доказанных теорем), $\vdash$ знак выводимости, $\oplus$ знак операции $\rightarrow$ $\Rightarrow$ или $\Leftrightarrow$. Секвенция 1-го типа указывает, что предложение B_2 выводится из B_1 с помощью аксиом Γ; секвенция 2-го типа указывает, что $B_1 \oplus B_2$ выводится из ничего, т. е. аксиома эта задана.

Имеем следующие правила вывода в секвенциальной форме:

Основные правила

R0. Подстановка (для переменных),

$$\frac{\vdash B_1 \oplus B_2}{\vdash B_1\theta \oplus B_2\theta},$$ где $\theta = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ — подстановка, осуществляющая соответствующую замену переменных в B_1 и B_2.

R1. $\Gamma B_1 \vdash B_2$, где $B_2 \subseteq B_1$.

R2. $$\frac{\Gamma B_1 \vdash B_2B_3}{\Gamma B_1 \vdash B_2};$$ $$\frac{\Gamma B_1 \vdash B_2B_3}{\Gamma B_1 \vdash B_3};$$

R3. $$\frac{\Gamma B_1 \vdash B_2 \quad \vdash B_2 \oplus B_3}{\Gamma B_1 \vdash B_3};$$

R4. $$\frac{\Gamma B_1 \vdash B_2 \quad \Gamma B_1 \vdash B_3}{\Gamma B_1 \vdash B_2B_3}$$

R5. $$\frac{\Gamma B_1 \vdash \overline{B}_2}{\Gamma B_1 \vdash \overline{B_3B_2}}$$

R6. $$\frac{\Gamma B_1 \vdash \overline{B}_2 \quad \vdash B_2 \oplus B_3}{\Gamma B_1 \vdash \overline{B}_3};$$

R7. $$\frac{\Gamma B_1 \vdash B_2B_3 \quad \vdash B_2 \oplus B_4B_5}{\Gamma B_1 \vdash B_3B_5};$$

R8. $$\frac{\Gamma B_1 \vdash \overline{B}_2 \quad \vdash B_2B_3 \oplus B_4}{\Gamma B_1 \vdash \overline{B_4B_5}}.$$

Правило R6 при знаке $\Rightarrow$ используется только в обратном направлении, т. е. от заключения к посылкам секвенции.

Правила 7 и 8 являются производными правилами в системе L_i. Правило 7 есть последовательность правил 4—3—4—2, а правило 8–последовательность правил 5—6. Правило 7 (или 8) разбивает вывод предложения B_3B_5 (или $\overline{B_4B_5}$) на две части: вывод B_3 (или $\overline{B_2}$) и элементарный вывод $\overline{B_5}$ (или B_4) с помощью подходящей аксиомы. Это, по существу, формально описанные эвристики «средства-цель» [11].

Технические правила

R9. Чистка

$$\frac{\Gamma B_1 \vdash B_2 B_3 B_2}{\Gamma И B_1 \vdash B_2 B_3};$$

R10. Перестановка

$$\frac{\Gamma B_1 \vdash B_2 B_3}{\Gamma B_1 \vdash B_3 B_2};$$

R11. Констатация противоречия

$$\frac{\Gamma B_1 \vdash B_2 \quad \Gamma B_1 \vdash \bar{B}_2}{\Gamma B_1 \vdash \varnothing};$$

R12.

$$\frac{\Gamma B_1 \vdash B_2 \quad \Gamma B_1 B_2 \vdash B_3}{\Gamma B_1 \vdash B_2 B_3};$$

Правило R12 также является производным в системе L_i. Действительно, из посылки $\Gamma B_1 B_2 \vdash B_3$ и R13 следует $\Gamma B_1 \vdash B_2 \Rightarrow B_3$*). Из 1-ой посылки и *) по R3 следует $\Gamma B_1 \vdash B_3$**). Из 1-ой посылки и **) по R4 следует $\Gamma B_1 \vdash B_2 B_3$.

Правила пополнения системы аксиом Γ

R13. Теорема дедукции

$$\frac{\Gamma B_1 \vdash B_2}{\Gamma \vdash B_1 \Rightarrow B_2}$$

R14.

$$\frac{\Gamma \vdash B_1 \Rightarrow B_2 \quad \Gamma \vdash B_2 \Rightarrow B_1}{\Gamma \vdash B_1 \Leftrightarrow B_2}.$$

Определение вывода. Выводом D_K из Γ и B_n ($\Gamma B_и \vdash D_K$) в системе L_i называется конечный список $E = E_1 \ldots E_\eta \ldots E_m$, в котором всякое предложение $E_\eta (1 \leq \eta \leq m)$ либо принадлежит Γ, либо входит в $B_И$, либо получено из предыдущих предложений списка E по одному из правил грамматики [12].

Теорема 1. Пусть в системе L_i имеет место вывод из предложения B предложения $D = \Phi_1 \Phi_2 \ldots \Phi_n$, где Φ_ξ — ППФ и $D \cap B = \varnothing$. Тогда существуют такие предложения $B_\eta = B'_\eta \theta_\eta$ и $C_\eta = C'_\eta \theta_\eta$, где $1 \leq \eta \leq K$, $B'_\eta \oplus C'_\eta$ — аксиома списка Γ и θ_η — подстановка, что в системе L_i доказуемо любое из из двух предложений:

1) $\nu_1 = (B_1 \oplus C_1) \oplus \big((B_2 \oplus C_2) \oplus ((\ldots) \oplus (B_K \oplus C_K) \ldots)\big)$;
2) $\nu_2 = (B_K \oplus C_K) \oplus \big((B_{K-1} \oplus C_{K-1}) \oplus ((\ldots) \oplus (B_1 \oplus C_1) \ldots)\big)$.

В принятых обозначениях теорема 1 формулируется так. Если $\Gamma B \vdash D$ и $D \cap B = \varnothing$, то: 1) $\Gamma \vdash \nu_1$ и 2) $\Gamma \vdash \nu_2$.

Выражения ν_1 и ν_2 представляют собой результат:

ν_1 — прямого доказательства,
ν_2 — обратного доказательства.

Доказательство теоремы I. Вначале рассмотрим случай прямого доказательства. Вывод нового предложения D из B ($D \cap B = \varnothing$) возможен только с помощью правил вывода, использующих аксиомы списка Γ. Такими правилами являются R0, R3 и R6 (R7 и R8-производные). Поэтому использование при выводе η-й аксиомы ($i \leq \eta \leq k$) $\Gamma_\eta = (B_\eta \oplus C_\eta)$ возможно только по следующей схеме:

1) вывод B_η (или $\overline{B}_\eta$) из B;
2) вывод $\Gamma_\eta \theta_\eta = B'_\eta \theta_\eta \oplus C'_\eta \theta_\eta = B_\eta \oplus C_\eta$ с помощью правила $R0$ и постановки θ_η (возможно пустой);
3) вывод C_η из B_η и $B_\eta \oplus C_\eta$ с помощью R3 (или $\overline{C}_\eta$ из $\overline{B}_\eta$ и $B_\eta \oplus C_\eta$ с помощью R6).

Рассмотрим указанную схему по шагам вывода.

Первый шаг

1) Определяется такая аксиома $\Gamma_1 = (B'_1 \oplus C'_1)$, для которой найдутся подстановка θ_1 и предложение $B^{(1)}$ такие, что $B^{(1)} = B_1 \cap B$ и $\Gamma_1 \theta_1 = B^1_1 \theta_1 \oplus C^1_1 \theta_1 = B_1 \oplus C_1$. Очевидно такая аксиома существует, иначе вывод D был бы невозможен.
2) Используется правило R3 (или R6) и выводится C_1 (или $\overline{C}_1$).

η-й шаг ($1 < \eta \leq k$)

1) Определяется такая аксиома $\Gamma_\eta = B^1_\eta \oplus C^1_\eta$, для которой найдутся подстановка θ_η и предложение $B^{(\eta)}$ (возможно пустое) такие, что $B^{(\eta)} = B_\eta \cap B$, $B_\eta \subseteq B^{(\eta)} C_1 \ldots C_{\eta-1}$ и $\Gamma_\eta \theta_\eta = B_\eta \oplus C_\eta$. Если вывод не закончен, то такая аксиома должна быть.

Очевидно на некотором η-м шаге ($i \leq \eta \leq k$) впервые будет выведено предложение $C^{(\eta)} \subseteq C_\eta$, тождественно равное $\Phi_\xi (1 \leq \xi \leq n)$ или некоторой их совокупности, т. е. $C^{(\eta)} = C_\eta \cap D$. Обозначим эту совокупность как D_1. Продолжая далее, в конце-концов получим вывод предложения $D = D_1 D_2 \ldots D_m$ ($m \leq n$) в виде упорядоченного (по выводу) списка предложений D_r ($1 \leq r \leq m$). При этом допущение B представится в виде $B = B^{(1)} B^{(2)} \ldots B^{(K)}$, где K — общее число шагов вывода.

Таким образом, на каждом шаге вывода выводились элементарная теорема $B_\eta \oplus C_\eta$ и $C_\eta (\Gamma B_\eta \vdash C_\eta)$. Добавляя на каждом шаге, начиная с первого, к списку Γ элементарную теорему $B_\eta \oplus C_\eta$ получим для последнего K-го

шага

$$\Gamma,\ (B_1 \oplus C_1),\ (B_2 \oplus C_2), \ldots, (B_{K-1} \oplus C_{K-1}),\ B_K \vdash C_K,$$

откуда многократным применением теоремы дедукции получим $\Gamma \vdash \nu_1$.

Рассуждая аналогично для случая обратного вывода, на каждом шаге также будем иметь $\Gamma B_\eta \vdash C_\eta$ и $B_\eta \oplus C_\eta$. Добавляя на каждом шаге, начиная с последнего, к списку Γ элементарную теорему $B_\eta \oplus C_\eta$, получим для первого шага

$$\Gamma, (B_K \oplus C_K),\ (B_{K-1} \oplus C_{K-1}, \ldots, (B_2 \oplus C_2),\ B_1 \vdash C_1,$$

Откуда многократным применением теоремы дедукции получим $\Gamma \vdash \nu_1$. Теорема I доказана.

Следствие теоремы 1. Если $\Gamma B \vdash D$, где D и B удовлетворяют условиям теоремы 1, то имеет место $\Gamma \vdash ((B \oplus D) \Rightarrow \nu_1)$, $\Gamma \vdash ((B \oplus D) \Rightarrow \nu_2)$.

Доказательство. Действительно, из $\Gamma B \vdash D$ по теореме дедукции следует $\Gamma \vdash (B \oplus D)$. Согласно теореме I имеет место $\Gamma \vdash \nu_1$ и $\Gamma \vdash \nu_2$, а, следовательно, $\Gamma, B \oplus D \vdash \nu_1$ и $\Gamma, B \oplus D \vdash \nu_2$. Теорема дедукции завершает доказательство.

Теорема 2. Пусть в системе L_i имеет место $\Gamma \vdash \nu_1$ или $\Gamma \vdash \nu_2$. Тогда существуют предложения $D = D_1 D_2 \ldots D_m$ и $B = B^{(1)} B^{(2)} \ldots B^{(K)}$ такие, что $D \cap B = \varnothing$ и $\Gamma \vdash (B \oplus D)$.

Доказательство очевидно и в работе не приводится.

Следствие теоремы 2. Если $\Gamma \vdash \nu_1$ и $\Gamma \vdash \nu_2$, то $\Gamma \vdash (\nu_1 \Rightarrow (B \oplus D))$ и $\Gamma \vdash (\nu_2 \Rightarrow (B \oplus D))$.

Доказательство аналогично доказательству следствия теоремы 1.

Из следствий теорем 1 и 2, в свою очередь, следует $\Gamma \vdash \nu_1 \Leftrightarrow (B \oplus D)$ и $\Gamma \vdash \nu_2 \Leftrightarrow (B \oplus D)$, откуда получаем $\Gamma \vdash \nu_1 \Leftrightarrow \nu_2$.

В дальнейшем нас не будет интересовать характер доказательства (прямой или обратный), поэтому опустим в выражениях ν_1 и ν_2 знак $\oplus$ вместе с частью скобок и будем писать

$$\nu = (B_1 \oplus C_1)(B_2 \oplus C_2) \ldots (B_K \oplus C_K)$$

как упорядоченный (слева направо) список элементарных теорем. Назовем последовательность ν в отличие от выражений ν_1 и ν_2 доказательством теоремы $B \oplus D$.

Если из доказательства v теоремы $B \oplus D$ удалить все элементарные статические теоремы, оставив только элементарные теоремы действия, то получим план реализации u теоремы $B \oplus D$. Очевидно,

$$u = (B_1 \to C_1)(B_2 \to C_2) \ldots (B_r \to C_r), \text{ где } r \leq K.$$

Синтаксис трансляционного языка

Алфавит

1. Знаки пунктуации, ( ) [ ] { };
2. Переменные i-го порядка (с индексами или без них) $x^i, x_1^i \ldots$;
3. ППФ i-го порядка (с индексами или без них) $g^i, g_1^i \ldots$;
4. Переменные j-го порядка (с индексами или без них) $x^j, x_1^j, \ldots$;
5. ППФ j-го порядка (с индексами или без них) $g^j, g_j^1, \ldots$;
6. Символы логических операций —, &, $\vee$, $\Rightarrow$, $\Leftrightarrow$;
7. Кванторы всеобщности $\forall$, существования J, единственности $J!$

Кванторы $\forall$, J и $J!$ являются вспомогательными символами, которые используются, во-первых, для сокращения записи соответственно конъюнкции, дизъюнкции и строгой дизъюнкции, во-вторых, для указания области определения соответствующих n-ок.

Выражения

Квазиформулы — КФ.

1. Каждая ППФ i-го порядка есть КФ
2. Если A — КФ, то $\overline{A}$ — КФ;
3. Если A и B — КФ, то $A \,\&\, B$, $A \vee B$ и $A \Rightarrow B$ — КФ;
4. Если A — КФ, x^i — *не единственная* свободная переменная в A и такая, что $x^i \in x^j$, то $(\forall x^i \in x^j) A$, $(J x^i \in x^j) A$, $(J! x^i \in x^j) A$ — КФ;
5. Никакие другие выражения не являются КФ.

Правильно построенные трансляционные формулы — ППФ.

1. Если A — КФ и x^i — *единственная* свободная переменная в A такая, что $x^i \in x^j$, то $(\forall x^i \in x^j) A$, $(B^j x^i \in x^j) A$, $(B^j! x^i \in x^j) A$ — ППФ;
2. Если A — ППФ, то $\overline{A}$ — ППФ;
3. Никакие другие выражения не являются ППФ.

Аксиомы

1. Если B^j — предложение j-го порядка с определенной семантикой в модели языка L_j, а A — ППТФ с определенной семантикой в модели языка L_{ij}, то $B^j \Leftrightarrow A$ — аксиома.

2. Никакие другие выражения не являются аксиомами.

Аксиомы языка L_{ij} также выражают пространственные, временные и функциональные отношения между элементами (действиями) j-го порядка и элементами (действиями) i-го порядка. Примеры. ВЫШЕ $(x_1^j, x_2^j, t^j) \Leftrightarrow$ $\Leftrightarrow (\forall x_2^i \in x_2^j)\,(Jx_1^i \in x_1^j)\,(Jt^i \in t^j)$ (ВЫШЕ (x_1^i, x_2^i, t^i)) БЫТЬ $(f^j, t^j) \Leftrightarrow (Jf^i \in f^j)$ $(Jt^i \in t^j)$ (БЫТЬ (f^i, t^i)), СУБЪЕКТ $(f^i w^j, t^j) \Leftrightarrow (Jf^i \in f^j)\ (Jw^i \in w^j)\ (Jt^i \in t^j)$ (СУБЪЕКТ $(f^i,\ w^i,\ t^i)$)

Грамматика

Пусть Γ — система аксиом языка L_{ij}. Пусть B^i — предложение i-го порядка и A — квазиформула. Тогда имеем следующие правила вывода в секвенциальной форме. Подстановка (для переменных).

T0. $\dfrac{\vdash A \Leftrightarrow B^j}{\vdash A\theta \Leftrightarrow B^j\theta}$, где θ — подстановка, осуществляющая соответствующую замену переменных.

T1. $\Gamma A \vdash A$;

T2. $\dfrac{\Gamma B^i \vdash A_1 \quad \Gamma B^i \vdash A_2}{\Gamma B^i \vdash A_2 \,\&\, A_2}$; $\dfrac{\Gamma B^i \vdash \overline{A}_1}{\Gamma B^i \vdash \overline{A_1 \,\&\, A_2}}$; $\dfrac{\Gamma B^i \vdash \overline{A}_2}{\Gamma B^i \vdash \overline{A_1 \,\&\, A_2}}$;

T3. $\dfrac{\Gamma B^i \vdash A_1}{\Gamma B^i \vdash A_1 \vee A_2}$; $\dfrac{\Gamma B^i \vdash A_2}{\Gamma B^i \vdash A_1 \vee A_2}$; $\dfrac{\Gamma B^i \vdash \overline{A_1} \quad \Gamma B^i \vdash \overline{A_2}}{\Gamma B_i^i \vdash \overline{A_1 \vee A_2}}$;

T4. $\dfrac{\Gamma B^i \vdash \overline{A_1}}{\Gamma B^i \vdash A_1 \Rightarrow A_2}$; $\dfrac{\Gamma B^i \vdash A_2}{\Gamma B^i \vdash A_1 \Rightarrow A_2}$; $\dfrac{\Gamma B^i \vdash A_1 \quad \Gamma B^i \vdash \overline{A_2}}{\Gamma B^i \vdash \overline{A_1 \Rightarrow A_2}}$;

Пусть $x^j = \{x_\xi^i \mid \mathrm{I} \leq \xi \leq n\}$, где x_ξ^i и x^j — переменные соответственно i-го и jго порядка. Тогда имеет место

T5. $\dfrac{\Gamma B^i \vdash A(x_1^i) \quad \Gamma B^i \vdash A(x_2^i) \quad \ldots\ldots \quad \Gamma B^i \vdash A(x_n^i)}{\Gamma B^i \vdash (\forall x^i \in x^j)\, A(x^i)}$; $\dfrac{\Gamma B^i \vdash \overline{A}(x_\xi^i)}{\Gamma B^i \vdash (\overline{\forall}\, x^i \in x^i)\, A(x^i)}$,

Т6. $\dfrac{\Gamma B^i \vdash A(x_\xi^i)}{\Gamma B^i \vdash (Jx^i \in x^j)\, A(x^i)}$; $\quad \dfrac{\Gamma B^i \vdash \overline{A}(x_1^i)\quad \Gamma B^i \vdash \overline{A}(x_2^i)\quad \ldots\ldots\ldots\ldots\ldots\ldots\quad \Gamma B^i \vdash \overline{A}(x_n^i)}{\Gamma B^i \vdash (\bar{J}x^j \in x^j)\, A(x^i)}$;

Т7. $\dfrac{\Gamma B^i \vdash A(x)_\xi^i \quad \Gamma B^i \vdash (\forall x^i \in x^j \setminus x_\xi^i)\, \overline{A}(x^i)}{\Gamma B^i \vdash (J!\, x^i \in x^j)\, A(x^i)}$; $\quad \dfrac{\Gamma B^i \vdash A(x_\xi^i) \quad \Gamma B^i \vdash (J\, x^i \in x\, j \setminus x_\xi^i)\, A(x^i)}{\Gamma B^i \vdash \bar{J}(!\, x^i \in x^j)\, A(x^i)}$;

$\dfrac{\Gamma B^i \vdash (\forall x^i \in x^j)\, \overline{A}(x^i)}{\Gamma B^i \vdash (J!\, x^i \in x^j)\, A(x^i)}$

Т8. $\dfrac{\Gamma B^i \vdash A \quad \vdash A \Leftrightarrow B^j}{\Gamma B_i \vdash B^j}$; Т9. $\dfrac{\Gamma B^i \vdash \overline{A} \quad \vdash A \Leftrightarrow B^j}{\Gamma B^i \vdash \overline{B}^j}$; Т10. $\dfrac{\Gamma B^i \vdash B_1^j \quad \Gamma B^i \vdash B_2^j}{\Gamma B^i \vdash B_1^j B_2^j}$; Т11. $\dfrac{\Gamma B^i \vdash \overline{B_1^j}}{\Gamma B^i \vdash \overline{B_1^j B_2^j}}$

В правилах Т0, Т8—Т11 A обозначает ППТФ.

Вывод. Выводом предложения B^j j-го порядка из предложения B^i i-го порядка ($j > i$) в системе L_{ij} является конечный список $E = E_1 \ldots E_m$, в котором:

— всякое выражение $E_\xi (1 \leq \xi \leq m - 2)$ является либо ППФ i-го порядка, либо получено из предыдущих выражений списка E по одному из правил трансляционной грамматики;

— $E_{m-2} = \widetilde{A}$;

— $E_{m-1} = A \Leftrightarrow B^j$;

— $E_m = \widetilde{B}^j$, где $\widetilde{A}$ — ППТФ и $\widetilde{A} \in \{A, \overline{A}\}$ и $\widetilde{B}^j \in \{B^j, \overline{B}^j\}$.

Структура вывода в системе L

Пусть имеет место:

— в системе L_i — доказательство $\nu^i = (B_1^i \oplus C_1^i) \ldots (B_l^i \oplus C_l^i)$ теоремы $B_{\text{И}}^i \oplus D_l^i$;

— в системе L_{ij} — для каждого элемента $\xi \in [1, K]$, выделяющего в ν^i часть ν_ξ^i так, что $\nu^i = \nu_1^i \ldots \nu_\xi^i \ldots \nu_K^i$, существуют выводы E'_ξ и E''_ξ, соответствующие $\Gamma_{ij} H_\xi^i \vdash B_\xi^j$ и $\Gamma_{ij} D_\xi^i \vdash C_\xi^j$, где $H_\xi^i \oplus D_\xi^i$ — теорема с доказательством ν_ξ^i, а $B_\xi^j \oplus C_\xi^j$ — элементарная теорема в ν^j,

— в системе L_j — доказательство $\nu^j = (B_1^j \oplus C_1^j) \ldots (B_{\text{К}}^j \oplus C_{\text{К}}^j$ теоремы $B_{\text{И}}^j \oplus D_{\text{К}}^j$.

Тогда совокупность $\nu = \{\nu^j, E, \nu^i\}$, где $E = \{E'_\xi, E''_\xi / 1 \leq \xi \leq K\}$, представляет в системе L полное доказательство теоремы $B^i_{\text{И}} \oplus D^j_{\text{К}}$, план реализации которой, в свою очередь, согласно определению, представляется в виде $u = \{u^j, u^i\}$.

Рассмотрим класс полных систем L.

Определение. Если всякий раз, когда в L_i разрешима теорема $B^i_{\text{И}} \oplus D^i_l$ с доказательством $\nu^i = (B^i_1 \oplus C^i_1) \ldots (B^i_1 \oplus C^i_l)$, а в L^i_j разрешимы выводы $\Gamma_{ij} B^i_{\text{И}} \vdash B^j_{\text{И}}$ и $\Gamma_{ij} D^i_l \vdash D^j_{\text{К}}$, следует, что в L_j разрешима теорема $B^j_{\text{И}} \oplus D^j_{\text{К}}$ с доказательством $\nu^j = (B^j_1 \oplus C^j_1) \ldots (B^j_K \oplus C^j_{\text{К}})$, то система L является полной.

Достаточным условием полноты системы L может служить наличие разбиения $\langle \nu^i \rangle$ для каждого доказательства ν. Разбиение $\langle \nu^i \rangle$ строится следующим образом.

Шаг 1. Из элементарных теорем доказательства, начиная с первой, последовательно формируются два таких предложения $H^i_\eta = B^1\text{-}B^2\text{-} \ldots B^\eta$ и D^i_1, что:

1) $B^\xi = B^i_{\text{И}} \cap B^i_\xi (1 \leq \xi \leq \eta) \cap D^i_1 = C^i_\eta \cap D^i_l$:

2) впервые (по номеру η) имеют место выводы

$$\Gamma_{ij} H^i_\eta \vdash B^j_1 \text{ и } \Gamma_{ij} D^i_1 \vdash C^j_1, \text{ где } B^j_1 \oplus C^j_1 = \Gamma^j_1 \Theta^j_1 \text{ и } \Gamma^j_1 \in \Gamma_j.$$

Полученное при этом значение η определяет первый элемент ν^i_1 разбиения $< \nu^i >$; этому элементу соответствует элементарная теорема $(B^j_1 \oplus C^j_1) \in \nu^j$. Если при этом $\eta = l$, то построение разбиения закончено, иначе переход к шагу 2.

Шаг 2. Переопределяются ν^i, $B^i_{\text{И}}$ и D^i_l, т. е. вместо ν^i длины l рассматривается его часть $\nu^i \backslash \nu^i_1$ длины $l - \eta$, вместо $B^i_{\text{И}} - B^i_{\text{И}} C^i_1 \ldots C^i_\eta$, вместо D^i_l — $D^i_l | D^i_1$. Затем переход к шагу 1. В итоге формируется разбиение $\langle \nu^i \rangle$, упорядоченной последовательности элементов которого соответствует доказательство ν^j теоремы $B^j_{\text{И}} \oplus D^j_{\text{К}}$.

Следствие полноты системы L. Пусть в L_j не разрешима теорема $B^j_{\text{И}} \oplus D^j_{\text{К}}$ при разрешимости в L_{ij} выводов $\Gamma_{ij} B^i_{\text{И}} \vdash B^j_{\text{И}}$ и $\Gamma_{ij} D^i_l \vdash D^j_{\text{К}}$. Тогда, если L — полная система, то в L_i не разрешима теорема $B^i_{\text{И}} \oplus D^i_l$.

Действительно, пусть теорема $B^i_{\text{И}} \Rightarrow D^i_e$ разрешима в системе L_i. Тогда из полноты L следует, что в L_j должна быть разрешима теорема $B^j_{\text{И}} \Rightarrow D^j_{\text{К}}$, что противоречит исходному условию.

Семантика и прагматика системы L

Мир $S = \{S_i, S_{ij}, S_j\}$ является моделью системы $L = \{L_i L_{ij} L_j\}$, если существует взаимно однозначная система отображений

$$\varrho = \{\varrho_i : L_i \to S_i\ ,\ \varrho_{ij} : L_{ij} \to S_i \times S_j\ ,\ \varrho_j, : L_j \to S_j\}$$

такая, что:

1) если w — элемент в L_i (или в L_j), то ϱw — объект в S_i (или в S_j);
2) если f — имя действия в L_i (или в L_j), то ϱf — действие в S_i (или в S_j);
3) если t — терм в L_i (или в L_j), то

$$\varrho t = \begin{cases} \varrho\alpha \text{ — значение константы в } S_i \text{ (или в } S_j), \\ \varrho y \text{ — значение характеристики в } S_i \text{ (или в } S_j), \\ \varrho\psi(\varrho t_1 \ldots,\ \varrho t)_n \text{ — значение функции в } S_i \text{ (или в } S_j); \end{cases}$$

4) если ψ — функциональная переменная в L_i (или в L_j), то $\varrho\psi$ — функция в S_i (или в S_j);
5) если q — пропозициональная переменная в L_i (или в L_j), то ϱq — истинное или ложное высказывание в S_i (или в S_j);
6) если g и $\bar{g}$ — предикатные переменные в L_i (или в L_j), то ϱg и $\varrho\bar{g}$ — одно из двух взаимно отрицающих n-местных отношений в S_i (или в S_j), как правило $1 \leq n \leq 2$;
7) если A — ППФ в L_i (или в L_j), то

$$\varrho A = \begin{cases} \varrho g(\varrho t_1, \ldots, \varrho t_n) & \text{— } n\text{-ка значений термов, связанных отношением } \varrho g \text{ в } S_i \text{ (или в } S_j), \\ \varrho g(px, \varrho t_1, \ldots, \varrho t_n) & \text{— } n\text{-ка значений характеристик либо объекта } \varrho w, \text{ либо действия } \varrho f \text{ в } S_i \text{ (или в } S_j), \\ \varrho g(\varrho x_1 \ldots, \varrho x t_n) & \text{— } n\text{-ка либо объектов } \varrho w, \text{ либо действий } \varrho f, \text{ либо того и другого в } S_i \text{ (или в } S_j); \end{cases}$$

8) если B — предложение в L_i (или в L_j), то ϱB — ситуация в S_i (или в S_j);
9) если C — аксиома (теорема) в L_i (или в L_j), то ϱC — либо статический, либо динамический закон в S_i (или в S_j);
10) если $B_\text{И} \Rightarrow D_\text{К}$ — доказываемая теорема в L_i (или в L_j), то $(\varrho B_\text{И}, \varrho D_\text{К})$ — задача в S_i (или в S_j);
11) если v — доказательство теоремы $B_\text{И} \Rightarrow D_\text{К}$ в L_i (или в L_j), то ϱv — решение задачи $(\varrho B_\text{И}, \varrho D_\text{К})$ в S_i (или в S_j);
12) если u — план реализации теоремы $B_\text{И} \Rightarrow D_\text{К}$ в L_i (или в L_j), то ϱu — реализация задачи $(\varrho B_\text{И}, \varrho D_\text{К})$ в S_i (или в S_j);

13) если A — ППТФ в L_{ij}, то $\varrho_{ij}A$ — n-местное отношение в $S_i \times S_j$, связывающее (всякую, хотя бы единственную, одну) n-ку элементов и действий низшего i-го порядка, принадлежащих соответственно элементам и действиям высшего j-го порядка, как правило, $n \geq 2$;
14) если C — аксиома в L_{ij}, то $\varrho_{ij}C$ — логический закон в $S_i \times S_j$, устанавливающий взаимно однозначное соответствие между n_1-местным отношением в S_i и n_2-местным отношением в S_j, как правило $n_2 < n_1$;
15) если E — вывод в L_{ij}, то $\varrho_{ij}E$ — либо обобщение ситуаций в S_i путем свертывания n_1-местных отношений в S_i в n_2-местные отношение в S_j, либо детализация ситуаций в S_j путем развертывания n_2-местных отшений в S_j в n_1-местные отношения в S_1;
16) если $B^i_{\text{И}} \Rightarrow D^j_{\text{К}}$ — доказываемая теорема в L_{ij}, то $(\varrho_i B^i_{\text{И}}, \varrho_j D^j_{\text{К}})$ — задача в $S_i \times S_j$, при этом $\varrho_i B^i_{\text{И}}$ — исходная ситуация в S_i, а $\varrho_j D^j_{\text{К}}$ — целевая, конечная ситуация в S_j.

В качестве конкретных моделей языка L рассматривались шахматный эндшпиль [13] и перемещение робота на местности [14].

Литература

1. *Поспелов Д. А.* Принципы ситуационного управления Известия АН СССР, Техническая кибернетика 1971, *2*.
2. *Клыков Ю. И.* Ситуационное управление большими системами. »Энергия«, М., 1974.
3. *Davis, M.*, Eliminating the irrelevant from mechanical proofs.
4. *Darlington, J.*, Automatic theorem proving with equality substitutions and mathematical induction, in Michie(ed), "Machine Intelligence 3", American Elsevier Publishing Co, Inc, New York, 1968.
5. *Lucham, D.*, Refinement theorems in resolution theory, Presented of the IRIA Symposium on Automatic Deduction, Versailles, France, 1969.
6. *Kowalski, R., Hayes, P.*, Semantic trees in automatic theorem proving, in B. Meltzer and D. Michi (eds) "Machine Intelligence 4", American Elsevier Publishing Co., Inc., New York, 1969.
7. *Robinson, G., Wos, L.*, Paramodulation and theorem-proving in first-order theories with equality, in B. Meltzer and D. Michi (eds) "Machine Intelligence 4", American Elsevier Publishing Co., Inc., New York, 1969.
8. *Robinson, J.*, Mechanizing higher order logic, in B. Meltzer and D. Michi (eds) "Machine Intelligence 4", American Elsevier Publishing Co., Inc., New York, 1969.
9. *Robinson, J.*, A note on mechanizing higher order logic, in B. Meltzer and D. Michi (eds) "Machine Intelligence 5", American Elsevier Publishing Co., Inc., New York, 1969.
10. *Robinson, J.*, An overview of mechanical theorem proving, 4 System Simposium, 1970.
11. *Ньюэлл А., Шоу Дж., Саймон Г.* Разновидность интеллектуального обучения и »вычислителя« для решения задач общего типа. В кн. »Самоорганизующиеся системы«, Мир., М., 1964.
12. *Клини С. К.* Математическая логика. Мир. М., 1973.
13. *Ефимов Е. И.* Эвристическая модель »Шахматный эндшпиль«. В кн. »Труды семинара МДНТП им. Дзержинского«, 1971.
14. *Ефимов Е. И.* Передвижение робота по местности путем прокладки маршрута на заранее формируемой им системе карт. Вопросы радиоэлектроники. Сер. Общетехническая, вып. 14, 1973.

Purposeful system behaviour logic

E. I. EFIMOV

(Moscow)

A purposeful system behaviour is a process of the decision of the problems which are put to it. The decision is realized as sequence of requisite actions directing to the achievement of the given purpose. The indicated problems are identical to that of situating management receiving broad expansion safely in investigation of complex systems [1, 2].

Trials are known deciding these problems on the basis of theory of automatic proving theorem using universal procedures of deduction-resolution and paramodulation within different modification limits of the first-oder logic.

These trials did not bring to the creation of effication algorithms of a proof search for some interesting practical problems. Therefore some modes deserve attention, for instance, that based on treatment of special management logic taking into consideration the peculiarity of a man decision.

Outgoing from the above, the language system including lower order language L_i, higher order language $L_j\,(j > i)$ and translation language L_{ij} serving for inner dialogue between language L_i and L_j is examined in the paper.

Syntax, semantics and pragmatics of language L are described. The definitions of deduction, proof and plan realization theorem are given in system L.

Е. И. Ефимов
Вычислительный центр АН СССР
СССР, Москва В—333,
ул. Вавилова, 40

Problems of Control and Information Theory, Vol. 5 (3), pp. 263—270 (1976)

NEW CLASS OF THE CONTROLLED SHIFT-REGISTERS. II

Z. GRODZKI

(Lublin)

(Received April 23, 1975)

This paper is a continuation of the author's previous paper [3] and is related to the Boolean controlled (k, m, p)-machines.

Further properties of the (k, m, p)-computation sets will be given. Then a problem with respect to which relations a set $D^{\infty} = \bigcup_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcup_{p=1}^{\infty} D_{k,m,p}$ is closed will be studied, where $D_{k,m,p}$ denotes the set of all (k, m, p)-computations.

At the very end a nondeterministic case of the controlled (k, m, p)-machines will be investigated and a few open problems will be put forward.

1. Preliminaries and definitions

We assume the reader to be familiar with the fundamental notions of [3]. The notations of [3] will be used here.

Let $M = \{0, 1\}$ and let k, m, p be the arbitrary (fixed) numbers.

Definition 1. An infinite sequence $T = t_1, t_2, \ldots \in M^{\infty}$ is said to be cyclic iff there are the numbers i, j such that for every $n \geq i$ we have $t_n = t_{n+j}$.

Let $T = t_1, t_2, \ldots \in M^{\infty}$ be a cyclic sequence.

Definition 2. By a threshold set associated with T (Notation: $\mathrm{Th}(T)$) we mean a set:

$$\{U : U = \Lambda \,\&\, \exists_{j\geq 1} \forall_{i\geq 1} (t_i = t_{i+j}) \vee \exists_{p\geq 1} \exists_{q\geq 1} \forall_{r>p} (U = T\,|_{1,p} \,\&\, t_r = t_{q+r}),$$

where Λ denotes the empty sequence.

Definition 3. A sequence $U \in \mathrm{Th}(T)$ (eventually empty) is said to be a threshold segment of T iff the following conditions are satisfied:

(1) if $\Lambda \in \mathrm{Th}(T)$ then $U = \Lambda$;

(2) if $\Lambda \notin \mathrm{Th}(T)$ then U is a nonempty shortest sequence in $\mathrm{Th}(T)$.

Definition 4. Let $U \in M^*$ be the threshold segment of T of the length i for $i \geq 0$.

A sequence $V \in M^* - \{\Lambda\}$ is said to be a cycle of T iff the following condition is satisfied:

(1) there is minimal $j \geq 1$ such that for every $q \geq 0$ we have

$$V = T\,|_{i+jq+1,\, i+j(q+1)}\,.$$

Remark 1. Let us see that for every cyclic sequence T its threshold segment and the cycle are defined in an unique way.

Example. Let us define a cyclic sequence $T = t_1, t_2, \ldots$ as follows:

$$t_i = 0 \text{ for } 1 \leq i \leq 4,\ t_{5+3j} = 1,\ t_{5+3j+1} = 0,\ t_{5+3j+2} = 0$$

for every $j \geq 0$.

Then $U = 0000$ is its threshold segment and $V = 100$ is its cycle.

Definition 5. Let us define a ternary relation $\oplus \subseteq M^\infty \times M^\infty \times M^\infty$ as follows:

$$\oplus(T,\ U,\ V) \Leftrightarrow \forall_{i \geq 1}\,(V\,|_{i,i} = T\,|_{i,i} + U\,|_{i,i})$$

where sign $+$ denotes the modulo 2 addition.

Instead of $\oplus(T, U, V)$ we shall write $V = T \oplus U$ and understand that V is obtained by means of the addition operation on T and U.

Definition 6. Let us define a ternary relation $o \subseteq M^\infty \times M^\infty \times M^\infty$ between cyclic sequences.

$o(T, U, V)$ iff the following conditions are satisfied:

(1) if W, X are the threshold segments of T and U respectively then WX is the threshold segment of V;

(2) if Y, Z are the cycles of T and U respectively then YZ is the cycle of V.

Instead of $o(T, U, V)$ we shall write $V = T \circ U$ and understand that V is obtained by means of the composition on T and U.

Definition 7. Let us define a binary relation $- \subseteq M^\infty \times M^\infty$ as follows:

$$-(T,\ U) \Leftrightarrow \forall_{i \geq 1}\ [(T\,|_{i,i} = 0\ \&\ U\,|_{i,i} = 1) \vee (T\,|_{i,i} = 1\ \&\ U\,|_{i,i} = 0)]\,.$$

Instead of $-(T, U)$ we shall write $U = \overline{T}$ and understand that U is obtained by means of the negation operation on T.

Definition 8. Let $T, U \in M^\infty$ be two cyclic sequences with the empty threshold segment.

Then U is said to be inverse to T (Notation: $U = \widehat{T}$) iff the cycles of both sequences are inverse one to another i.e. if $v_1, v_2, \ldots, v_{n-1}, v_n$ is the cycle of T then $v_n v_{n-1}, \ldots, v_2, v_1$ is the cycle of U.

2. The properties of the (k, m, p)-computations

The properties of singular (k, m, p)-computations will be investigated, in particular a problem how the maximal cycle of these sequences will be solved.

At the very end the properties of the set $D^\infty = \bigcup_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcup_{p=1}^{\infty} D_{k,m,p}$, where $D_{k,m,p} \subseteq M^\infty$ denotes the set of all (k, m, p)-computations will be given.

Theorem 1. Every (k, m, p)-computation $T \in M^\infty$ is a cyclic sequence.

Proof. Let $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$ be an arbitrary (k, m, p)-machine and let $T \in M^\infty$ be its computation. Since for the set M there are 2^k k-element variations with repetitions then there are two sequences $T|_{ip+1,(i+1)p}$ and $T|_{jp+1,(j+1)p}$ $(i \neq j)$ such that $\psi(T|_{ip+1,ip+k}, ip+1) = \psi(T|_{jp+1,jp+k}, jp+1)$ and $T|_{r,r+k-1} = T|_{s,s+k-1}$ for some numbers r, s

$$(ip + 1 \leq r \leq (i+1)p,\ jp + 1 \leq s \leq (j+1)p).$$

It is possible to prove by mathematical induction that $T|_{r+n,r+k+n-1} = T|_{s+n,s+k+n-1}$ for every $n \geq 0$. It follows from Definition 1 that T is cyclic. Q.E.D.

Lemma 1. Let $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$ $(k \geq 1, m > 1, p \geq 1)$ be an arbitrary (k, m, p)-machine and let $T \in M^\infty$ be its computation.

If the cycle of T is of the maximal length then there is a permutation $\varphi_{i_1}, \ldots, \varphi_{i_m}$ such that the following condition is satisfied:

$$\psi(T_{pq+1,pq+k}, pq+1) = \varphi_{i_j} \Leftrightarrow q \equiv j \pmod{m}. \tag{1}$$

Remark 2. Condition (1) means that the transition functions of $A_{k,m,p}$ are used cyclically in a sequence $\varphi_{i_1}, \ldots, \varphi_{i_m}$ for any permutation $\varphi_{i_1}, \ldots, \varphi_{i_m}$ i.e. $\psi(T|_{1,k}, 1) = \varphi_{i_1}$, $\psi(T|_{mp+1,mp+k}, mp+1) = \varphi_{i_m}$ and so on.

Remark 3. For the simplicity a (k, m, p)-machine as above will be called a simple (k, m, p)-machine and will be denoted as $\langle M, \varphi_{i_1}, \ldots, \varphi_{i_m} \rangle$ (the control ψ is omitted).

The proof of Lemma 1 is obvious and will be omitted.

Lemma 2. Let $A_{k,m,1} = \langle M, \varphi_1, \ldots, \varphi_m \rangle$ be a simple $(k, m, 1)$-machine. Let $T \in M^\infty$ be an arbitrary computation of $A_{k,m,1}$ and let U, $V \in M^*$ be its threshold segment and the cycle.

Then we have:

(1) $$1 \leq 1(V) \leq 2^k m;$$

(2) $$\text{if } 1(V) = 2^k m \text{ then } 1(U) = 0.$$

Lemma 3. Let $A_{k,m,1} = \langle M, \varphi_1, \ldots, \varphi_m \rangle$ be a simple $(k, m, 1)$-machine.

If there is at least one computation T of $A_{k,m,1}$ with the cycle length of $2^k m$ then all its computations have the cycles of the same length.

The proofs of Lemmas 2 and 3 have been given in [2].

Theorem 2. Let $T \in M^\infty$ be an arbitrary (k, m, p)-computation and let U, V be its threshold segment and the cycle respectively. Then we have:

(1) $$1 \leq 1(V) \leq 2^k mp;$$

(2) $$\text{if } 1(V) = 2^k mp \text{ then } 1(U) = 0.$$

Proof. Let $T \in M^\infty$ be a (k, m, p)-computation as above. Let us see that for the set M there are 2^k k-element variations with repetitions. Let W_i be such a variation that for the maximal T cycle we have $T|_{ip+1, ip+k} = W_{i+1}$ and $T|_{ip+1, (i+1)p+k} \neq T|_{jp+1, (j+1)p+k}$ for all i, j, $0 \leq i < j < 2^k$.

It is clear that for this case we have $1(V) = 2^k mp$ and $1(U) = 0$. Therefore conditions (1) and (2) are held. Q.E.D.

Corollary 1. Let $E \subset M^\infty$ be a (k, m, p)-computation set.

If there is at least one sequence $T \in E$ with the cycle length of $2^k mp$ then all sequences of E have the cycles of the same length.

Theorem 3. The set D^∞ is closed with respect to the following operations:

(1) addition
(2) composition
(3) negation
(4) inversion.

Proof. Let T, U be two cyclic sequences and let V, W (X, Y) be the threshold segment and the cycle of T (U) respectively. Let be $1(V) = i$, $1(W) = j$ ($i \geq 0$, $j > 1$) and $1(X) = r$, $1(Y) = s$($r \geq 0$, $s > 1$).

Let be $Z = T \oplus U$. Then there are the numbers p_1 and p_2 such that $i + jp_1 = i + jp_2$. Let us denote these numbers by q_1 and let q_2 be the least common multiple of j and s. Then Z is cyclic with $Z|_{1,q_3}$ as the threshold segment and $Z|_{q_3+1,\, q_4}$ as the cycle for some numbers q_3, q_4, $q_3 \leq q_1$, $q_4 \leq q_2$.[1] It follows from Theorem 1 that $Z \in D^\infty$.

For the cases (2)–(4) the proof follows directly from Definitions 6–8 and Theorem 1. Q.E.D.

3. The properties of the (k, m, p)-computation sets

A few properties of the (k, m, p)-computation sets will be given.

Theorem 4. A nonempty set $E \subseteq M^\infty$ is a (k, m, p)-computation set iff the following conditions are satisfied:

(1) $$\forall_{T \in M^k} \exists_{U \in E} (T = U|_{1,k});$$

(2) $$\forall_{T \in E} \forall_{i \geq 1} (T|_{ip+1} \in E);$$

(3) there is a set $F = \mathrm{div}_{k,p}(E)$ such that $\mathrm{card}(F) = m$, where $\mathrm{div}_{k,p}(E)$ denotes a (k, p)-division of E.

The proof of Theorem 4 has been given in [2].

Theorem 5. Let $A_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m \rangle$ and $A'_{k,m,p} = \langle M, \pi_1, \ldots, \pi_m \rangle$ be the simple controlled (k, m, p)-machines. Let E and F be their computation sets.

If for their transition functions the following condition is satisfied:

(1) $$\forall_{i \leq m} \forall_{(u_1, \ldots u_k) \in M^k} \varphi_i(u_1, \ldots, u_k) = \neg\, \pi_i(\neg\, u_1, \ldots, \neg\, u_k)]$$

then we have $E = \bar{F}$, where $\bar{F}$ is the set obtained from F by replacing 0 with 1 and conversely in all its sequences.

The inverse statement is not true,

Proof. Let us assume that condition (1) of Theorem 5 is satisfied. We are going to prove that $E = \bar{F}$.

Let $V = v_1, v_2, \ldots$ be an arbitrary sequence of E and let $W = w_1, w_2, \ldots$ be such a sequence of F that $v_i = \neg\, w_i$ for $1 \leq i \leq k$. It can be proved by means of mathematical induction that $v_j = \neg\, w_j$ for every $j > k$. Since V is an arbitrary sequence of E then we have $E = \bar{F}$.

To prove that the inverse of the statement of Theorem 5 is not true, the reader is able to give a respective counter-example. Q.E.D,

[1] If $i = r = 0$ then the empty sequence is the threshold segment of Z.

4. Final conclusions

1. Let us consider a simple $(k, m, 2^k)$-machine $A_{k,m,2^k} = \langle M, \varphi_1, \ldots, \varphi_m \rangle$ such that all its transition functions are linear. Additionally, let every k-machine $A_k = \langle M, \varphi_i \rangle$ for $1 \leq i \leq m$ has all computations with the cycle length of 2^k.[1] For this case there is known the distribution of 0's and 1's in every restricted sequence $T\,|_{n2^k+1,\,(n+1)2^k}$ for $n \geq 0$, where T is a computation of $A_{k,m,2^k}$ (see [1]).

2. If we put $p \geqslant 2^k$ (p much more than 2_k) and if all transition functions of a (k, m, p)-machine are Boolean then the cycle of all its computations is long. If p increases then the length of the cycle increases too. The technical realization of such (k, m, p)-machines is very simple but the use is very broad (cryptology, communication system, simulation of the deterministic or nondeterministic processes).

5. The nondeterministic controlled (k, m, p)-machines

In this section we shall introduce (formally) the notion of a nondeterministic controlled (k, m, p)-machine which is the generalization of Kwasowiec's [4] as well as Meznik's [5] machine.

A necessary and sufficient condition for a set $E \subseteq M^\infty$ to be nondeterministic (k, m, p)-computation set will be given.

Definition 10. By a nondeterministic controlled (k, m, p)-machine $NA_{k,m,p}$ we mean an $(m + 2)$-tuple $\langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$, where every φ_i for $1 \leq i \leq m$ is a total function of M^k into M and ψ is a total function of $M^k \times H$ into a set of all nonvoid subsets of $\Phi = \{\varphi_1, \ldots, \varphi_m\}$ i.e. ψ: $M^k \times H \to 2^\Phi$ where $H = \{i \in N : i = jp + 1;\ j = 0, 1, 2, \ldots\}$.

Every φ_i is called a transition function and ψ — a control of $NA_{k,m,p}$: We shall assume that $\varphi_i \neq \varphi_j$ for $1 \leq i < j \leq m$.

Definition 11. Let $NA_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi_m, \psi \rangle$ be a (k, m, p)-machine as above and let $\Phi = \langle \varphi_1, \ldots, \varphi_m \}$.

An infinite sequence $T = t_1, t_2, \ldots$ is said to be computation of $NA_{k,m,p}$ iff the following condition is satisfied:

$$\forall_{n \geq 0}\, \forall_{(\varphi_{i_1}, \ldots, \varphi_{i_m}) \subseteq \Phi}\ [\psi(t_{np+1}, \ldots, t_{np+k}, np + 1) =$$

$$= \{\varphi_{i_1}, \ldots, \varphi_{i_j}\} \Rightarrow \exists\ \varphi_{i_q} \in \{\varphi_{i_j}, \ldots, \varphi_{i_j}\}\ \forall\ r$$

$$(np + 1 \leq r \leq (n + 1)\, p \Rightarrow t_{k+r} = \varphi_{i_q}(t_r, \ldots, t_{k+r-1})].$$

[1] An effective method of verifying if a linear k-machine $A_k = < M, \varphi >$ has all computations with the cycle length of 2^k or not has been shown in [1].

Theorem 6. A nonempty set $E \subseteq M^{\infty}$ is a nondeterministic (k, m, p)-computation set iff the following conditions are held:

(1) $\forall_{T \in M^k} \exists_{U \in E} (T = U|_{1,k})$;

(2) $\forall_{T \in E} \forall_{i \geq 1} (T|_{ip+1} \in E)$;

(3) $\forall_{X,Y \in E} \forall_{i \geq 0} \forall_{j \geq 0} [X|_{ip+1,(i+1)p+k} = Y|_{jp+1,(j+1)p+k} \Rightarrow X|_{1,(i+1)p+k}\, Y|_{(j+1)p+k+1} \in E]$;

(4) $\forall_{X,Y \in E} \forall_{i \geq 1} [i = np + 1 \;\&\; n \geq 0 \;\&\; X|_{i,i+k-1} = Y|_{i,i+k-1} \Rightarrow S_i(X) = S_i(Y)]$;

where $S_i(Z)$ denotes a set $\{T \in M^{k+p}\colon \exists_{U \in E} (T = U\,|_{i,i+k+p})\}$;
(5) there is a set $G = \mathrm{div}_{k,p}(E)$ such that $\mathrm{card}(G) = m$.

Proof. Necessity. Let us suppose that for the set E conditions (1)–(5) of Theorem 6 are satisfied. Let be $G = \mathrm{div}_{k,p}(E) = \{G_1, \ldots, G_m\}$.

For every $(t_1, \ldots, t_{k+1}) \in M^{k+1}$ and $1 \leq i \leq m$ let us define the total function $\varphi_i : M_k \to M$ as follows: $\varphi_i(t_1, \ldots, t_k) = t_{k+1}$ iff there is $T \in G_i$ and $j \leq 1(T) - k$ such that $t_1, \ldots, t_{k+1} = T|_{j,j+k}$, otherwise we put $\varphi_i(t_1 \ldots, \ldots, t_k) = 1$. Since every G_i, $1 \leq i \leq m$ is k-homogeneous set then φ_i is the function.

Let us define now a total function $\psi : M^k \times H \to 2^{\Phi}$ where $H = \{i \in N : i = jp + 1;\ j = 0, 1, 2, \ldots\}$ and $\Phi = \{\varphi_1, \ldots, \varphi_m\}$ as follows: $\psi(U, ip + 1) = \{\varphi_{i_1}, \ldots, \varphi_{i_p}\}$ iff there is a nonempty set $F = \{T_{i_1}, \ldots, T_{i_p}\}$ with the maximal cardinality such that $T_{i_j} \in G_{i_j}$ and $T_{i_j}|_{r,r+k+1} = U$ for every j, $1 \leq j \leq p$ and some $r \leq 1(T_{i_j}) - k$. If there is no a set F as above then we put $\psi(U, ip + 1) = \emptyset$. It follows from condition (4) that ψ is a function.

Let we put $NA_{k,m,p} = \langle M, \varphi_1, \ldots, \varphi, \psi \rangle$. It follows from the construction and conditions (1)–(5) that $E = C_{NA_{k,m,p}}$.

Proof of sufficiency is obvious. Q.E.D.

Remark 4. It is possible to give an equivalent definition of a nondeterministic controlled (k, m, p)-machine supported on the basis of graph theory. If we assume that the states of $NA_{k,m,p}$ are the nodes of the graph then with every initial state of $NA_{k,m,p}$ is associated an infinite oriented graph. Then $NA_{k,m,p}$ can be defined by means of a set of 2^k infinite oriented graphs.

However a rigorous definition of a nondeterministic controlled (k, m, p)-machine supported on the basis of graph theory has been not given here, the reader is able to give this definition easily.

6. Open problems

We shall put forward a few problems for the class of nondeterministic (k, m, p)-machines.

1. The synthesis problem can be stated as follows: for the given set $E \subseteq M^{\infty}$ we have to construct (in an effective way) a nondeterministic controlled (k, m, p)-machine $NA_{k,m,p}$ such that $E = C_{NA_{k,m,p}}$ (if such exists).

2. It is useful to solve all problems included in this paper and in [3].

3. What is the relation between the nondeterministic (k, m, p)-machines and the class of finite automaton.

4. According to Remark 4 a nondeterministic controlled (k, m, p)-machine can be defined by means of a set of 2^k oriented graphs.

An open problem is to give set-theoretical as well as topological properties of such sets of graphs.

References

1. *Golomb, S. W.*, Shift-register sequences, San Francisco Holden-Day (1967), 24—89.
2. *Grodzki, Z.*, The controlled shift-registers. Elektronische Informationsverarbeitung und Kybernetik **3** (1975)), 143-150.
3. *Grodzki, Z.*, New class of the controlled shift-registers. I. Problems of Control and Information Theory **5,** 2 (1976). 181-188
4. *Kwasowiec, W.*, Relational machines, Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. et Phys. **18** (1970) 545—549
5. *Meznik, I.*, G-machines and generable sets, Information and Control **20** (1972), 499—509.

Новый класс управляемых регистров сдвига. II

З. ГРОДЗКИ

(Люблин)

Настоящая статья является продолжением работы автора [3] и касается булевых управляемых (k, m, p)-регистров сдвига.

В статье рассматриваются далънейшие свойства множеств $(k, m, p,)$-вычислени Затем изучаются задачи, по отношению к которым множество $D^{\infty} = \bigcup_{k=1}^{\infty} \bigcup_{m=1}^{\infty} \bigcup_{p=1}^{\infty} D_{k,m,p}$ является замкнутым. В конце статьи изучается недетерминированный случай управляемых (k,m,p)-регистров сдвига и ставится несколько задач, которые должны быть решены в будущем.

Z. Grodzki
Higher Engineering School
20–950 Lublin
Dabrowskiegom 13
Poland

Problems of Control and Information Theory, Vol. 5 (3), pp. 271–284 (1976)

NECESSARY CONDITIONS FOR THE OPTIMALITY IN ABSTRACT OPTIMUM CONTROL PROBLEMS WITH NONSCALAR-VALUED PERFORMANCE CRITERION

B. LANTOS

(Budapest)

(Received October 27, 1975)

The principal aim of this paper is to give necessary conditions of attaining optimum (minimum or infimum) for constrained nonscalar-valued functions. The range of the function is a partially ordered space. The method of Dubovitskij–Milyutin–Girsanov for scalar-valued functions is extended to the case in which the range of the function is a partially ordered linear topological space and the function has a minimum. A reasonably complete theory of the necessary conditions for the optimum (minimum or infimum) is detailed for the case in which the function is Frechet differentiable and its range is a partially ordered reflexive Banach space.

$\in$	element of
$\notin$	not element of
$\therefore$	implies
$\leftrightarrow$	if and only if
$\emptyset$	empty set
τ	topology
A°	interior of A
$\overline{A}$	closure of A
A^c	complement of A
R^n	finite-dimensional normed space
$F : E_1 \to E_2$	mapping from E_1 into E_2
$D(F)$	domain of F
$R(F)$	range of F
$F'(x)$	Frechet derivative of F
E'	continuous linear functionals defined on E
$\mathfrak{L}(E_1 \to E_2)$	linear operators from E_1 into E_2
$\mathfrak{B}(E_1 \to E_2)$	bounded linear operators from E_1 into E_2
$\langle x, y \rangle$	inner product
K, P	cones (with vertex 0)

$K^* = \{f \in E' : f(K) \geq 0\}$ polar cone of K

$A^* \in \mathfrak{L}(E_2' \to E_1')$ adjoint of $A \in \mathfrak{L}(E_1 \to E_2)$

$\mathrm{Ker} A = \{x \in E_1 : Ax = 0\}$ null space of $A \in \mathfrak{L}(E_1 \to E_2)$

$M(K, P) = \{A \in \mathfrak{B}(E_1 \to E_2) : A(K) \subset P\}$ maximal set of bounded linear operators mapping $K \subset E_1$ into $P \subset E_2$

$F \circ G$ composition of the functions F and G such that $F \circ G(x) = (FG(x))$

Introduction

In many optimization problems the quality of the process cannot be characterized by a single scalar-valued optimality criterion, because the user is simultaneously interested in several cost functionals. The scalar-valued cost functionals can be reduced to a single vector-valued performance criterion.

The linear state estimation problem leads also to an optimum problem with nonscalar-valued performance criterion, if the covariance matrix of the error between the state and its estimation must be infimum. In this case the performance criterion is matrix-valued.

Optimization problems with nonscalar-valued performance criteria are studied in the paper. The meaning of "better than" has to be defined, which is done by a partial-order relation. The partial ordering is defined by a positive cone. Like the minimization of a scalar-valued function, the constraints are approximated by their admissible or tangential directions (method of Dubovitskij–Milyutin–Girsanov, [1]). For the decreasing directions of a nonscalar-valued function a new definition is introduced. The necessary conditions of the optimum are summarized in three theorems. These theorems establish a relation between the method of Dubovitskij–Milyutin–Girsanov [1], the results of Neustadt [2] and the results of Ritter [4].

The special case of the problem studied here, when the range of the performance criterion is a finite-dimensional partially ordered normed space, was considered by Athans and Geering [5, 6].

The proofs of the lemmas in the Appendix of this paper can be found in the author's dissertation.

Partial ordering

The partial ordering on a set is a reflexive, antisymmetric and transitive relation. If the set is a linear topological space, then it will be supposed that the partial ordering is given by a closed and convex cone having a nonempty interior.

Definition 1. Let (E, τ) be a linear topological space, let $P \subset E$ be a closed and convex cone such that $P^\circ \neq \emptyset$, $0 \in P$ and $\pm z \in P \Rightarrow z = 0$. We say that $x \geq y$ if $x, y \in E$ and $x - y \in P$. A linear topological space with a relation $\geq$ defined in this way is said to be a partial ordered linear topological space. Notation: $(E, \tau, \geq)$. Since $x \in P \Leftrightarrow x \geq 0$, the cone P will be called the positive cone (defining the relation $\geq$).

Example 1. If $P = \{x = (x_1, \ldots, x_n) \in R^n : x_i \geq 0, \quad i = 1, \ldots, n\}$, then P is a positive cone in R^n and so P defines a partial ordering in R^n. Notation: $(R^n, \geq)$. If $\| x \| = \| y \| = 1$ and $x, y \in P$, then $\| x + y \| \geq 1$.

Example 2. Let H be a Hilbert space. If $E = \{A \in \mathfrak{B}(H \to H)$: A is self-adjoint$\}$ and $P = \{A \in E : \langle Ax, x\rangle \geq 0$ for all $x \in H\}$, then $E \subset \mathfrak{B}(H \to H)$ is a closed subspace $\Rightarrow E$ is a Banach space and $P \subset E$ is a positive cone in E. Hence P defines a partial ordering in E. Notation: $(E, \geq)$. If $\|A\| = \|B\| = 1$ and $A,\ B \in P$, then $\|A + B\| \geq 1$.

Exampie 3. Notation is as in Example 2. Let $H = R^n$ (with fixed orthonormal basis). Then $\mathfrak{B}(R^n \to R^n)$ can be identified with the set of $n \times n$ matrices and similarly E with the set of symmetric $n \times n$ matrices. Then P is the set of positive semidefinite symmetric $n \times n$ matrices. The positive cone P defines a partial ordering in the Banach space of the symmetric $n \times n$ matrices. Notation: $(M^s_{n\times n}, \geq)$. $M^s_{n\times n}$ can be considere as an $\frac{n(n+1)}{2}$-dimensional subspace of the normed space R^{n^2} (without inner product).

Remark. If $(E, \tau, \geq)$ is a partially ordered linear topological space, then $x \geq y$ and $z \in E \Rightarrow x + z \geq y + z$.

Necessary conditions of optimality

Definition 2. Let $(E, \tau, \geq)$ be a partially ordered linear topological space, $Q \subset E$ and $x_0 \in Q$. We say that

1) $x_0 = \max Q$, if there does not exist any $x \in Q$ such that $x \geq x_0$ and $x \neq x_0$,

2) $x_0 = \min Q$, if there does not exist any $x \in Q$ such that $x_0 \geq x$ and $x \neq x_0$,

3) $x_0 = \sup Q$, if $x_0 \geq x$ for all $x \in Q$,

4) $x_0 = \inf Q$, if $x \geq x_0$ for all $x \in Q$.

In general, $\max Q$ and $\min Q$ are not unique, because the partial ordering is usually not a linear ordering (Q may have elements which are not comparable). $\operatorname{Sup} Q$ and $\inf Q$ are always unique (supposed that they exist).

Definition 3. Let (E, τ) be a linear topological space, let $(E_0, \tau_0, \geq)$ be a partially ordered linear topological space and let $F_0 : E \to E_0$ be a mapping. We say that $h \in E$ is a decreasing direction of the function F_0 in the point $x_0 \in E$, if there exist a neighbourhood V of $0 \in E$ and a real number $\varepsilon_0 > 0$ such that

$$F_0(x_0) \geq F_0(x_0 + \varepsilon\bar{h}) \quad \text{and} \quad F_0(x_0) \neq F_0(x_0 + \varepsilon\bar{h}) \tag{1}$$

for all $h \in h + V$ and $0 < \varepsilon < \varepsilon_0$.

Remark. If the partial ordering $\geq$ in the linear topological space (E_0, τ_0) is defined by the positive cone P_0, then

$$F_0(x_0) \geq F_0(x_0 + \varepsilon\bar{h}) \Leftrightarrow F_0(x_0) - F_0(x_0 + \varepsilon\bar{h}) \in P_0 . \tag{2}$$

Definition 4. Let (E, τ) be a linear topological space, let $Q \subset E$ and $x_0 \in E$. We say that $h \in E$ is an admissible direction of the set Q in the point x_0, if there exists a neighbourhood V of $0 \in E$ and a real number $\varepsilon_0 > 0$ such that

$$x_0 + \varepsilon\bar{h} \in Q \tag{3}$$

for all $\bar{h} \in h + V$ and $0 < \varepsilon < \varepsilon$

Remark. Let K be the set of admissible directions of the set Q in the point x_0, then K is an open cone. If $x_0 \in Q^0$, then $K = E$. If $x_0 \notin \bar{Q}$, then $K = \emptyset$.

Definition 5. Let (E, τ) be a linear topological space, let $Q \subset E$ and $x_0 \in E$. We say that $h \in E$ is a tangential direction of the set Q in the point x_0, if there exist a real number $\varepsilon_0 > 0$ and a mapping $r : (0, \varepsilon_0) \to E$ such that

$$x(\varepsilon) = x_0 + \varepsilon h + r(\varepsilon) \in Q \tag{4}$$

for all $0 < \varepsilon < \varepsilon_0$, furthermore if V is an arbitrary neighbourhood of $0 \in E$, then there exists a real number $0 < \eta_0(V) \leq \varepsilon_0$ such that

$$\frac{1}{\varepsilon} r(\varepsilon) \in V \tag{5}$$

for all $0 < \varepsilon < \eta_0(V)$.

Remark. Let K be the set of tangential directions of the set Q in the point x_0, then K is a cone.

Theorem 1. Let (E, τ) be a linear topological space, let $(E_0, \tau_0, \geq)$ be a partially ordered linear topological space, let $Q = \bigcap_{i=1}^{n+1} Q_i \subset E$, $x_0 \in Q$ and let $F_0 : E \to E_0$ be a mapping. Let S_0 denote the set of decreasing directions of the function F_0, let S_i denote the set of admissible directions of the set Q_i, $i = 1, \ldots, n$ and let S_{n+1} be the set of tangential directions of the set Q_{n+1} (understood in the point x_0 respectively). Suppose

1) there exist open and convex cones K_i such that $\emptyset \neq K_i \subset S_i$ for $i = 0, \ldots, n$,

2) there exists a convex cone K_{n+1} such that $\emptyset \neq K_{n+1} \subset S_{n+1}$,

3) there is neighbourhood U of x_0 such that

$$\min \{F_0(x) : x \in Q \cap U\} = F_0(x_0)\,. \tag{6}$$

Then there are continuous linear functionals $f_i \in K_i^*$, $i = 0, \ldots, n+1$ such that

$$\text{i)} \quad f_0 + f_1 + \ldots + f_{n+1} = 0, \tag{7}$$

ii) $f_0 \neq 0$ for at least one i.

Proof. It will be proved

$$\bigcap_{i=0}^{n+1} K_i = \emptyset\,.$$

On the contrary, suppose $h \in \bigcap_{i=0}^{n+1} K_i$. It follows from (1) and (3) that there are a neighbourhood V of $0 \in E$ and a real number $\varepsilon_1 > 0$ such that

$$F_0(x_0) \geq F_0(x_0 + \varepsilon\bar{h}) \text{ and } F_0(x_0) \neq F_0(x_0 + \varepsilon\bar{h}),$$

$$x_0 + \varepsilon\bar{h} \in \bigcap_{i=1}^{n} Q_i \tag{9}$$

for all $\bar{h} \in h + V$ and $0 < \varepsilon < \varepsilon_1$. Since $h \in K_{n+1} \subset S_{n+1}$, it follows from (4) and (5) that there are real numbers $\varepsilon_2 > 0$ and $0 < \eta_2(V) \leq \varepsilon_2$ and a mapping $r : (0, \varepsilon_2) \to E$ such that

$$x(\varepsilon) = x_0 + \varepsilon h + r(\varepsilon) \in Q_{n+1}\,,$$

$$\frac{1}{\varepsilon} r(\varepsilon) \in V \tag{10}$$

for all $0 < \varepsilon < \eta_2(V)$. Let $\varepsilon_0 = \min \{\varepsilon_1, \eta_2(V)\}$, then it follows from (9) and (10) that

$$\bar{h}(\varepsilon) \in h + V\,,$$

$$x(\varepsilon) = x_0 + \varepsilon\bar{h}(\varepsilon) \in Q\,, \tag{11}$$

$$F_0(x_0) \geq F_0(x_0 + \varepsilon\bar{h}(\varepsilon)) \text{ and } F_0(x_0) \neq F_0(x_0 + \varepsilon\bar{h}(\varepsilon))$$

for all $0 < \varepsilon < \varepsilon_0$ and $\bar{h}(\varepsilon) = h + \frac{1}{\varepsilon} r(\varepsilon)$. But for ε sufficiently small $x_0 + \varepsilon\bar{h}(\varepsilon) \in Q \cap U$ is also valid and so (11) contradicts to (6) and to the definition of the minimum. Hence $\bigcap_{i=0}^{n+1} K_i = \emptyset$ and (7) follows from Lemma 5.11 (Dubovitskij–Milyutin) in [1].

Theorem 2. Suppose the next conditions are satisfied:

1) E and E_i are Banach spaces, $i = 0, \ldots, n + k$;

2) $P_i \subset E_i$ is a closed and convex cone, $0 \notin P_i^0 \neq \emptyset$ and P_i (as a positive cone) defines a partial ordering $\geq$ in the Banach space E_i, $i = 0, \ldots, n$;

3) $F_i : E \to E_i$ is a mapping which has a Frechet derivative $F_i'(x_0)$ in the point x_0, $i = 0, \ldots, n$;

4) $F_i : E \to E_i$ is a mapping which is continuously Frechet differentiable in a neighbourhood of x_0 and for which $R(F_i'(x_0))$ is closed in E_i, $i = n + 1, \ldots, n + k$;

5) $A \subset E$ is a convex set and $A^\circ \neq \emptyset$;

6) $Q = \{x \in E : -F_i(x) \geq 0,\ i = 1, \ldots, n;\ F_i(x) = 0,\ i = n + 1, \ldots, n + k;\ x \in A\}$

is a constraint, $x_0 \in Q$ and there exists a neighbourhood U of x_0 such that

$$\min \{F_0(x) : x \in Q \cap U\} = F_0(x_0). \tag{12}$$

Then there are continuous linear functionals $f_i \in E_i'$, $i = 0, \ldots, n + k$ such that

i) $f_i \in P_i^*$ i.e. $f_i(P_i) \geq 0$, $i = 0, \ldots, n$,

ii) with the notation $\varphi = \sum_{i=0}^{n+k} f_i \circ F_i'(x_0)$ the inequality $\varphi(x) \geq \varphi(x_0)$ holds for all $x \in A$,

iii) $f_i \neq 0$ for at least one i,

iv) if $i \in \{1, \ldots, n\}$ and $-F_i(x_0) \in P_i^0$, then $f_i = 0$,

v) specially if the system

$$\begin{aligned} &R(F_i'(x_0)) = E_i, \qquad i = n + 1, \ldots, n + k, \\ &R(-F_i'(x_0)) \cap P_i^0 \neq \emptyset, \\ &i \in \{1, \ldots, n\} \text{ and } -F_i(x_0) \notin P_i^0 \Rightarrow R(-F_i'(x_0)) \subset P_i^0 \neq \emptyset \end{aligned} \tag{13}$$

can also be satisfied, then $f_i \circ F_i'(x_0) \neq \emptyset$ for at least one i; if in addition the system

$$\begin{aligned} &F_i'(x_0)(x - x_0) = 0, \qquad i = n + 1, \ldots, n + k, \\ &i \in \{1, \ldots, n\} \text{ and } -F_i(x_0) \notin P_i^0 \Rightarrow -F_i'(x_0)(x - x_0) \in P_i^0 \\ &x \in A^0 \end{aligned} \tag{14}$$

has also a solution in x, then $f_0 \circ F_0'(x_0) \neq 0$.

Proof. Case I. Suppose that there is an $i \in \{0, \ldots, n\}$ such that $i \neq 0$, $-F_i(x_0) \notin P_i^0$ and $R(-F_i'(x_0)) \cap P_i^0 = \emptyset$, or $i = 0$ and $R(-F_i'(x_0)) \cap P_0^0 = \emptyset$.

Since $R(-F_i'(x_0))$ is a subspace it is a convex set. By Theorem 3.3 in [1] $R(-F_i'(x_0))$ and P_i can be separated, i.e. there is a nonzero $f_i \in E_i'$ such that $f_i(R(-F_i'(x_0))) = 0 \leq f_i(P_i)$. Hence $f \circ F_i'(x_0) = 0$ and $f_i \in P_i^*$. Let $f_j = 0$ for $j \neq i$, then the theorem is valid.

Case II. Suppose that there exists an $i \in \{n+1, \ldots, n+1\}$ such that $R(F_i'(x_0)) \neq E_i$.

Let $y_i \in E_i \setminus R(F_i'(x_0))$. By Theorem 3.4 in [1] the closed subspace $R(F_i'(x_0))$ and y_i can be strictly separated, i.e. there is a nonzero $f_i \in E_i'$ such that $f_i(R(F_i'(x_0))) = 0 > f_i(y_i)$. Hence $f_i \circ F_i'(x_0) = 0$. Let $f_j = 0$ for $j \neq i$. Then the functionals $f_0, \ldots, f_{n+k}$ satisfy the theorem.

Case III. Suppose $R(F_i'(x_0)) = E_i$, $i = n+1, \ldots, n+k$ and the set $R(\Phi)$ is a proper subset of $E_{n+1} \times \ldots \times E_{n+k}$, where $\Phi = (F_{n+1}'(x_0), \ldots, F_{n+k}'(x_0))$. It follows from Theorem 3.4 in [1] and from the proof of Lemma 2.4 in [4] that there are functionals $f_i \in E_i'$, $i = n+1, \ldots n+k$ such that $\sum_{i=n+1}^{n+k} f_i \circ F_i'(x_0) = 0$ and $f_i \circ F_i'(x_0) \neq 0$ for at least one i. Let $f_j = 0$, $j = 0, \ldots, n$, then the theorem is valid.

Case IV. Suppose none of cases I, II and III is valid. We introduce the following notations:

$$Q_0 = \{x \in E : F_0(x_0) - F_0(x) \in P_0\}, \tag{15}$$

$$Q_i = \{x \in E : -F_i(x) \in P_i\}, \; i = 1, \ldots, n, \tag{16}$$

$$Q_{n+1} = \{x \in E : F_i(x) = 0, i = n+1, \ldots, n+k\}. \tag{17}$$

$$I = \{i \in \{1, \ldots, n\} : -F_i(x_0) \notin P_i^0\}. \tag{18}$$

Then the constraint Q has the form

$$Q = \bigcap_{i=1}^{n+1} Q_i \cap A. \tag{19}$$

1) If $i \in \{1, \ldots, n\} \setminus I$, then $x_0 \in Q_i^0$ and so the cone $K_a^{(i)}$ of admissible directions of the set Q_i in the point x_0 is $K_a^{(i)} = E \Rightarrow K_a^{(i)*} = \{0 \in E'\}$. If $i \in I$, then $R(-F_i'(x_0)) \cap P_i^0 \neq \emptyset$. By Lemma 1A (in the Appendix) the cone $K_a^{(i)}$ of admissible directions of the set Q_i in the point x_0 contains the nonempty open and convex cone K_i, where

$$K_i = \{h \in E : -F_i'(x_0)h \in P_i^0\} \subset K_a^{(i)}. \tag{20}$$

Since in Banach spaces the closed and convex sets are the same in the norm topology and in the weak topology, it follows from $\overline{P_i^0} = P_i$ and from Lemma

5.3 in [1] that $(P_i^0)^* = P_i^*$. Furthermore it follows from Theorem 10.4 (Minkowski–Farkas) in [1] that

$$K_i^* = (-F_i'(x_0))^* (P_i^0)^* = (-F_i'(x_0))^* P_i^*. \tag{21}$$

Hence $\varphi_i \in K_i^* \Leftrightarrow$ there is a functional $f_i \in P_i^*$ such that $\varphi_i = -f_i \circ F_i'(x_0)$.

2) Let K_t^{n+1} denote the cone of tangential directions of the set Q_{n+1} in the point x_0. By Theorem 9.1 (Lusternik) in [1]

$$K_t^{(n+1)} = \bigcap_{i=n+1}^{n+k} \operatorname{Ker} F_i'(x_0)\,. \tag{22}$$

Since $K_t^{(n+)}$ is a subspace, so if $\varphi_{n+1} \in K_t^{(n+1)*}$, then $\varphi_{n+1}(K_t^{(n+1)}) = 0$ and

$$\bigcap_{i=n+1}^{n+k} \operatorname{Ker} F_i'(x_0) \subset \operatorname{Ker} \varphi_{n+1} \tag{23}$$

By Lemma 2.2 in [4] and by Theorem 3.1 (Hahn–Banach) in [1] there exist functionals $f_i \in E_i'$, $i = n+1, \ldots, n+k$ such that

$$\varphi_{n+1} = \sum_{i=n+1}^{n+k} f^i \circ F_i'(x_0)\,. \tag{24}$$

Inversely, if $f_i \in E_i'$, $i = n+1, \ldots, n+k$, then

$$\varphi_{n+1}(x) \overset{\text{def}}{=} \sum_{i=n+1}^{n+k} f_i \circ F_i'(x_0)\,(x) = \sum_{i=n+1}^{n+k} f_i(0) = (0) \tag{25}$$

for all $x \in \bigcap_{i=n+1}^{n+k} \operatorname{Ker} F_i'(x_0)$ and so $\varphi_{n+1} \in K_t^{(n+1)*}$. Hence

$$K_t^{(n+1)*} = \sum_{i=n+1}^{n+k} (F_i'(x_0))^* E_i'\,. \tag{26}$$

3) Since $R(-F_0'(x_0)) \cap P_0^0 \neq \emptyset$, it follows from Lemma 1A that the cone $K_a^{(0)}$ of the admissible directions of the set Q_0 in the point x_0 contains the nonempty open and convex cone K_0, where

$$K_0 = \{h \in E : -F_0'(x_0)h \in P_0^0\} \subset K_a^{(0)}. \tag{27}$$

Let $h \in K_0$, then by (3) there are real numbers $\delta_0 > 0$ and $\varepsilon_0 > 0$ such that

$$x_0 + \varepsilon\bar{h} \in Q_0\,, \tag{28}$$

$$F_0(x_0) - F_0(x_0 + \varepsilon\bar{h}) \in P_0 \Leftrightarrow F_0(x_0) \geq F_0(x_0 + \varepsilon\bar{h})$$

for all $\| \bar{h} - h \| < \delta_0$ and $0 < \varepsilon < \varepsilon_0$. Since P_0^0 is an open cone, it follows from $-F_0'(x_0)\, h \in P_0^0$ and from $0 \notin P_0^0$ that for sufficiently small $\delta_0 > 0$ and $\varepsilon_0 > 0$

$$-F_0(x_0 + \varepsilon\bar{h}) + F_0(x_0) \in P_0^0,$$
$$F_0(x_0) \neq F_0(x_0 + \varepsilon\bar{h}). \tag{29}$$

Let S_0 denote the set of decreasing directions of the function F_0 in the point x_0, then by (1), (28) and (29)

$$K_0 \subset S_0, \tag{30}$$

where K_0 is a nonempty open and convex cone. By Theorem 10.4 (Minkowski–Farkas) in [1]

$$K_0^* = (-F_0'(x_0))^*\, P_0^*. \tag{31}$$

4) Let $K_a^{(A)}$ denote the cone of admissible directions of the set A in the point x_0. By Theorem 8.2 in [1]

$$K_a^{(A)} = \{\lambda(x - x_0) : x \in A^0,\ \lambda > 0\} \neq \emptyset \tag{32}$$

and the cone $K_a^{(A)}$ is open and convex. By Theorem 10.5 in [1]

$$K_a^{(A)*} = \{\varphi \in E' : \varphi(x) \geq \varphi(x_0) \text{ for all } x \in A\}. \tag{33}$$

5) By (20), (22), (30) and (32) Theorem 1 can be applied. Hence there are functionals $\varphi_0 \in K_0^*$; $\varphi_i \in K_i^*$, $i \in I$; $\varphi_{n+1} \in K_l^{(n+1)*}$ and $\varphi \in K_a^{(A)*}$ such that

i) $\varphi_0 + \sum_{i \in I} \varphi_i + \varphi_{n+1} + \varphi = 0 \Rightarrow \varphi = -\varphi_0 - \sum_{i \in I} \varphi_i - \varphi_{n+1}$, (34)

ii) $\varphi_i \neq 0$ for at least one i.

Now by (31), (21) and (26) the theorem follows from (34) and (33).

6) Let

$$\bigcap_{i \in I} K_i \cap K_a^{(A)} \cap K_l^{(n+1)} \neq \emptyset,$$

then by Remark 3 on page 41 in [1], $f_0 \circ F_0'(x_0) \neq 0$. In this case the system

$$-F_i'(x_0)\, h \in P_i^0, \quad i \in I,$$
$$h = \lambda(x - x_0), \quad x \in A^0, \quad \lambda > 0, \tag{35}$$
$$F_i'(x_0)\, h = 0, \quad i = n+1, \ldots, n+k$$

can be solved. Since P_i^0 is a cone and $F_i'(x_0)$ is a linear operator, it follows from $\lambda > 0$, that (35) can be solved if and only if the system

$$\begin{aligned} -F_i'(x_0)\,(x - x_0) &\in P_i^0, \qquad i \in I, \\ x &\in A^0, \\ F_i'(x_0)\,(x - x_0) &= 0, \qquad i = n+1, \ldots, n+k \end{aligned} \tag{36}$$

has a solution in x. Thus the theorem is proved.

Theorem 3. Suppose that the next conditions are satisfied.

1) E and E_i are Banach spaces, $i = 1, \ldots, n+k$, E_0 is a reflexive Banach space;

2) $P_i \subset E_i$ is a closed and convex cone, $0 \notin P_i^0 \neq \emptyset$ and P^i (as a positive cone) defines a partial ordering $\geq$ in the Banach space E_i, $i = 0, \ldots, n$; furthermore there is a real number $\delta > 0$ such that for all $y_1, y_2 \in P_0$ and $\| y_1 \| = \| y_2 \| = 1$ it is $\| y_1 + y_2 \| \geq \delta$;

3) $F_i : E \to E_i$ is a mapping which has a Frechet derivative $F_i'(x_0)$ in the point x_0, $i = 0, \ldots, n$ and for which $R(F_i'(x_0))$ is closed in E_i, $i = 1, \ldots, n$;

4) $F_i : E \to E_i$ is a mapping which is continuously Frechet differentiable in a neighbourhood of x_0 and for which $R(F_i'(x_0))$ is closed in E_i, $i = n+1, \ldots, n+k$;

5) $A \subset E$ is a convex set and $A^0 \neq \emptyset$;

6) $Q = \{x \in E : -F_i(x) \geq 0,\ i = 1, \ldots, n;\ F_i(x) = 0,\ i = n+1, \ldots, n+k;\ x \in A\}$ is a constraint, $x_0 \in Q$ and there exists a neighbourhood U of x_0 such that

$$\inf \{F_0(x) : x \in Q \cap U\} = F_0(x_0)\,. \tag{37}$$

Then there are linear mappings $T_i \in \mathfrak{L}(E_i \to E_0)$, which are continuous on $R(F_i'(x_0))$, $i = 0, \ldots, n+k$ and for which

i) $T_0 y_0 \in P_0$ i.e. $T_0 y_0 \geq 0$ for all $y_0 \in P_0$; furthermore $T_i y_i \in P_i$ i.e. $T_i y_i \geq 0$ for all $y_i \in R(F_i'(x_0)) \cap P_i$, $i = 1, \ldots, n$,

ii) with the notation $T = \sum_{i=0}^{n+k} T_i \circ F_i'(x_0)$ the inequality $Tx \geq Tx_0$ i.e. $Tx - Tx_0 \in P_0$ holds for all $x \in A$,

iii) $T_i \neq 0$ for at least one i,

iv) if $i \in \{1, \ldots, n\}$ and $-F_i(x_0) \in P_i^0$, then $T_i = 0$,

v) specially if the system (13) can also be satisfied, then $T_i \circ F_i'(x_0) \neq 0$ for at least one i; if in addition the system (14) has also a solution in x, then $T_0 = I$ is the identity operator.

Proof. We will use the notations of Theorem 2. Let $0 \neq y \in P_0$ and

$$T_i \overset{\text{def}}{=} yf_i, \qquad i = 0, \ldots, n + k, \tag{38}$$

then with the exception of Case IV/6 i.e.

$$\bigcap_{i \in I} K_i \cap K_a^{(A)} \cap K_t^{(n+1)} \neq \emptyset, \tag{39}$$

Theorem 3 follows from Theorem 2. Hence for proving Theorem 3 we can suppose that (39) is satisfied.

First it will be proved $h \in \bigcap_{i \in I} K_i \cap K_a^{(A)} \cap K_t^{(n+1)} \Rightarrow F_0'(x_0)\, h \notin P_0$. On the contrary, suppose $h \in \bigcap_{i \in I} K_i \cap K_a^{(A)} \cap K_t^{(n+1)}$ but $F_0'(x_0)h \notin P_0$. Since the cone P_0^c is open and $0 \notin P_0^c$, there are real numbers $\delta_1 > 0$ and $\varepsilon_1 > 0$ such that

$$F_0(x_0 + \varepsilon \bar{h}) - F_0(x_0) \notin P_0 \tag{40}$$

for all $\| \bar{h} - h \| < \delta_1$ and $0 < \varepsilon < \varepsilon_1$. By Definition 4 there exist real numbers $\delta_2 > 0$ and $\varepsilon_2 > 0$ such that

$$x_0 + \varepsilon \bar{h} \in \bigcap_{i=1}^{n} Q_i \cap A \tag{41}$$

for all $\| \bar{h} - h \| < \delta_2$ and $0 < \varepsilon < \varepsilon_2$. Let $\delta_0 = \min \{\delta_1, \delta_2\}$. By Definition 5 there are real numbers $\varepsilon_3 > 0$ and $0 < \eta_3(\delta_0) \leq \varepsilon_3$ and a mapping $r : (0, \varepsilon_3) \to E$ such that

$$x(\varepsilon) = x_0 + \varepsilon h + r(\varepsilon) \in Q_{n+1},$$

$$\left\| \frac{1}{\varepsilon} r(\varepsilon) \right\| < \delta_0$$

for all $0 < \varepsilon < \eta_3(\delta_0)$. Let $\varepsilon_0 = \min \{\varepsilon_1, \varepsilon_2, \eta_3(\delta_0)\}$ then

$$x(\varepsilon) = x_0 + \varepsilon \bar{h}(\varepsilon) \in Q \tag{43}$$

for all $0 < \varepsilon < \varepsilon_0$, $\bar{h}(\varepsilon) = h + \dfrac{1}{\varepsilon} r(\varepsilon)$ and $\| h(\varepsilon) - h \| < \delta_0$. But for ε sufficiently small $x_0 + \varepsilon \bar{h}(\varepsilon) \in Q \cap U$ is also valid and so (43) and (40) contradict to (37) and to the definition of the infimum. This contradiction proves $F_0'(x_0)h \in P_0$. Hence

$$F_0'(x_0) \in M\Big(\bigcap_{i \in I} K_i \cap K_a^{(A)} \cap K_t^{(n+1)}, P_0\Big). \tag{44}$$

It follows from Theorem 4.9 (Ritter) in [3] that there are linear operators $G_i \in M(K_i, P_0)$, $i \in I$; $T \in M(K_a^{(A)}, P_0)$ and $G_t \in M(K_t^{(n+1)}, P_0)$ such that

$$F_0'(x_0) = \sum_{i \in I} G_i + T + G_t. \tag{45}$$

Since $G_i(K_i) \subset P_0$ and by (20) $K_i = \{h \in E : -F_i'(x_0)h \in P_i^0\}$, it follows $\{h \in E : -F_i'(x_0)h \in P_i^0\} \subset \{h \in E : G_i h \in P_i\}$. Hence by Lemma 2A there is a linear operator $T_i \in \mathfrak{L}(E_i \to E_0)$ such that $G_i = -T_i \circ F_i'(x_0)$ and T_i is continuous on $R(F_i'(x_0))$ and $T_i\hat{y}_i \in P_0$ for all $\hat{y}_i \in R(-F_i'(x_0)) \cap P^0$. Let $y_i \in R(-F_i'(x_0)) \cap P_i$. Since P_i is a convex cone and $R(F_i'(x_0))$ is a subspace, follows

$$y_i + \lambda_i\hat{y} \in R(F_i'(x_0)) \cap P_i^0 \Rightarrow T_i(y_i + \lambda\hat{y}_i) \in P_0 \tag{46}$$

for all $\lambda > 0$. Since P_0 is closed and T_i is continuous on $R(F_i'(x_0))$, it follows from (46)

$$T_i y_i = \lim_{\lambda \downarrow 0} T_i(y_i + \lambda\hat{y}_i) \in P_0. \tag{47}$$

Because $K_l^{(n+1)} = \bigcap_{i=n+1}^{n+k} \operatorname{Ker} F_i'(x_0)$ is a subspace and $G_l(K_l^{(n+1)}) \subset P_0$, so if $x \in K_l^{(n+1)}$, then $\pm G_l x \in P_0$ and according to the condition 2) in the theorem, $G_l x = 0$. Hence

$$\bigcap_{i=n+1}^{n+k} \operatorname{Ker} F_i'(x_0) \subset \operatorname{Ker} G_l \tag{48}$$

and by Lemma 2.2 in [4] there are linear operators $T_i \in \mathfrak{L}(E_i \to E_0)$, T_i is continuous on $R(F_i'(x_0))$, $i = n+1, \ldots, n+k$ such that

$$G_l = -\sum_{i=n+1}^{n+k} T_i \circ F_i'(x_0). \tag{49}$$

Since $K_a^{(A)} = \{\lambda(\hat{x} - x_0) : \hat{x} \in A^0 \text{ and } \lambda > 0\}$ and $T(K_a^{(A)}) - P_0$, thus

$$T(\lambda(\hat{x} - x_0)) = \lambda(T\hat{x} - Tx_0) \in P_0 \Rightarrow T\hat{x} - Tx_0 \in P_0 \tag{50}$$

for all $\hat{x} \in A^0$ and $\lambda > 0$. Let $x \in A$, then $(1-\alpha)x + \alpha\hat{x} \in A^0$ for all $0 < \alpha < 1$. Hence $T((1-\alpha)x + \alpha\hat{x} - x_0) \in P_0$ follows from (50). Since P_0 is closed and T is continuous, thus

$$Tx - Tx_0 = \lim_{\alpha \downarrow 0} T((1-\alpha)x + \alpha\hat{x} - x_0) \in P_0 \tag{51}$$

is satisfied. Now the theorem follows from (45), (47), (49) and (51).

Remark. The case $U = E$ in Theorems 1, 2 and 3 gives necessary conditions for global optimality.

Appendix

Lemma 1A. Let E and E_1 be linear normed spaces, let $P_1 \subset E_1$ be a closed and convex cone for which $0 \notin P_1^0 \neq \emptyset$, let $F : E \to E_1$ be a mapping which has a Frechet derivative $F'(x_0)$ in the point x_0 and let $Q = \{x \in E :$

$F(x) \in P_1\}$ and $x_0 \in Q$ and $F(x_0) \notin P_1^0$. Suppose that there is a $h_0 \in E$ such that $F'(x_0)h_0 \in P_1^0$. Let $K = \{h \in E : F'(x_0)\, h \in P_1^0\}$ and let K_a denote the cone of admissible directions of the set Q in the point x_0. Then $K \subset K_a$ and K is an open and convex cone.

Remark. $K \subset \widehat{K} \overset{\text{def}}{=} \{h \in E : F'(x_0)\, h \in P_1^0 - \{\lambda F(x_0) : \lambda \geq 0\}\} \subset K_a$ is also valid, hence Theorem 2 and Theorem 3 can be sharpened. If $E_1 = R^1$, then $K = \widehat{K}$.

Lemma 2A. Let E, E_0 and E_1 be linear normed spaces, let $A \in \mathfrak{B}(E \to E_1)$ and $B \in \mathfrak{B}(E \to E_0)$ be linear operators, let $P_0 \subset E_0$ and $P_1 \subset E_1$ be convex cones such that $\pm y \in P_0 \Rightarrow y = 0$ and let $0 \in P_1$ or let P_0 be closed.

1) If $\operatorname{Ker} A \subset \{x \in E : Bx \in P_0\}$, then there exists a linear operator $T \in \mathfrak{L}(E_1 \to E_0)$ such that $B = T \circ A$. If E and E_1 are Banach spaces and $R(A)$ is closed in E_1, then T is continuous on $R(A)$.

2) If $\{x \in E : Ax \in P_1\} \subset \{x \in E : Bx \in P_0\}$, then there exists a linear operator $T \in \mathfrak{L}(E_1 \to E_0)$ such that $B = T \circ A$ and $Ty \in P_0$ for all $y \in R(A) \cap \cap P_1$. If E and E_1 are Banach spaces and $R(A)$ is closed in E_1, then T is continuous on $R(A)$.

References

1. *Гирсан И. В.*, Лекции по математической теории экстремальных задач. Изд. Московского Университета, 1970.
2. *Neustadt, L* A General Theory of Extremals. Journal of Computer and System Sciences **3** (19 9) 57–92.
3. *Ritter, K.*, Optimization Theory in Linear Spaces. I. Mathematische Annalen **182** (1969) 189–206.
4. *Ritter, K.*, Optimization Theory in Linear Spaces. III. Mathematische Annalen **184** (1970) 133–154.
5. *Athans, M., Geering, H. P.*, Necessary and Sufficient Conditions for Differentiable Nonscalar-Valued Functions to Attain Extrema. IEEE Transactions on Automatic Control **18** (1973) 132–139.
6. *Geering, H. P., Athans, M.*, The Infimum Principle. IEEE Transactions on Automatic Control **19** (1974) 485–494.

Необходимые условия оптимальности по нескалярному критерию в абстрактных проблемах оптимального управления

Б. ЛАНТОШ

(Будапешт)

В статье изучаются необходимые условия оптимума нескалярной функции при наличии ограничений. Область значений функции является частично упорядоченным множеством. Метод Дубовицкого—Милютина—Гирсанова для скалярной функции распространен на случай, когда область значений является частично упорядоченным линейным топологическим пространством и функция достигает минимума.

Рационально полная теория необходимых условий оптимума разработана для случая, когда функция дифференцируема по Фреше, и область ее значений является частично упорядоченным рефлексивным банаховым пространством.

B. Lantos
Department of Process Control
Technical University of Budapest
H–1111 Budapest, XI. Műegyetem rakpart 9.
Hungary

THE IVth IFAC SYMPOSIUM ON IDENTIFICATION AND SYSTEM PARAMETER ESTIMATION

The Symposium will be held in the USSR, Tbilisi, between September 21 and 27, 1976. It is organized by the USSR National Committee of Automatic Control, the Academy of Sciences of the USSR, the Academy of Sciences of the Georgian SSR within the framework of IFAC.

The members of the International Program Committee are: K. J. Åström (Sweden), A. V. Balakrishnan (USA), J. G. Balchen (Norway), S. P. Bingulac (Yugoslavia), F. Csáki (Hungary), P. Eykhoff (The Netherlands), R. Isermann (FRG), G. Koch (Italy), V. Peterka (Czechoslovakia), N. S. Raibman (USSR), J. E. Richalet (France), A. P. Sage (USA), Y. Sawaragi (Japan), V. Strejc (Czechoslovakia), A. Sydow (GDR).

The main aim of the IVth Symposium on Identification and System Parameter Estimation is to discuss and generalize the most recent advances in theory and applications of techniques for designing mathematical models and system parameter estimation.

The Symposium comprises the following topics: techniques for identification and parameter estimation of automatic control systems of various classes; choice of model structures; choice of technical means; problems of observability; identification and state estimation; problems of application of identification and parameter estimation techniques for various branches of industry, in biological, medical, and economical systems, in aeronautics and space research, etc.

A special topic will be the estimation of the practical problems results. In addition to the accepted papers in several sessions listed below, the program includes a number of lectures on statistical and engineering foundations of identification and parameter estimation; optimal techniques; foundations of linear and nonlinear stationary and nonstationary systems identification and other modern identification methods. Invited speakers are Professors P. Eykhoff, K. J. Åström, R. Isermann, A. V. Balakrishnan, V. Strejc and M. A. Krasnoselsky.

Some actual examples of practical applications of the identification techniques will be considered in detail.

The program will provide both for discussions on the papers and *round table discussions*.

The *official languages* of the Symposium are Russian and English.

The detailed time-table of the scientific program is shown in the Table enclosed which contains the tutorials, the survey papers, the round table discussion and the other scientific sessions entitled: Least-Squares Technique, Maximum Likelihood Technique, Model Structure Determination, Comparison of Identification Methods, Feedback Systems, Nonlinear Systems, Multivariable Systems, Distributed Parameter Systems, On-line Identification, Adaptive Systems, Case Studies, Biological and Medical Systems, Chemical-Physical Processes, Metallurgical Processes, Power Systems, Vehicles, Ecological and Environmental Systems, Various Processes.

F. Cs.

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

	20th September Monday	21st September Tuesday	22nd September Wednesday	23rd September Thursday	24th September Friday	25th September Saturday	26th September Sunday	27th September Monday	28th September Tuesday
8 — 9	ARRIVING REGISTRATION								
9 — 10		SURVEY PAPERS	Comparison of Identification Methods	Comparison of Identification Methods	TECHNICAL VISITS AND EXCURSIONS	Round-Table Discussion			
10 — 11			Power Systems	Vehicles		Ecological and Environmental Systems			
11 — 12			Least-Squares Technique	Maximum Likelihood Technique		Feedback Systems			
12 — 13			Multivariable Systems	Distributed Parameter Systems					
13 — 14		OPENING SESSION	Case Studies	Case Studies		Case Studies			
14 — 15			Biological and Medical Systems	Biological and Medical Systems		Comparison of Identification Methods			
15 — 16			Chemical — Physical Processes	Model Structure Determination		Metallurgical Processes			
16 — 17			Adaptive Systems	Adaptive Systems		On-line Identification			
17 — 18		TUTORIALS	SURVEY PAPERS	SURVEY PAPERS		Nonlinear Systems			
						Round-Table Discussion			
						Economic Systems			
						Test Signals			
						SURVEY PAPERS	Case Studies		
							Comparisonof Identification Methods		
							Ecological and Environmental Systems		
							Various Processes		
							Nonlinear Systems		
							Closing Session	Round-Table Discussion	Round-Table Discussion
								TUTORIALS	
	DEPARTURE								

BOOK REVIEW

K. A. ABGARYAN, Matrix and Asymptotic Methods in the Theory of Linear Systems. Nauka, Moscow, 1973, 432 pp. (in Russian)

The book treats the application of matrix calculus and the asymptotic integration of linear differential equations in the theory of controllable and uncontrollable processes. It consists of 16 chapters and an Appendix. The volume may be divided into three major parts.

Part 1 (Chapters 1—5) is devoted to the fundamentals of matrix algebra, dealing especially with the details of matrix calculus utilized in the theory of linear systems. Here is given a well-founded background of linear spaces as an essential prerequisite. The material in this part permits full understanding without any preliminary knowledge of matrix theory on the part of the reader.

Part 2 (Chapters 6—12, and the Appendix) explains the application of asymptotic and matrix algebra methods in the theory of linear systems. Besides the general theory there is shown how to decompose the original system of differential equation into independent subsystems. The schemes of approximate solutions are also given. The main attention is devoted to the non-stationary systems. Along with the non-autonomous systems of differential or integro-differential equations the author presents a system substituting the time variable t by the so-called slow time variable $= t$, where is a parameter, giving then the formal solution of the second system to a series of the parameter. The two systems coincide when $= 1$. So the approximate solution of the initial system may be obtained as a sum of the first terms in the series. In this part the ideas of N. M Krylov and N. N. Bogolyubov, on the asymptotic integration of differential equations by introducing the conception of slow time, find and expression.

Part 3 deals with the stability theory of the processes. It presents the formulation of the task of process stability in the given time interval, by constructing the stability criteria for the solution of the linear systems and with a certain deviation from the theme, also for the solution of non- linear systems by linear approximation.

At the end of the book a short list of references is given. This work may be useful for students, engineers and scientists.

Tivadar Kovács

EBERHARD HOFFER, REINHART LUNDERSTÄDT: Numerische methoden der Optimierung (Numerical Methods of Optimization). R. Oldenburg Verlag, Munich and Wien, 1975, 231 pp., 43 figs. and 12 tables (in German)

The book is a numerical approach to the optimal control of linear deterministic processes. The subject is treated with the aspects of applications in the foreground. A number of flow-charts makes clearer the presented computational procedures.

The volume consists of six chapters.

The first chapter is an introduction to the optimization problems. After overviewing the static and the dynamic optimization problems, the variational approach with constrained and unconstrained control is presented. The chapter concludes with the formalization of the optimum control tasks by the help of the maximum principle.

The second chapter deals with direct methods. It presents the gradient methods for the Mayer and Bolza problems. The second variation method is used for solving the Bolza problem with and without terminal conditions. One of the nine examples in this chapter deals with the optimum control of a travelling crane.

The third chapter presents the Newton–Raphson method and its improved versionse As a further indirect method the quasilinearization techniques are shown, too. There are three examples in this chapter of which one is for a nuclear reactor, and one is for a chemical reactor.

The fourth chapter is devoted to the time optimal control of linear systems. The main subjects discussed here are as follows: an analytical solution by the help of the maximum principle, the Neustadt method, a geometric interpretation, the Neustadt theorem and two examples.

The fifth chapter is an Appendix. It summarizes a number of notations, operations, properties and theorems concerning vectors, matrices and linear systems and concerning the calculus of variations and the maximum principle. A numerical method for solving ordinary differential equations is also shown in this chapter.

The final chapter is an ample guide to the literature.

The book is an original one in this topic. Besides the mathematical treatment the practical aspects are always taken into consideration. The book may be very useful for engineers, scientists and students.

Tivadar Kovács

RAIBMAN, N. S., CHADEEV, V. M.: Modelbuilding of Industrial Processes (in Russian). Original title: Моделирование промышленных процессов (Moscow, 1975, 375 pp., 231 references)

The book deals with a new field of the control theory, the methods of modelbuilding of industrial processes and with the theory of identification.

The book is devoted to the results of the authors regarding the development of correlation techniques and their application in the theory of process identification. The book consists of nine chapters. Each of them starts with a short introduction summarizing the former results and stating the problems to be investigated. The first chapter considers identification problems reviewing the different mathematical models, frequently used in the control theory, of the processes.

The second chapter investigates the identification of nonmemory (nondynamic, i.e. static) processes and gives the technical fields where this description method can be applied. It shows some examples for single and multivariable static processes.

The third chapter discusses the identification of linear dynamic processes. The most attention is devoted to the type (or structure identification where the values of correlation functions are calculated from the input and output signals and the structure is selected. The parameters of the model are determined by comparing the correlation functions computed from the data with the typical ones in the tables. The identification tables of different types of basic elements are given in this chapter.

In the fourth chapter the identification of nonlinear processes is treated. In order to extend the correlation theory the authors introduce the dispersion function by means of which the dispersion equation of the identification can be obtained and it makes possible to use the same calculation methods as in the classical linear correlation theory.

In the fifth chapter nonstationary processes are investigated and an optimal, one-step algorithm is given for the model parameters estimation. The convergency "limit" possibilities and the fitting properties of the algorithm are also investigated and the maximum speed of parameter changes is determined, as well, under which the parameter changes can be followed during the identification.

In the sixth chapter a so-called "basis" identification method is suggested whose memory deepness is less than the dimension number of the process which makes possible to reduce the memory and speed requirements of computers.

The seventh chapter is devoted to the investigation of the problems of analog adaptive modelbuilding which can be well used on the lower levels of a hierarchical control system.

In the eighth chapter a method is suggested for the direct estimation of the degree of model adequacy by the information and dispersion measure introduced by the authors.

In the ninth chapter the authors show a real technical example to illustrate the necessity of realization of special computers whose task is to identify the process.

The book is written on high scientific-technical level and the basic theories are illustrated by many examples well chosen from different industrial branches. Most of the presented methods have also been tried in the practice (machine industry, metallurgy, petroleum industry and in the biology, physiology, medical siences).

The book is written first of all for research workers and engineers dealing with process control and identification, but it can be well used as a text-book for students in the high-level educations, too.

F. Cs.

Printed in Hunngary
AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY Vol. 5 (3)
ENGLISH TRANSLATION OF THE COMMUNICATIONS IN RUSSIAN

DECODING AND DIOPHANTINE APPROXIMATIONS

V. D. GOPPA

(Moscow)

A new approach to the decoding of (L, g)-codes is developed regarding them as a problem of the best rational approximation in g-adic topologies. For (L, g)-codes, an algorithm is proposed for decoding within the constructive boundary similar to Mahler's algorithm for the rational approximation of g-adic numbers.

1. Valuation and pseudovaluation of fields

Let k be an arbitrary field. Any real-valued function over field K is referred to as its normalization if it has the following properties [1]:

(1): $w(0) = 0$, $w(a) > 0$ for $a \neq 0$;

(2): $w(a + b) \leq w(a) + w(b)$;

(3): $w(ab) = w(a)w(b)$

for any $a, b \in K$.

If, instead of condition (3), a weaker condition (3′):

$$w(ab) \leq w(a)\, w(b)$$

is met, $w(a)$ is referred to as pseudovaluation of K.

The following properties of $w(a)$ may be readily deduced from these definitions:

$$w(-a) = w(a)$$
$$w(a) - w(b) \leq w(a - b) \leq w(a) + w(b)$$
$$|\, w(a) - w(b)\, | \leq w(a - b)$$
$$w(\pm 1) = 1 \text{ (if } w \text{ is a norm)}$$
$$w(\pm 1) \geq 1 \text{ (if } w \text{ is a pseudonorm).}$$

Every field has at least one normalization, namely a trivial one:

$$w_0(a) = \begin{cases} 0 \text{ if } a = 0 \\ 1 \text{ if } a \neq 0. \end{cases}$$

For the trivial valuation, the triangle inequality takes a stronger form:

$$w_0(a+b) \leq \max\,(w_0(a),\ w_0(b)) \tag{2'}$$

(ultrametric inequality).

Any valuation/pseudovaluation is referred to as non-Archimedean if inequality (2′) holds for it, otherwise it is referred to as Archimedean

Thus, valuation $|\,a\,|$ of the field of rational number is Archimedean.

In addition to $|\,a\,|$ and $w_0(a)$, field Q also has infinite number of p-adic normalizations $|\,a\,|_p$ defined as follows:

(i): $|\,0\,|_p = 0$

(ii): if $a \neq 0$, it may be represented in a unique way as $a = p^n \frac{r}{s}$ where r and s cannot be divided by p, and n is an integer. Assume that

$$|\,a\,|_p = p^{-n}.$$

It is easy to verify that $|\,a\,|_p$ is a non-Archimedean valuation.

If q is an integer, then

$$|\,q\,|_p = \begin{cases} \frac{1}{p} & \text{if } q = p \\ 1 & \text{if } q \neq p. \end{cases}$$

This implies that p-adic valuations corresponding to different prime numbers p are different.

Thus, Q has infinite number of different valuations $|\,a\,|$, $w_0(a)$ and $|\,a\,|_p$ for different p.

These valuations are related for all $a \in Q$ through a fundamental identity: $|\,a\,| \cdot \prod_p |\,a\,|_p = w_0(a)$ which is equivalent to the fundamental number-theoretic theorem on the uniqueness of factorization of integers into prime numbers.

Let now K be an arbitrary field, and $K(x)$ be its simple transcendental expansion (i.e. the set of all rational functions of x). $K(x)$ is a field of quotients of ring $K[x]$.

Denote by p an irreducible polynomial in $K[x]$ and choose a real constant $0 < \theta < 1$. If $r \neq 0$ is an arbitrary element of $K[x]$, degree of r will be denoted as deg r, and order of r with respect to p (i.e. the greatest integer m such that p^m divides r) will be denoted as ord $_p r$.

Each element $a \neq 0$ of field $K(x)$ may be written uniquely as $a = \frac{r}{S}$ where r and S are mutually prime polynomials, and S is a nonzero polynomial with leading coefficient equal to 1. Denote

$$|| a || = \begin{cases} 0 & \text{if } a = 0 \\ \theta^{\deg S - \deg r} & \text{if } a \neq 0 \end{cases}$$

$$|| a ||_p = \begin{cases} 0 & \text{if } a = 0 \\ \theta^{\operatorname{ord}_p r - \operatorname{ord}_p S} & \text{if } a \neq 0. \end{cases}$$

It is easy to verify that both of these functions are non-Archimedean valuations of field $K(x)$, and that, for all prime polynomials p an infinite system of various valuations related through the following identity

$$|| a || \cdot \prod_p || a ||_p^{\deg p} = w_0(a), \quad a \in K(x)$$

is formed by $w_0(a)$, $|| a ||$ and $|| a ||_p$.

Each valuation generates in the field a topology where the base of the vicinity of zero consists of spheres

$$V_w^{\varepsilon} = \{a \in K : w(a) \leq \varepsilon\}, \ \varepsilon > 0 .$$

By means of the well-known Cauchy technique, the normalized field may be embedded into the complete field where the initial field will be compact everywhere. If two norms define the same topological structure, these norms lead to the same augmentations. Such norms are called equivalent. For example, from any norm $w(a)$ a new one, $w^*(a) = (w(a))^{\lambda}$ may be obtained, $0 < \lambda < f$, which is evidently equivalent to the initial norm.

All valuations neighborhoods of field Q are exhausted to within the equivalence by norms $w_0(a)$, $| a |$ and $| a |_p$ (Ostrovskii's theorem [2]), and lead to different augmentations: Q, R (field of real numbers), and Q_p (field of p-adic numbers), respectively.

2. g-adic pseudovaluation of field $K(x)$

Let $g \in K[x]$where K is an arbitrary field. Assume that polynomial g has the following factorization into irreducible factors:

$$g = p_1^{e_1} p_2^{e_2} \ldots p_r^{e_r}.$$

Fix numbers $\lambda_1, \lambda_2, \ldots, \lambda_r$ so that

$$|| g ||_{p_1}^{\lambda_1} = || g ||_{p_2}^{\lambda_2} = \ldots = || g ||_{p_2}^{\lambda_2} = \theta < 1$$

i.e. take $\lambda_i = \frac{1}{e_i}$, $i = 1, 2, \ldots, r$.

Define now

$$|| a ||_g = \max (|| a ||_{p_1}^{\lambda_1}, \ldots, || a ||_{p_r}^{\lambda_r}) .$$

$|| a ||_g$ may be readily demonstrated to be a non-Archimedean pseudonorm over $K(x)$.

Sphere $V_g^1 = \{a \in K(x) : || a ||_g \leq 1\}$ is a ring.

Indeed, if $a, b \in V_g^1$ then

$$|| a \pm b ||_g \leq \max (|| a ||_g, || b ||_g) \leq 1 ,$$

$$|| ab ||_g \leq || a ||_g || b ||_g \leq 1 .$$

Let $a \in K(x)$ be represented in an irreducible form as $a = \frac{r}{s}$. If s has a common divisor with g (say, p_i^l) then $|| a ||_{p_i} = \theta^{-l} > 1$ because $0 < 1$, and therefore $|| a ||_g > 1$. If s and g have no common factors, then

$$|| a ||_g = \max (|| r ||_g, 1) \leq 1 .$$

Thus, ring V_g^1 involves all rational functions whose denominators have no common factors with g:

$$V_g^1 = \left\{ \frac{r}{s} \in K(x) : (s, g) = 1 \right\} .$$

If A is a ring (commutative with 1), any subset $s \subset A$ containing 1 and closed with respect to multiplication is referred to as multiplicative subset in A. The binary relation $\equiv$ over $A \times S : (a, s) \equiv (b, t) \Leftrightarrow (at - bs)\, u = 0$ for some $u \in S$ is that of equivalence. If a/s denotes equivalence class of pair (a, s), and $S^{-1}A$ denotes a set of equivalence classes, one can introduce a ring structure by defining addition and multiplication of "fractions" a/s through the following formulae:

$$(a/s) + (b/t) = (at + bs)/st,$$

$$(a/s)\,(b/t) = (ab/st) .$$

Ring $S^{-1}A$ is called the ring of quotients of A with respect to S [3] and is a generalization of the well-known construction of the quotient field.

Evidently, if one takes $A = K[x]$, $S_g = \{f \in A : (f, g) = 1\}$, then

$$V_g^1 = S_g^{-1} K[x] .$$

Further, consider a sphere

$$V_g^\theta = \{a \in K(x) : || a ||_g \leq \theta\}.$$

Evidently, $V_g^\theta \subset V_g^1$. If $a, b \in V_g^\theta$, $c \in V_g^1$, then

$$|| a \pm b ||_g \leq \max (|| a ||_g, || b ||_g) \leq \theta ,$$

$$|| ac ||_g \leq || a ||_g \cdot || c ||_g \leq \theta \cdot 1 = \theta .$$

Thus, V_g^θ is an ideal of ring V_g^1. Since $|| g ||_g = \theta$, $g \in V_g^\theta$.

Further,

$$\left\| \frac{a}{g} \right\|_g = \max \left(\left\| \frac{a}{g} \right\|_{p_1}^{\lambda_1}, \ldots, \left\| \frac{a}{g} \right\|_{p_r}^{\lambda_r} \right) = \max \left(\frac{\| a \|_{p_1}^{\lambda_1}}{\theta}, \ldots, \frac{\| a \|_{p_r}^{\lambda_r}}{\theta} \right) = \frac{1}{\theta} \| a \|_g ,$$

and it follows that

$$a \in V_g^\theta \Leftrightarrow \frac{a}{g} \in V_g^1 .$$

Thus, V_g^θ consists of all products $a = a'g$ where $a' \in V_g^1$, i.e. V_g^θ coincides with the principal ideal (g) of ring V_g^1.

For the case where g is a prime polynomial, ideal V_g^θ is maximal, and the unique maximal ideal of ring V_g^1. It coincides in this case with the set all nou-units of ring V_g^1. Thus, V_g^1 is a local ring in this instance.

3. Algebraic codes. Ultrametric majorization of Hamming metric

Let E^n be a vector space of dimension n over a Galois field $GF(q)$ where q is a power of a prime number. Assume that some basis is fixed over this subspace. The number of non-zero coordinates of vector $x \in E^n$ with respect to this basis is referred to as Hamming weight $|| x ||_H$ of vector x. The Hamming weight is an Archimedean norm over E^n.

Subspace $C \subset E^n$ is referred to as algebraic (linear) code [4]. Ratio k/n, where k is the dimension of C, is referred to as transmission rate, and $d = \min_{x \in C - \{0\}} || x ||_H$ is referred to as code distance or weight.

For code C decoding is to find a vector with the least Hamming weight in coset $x + C$. Take $L = GF(q^m)$, $m > 0$, $n = q^m$.

Mapping

$$x = (a_1, a_2, \ldots, a_n) \xrightarrow{\varphi} \sum_{i=1}^{n} \frac{a_i}{z - \alpha_i} = \xi_x(z) = \frac{\psi(z)}{f(z)} ,$$

where $\{\alpha_i\}$ are all elements of field L, is linear, injective and translates Hamming weight $|| x ||_H$ into the fraction $\xi_x(z)$ denominator degree.

This mapping, called rational representation of codes [5], enables majorization of Hamming metric by non-Archimedean metrics of rational

function field $L(z)$. Indeed, let $g \in L(z)$ have no roots in L, and R denote image of E^n for mapping φ. Then $R \subset V_g^1$, $||\xi_x(z)||_g = ||p\psi(z)||_g$, $||f(z)|| = \theta^{-\deg f(z)} = \theta^{-||x||_H}$, and if $||\xi_x(z)||_g$ is small, $||f(z)||$ and, consequently, $||x||_H$ are great.

Take intersections of ideals V_g^θ with space R as codes. These are linear codes consisting of fractions whose numerators may be divided by polynomial g ((L, g)-codes [5]).

Quotient space R/V_g^θ consists of polynomials of degree $< r = \deg g$. Each element $S(z)$ of this space is called syndrome. Decoding of (L, g)-codes is then as follows: for each $S(z) \in R/V_g^\theta$ find fraction $\dfrac{\psi(z)}{f(z)} \in R$ such that

$$\left\| \frac{\psi(z)}{f(z)} - S(z) \right\|_g \leq \theta$$
$$||f(z)|| \text{ is minimal.} \tag{1}$$

This formalization of the decoding problem has very close resemblance to the problems in the Diophantine approximation theory [1].

A classical problem of this kind first formulated by Huygens is to find a rational number $\dfrac{p}{q}$ with as small as possible $|q|$ for which $\left|\dfrac{p}{q} - \alpha\right| \leq \varepsilon$ where α is an arbitrary real number.

An effective algorithm is known to exist solving this problem by means of continued fractions and finding approximation of $\dfrac{p}{q}$ with $|q| < \dfrac{1}{\sqrt{\varepsilon}}$.

Mahler [1] formulated a g-adic analog of this problem and extended the continued fraction algorithm to this case. For any g-adic number α, his algorithm finds rational number $\dfrac{p}{q}$ such that

$$\left| \frac{p}{q} - \alpha \right|_g \leq \varepsilon,$$
$$\max(|p|, |q|) < \frac{1}{\sqrt{\varepsilon}}. \tag{2}$$

Mahler's algorithm is applicable also to find a rational function satisfying condition (1).

Note, that for any polynomial S there is a unique fractional function $\frac{P}{Q}$ such that

$$\frac{P}{Q} - S \equiv 0 \pmod{g},$$

$$\deg Q \leq \frac{1}{2} \deg g, \deg P < \frac{1}{2} \deg g.$$

Mahler's algorithm finds such a pair of polynomials by means of a decomposition into continued fraction.

4. Polynomial continued fractions

Let u and v be two polynomials over an arbitrary field K, $\deg u \geq \deg v$. Then, there is a unique representation

$$u = q_0 v + r_1, \quad q_0, r_1 \in K[z], \quad \deg r_1 < \deg v.$$

Divide by v and denote by $\alpha_1 = \frac{v}{r_1}$:

$$\frac{u}{v} = \alpha_0 = q_0 + \frac{1}{\alpha_1}.$$

Evidently,

$$|| \alpha_1 || > 1, \quad || \alpha_0 || = || q_0 ||.$$

By applying the same decomposition to α_1 and so on, obtain decomposition of α_0 into continued fraction

$$\alpha_0 = q_0 + \frac{1}{q_1} \cdots \frac{1}{\alpha_n}. \tag{3}$$

Fraction α_n is called n-residue of fraction α_0. This decomposition may be continued until a step of the Euclidean algorithm giving the zero residue.

For any continued fraction

$$\alpha_0 = [q_0, q_1, \ldots, q_m]$$

convergents $\frac{P_k}{Q_k}$ are defined through formulae

$$\begin{Bmatrix} P_{-1} = 1 \\ Q_{-1} = 0 \end{Bmatrix}, \quad \begin{Bmatrix} P_0 = q_0 \\ Q_0 = 1 \end{Bmatrix}, \quad \begin{Bmatrix} P_k = q_k P_{k-1} + P_{k-2} \\ Q_k = q_k Q_{k-1} + Q_{k-2} \end{Bmatrix}, \quad k = 1, 2, \ldots, m. \tag{4}$$

Identity

$$P_{k-1} Q_k - P_k Q_{k-1} = (-1)^k \tag{5}$$

shows that all convergents are irreducible.

For α_0, there is the following expression

$$\alpha_0 = \frac{P_{k-1}\alpha_k + P_{k-2}}{Q_{k-1}\alpha_k + Q_{k-2}} \tag{6}$$

which might be easily proved by induction.

It follows from (5) and (6) that

$$\alpha_0 - \frac{P_{k-1}}{Q_{k-1}} = \frac{(-1)^{k-1}}{Q_{k-1}(Q_{k-1}\alpha_k + Q_{k-2})} . \tag{7}$$

5. Mahler's algorithm

Let g, S be arbitrary polynomials, $\deg S < \deg g$. A pair of polynomials P, Q is to be found such that

$$|| QS - P ||_g \leq \theta \tag{8}$$

with small $|| Q ||$, $|| P ||$. Note, that inequality (8) is equivalent to comparison $QS - P \equiv 0 \pmod{g}$. Here P and Q are not necessarily coprime, but we shall assume that their common divisors are necessarily divisors of g. Decompose $\alpha_0 = \dfrac{S}{g}$ into continued fraction

$$\alpha_0 = [q_1, q_2, \ldots, q_m].$$

Let $\dfrac{R_k}{Q_k}$ be convergent for α_0. Take

$$P_k = S\,Q_k - gR_k. \tag{9}$$

Since Q_k and R_k are coprime, common divisor of P_k and Q_k is necessarily divisor of g. Moreover, pair (P_k, Q_k) meets condition (8).

It follows further from (7) that

$$\frac{S}{g} - \frac{R_k}{Q_k} = \frac{(-1)^k}{Q_k(Q_k\alpha_{k+1} + Q_{k-1})}$$

and therefore

$$P_k = \frac{(-1)^k g}{Q_k\alpha_{k+1} + Q_{k-1}} , \qquad k = 0,\ 1,\ \ldots,\ m-1. \tag{10}$$

Further, $\alpha_{k+1} = q_{k+1} + \beta$, $|| \beta || < 1$. Therefore, $|| Q_k\alpha_{k+1} + Q_{k-1} || = || (Q_kq_{k+1} + Q_{k+1}) + Q_k\beta || = || Q_{k+1} ||$ because $\deg Q_{k+1} \geq \deg Q_k$ and β is a regular fraction. Thus,

$$|| P_k || \cdot || Q_{k+1} || = ||g ||, \quad k = 0, 1, \ldots, m-1.$$

i.e.

$$\deg P_k + \deg Q_{k+1} = \deg g,\ k = 0, 1, \ldots, m-1. \tag{11}$$

Consider now two cases. If $\deg Q_m > \frac{1}{2} \deg g$, then there will be number j, $0 \leq j \leq m-1$ such that $\deg Q_j \leq \frac{1}{2} \deg g < \deg Q_{j+1}$ because

$$0 = \deg Q_0 < \deg Q_1 < \ldots < \deg Q_m.$$

For such a number j, inequality $\deg P_j < \frac{1}{2} \deg g$ will also hold by virtue of (11).

If $\deg Q_m \leq \frac{1}{2} \deg g$, this implies that $\deg (S, g) \geq \frac{1}{2} \deg g$ because $\frac{R_m}{Q_m} = \frac{S}{g}$

However, $Q_m \cdot S = R_m \cdot g = 0 (\mathrm{mod}\, g)$, and we can choose $P_m = 0$:

$$|| Q_m S - 0 ||_g \leq 0, \ \deg Q_m \leq \frac{1}{2} \deg g.$$

Thus, an effective algorithm has been constructed finding a unique pair of polynomials P and Q such that

$$QS - P \equiv 0 \ (\mathrm{mod}\, g)$$

$$\deg Q \leq \frac{1}{2} \deg g,$$

$$\deg P < \frac{1}{2} \deg g$$

for any polynomials S and g.

6. Decoding of (L, g)-codes

Formulate an algorithm for decoding (L, g)-codes based on that of Mahler and correcting t errors for $\deg g = 2t$.

1. Compute syndrome $S(z)$:

$$S(z) = \sum_{i=1}^{n} a_i \frac{g^{-1}(\alpha_i)\, g(z) - 1}{z - \alpha_i}$$

where $\{a_i\}$ are coordinates of output vector, $a_i \in GF(q)$, and $\{\alpha_i\}$ are elements of field $GF(q^n)$ used to enumerate the coordinates.

2. Decompose $\frac{S(z)}{g(z)}$ into a continued fraction concurrently with computation of convergents $\frac{P_k}{Q_k}$. Stop calculations as soon as deg Q_k becomes greater than t.

Polynomial Q_{k-1} is error locator.

3. Compute $P_{k-1} = Q_{k-1} S(z) - g(z) R_{k-1}$

Fraction $\frac{P_{k-1}}{Q_{k-1}}$ is that we are looking for.

Illustrate the algorithm with two examples.

1. The binary BCH-code of lenght 31 correcting 3 errors is a (L, g)-code with $g(z) = z^6$, $L = GF(2^5) - \{0\}$.

Enumerate all word positions by powers α^i of the root of primitive polynomial $x^5 + x^2 + 1$.

Assume that the output vector has ones in positions corresponding to α^0, α^1 and α^4. Then

$$S(z) = \frac{z^6 - 1}{z - 1} + \frac{\alpha^{-6} z^6 - 1}{z - 2} + \frac{\alpha^7 z^6 - 1}{z - \alpha} = \alpha^{20} z^5 + \alpha^{10} z^4 + \alpha^{22} z^3 + \\ + \alpha^{10} z^2 + \alpha^{11} z + \alpha^{21}$$

$$\frac{S(z)}{z^6} = [0, \alpha^{11} z + \alpha, \alpha^2 z + \alpha^{-7}, \alpha^8 z + \alpha, \alpha^{20} z + \alpha^6, \alpha^{20} z^2 + \alpha^{15}]$$

q_k		0	$\alpha^{11} z + \alpha$	$\alpha^2 z + \alpha^{-7}$	$\alpha^8 z + \alpha$
R_k	1	0	1	$\alpha^2 z + \alpha^{-7}$	$\alpha^{10} z^2 + \alpha^6 z + \alpha^{21}$
Q_k	0	1	$\alpha^{11} z + \alpha$	$\alpha^{13} z^2 + \alpha^{21} z + \alpha^{21}$	$\alpha^{21} z^3 + \alpha^7 z^2 + \alpha^{16} z + \alpha^{16}$

It is easy to verify that

$$Q_3 = \alpha^{21} z^3 + \alpha^7 z^2 + \alpha^{16} z + \alpha^{26} = \alpha^{21}(z - 1)(z - \alpha)(z - \alpha^4),$$

$$P_3 = Q_3 S - R_3 z^6 = \alpha^{21} z^2 + \alpha^{16}$$

(note that for binary codes R_k need not be computed).

2. Let $L = GF(2^5)$, $g = (z^3 + z + 1)^2$ (irreducible code), errors are in positions corresponding to 0 and 1, i.e. error locator is as follows: $(z + 1)z$.

Compute syndrom:

$$S(z) = (z^5 + z) + (z^5 + z^4 + z^3 + z^2) = z^4 + z^3 + z^2 + z$$

$$\frac{S(z)}{g(z)} = [0, z^2 + z, z^4 + z^3 + z^2 + z]$$

q_k		0	$z^2 + z$	$z^4 + z^3 + z^2 + z$
R_k	1	0	1	$z^4 + z^3 \pm z^2 + z$
Q_k	0	1	$z^2 + z$	$z^6 + z^2 + 1$

$$Q_2 = z^2 + z = (z + 1)z$$

$$P_2 = Q_2 S - R_2 g = (z^2 + z)(z^4 + z^3 + z^2 + z) - (z^6 + z^2 + 1) = 1.$$

7. Conclusions

Complexity of the proposed algorithm for decoding of (L, g)-codes is about the same as that of Berlekamp's algorithm [4]. One iteration in Berlekamp's algorithm involves multiplication of two polynomials, that of the continued fraction algorithm involves division with residue of one polynomial by another and finding a convergent (multiplication of two polynomials). Although one iteration of Berlekamp's algorithm is simpler, their total number is fixed ($= 2t$) where t is the number of corrected errors; in the continued fraction algorithm the number of iterations is variable, always $\leq t$, and, for many a syndrome, is strictly less than t.

Moreover, division in the finite field usually is also done by decomposition into a continued fraction. Application of the same algorithm (with only minor differences) both to division and decoding offers certain degree of hardware standardization, thus facilitating design of communication systems.

Thus, several effective algorithms for decoding (L, g)-codes are presently known. All of them correct errors only within the constructive boundary. At the same time, it is known that code generating polynomial g may be selected such that the corresponding code would lie in Hilbert's boundary, i.e for the majority of codes, the distance between the constructive and the actual boundaries is large. Decoding of the total number of errors is, with such codes, a difficult problem of Diophantine approximations with constraints since one has to approximate a given polynomial with respect to the g-adic norm by fractions of a special form, denominator of each of these fractions falling

into linear factors in fixed field $GF(q^m)$. Hopefully, some of the methods developed for decoding within the constructive boundary (or the ideas underlying them) would prove to be an initial step towards solution of one of the most important problems in the coding theory, that of decoding in Hilbert's boundary.

References

1. *Mahler, K.*, Lectures on Diophantine Approximations. Part I: g-adic numbers and Roth's theorem. University of Notre Dame, 1961.
2. *Borevich, Z. I., Shafarevich,* I. R., Theory of Numbers, Nauka Publ., Moscow, 1972 (in Russian).
3. *Atjah, M., MacDonald, I.*, Introduction to commutative algebra. Mass., 1969.
4. *Berlekamp, E.*, Algebraic Coding Theory, Mc Graw-Hill Book Company, 1968.
5. *Goppa, V. D.*, Rational Code Representation and (L, g)-codes. "Problemi peredachi informatsii", 1971 **7,** 3, pp. 41—49 (in Russian).

PURPOSEFUL SYSTEM BEHAVIOUR LOGIC

E. I. EFIMOV

(Moscow)

Two problems are to be solved to construct the theory of purposeful systems, (*i*) development of a semiotic axiomatic-deductive logic of control, and (*ii*) development of a semiotic purposeful system based on the language of this logic. Only the first of the two problems, that one which related to the purposeful system behaviour logic is covered by this paper. A system of languages is constructed involving languages of various orders interrelated by translation languages. Syntax, semantics and pragmatics of such a system are demonstrated, and the notion of deduction is given.

A rather typical act of the purposeful system behaviour is the formation of the solution of the following problem. An initial situation and a system objective are given. The latter, generally, represents only some of the desired properties of the final situation. The system has to form a solution of the problem of reaching this objective as a sequence of acts and to realize it. This problem belongs to the situational control which recently found application to the study of complex systems [1, 2].

As opposed to the problems solved previously by the situational control, in the problem above solution is not selected from an a *priori* prescribed set, but rather is formed a new. Thus, the system has to work out a sequence of acts depending on the objective and the initial situation, this sequence later on could be used in similar problems as a ready-made solution.

Formation of particular problems of system behaviour will be represented as automatic proof of theorems in some logic systems. Two approaches are possible here, the first one is to develop effective methods of deductive inference within the framework of somewhat modified formal systems of the mathematical logic. Numerous American publications on various modifications of the theorem proving algorithms based on resolution or paramodulation methods [3–9] are examples for is way. For the time being, these alrorithms work in the first-order logics. As the American researchers themselves admit [9, 10], this appreciably decreases the effectiveness of theorem-proving because of the excessively detailed deduction which is cumbersome and unusual to the man. It should be also noted, that the methods of selective theorem

proving require various rules of thumb. Attempts to work out general resolution or paramodulation methods for higher-order logics did not result in any appreciable progress as yet [10].

The second approach attempts to construct a special logic, that of control on the basis of characteristics of control and human decision taking that compel us to give up the traditional ways to constructing *formal* axiomatic-deductive systems of the mathematical logic in favour of control logic as a dynamic semiotic axiomatic-deductive system. This means that such a logic is essentially related to the model (interpretation) and its dynamics through the axiomatics and that the notion of the "formula of general meaning" loses here its universal sense.

It follows from the above-said that the problem of consistency and completeness of a semiotic axiomatic-deductive system should be solved for a fixed model.

This paper considers one of the possible ways within the framework of the second approach where the control logic is represented as some linguistic system L built around two major points:

1. The basic philosophy of the situational control is the description of situations through states and connections (relations) of environmental elements and their actions.
2. Description is done in the language of the mathematical logic with the aim of expanding it further so as to allow for the features of the purposeful system behaviour problems.

In accordance to the peculiarities of human decision taking, it is sufficient to assume that language L includes language L_i of lower, ith order, language L_j of higher, jth order and translation language L_{ij} for internal dialogue between languages L_i and L_j.

Syntax of language L

Syntax of languages L_i and L_j

As both languages are syntactically identical, index j is temporarily omitted.

Alphabet

1. Punctuation marks: ( ) [ ] { };
2. Element names (with or without indexes) w, w_1, . . . ;
3. Action names (with or without indexes) f, f_1 . . . ;

4. Variables of value with respect to jth attribute (with or without indexes) y^j, y_1^j, . . . ;

5. Functional variables (with or without indexes) ψ, ψ_1, . . . ;

6. Constants (with or without indexes) α_1 α_2, . . . ;

7. Predicative variables and constants (with or without indexes) g, g_1, . . . ;

8. Propositions (with or without indexes) q, q_1, . . . ;

9. Symbols of logic operations: — (negation), $\Rightarrow$ (implication), $\rightarrow$ (strict implication), $\leftrightarrow$ (equivalence).

Expressions

Terms:

1. Each symbol α and y is a term;
2. If t_1, . . . , t_n are qualitative terms, $\psi(t_1, \ldots, t_n)$ is a term;
3. No other expression is a term.

Individual variables:

1. Each variable term is an individual variable;
2. Each symbol w and f is an individual variable;
3. No other expression is an individual variable.

Correctly constructed formulae (CCF):

1. Each letter q is CCF;
2. If $t_1, \ldots, \bar{t}_n$ are terms and t_n is time, $g(t_1, \ldots, t_n)$ is CCF;
3. If $x \in \{w, f\}$, $t_1, \ldots, t_n$ are terms and t_n is time, $g(x_1, \ldots, x, t_1)$ is CCF;
4. If $x_1, \ldots, x_n \in \{w, f\}$ are terms and t is time, $g(x_1, x_2, \ldots, x_n, t)$ is CCF;
5. If A is CCF, $\bar{A}$ is CCF;
6. No other expression is CCF.

Sentences:

1. Each CCF is a sentence.
2. If B_1 and B_2 are sentences, B_1B_2 and B_2B_1 are sentences.
3. If B is a sentence, $\bar{B}$ is a sentence.
4. No other expression is a sentence.

Axioms

We shall distinguish two kinds of axioms, statical and dynamical (action) axioms. The former enables deduction within the framework of one temporal cross-section of the behaviour model, the latter involves different temporal

cross-sections. In a model of language L_i, let A and B be sentences with certain semantics. Then the following holds:

a) statical axiom $A \Rightarrow B$ if and only if A expresses sufficient, rather than necessary premises for inference B;

b) statical axiom $A \Rightarrow B$ if and only if A and B express sufficient and necessary premises for each other;

c) action axiom $A \rightarrow B$ if and only if A expresses both initial, and necessary and sufficient conditions for application of an action related to time t_1, and B expresses the result of this action related to time t_2, $t_2 > t_1$.

No other expression is an axiom.

Practically, it is convenient to use words as predicative constants and variables g. For example, constant BEFORE in CCF BEFORE (x_1x_2) (x_1 before x_2) or variable PLACE in CCF PLACE (x_1x_2) (x_1 has x_2 as a place). In the latter case constants ON, IN, INSIDE, etc. may be values of the variable PLACE.

Axioms in system L_i reflect spatial, temporal and functional relations between elements and actions in situations. Examples below are of a spatial axiom:

PLACE (w_1, w_2, t) SUBJECT $(f, w_1, t) \Leftrightarrow$ PLACE (f, w_2, t); of temporal one:
WAS $(x, t) \Rightarrow$ WAS $(x, t + \Delta t)$ where $\Delta t \geq 0$;
and of functional axiom (for chess):
ON (w, x_1, t) MOVE (w, t) ACCESS $(w, x_2, t) \rightarrow$ ON $(w, x_2, t + 1)$.

Grammar

Two types of sequences will be used as meta-expressions, $GB_1 \vdash B_2$ and $\vdash B_1 \oplus B_2$, where B_1 and B_2 are sentences (not necessarily different), G is a list of axioms (including theorems proved before) $\vdash$ is the deducibility sign, and $\oplus$ denotes operations $\rightarrow \Rightarrow$, or $\Leftrightarrow$. The sequence of the first type indicates that sentence B_2 is deduced from B_1 by means of axioms G; that of the second type indicates that $B_1 \oplus B_2$ is not deduced from anything, i.e. that this axiom is given.

The deduction rules in the sequential form are as follows:

Basic rules

R 0. Substitution (for variables)

$$\frac{\vdash B_1 \oplus B_2}{\vdash B_1\theta \oplus B_2\theta},$$

where $\theta = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ is substitution performing appropriate changes of variables in B_1 and B_2.

R 1. $GB_1 \vdash B_2$, where $B_2 \subseteq B_1$.

R 2. $\dfrac{GB_1 \vdash B_2 B_3}{GB_1 \vdash B_2}$; $\quad \dfrac{GB_1 \vdash B_2 B_3}{GB_1 \vdash B_3}$;

R 3. $\dfrac{\begin{array}{l} GB_1 \vdash B_2 \\ \vdash B_2 \oplus B_3 \end{array}}{GB_1 \vdash B_3}$;

R 4. $\dfrac{\begin{array}{l} GB_1 \vdash B_2 \\ GB_1 \vdash B_3 \end{array}}{GB_1 \vdash B_2 B_3}$;

R 5. $\dfrac{GB_1 \vdash \bar{B}_2}{GB_1 \vdash \overline{B_3 B_2}}$;

R 6. $\dfrac{\begin{array}{l} GB_1 \vdash \overline{B_2} \\ \vdash B_2 \oplus B_3 \end{array}}{GB_1 \vdash \bar{B}_3}$;

R 7. $\dfrac{\begin{array}{l} GB_1 \vdash B_2 B_3 \\ \vdash B_2 \oplus B_4 B_5 \end{array}}{GB_1 \vdash B_3 B_5}$;

R 8. $\dfrac{\begin{array}{l} GB_1 \vdash \bar{B}_2 \\ \vdash B_2 B_3 \oplus B_4 \end{array}}{GB_1 \vdash \overline{B_4 B_5}}$;

Rule R 6 with sign $\rightarrow$ is used in the reverse direction only, i.t. from the sequence conclusion to its premises.

Rules R 7 and R 8 in system L_i are derivative ones. Rule R 7 is a sequence of rules 4–3–4–2, and rule R 8 is a sequence of rules 5–6. Rule 7 (or 8) divides deduction of sentence $B_3 B_5$(or $\overline{B_4 B_5}$) into two parts, deduction of B_3 (or $\bar{B}_2$) and elementary deduction of B_5 (or $\bar{B}_4$) by means of an appropriate axiom. They are, in essence, formally described heuristics "means — objective" [11].

Technical rules

R 9. Purge $\dfrac{GB_1 \vdash B_2 B_3 B_2}{GB_1 \vdash B_2 B_3}$;

R 10. Permutation $\dfrac{GB_1 \vdash B_2 B_3}{GB_1 \vdash B_3 B_2}$;

R 11. Contradiction statement $\dfrac{\begin{array}{l} GB_1 \vdash B_2 \\ GB_1 \vdash B_2 \end{array}}{GB_1 \vdash \varnothing}$;

R 12. $\dfrac{\begin{array}{l} GB_1 \vdash B_2 \\ GB_1 B_2 \vdash B_3 \end{array}}{GB_1 \vdash B_2 B_3}$;

Rule R 12 in system L_i is derivative as well.

Indeed, it follows from premise $GB_1 B_2 \vdash B_3$ and R 13 that $GB_1 \vdash B_2 \Rightarrow B_3^{*)}$ and from the first premise and $*)$ with respect to R 3, $GB_1 \vdash B^{3**)}$ follows from the first premise and $**) GB_1 \vdash B_2 B_3$ with respect to R 4.

Augmentation rules for the system of axioms G

R 13. Deduction theorem

$$\frac{GB_1 \vdash B_2}{G \vdash B_1 \Rightarrow B_2}$$

R 14.

$$\frac{G \vdash B_1 \Rightarrow B_2 \quad G \vdash B_2 \Rightarrow B_1}{G \vdash B_1 \Leftrightarrow B_2}.$$

Definition of deduction. In system L_i, deduction D_k from G and B_n ($GB_n \vdash D_k$) is finite list $E = E_1 \ldots E_\eta \ldots E_m$, where each sentence $E_\eta (1 \leq \eta \leq m)$ either belongs to G, or is included into B, or is obtained from previous sentences of list E by means of one of grammatical rules [12].

Theorem 1. Let deduction of sentence $D = \Phi_1\Phi_2 \ldots \Phi_n$ from sentence B take place in system L_i, where Φ_n is CCF and $D \cap B = \varnothing$. Then, there are sentences $B_\eta = B'_\eta\theta_\eta$ and $C_\eta = C'_\eta\theta_\eta$ ($1 \leq \eta \leq K$, $B'_\eta \oplus C'_\eta$ is axiom from list G, and θ_η is substitution) such that any of two following sentences are provable in system L_i:

1) $$v_1 = (B_1 \oplus C_1) \oplus \big((B_2 \oplus C_2) \oplus \big((\ldots) \oplus (B_k \oplus C_k) \ldots\big)\big);$$

2) $$v_2 = (B_k \oplus C_k) \oplus \big((B_{k-1} \oplus C_{k-1}) \oplus ((\ldots) \oplus (B_1 \oplus C_1) \ldots)\big).$$

With the notation accepted, Theorem 1 is formulated as follows: if $GB \vdash D$ and $D \cap B = \varnothing$, then 1) $G \vdash v_1$ and 2) $G \vdash v_2$. Expressions v_1 and v_2 are the results of direct (v_1) and reverse (v_2) proofs.

Proof of Theorem 1. Consider fist the case of the direct proof. A new sentence D may be deduced from $B(D \cap B = \varnothing)$ only by means of deduction rules making use of the list G axioms. These rules are R 0, R 3 and R 6 (R 7 and R 8 are derivatives). Therefore, ηth axiom ($1 \leq \eta \leq k$) may be used in deduction only according to the following plan:

1) deduction of B_η (or $\bar{B}_\eta$) from B;

2) deduction of $G_\eta\theta_\eta = B'_\eta\theta_\eta \oplus C'_\eta\theta_\eta = B_\eta \oplus C_\eta$ by means of rule R 0 and substitution θ_η (possibly empty);

3) deduction of C_η from B_η and $B_\eta \oplus C_\eta$ by means of R 3 (or $\bar{C}_\eta$ from $\bar{B}_\eta$ and $B_\eta \oplus C_\eta$ by means of R 6).

Consider each step of this plan.

Step 1

1) Axiom $G_1 = (B'_1 \oplus C'_1)$ is defined such that for it there are substitution θ_1 and sentence $B^{(1)}$ such that $B^{(1)} = B_1 \cap B$ and $G_1\theta_1 = B'_1\theta_1 \oplus C'_1\theta_1 =$

$= B_1 \oplus C_1$. Evidently, such an axiom exists, otherwise deduction D would be impossible.

2) C_1 (or $\bar{C}_1$) is deduced by means of rule R 3 (or R 6).

Step η. $(1 \leq \eta \leq k)$

1) Axiom $G_\eta = B'_\eta + C'_\eta$ is defined such that there are for it substitution θ_η and sentence $B^{(\eta)}$ (possibly empty) such that $B^{(\eta)} = B_\eta \cap B$, $B_\eta \subseteq B^{(\eta)} C_1 \ldots C_{\eta-1}$ and $G_\eta \theta_\eta = B_\eta \oplus C_\eta$. If the deduction is not completed, such an axiom should exist.

Evidently, at some ηth step $(1 \leq \eta \leq k)$ for the first time sentence $C^{(\eta)} \subseteq C_\eta$ will be deduced identical to Φ_ξ $(1 \leq \xi \leq n)$ or to some set of them, i.e. $C^{(\eta)} = C_\eta \cap D$. Denote this set as D_1. Continuing this process, obtain eventually deduction of sentence $D = D_1 D_2 \ldots D_m$ $(m \leq n)$ in the form of ordered (with respect to deduction) list of sentences D_r $(1 \leq r < m)$. In doing so, assumption B will be represented as $B = B^{(1)} B^{(2)} \ldots B^{(k)}$ where k is the total number of deduction steps.

Thus, elementary theorem $B_\eta \oplus C_\eta$ and $C_\eta (G B_\eta \vdash C_\eta)$ was deduced at each deduction step. By augmenting list G at each step, beginning from the first one, by elementary theorem $B_\eta \oplus C_\eta$, obtain at the last, kth step

$$G, (B_1 \oplus C_1), (B_2 \oplus C_2), \ldots, (B_{k-1} \oplus C_{k-1}), B_k \vdash C_k .$$

Hence, obtain $G \vdash v_1$ by multiple application of the deduction theorem.

Similar reasoning for the case of reverse deduction gives at each step $G B_\eta \vdash C_\eta$ and $B_\eta \oplus C_\eta$. By adding at each step, beginning from the last one, an elementary theorem $B_\eta \oplus C_\eta$ to list G, obtain for the first step

$$G, (B_k \oplus C_k), (B_{k-1} \oplus C_{k-1}), \ldots, (B_2 \oplus C_2), B_1 \vdash C_1 .$$

Hence, obtain $G \vdash v_2$ by multiple application of deduction.

Corollary of Theorem 1. If $GB \vdash D$, where D and B meet conditions of Theorem 1, the following holds:

$$G \vdash ((B \oplus D) \Rightarrow v_1) ,$$

$$G \vdash ((B \oplus D) \Rightarrow v_2) .$$

Proof. Indeed, in virtue of the deduction theorem it follows from $GB \vdash D$ that $G \vdash (B \oplus D)$. According to Theorem 1, $G \vdash v_1$ and $G \vdash v_2$ hold, and, consequently, $G, B \oplus D \vdash v_1 G$, $B \oplus D \vdash v_2$ hold, too. The deduction theorem completes the proof.

Theorem 2. Let $G \vdash v_1$ or $G \vdash v_2$ be true in system L_i. Then there are sentences $D = D_1 D_2 \ldots D_m$ and $B = B^{(1)} B^{(2)} \ldots B^{(k)}$ such that $D \cap B = \varnothing$ and $G \vdash (B \oplus D)$. The proof is evident.

Corollary of Theorem 2. If $G \vdash v_1$ and $G \vdash v_2$, then $G \vdash (v_1 \Rightarrow (B \oplus D))$ and $G \vdash (v_1 \Rightarrow (B \oplus D))$.

Its proof follows the lines of that for corollary of Theorem 1.

In its turn, it follows from corollaries of Theorems 1 and 2 that $G \vdash v_1 \Leftrightarrow (B \oplus D)$ and $G \vdash v_2 \Leftrightarrow (B \oplus D)$, and this leads to $G \vdash v_1 \Leftrightarrow v_2$.

Below, we shall not be concerned with the type of proof (whether direct or reverse), therefore we omit the sign $\oplus$ and some of the parentheses in v_1 and v_2 are written

$$v = (B_1 \oplus C_1)\,(B_2 \oplus C_2) \ldots (B_K \oplus C_K)$$

as ordered from the left to the right list of elementary theorems. As distinguished from v_1 and v_2, sequence v will be referred to as proof of theorem $B \oplus D$.

If in proof v of theorem $B + D$ all elementary static theorems are eliminated and only elementary action theorems are left, we obtain realization plan u of theorem $B \oplus D$. It is evident that

$$u = (B_1 \rightarrow C_1)\,(B_2 \rightarrow C_2) \ldots (B_2 \rightarrow C_2)\,,$$

where $r \leq k$.

Syntax of the translation language

Alphabet

1. Punctation marks: ( ) [ ] { }
2. Variables of ith order (with or without indexes) x_i, $x_i^1 \ldots$;
3. CCF of ith order (with or without indexes) g^i, $g_1^i \ldots$;
4. Variables of jth order (with of without indexes) x^j, x_1^j, $\ldots$;
5. CCF of jth order (with or without indexes) g^j, g_1^j, $\ldots$;
6. Symbols of logic operations $-$, &, V, $\Rightarrow$, $\Leftrightarrow$
7. Quantors of generality $\forall$, existence J and uniqueness $J!$

Quantors $\forall$, J and $J!$ are auxiliary symbols used, first, to reduce writing of conjunction, disjunction and exclusive disjunction, respectively, and, second, to indicate the definition domains of corresponding n-tuples.

Expressions

Quasiformulae (QF)

1. Each CCF of ith order if QF;
2. If A is QF, $\bar{A}$ is QF;
3. If A and B are QF's, then $A \,\&\, B$, $A \vee B$ and $A \Rightarrow B$ are QF's;
4. If A is QF and if x_i is *not* a *unique* free variable in A such that $x^i \in x^j$ then $(\forall x^i \in x^j)\, A$, $(Jx^i \in x^j)\, A$, $(Jx^i \in x^j)\, A$ are QF;
5. No other expression is QF.

Correctly constructed translation formulae (CCTF)

1. If A is QF and if x^i is a *unique* free variable in A such that $x^i \in x^j$, then $(\forall x^i \in x^j)\, A$, $(Jx^i \in x^j)\, A$, $(J!\, x^i \in x^j)\, A$ are CCTF's;
2. If A is CCTF, $\bar{A}$ is CCTF;
3. No other expression is CCTF.

Axioms

1. If B^j is sentence of jth order with definite semantics in the model of language L_j, and if A is CCTF with a definite semantics in the model of language L_{ij}, then $B^j \Rightarrow A$ is axiom.
2. No other expression is an axiom.

The axioms of language L_{ij} give also spatial, temporal and functional relations between elements (actions) of jth order and elements (actions) of ith order. Examples are as follows:

ABOVE $(x_1^j, x_2^j, t^j) \Leftrightarrow (\forall x_2^i \in x_2^j)\,(Jx_1^i \in x_1^j)\;\; (Jt^i \in t^j)$ ABOVE $(x_1^i, x_2^i, t^i))$ BE $(f^j, t^j) \Leftrightarrow (Jf^i \in f^j)\,(J^i \in t^j)$ BE $(f^i, t^i))$

SUBJECT $(f^j, w^j, t^j) \Leftrightarrow (Jf^i \in f^j)\,(Jw^i \in w^j)\,(Jt^i \in t^j)$ SUBJECT $(f^i, w^i, t^i))$

Grammar

Let G be the system of axioms of language L_{ij}.

Let B^j be sentence of ith order and A be a quasiformula. Then we have the following deduction rules in the sequential form.

T 0. Substitution (for variables)

$$\frac{\vdash A \Leftrightarrow B^j}{\vdash A\theta \Leftrightarrow B^j\theta}\,,$$

where θ is a substitution performing appropriate change of variables.

T 1. $GA \vdash A$;

T 2. $\dfrac{GB^i \vdash A_1 \quad GB^i \vdash A_2}{GB^i \vdash A \,\&\, A_2}$; $\dfrac{GB^i \vdash \bar{A}_1}{GB^i \vdash A_1 \,\&\, A_2}$; $\dfrac{GB^i \vdash \bar{A}_2}{GB^i \vdash A_1 \,\&\, A_2}$;

T 3. $\dfrac{GB^i \vdash A_1}{GB^i \vdash A_1 \vee A_2}$; $\dfrac{GB^i \vdash A_2}{GB^i \vdash A_1 \vee A_2}$; $\dfrac{GB^i \vdash \bar{A}_1 \quad GB^i \vdash \bar{A}_2}{GB^i \vdash \overline{A_1 \vee A_2}}$;

T 4. $\dfrac{GB^i \vdash \bar{A}_1}{GB^i \vdash A_1 \Rightarrow A_2}$; $\dfrac{GB^i \vdash A_2}{GB^i \vdash A_1 \Rightarrow A_2}$; $\dfrac{GB \vdash A_1 \quad GB^i \vdash \bar{A}_2}{GB^i \vdash \overline{A_1 \Rightarrow A_2}}$.

Let $x^j = \{x^i_\xi \,/\, 1 \leq \xi \leq n\}$ where x^i_ξ and x^j are variables of, respectively, ith and jth orders. Then the following holds:

T 5. $\dfrac{GB^i \vdash A(x^i_1) \quad GB^i \vdash A(x^i_2) \quad \ldots\ldots\ldots\ldots \quad GB^i \vdash A(x^i_n)}{GB^i \vdash (\forall x^i \in x^j)\, A(x^i)}$; $\dfrac{GB^i \vdash \bar{A}(x^i_\xi)}{GB^i \vdash (\forall x^i \in x^j)\, A(x^i)}$;

T 6. $\dfrac{GB^i \vdash A(x^i_\xi)}{GB^i \vdash (Jx^i \in x^i)\, A(x^i)}$; $\dfrac{GB^i \vdash \bar{A}(x^i_1) \quad GB^i \vdash A(x^i_2) \quad \ldots\ldots\ldots\ldots \quad GB \vdash A(x^i_n)}{GB^i \vdash (Jx^i \in x^j)\, A(x^i)}$;

$\dfrac{GB^i \vdash A(x^i_\xi) \quad GB^i \vdash (Jx^i \in x^j \setminus x^i_\xi)\, A(x^i)}{GB^i \vdash (\bar{J}\,! \in x^j)\; A(x^i)}$

T 7. $\dfrac{GB^i \vdash A(x^i_\xi) \quad GB^i \vdash (\forall x^j \in x^j \setminus x^i_\xi)\, \bar{A}(x^i)}{GB^i \vdash (J!\, x^i \in x^j)\, A\,(x^i)}$; $\dfrac{GB^i \vdash A(x^i_\xi) \quad GB^i \vdash (Jx^i \in x^j \setminus x^i_\xi)\, A(x^i)}{GB^i \vdash (\bar{J}\,!\, x^i \in x^j)\, A(x^i)}$;

T 8. $GB^i \vdash A$ T 9. $GB^i \vdash \bar{A}$ T 10. $GB^i \vdash B^j$ T 11.

$$\frac{\dashv A \Leftrightarrow B^j}{GB^i \vdash B^j}; \qquad \frac{\vdash A \Leftrightarrow B^j}{GB^i \vdash \bar{B}^j}; \qquad \frac{GB^i \vdash B_2^j}{GB^i \vdash B_1^j B_2^j}; \qquad \frac{GB^i \vdash \bar{B}_1}{GB^i \vdash \overline{B_1^j B_2^j}}.$$

In rules T 0, T 8 to T 11, A stands for CCTF.

Deduction. In system L_{ij} deduction of sentence B^j of jth order from sentence B^i of ith order ($j > i$) is a finite list $F = E_1 \ldots E_m$ where:

— every expression $E_\xi (1 \leq \xi < m - 2)$ is either ith order CCF, or is obtained from preceeding expressions of list E with respect to one of the rules of the translation grammar;

— $E_{m-2} = -\widetilde{A}$

— $E_{m-1} = A \Leftrightarrow B^j$;

— $E_m = \widetilde{B}^j$, where $\widetilde{A}$ is CCTF, $\widetilde{A} \in \{A, \bar{A}\}$ and $\widetilde{B}^j \in \{B^j, \bar{B}^i\}$.

Deduction structure in the system L

Let the following hold:

— in system L_i $v^i = (B_1^i \oplus C_1^i) \ldots (B_l^i \oplus C_l^i)$ be the proof of theorem $B_n^i \oplus D_l^i$;
— in system L_{ij}, for each element $\xi \in [1, K]$ separating in v^i part v_ξ^i so that $v^i = v_1^i \ldots v_\xi^j \ldots v_k^i$ there be deductions E_ξ^i and E_ξ'' corresponding to $G_{ij} H_\xi^i \vdash B_\xi^i$ and $G_{ij} D_\xi^i \vdash C_\xi^j$, where $H_\xi^i \oplus D_\xi^i$ is a theorem with proof v_ξ^i, and $B_\xi^j \oplus C_\xi^i$ is an elementary theorem in v_j;
— in system L_j, $v_j = (B_1^j + C_1^j) \ldots (B_k^j + C_k^j)$ be proof of theorem $B_n^j + D_k^j$.

Then the totality $v = \{v^j, E, v^i\}$ (where $E = \{E_\xi', E_\xi'' \mid 1 \leq \xi \leq K\}$ is the complete proof of theorem $B_n^i \oplus D_k^i$ in system L; in its turn, the plan of the realization of the theorem is, by definition, representable in the form of $u = \{u^j, u^i\}$.

Consider the class of complete L.

Definition. If theorem $B_n^j \oplus D_l^j$ with proof $v^j = (B_1^j \oplus C_1^j) \ldots (B_K^j + C_K^j)$ is solvable in L_i each time as theorem $B_n^i \oplus D_l^i$ with proof $v^i = (B_l^j \oplus C_l^j) \ldots$ $\ldots (B_l^j \oplus C_l^j)$ is solvable in L_i and deductions $G_{ij} B_n^i \vdash B_n^j$ and $G_{ji} D_l^i \vdash D_k^j$ are solvable in L_{ij}, system L is complete.

Existence of decomposition $\langle v^i \rangle$ for each proof v may serve as a sufficient completeness condition for system L. Decomposition $\langle v^i \rangle$ is constructed as follows.

Step 1. From the elementary theorems of proof, beginning with the first one, two sentences $H^i_\eta = B^{(1)}B^{(2)} \ldots B^{(\eta)}$ and D^i_1 are formed consecutively such that

1) $$B^{(\xi)} = B^i_n \cap B^i_\xi \ (1 \leq \xi \leq n) \text{ and } D^i_1 = C^i_\eta \cap D^i_l;$$

2) for the first time (with respect to number η) deductions $G_{ij}H^i_\eta \vdash B^j_1$ and $G_{ij}D^i_1 \vdash C^j_1$ take place where $B^i_1 \oplus C^j_1 = G^j_1\theta^j_1$ and $G^j_1 \in G_j$.

The value of η obtained here defines the fist element v^i_1 of decomposition $\langle v^i \rangle$, with elementary theorem $(B^j_1 \oplus C^j_1) \in v^j$ corresponding to this element. If $\eta = 1$, construction of decomposition is completed, otherwise goes to Step 2.

Step 2. v^i, B^i_n and D^i_l are defined, i.e. instead of v^i of length l its part v^i/v^i_1 of length $l - \eta$ is considered, instead of B_n and D, respectively, $B^i_nC^i_1 \ldots C^i_\eta$ and $D^i_e \backslash D^i_1$ are considered. Next go to Step 1. As the result, decomposition $\langle v^i \rangle$ is formed, to whose ordered sequence of elements proof v^j of theorem $B^j_n \oplus D^j_k$ corresponds.

Consequence of system L completeness. Let theorem $B^j_n \oplus D^j_K$ be unsolvable in L_j, deductions $G_{ij}D^i_l \vdash D^j_K$ and $G_{ij}B^i_n \oplus B^j_n$ being solvable in L_{ij}. Then, if L is a complete system, theorem $B^i_n \oplus D^i_e$ is not solvable in L_i.

Indeed, let theorem $B^i_n = D^i_l$ be solvable in system L_i. It follows then from the completeness of L that theorem $B^i_n \oplus D^j_K$ should be solvable in L_j, this being in contradiction with the initial condition.

Semantics and pragmatics of system L

World $S = \{S_i, S_{ij}, S_j\}$ is a model of system $L = \{L_i, L_{ij}, L_j\}$, if there is a one-to-one system of mappings

$$\varrho = \{\varrho_i : L_i \rightarrow S_i, \varrho_{ij} : L_{ij} \rightarrow S_i \times S_j, \varrho_j : L_j \rightarrow S_j\}$$

such that

1) if w is an element in L_i (or in L_j), $\varrho\Psi$ is an object in S_i (or in S_j);
2) if f is the name of an action in L_i(or in L_j), ϱf is an action in S_i (or in S_j);
3) if t is term in L_i(or in L_j), then

$$\varrho t = \begin{cases} \varrho\alpha \text{ — value of constant in } S_i \text{ (or in } S_j) \\ \varrho y \text{ — value of characteristic in } S_i \text{ (or in } S_j) \\ \varrho\varphi(\varrho t_1 \ldots, \varrho t_n) \text{ — value of function in } S_i \text{ (or in } S_j). \end{cases}$$

4) if n is functional variable in L_i (or in L_j), ϱ_n is a function in S_i (or in S_j);
5) if q is a propositional variable in L_i(or in L_j), ϱq is a true or false proposition in S_i (or in S_j);

6) if g and $\bar{g}$ are predicate variables in L_i (or in L_j), then ϱg and $\varrho\bar{g}$ are mutually contradicting n-place relations of which only one occurs in S_i (or in S_j), as a rule $1 \leq n \leq 2$;

7) if A is CCF in L_i (or in L_j), then

$$\varrho A = \begin{cases} \varrho g(\varrho t_1, \ldots, \varrho t_n) \text{ is } n\text{-tuple of values of terms related by } \varrho g \text{ in } S_i \text{ (or in } S_j), \\ \varrho g(\varrho x, \varrho t_1, \ldots, \varrho t_n) \text{ is } n\text{-tuple of values of characteristics either of} \\ \qquad \text{object } \varrho w, \text{ or action } \varrho f \text{ in } S_i \text{ (or in } S_j) \\ \varrho g(\varrho x_1, \ldots, \varrho x, t_n) \text{ is } n\text{-tuple of either objects } \varrho w, \text{ or actions } \varrho f, \text{ or of} \\ \qquad \text{both in } S_i \text{ (or in } S_j); \end{cases}$$

8) if B is a proposition in L_i (or in L_j), ϱB is a situation in S_i (or in S_j);

9) if C is an axiom (theorem) in L_i (or L_j), then ϱC is either statical, or dynamical law in S_i (or in S_j);

10) if $B_n \Rightarrow D_K$ is a theorem provable in L_i (or in L_j), $(\varrho B_n, \varrho D_K)$ is a problem in S_i (or in S_j);

11) if v is the proof of theorem $B_n \Rightarrow D_K$ in L_i (or in L_j), ϱv is the solution of problem $(\varrho B_n, \varrho D_K)$ in S_i (or in S_j);

12) if u is a plan for realization of theorem $B_n \Rightarrow D_K$ in L_i (or in L_j), ϱu is realization of problem $(\varrho B_n, \varrho D_K)$ in S_i (or in S_j);

13) if A is CCTF in L_{ij}, $\varrho_{ij}A$ is an n-place relation in $S_i \times S_j$, relating (any, even a single) n-tuple of elements and actions of lower, ith order belonging, repectively, to elements and actions of higher, jth order, as a rule $n \geq 2$;

14) if C is axiom in L_{ij}, $\varrho_{ij}C$ is a logical law in $S_i \times S_j$ establishing the one-to-one relation between the n-place relation in S_i, and the n_2-place relation in S_j, as a rule $n_2 < n_1$;

15) if E is deduction in L_{ij}, ϱ_{ij} E is either a generalization of situations in S_i by means of convolution of n_1-place relations in S into n_2-place relations in S_j, or detailing of situations in S_j by developing n_2-place relations in S into n_1-place relations in S_i;

16) if $B_n^i \Rightarrow D_k^j$ is a theorem provable in L_{ij}, then $(\varrho_i B_n^i, \varrho_j D_j^j)$ is a problem in $S_i \times S_j$, in this case $\varrho_i B_n^i$ is an initial situation in S_i, and $\varrho^j D$ is the ${}^j_K$target, final situation in S_j.

Conclusion

End-game in chess [13] and robot movements over terrain [14] were considered as particular models of the language L.

References

1. *Pospelov, D. A.*, "Principles of Situational Control," in Izvestiya AN SSSR. Tekhnicheskaya Kibernetika, 2, 1971 (in Russian).
2. *Klykov, Yu. I.*, Situational Control of Large Systems, Energiya Publ., Moscow, 1974 (in Russian).
3. *Davis, M.*, Eliminating the irrelevant from mechanical proofs.
4. *Darlington, J.* Automatic theorem proving with equality substitutions and mathematical induction, in *Michie* (ed.), "Machine intelligence 3", American Elsevier Publishing 60., Inc., New York, 1968.
5. *Lucham, D.*, Refinement theorems in resolution theory, IRIA Symposium on Automatic Deduction, Versailles, France, 1965.
6. *Kowalski, R., Hayes,* P., Semantic trees in automatic theorem proving, in *Meltzer, Michi* (eds.), "Machine Intelligence 4", American Elsevier Publishing Co., Inc., New York, 1969.
7. *Robinson, J.*, A note on mechanizing higher order logic, in *Meltzer, Michi* (eds.), "Machine Intelligence 4", American Elsevier Publishing Co., Inc., New York, 1969.
8. *Robinson, G., Wos,* L., Paramodulation and theorem-proving in first-order theories with equality, in *Meltzer, Michi* (eds.), "Machine Intelligence 4", American Elsevier Publishing Co., Inc., New York, 1969.
9. *Robinson, J.*, A note on mechanizing higher order logic, in *Meltzer, Michi* (eds.), "Machine Intelligence 5", American Elsevier Publishing Co., Inc., New York, 1969.
10. *Robinson, J.*, An overview of mechanical theorem proving, 4th System Symposium 1970.
11. *Newell, A., Shaw, J. C., Simon, H. A.*, in "Self-Organizing Systems", Pergamon Press, 1960.
12. *Klini, S. K.*, Matematicheskaya logika. "Mir", Moscow, 1964 (in Russian)
13. *Efimov, E. I.*, A Heuristic Model of Chess End—Game, in "Trudy Seminara MDNTP im. Drerzhinskogo, 1971 (in Russian).
14. *Efimov, E. I.*, "Robot Moving Over Terrain by Tracing Its Route on a System of Maps Previously Formed by Itself", Voprosy Radiotekhniki, seriya Obschchetekhnicheskaya, **14,** 1973 (in Russian).

Printed in Hungary
AKADÉMIAI KIADÓ BUDAPEST

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

H—1111 Budapest XI., Goldmann György tér 3, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4—5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results' Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary — possibly in Russian if the paper is in English and *vice versa* — should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10—15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которых должен содержать рисунки, таблицы и литературу) направляются

Е. Д. Теряеву — Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский Проспект, 14, Москва В-71, СССР

или

Я. Кочишу — Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета
H—1111 Будапешт, XI, площадь Голдманн Д., 3,
Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре.

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4—5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру .

Статье должна предшествовать аннотация объемом 50—100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме — на русском языке, если статья написана на английском, и наоборот — должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем резюме не должен превышать 10—15% объема статьи.

Авторам высылается корректура статьи, которую они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Гоппа В. Д. Декодирование и диофантовы приближения (*Goppa, V. D.:* Decoding and diophantine approximations) 195

Raczyński, S.: A global optimizaton procedure for dynamic systems (*Рачински С.* Метод глобальной оптимизации динамических систем) 207

Fritz, J.: Two optimization methods in nearest neighbor pattern classification (*Фриц Й.* Два способа оптимизации при использовании метода классификации по правилам «ближайших соседей») 227

Donkov, D., Glushkov, L.: Simulation models of computing systems with two different memory structures (*Донков Д., Глушков Л.* Модели вычислительных систем с двумя различными стручтурами памяти) 239

Ефимов Е. И. Логика поведения целенаправленных систем (*Efimov, E. I.:* Purposeful system behaviour logic) 247

Grodzki, Z.: New class of the controlled shift-registers. II (*Гродзки З.* Новый класс управляемых регистров сдвига. II) 263

Lantos, B.: Necessary conditions for the optimality in abstract optimum control problems with nonscalar-valued performance criterion (*Лантош Б.* Необходимые условия оптимальности по нескалярному критерию в абстрактных проблемах оптимального управления) 271

The IVth IFAC Symposium on Identification and System Parameter Estimation 285

Book Review 287

316.920

VOL. 5 * NUMBER 4
TOM ТОМ НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1976

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

is an international quarterly sponsored jointly by the Presidium of the Academy of Sciences of the USSR and of the Hungarian Academy of Sciences. It offers publicity for original papers and short communications of the following topics:

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

— theory and methods of adaptation, learning, identification and pattern recognition

— methods of information processing; application of digital computers in control and communication systems

— new physical principles in constructing technical devices for automation, control and information processing

The six issues published per year make up a volume of some 480 pp.

While this quarterly is mainly a publication forum of the research results achieved within the USSR and Hungary, also papers of international interest from other countries are welcome.

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

H-1389 Budapest 62, P.O.B. 149, Hungary

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

H-1363 Budapest P.O.B. 24, Hungary

Академия наук СССР и Академия наук ВНР совместно издают журнал

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал публикует материалы о новых результатах научных исследований, разработок и краткие сообщения о достижениях в следующих областях науки:

— общая теория процессов управления и теория систем
— теория автоматов
— теория информации
— исследование операций; теория сложных систем
— теория управления экономическими системами; модели систем
— теория и методы адаптации, обучения, идентификации и распознавания образов

— методы обработки информации; применение вычислительных машин в системах управления и передачи информации

— новые физические принципы создания технических средств автоматики, управления и передачи информации

Журнал быхогит шесть раз в год общим объёмом 480 печатных страниц.

Журнал публикует статьи о научных достижениях обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

1.-1389 Будапешт 62, п. о. 149, Венгрия

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии наук Венгрии

H-1363 Будапешт п. о. 24, Венгрия

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)
F. CSÁKI (Budapest)

DEPUTY EDITORS

V. S. PUGACHEV (Moscow)
V. I. SIFOROV (Moscow)
S. CSIBI (Budapest)

COORDINATING EDITORS

M. A. GAVRILOV (Moscow)
L. KALMÁR (Szeged)

S. V. EMELYANOV (Moscow)
I. CSISZÁR (Budapest)

N. N. KRASOVSKII (Sverdlovsk)
A. PRÉKOPA (Budapest)

E. P. POPOV (Moscow)
L. VARGA (Budapest)

E. D. TERYAEV (Moscow)
J. KOCSIS (Budapest)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)
Ф. ЧАКИ (Будапешт)

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Ш. ЧИБИ (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

М. А. ГАВРИЛОВ (Москва)
Л. КАЛМАР (Сегед)

С. В. ЕМЕЛЬЯНОВ (Москва)
И. ЧИСАР (Будапешт)

Н. Н. КРАСОВСКИЙ (Свердловск)
А. ПРЕКОПА (Будапешт)

Е. П. ПОПОВ (Москва)
Л. ВАРГА (Будапешт)

Е. Д. ТЕРЯЕВ (Москва)
Я. КОЧИШ (Будапешт)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES

BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 5 (4), pp. 291–302 (1976)

GAME-THEORETIC CONTROL UNDER INCOMPLETE PHASE-STATE INFORMATION

N. N. KRASOVSKII

(Sverdlovsk)

(Received December 10, 1975)

The paper deals with a problem of control under the action of either a conflict-type or an uncertain disturbance, the situation being complexified by incomplete information on the realizations of the state-space variables of the system. The initial problem is formalized as one of approach or evasion in an appropriate differential game. The role of the positional element that constitutes a basic variable for constructing a closed-loop control procedure is now performed by the realization of a certain informational domain that includes the unknown actual state-space vector with no further information on the latter being available. An effective computer-implementable solution procedure is proposed. The paper is related to investigations [1, 2].

1. Here we take the definitions and notations of monograph [1] with no further reference. We shall introduce these notions by spacing the corresponding terms.

Consider a dynamic system described by the differential equation

$$\dot{x} = f(t, x, u, v) \tag{1.1}$$

where x is the phase vector for the controlled plant; u is the regulating force considered as the control for the first player-ally; v is the vector of uncertain disturbances that may reflect an organized counteraction. The vector v is being considered as a control for the secondp layer-opponent in an appropriate differential game. Admissible values of u and v are restricted by the inclusions $u \in \mathcal{P}$, $v \in \mathcal{Q}$ where the bounded and closed sets $\mathcal{P}$ and $\mathcal{Q}$ reflect the abilities of the regulator U and the device V that either generates the uncertain disturbance or directs the counteraction. We shall now discuss two problems for the first player.

The problem of approach consists in constructing a regulation law U, that would ensure the approach of the phase vector $x(t)$ to a given closed set $\mathfrak{M}$ in the space $\{x\}$ no further than at a given moment of time ϑ. In other words one is to ensure the inclusion

$$x(\tau) \in \mathfrak{M}, \qquad \tau \leq \vartheta. \tag{1.2}$$

The problem of e v a s i o n consists in constructing a regulation law U that would ensure the evasion of the phase vector $x(t)$ from set $\mathfrak{M}$ throughout the interval ending at ϑ. In other words one is to ensure the condition

$$x(t) \notin \mathfrak{M}, \qquad t_0 \leq t \leq \vartheta. \tag{1.3}$$

A special feature of the problem is that the first player that develops the control u for each moment of time t, $(t_0 \leq t \leq \vartheta)$ has no information whatever on both the future realization $v(\tau)$ (either for $\tau \geq t$ or $\tau > t$) of the disturbance and also on the precise value of the state space vector $x(t)$ for the current moment of time. At each moment t the first player is informed of but a realization $G(t)$ for a generally nonminor i n f o r m a t i o n a l domain $G(t)$ that includes the nonavailable precise value of the actual state-space vector $x(t)$. Hence the closed-loop regulator U is ought to generate for each $t \in (t_0, \vartheta)$ a control force realization $u(t)$ based on information on the realization $G(t)$. But in the latter case it is convenient to formalize the given problem of control for the motion $x(t)$ (1.1) as one of controlling the informational domain $G(t)$. Thus we arrive at a problem of approach that requires the inclusion

$$G(\tau) \in \mathfrak{M}, \qquad \tau \leq \vartheta \tag{1.4}$$

and at a problem of evasion that requires the condition

$$G(t) \cap \mathfrak{M} = \varnothing, \qquad t_0 \leq t \leq \vartheta. \tag{1.5}$$

(The symbol $\varnothing$ stands for an empty set. Hence (1.5) denotes that the sets $G(t)$, $(t_0 \leq t \leq \vartheta)$ have no intersection with $\mathfrak{M}$.)

The variation of sets $G(t)$ in time is determined by the dynamics of system (1.1) and also by the nature of the information accessible for the first player-ally. We shall characterize the possible cases for the problem.

In each case the first player-ally is informed on the realization $G(t)$ for each t while the second player may perhaps be informed by the realization $x(t)$. Here in Case I the first player receives no information on the realization $v(t)$ for the control of the opponent, while however the second one may perhaps be informed of the control realization $u(t)$ for the ally. In Case II the first player receives no information on $v(t)$ while the second one is noninformed on $u(t)$. In Case III the first player may use $v(t)$ for the computation of $u(t)$, but it may not use the h i s t o r y $v(\tau)$ $(\tau < t)$ of the opponent's control for a more precise estimation of the domain $G(t)$. Finally in Case IV the first player may use the values $v(\tau)$ $(\tau \leq t)$ both for computing $u(t)$ and for estimating $G(t)$.

In all the cases taken into consideration the control procedure for the first player within a minor semi-interval $t_* \leq t < t^*$ is based on the concept

of a respective program, which serves for the construction of the analogies of p u r e strategy (I), m i x e d strategy (II) and c o u n t e r s t r a t e g y (III, IV) as taken for the theory of closed-loop differential games with complete information [1]. In each case we start by defining the e l e m e n t a r y program $\Pi_э(t_*, t^*)$, which in Case I is given by a piecewise constant function $u(t) \in \mathcal{S}$, in Case II by probability measure $\mu_t(du)$ defined on set $\mathcal{S}$ and piecewise constant in t, in Cases III and IV by a function $u(t, v) \in \mathcal{S}$ piecewise constant in t. An elementary program is defined as an operation that constructs a set of points (in Cases I–IV)

$$G_{\mathrm{I}}[t^*, t_*, x_*, \Pi_э] = \{x: x = x^* + \int_{t_*}^{t^*} f\big(t, x[t], u[t], v(t)\big)\, dt;\ v(t) \in \mathcal{Q}\}^* \tag{1.6}$$

$$G_{\mathrm{II}}[t^*, t_*, x_*, \Pi_э] = \{x: x = x^* + \int_{t_*}^{t^*} \int_{\mathcal{S}\times\mathcal{Q}} f\big(t, x(t), u, v\big)\, \mu_t(du)\, \nu_t(dv)\, dt;\ \nu_t(dv)\}^*, \tag{1.7}$$

$$G_{\mathrm{III,IV}}[t^*, t_*, x_*, \Pi_э] = \{x: x = x^* + \int_{t_*}^{t^*} f\big(t, x(t), u(t, v(t)), v(t)\big)\, dt;\ v(t) \in \mathcal{Q}\}^*. \tag{1.8}$$

In Cases I and III the $v(t)$ of (1.6) and (1.8) runs over all admissible piecewise constant functions $v(t) \in \mathcal{Q}$. In Case II $\nu_t(dv)$ of (1.7) runs over all the probabilistic measures $\nu_t(dv)$ on $\mathcal{Q}$ piecewise constant in t. In (1.6)–(1.8) $x(t)$ is the solution of the corresponding integral equation where $x(t_*) = x_*$, $x(t^*) = x$; the symbol $\{\ldots\}^*$ stands for the closure of set $\{\ldots\}$. T h e p r o g r a m $\Pi(t_*, t^*)$ is defined as an operation that constructs either the set G_j or the closed convex hull of set G_j (the latter being closed in the Euclidean metric of space $\{x\}$) or the set that is a member of the closed convex hull of the previous variety of sets (as taken closed in the Hausdorff metric for the variety).

The fixed domain $G(t_*)$ and the program $\Pi(t_*, t^*)$ determine t h e d o m a i n o f a t t a i n a b i l i t y

$$G_j(t^*, t_*, G[t^*], \Pi) = \{x: x = x^*, x_* \in G[t_*],\ x^* \in G_j[t^*, t_*, x^*, \Pi]\}. \tag{1.9}$$

(The index $j = \mathrm{I}, \ldots, \mathrm{IV}$ in the notation for the admissible domain G is further omitted.) According to [1] we assume that under the realization of domain $G[t_*]$ and the action of program $\Pi[t_*, t^*]$, it is only the domain

$$G[t^*] \subset G(t^*, t_*, G[t_*], \Pi) \tag{1.10}$$

that may be realized at moment $t = t^*$.

Moreover, we assume that the realization of domain $G(t^*)$ must also satisfy the inclusion

$$G[t^*] \subset G_*[t^*], \tag{1.11}$$

where $G_*[t^*]$ is the realization of the estimating domain that is obtained as a result of an observation of system (1.1) throughout the semi-interval $t_* < t \leq \leq t^*$. The inclusion (1.10) for moment $t = t^*$ takes into consideration both the past realization $G[t_*]$ and the abilities of the opponent $v(t)$ at $t_* \leq t < t^*$. According to these features it turns out that at moment $t = t^*$ it may be only the state $x[t^*] = x^*$, unknown for the first player, that may be realized. Here $x_* \in G[t_*]$ and $x^* \in G_j(t^*, t_*, x_*, \Pi)$. Apart from this there may be an additional contraction of set $G(t^*)$ as compared to $G(t^*, t_*, G[t_*], \Pi)$ (1.9) due to additional information available at the time-interval $t_* < t \leq t^*$. The latter information includes the estimate (1.11) received through the new measurements made till the moment $t = t^*$.

Hence we assume that the first of the players forms its control action on base of the feedback principle while working in terms of a discrete-time scheme with s u b d i v i s i o n $\Delta\{\tau_i\}$ of the time axis formed of sufficiently small s t e p s

$$\delta(\Delta) = \max_i (\tau_{i+1} - \tau_i). \tag{1.12}$$

The program $\Pi[\tau_i, \tau_{i+1})$ is chosen at each moment of time $t = \tau_i$ for the entire future semi-interval $\tau_i \leq t < \tau_{i+1}$ on base of realization $G[\tau_i]$ and perhaps on base of the realization at $t = \tau_i$ of some additional variables worthy to be introduced into the control device for the first player. Therefore the desired strategy U, that governs the control law must determine the programs $\Pi(\tau_i, \tau_{i+1})$ as well as the variation of the additional variables mentioned above. Take the symbol $\mathfrak{M}^{(\varepsilon)}$ to denote the ε-neighbourhood of the target set $\mathfrak{M}$. The problems under consideration may be formulated as follows.

Problem 1.1 of approach. Given are the moment of time ϑ and the initial state $G[t_0]$ $(t_0 < \vartheta)$. One is to specify a strategy U, that would ensure for any value of $\varepsilon > 0$ the encounter

$$G(\tau) \subset \mathfrak{M}^{(\varepsilon)} \tag{1.13}$$

whatever are the possible realizations $G[t]$ satisfying conditions (1.10) for $(t_* = \tau_i,\ t^* = \tau_{i+1},\ \tau_0 = t_0,\ i = 1, 2, \ldots)$ and (1.11) (for $t^* = \tau_{i+1}$), provided the step $\delta(\Delta)$ (1.12) is sufficiently small.

Problem 1.2 of evasion. Given are the moment ϑ and the initial state $G[t_0]$ $(t_0 < \vartheta)$. One is to specify a strategy U that would ensure for a certain $\varepsilon > 0$ the evasion

$$G(t) \not\subset \mathfrak{M}^{(\varepsilon)} \tag{1.14}$$

whatever are the possible realizations $G[t]$ satisfying conditions (1.10) and (1.11) (provided the step $\delta(\Delta)$ is sufficiently small).

2. We shall seek for the solutions of problems 1.1 and 1.2 within the control scheme with a l e a d e r. Assume that the formation of realizations $G_*[t^*]$ for the estimating domain of (1.11) received due to observation of system (1.1) within the semi-interval $[t_*, t^*]$ may be described by means of a set of domains $G^*[t, y]$ majorizing the former from above and also with the aid of an additional variable y. This is as follows. Assume the value $y[t_*]$ has realized. For any future moment t^* the value $y[t_*]$ and the program $\Pi[t_*, t^*]$ determine the set of possible e s t i m a t i n g closed domains $\{G_*[t^*];\ y[t_*],\ \Pi[t_*, t^*)\}$ and m a j o r i z i n g closed domains $\{G^*[t^*, y[t^*]];\ y[t_*],\ \Pi(t_*, t^*)\}$. It is assumed that at time t^* it is any of the estimating sets $G_*[t^*]$ that may arrive and will intersect with the domain of attainability $G(t^*, t_*, G^*[t_*, y[t^*]], \Pi)$, where $\Pi[t_*, t^*)$ is the program designated for the semi-interval $[t_*, t^*)$. The given control process formalization leaves the choice of set $G_*[t^*]$ among the variety of sets $\{G_*[t^*];\ y[t_*],\ \Pi[t_*, t^*)\}$ to the second player-o p p o n e n t. The result of observation are hence uncontrollable in advance by the first player-a l l y, whose regulation law is aimed at the solution of problem 1.1 (or problem 1.2). The choice of realization $y[t^*]$ is left to the first player-ally as the value $y[t]$ constitutes one of the additional variables, formed in the g u i d e model included in the regulation scheme. The choice of $y[t^*]$ is restricted by the conditions

$$\left[G\left(t^*, t_*, G^*[t_*, y[t_*]], \Pi\right) \cap G_*[t^*]\right] \subset G^*[t^*, y[t^*]] \tag{2.1}$$

and

$$\| y[t^*] - y[t_*] \| \leq \beta(t^* - t_*)\,, \tag{2.2}$$

where β is a sufficiently large constant. (The symbol $\| x \|$ stands for the Euclidean norm of vector x.) The fulfilment of conditions (2.1) and (2.2) whose realization is demanded for the values t^* close to t_*, requires that there would exist at least one value of $y[t^*]$ that would satisfy relations (2.1) and (2.2). The introduction of the majorizing domains $G^*[t, y]$ gives but a rough estimate of the accessible information but on the other hand it simplifies the solution, reducing the latter to a finite dimensional control problem for a vector variable $y[t]$ rather than to one of controlling the domains $G[t]$. We further assume that the system parameters of (1.1) and the selection of the varieties $\{G_*[t^*];\ y[t_*],\ \Pi[t_*, t^*)\}$ and $\{G^*[t^*, y[t^*]];\ y[t_*],\ \Pi[t^*, t^*)\}$ are such that on the one hand the conditions (2.1) and (2.2) would hold while on the other there would exist sufficiently regular solutions for the equations and extremal problems discussed in the sequel. These assumptions are obviously of moderate restriction on the system.

Thus the vector $y[t]$ is one of the auxiliary variables mentioned in Section 1. Another of these variables is given by a certain vector $w[t]$, whose type

of variation over time is defined below. Therefore the motion of the closed-loop regulation system is characterized by the triple $\{G[t],\ w[t],\ y[t]\}$ where the realization of set $G(t)$ describes the transient current state of the controlled p l a n t (1.1) while vectors $w[t], y[t]$ describe the current state of the g u i d e model system, included in the regulator. The variation of $G[t]$ in time satisfies conditions (1.10) and (1.11); the variation of $y[t]$ will be at first restricted by conditions (2.1) and (2.2) while the variation of $w[t]$ will be defined as follows. Assume the program $\Pi[t_*, t^*]$ is selected and the vector s is fixed. Compute the values

$$\sigma(t^*, t_*, y[t_*], s, \Pi) = \max_{G_*[t^*]} \min_{y[t^*]} s'(y[t^*] - y[t_*]) \tag{2.3}$$

under restrictions (2.1) and (2.2)

$$\sigma(t_*, y[t_*], s, \Pi) = \limsup_{t^* \to t_*} \frac{\sigma(t^*, t_*, y[t_*], s, \Pi)}{t^* - t_*} \tag{2.4}$$

and

$$\sigma^0(t, y, s) = \sigma(t, y, s, \Pi^s) = \min_{\Pi} \sigma(t, y, s, \Pi). \tag{2.5}$$

The variation of $w[t]$ in the interval $[\tau_i, \tau_{i+1}]$ is subjected to the contingent equation

$$\dot{w}(t) \in Y_\beta(t, w(t), s) \tag{2.6}$$

where $Y_\beta(t, w, s)$ is the intersection of the ball $\| y \| \leq 2\beta$ and the semispace $Y(t, w, s)$ in $\{y\}$ defined by inequality

$$Y(t, w, s) = \{y: s'y \geq \sigma^0(t, y, s)\}. \tag{2.7}$$

$s = y[\tau_i] - w[\tau_i]$ and the prime stands for the transpose.

Thus the required strategy $U\{(\Pi(\tau_i, \tau_{i+1});\ w[\tau_i],\ y[\tau_i]);\ (w[\tau_{i+1}];\ w[\tau_i],\ y[\tau_i]);\ (y[\tau_{i+1}];\ w[\tau_{i+1}],\ G[\tau_{i+1}])\}$ will define the program $\Pi[\tau_i, \tau_{i+1})$ that generates the motion $G[t]$ of (1.10) and (1.11) on the basis of the actuated realization $\{w[\tau_i],\ y[\tau_i]\}$. It will also compute the value $w[\tau_{i+1}]$ and further on-the value $y[\tau_{i+1}]$ on the basis of $\{w(\tau_{i+1}),\ G[\tau_{i+1}]\}$. The part of the second player-opponent that governs the disturbance in the given process formalization is to select the realization $G(\tau_{i+1})$ within the restrictions (1.10) and (1.11).

3. The significance of the model described by vectors $w(t)$ and $y(t)$ for the initial problems of approach and evasion under incomplete information is determined by the following features. Construct the sets $\mathfrak{M}_1 = \{\mathfrak{M}_1(t),\ t_0 \leq t \leq \vartheta\}$ and $\mathfrak{M}_2 = \{\mathfrak{M}_2(t),\ t_0 \leq t \leq \vartheta\}$, where

$$\mathfrak{M}_1(t) = \{x: G^*[t, x] \subset \mathfrak{M}\} \tag{3.1}$$

$$\mathfrak{M}_2(t) = \{x: G^*[t, x] \cap \mathfrak{M} = \varnothing\}. \tag{3.2}$$

Form the equation

$$\dot{w} = p, \quad p \in Y_\beta(t, w, s) \tag{3.3}$$

where p and s are the control actions for the first and the second player, the vector s attaining arbitrary values.

Consider the function σ^0 (2.5) to be continuous and Lipschitz in y. The following assertion based on rather moderate and natural assumptions on the properties of the varieties $\{G_*[t^*];\ y[t_*],\ \Pi[t_*, t^*)\}$, $\{G^*[t^*, y[t^*]];\ y[t_*],\ \Pi[t_*, t^*)\}$ is true.

Assume the problem of approach $w[\tau] \in \mathfrak{M}_1(\tau)$, $\tau \leq \vartheta$ for the first player in the closed-loop m i n m a x differential game with complete information [1] for system (3.3) is soluble by means of a certain c o u n t e r s t r a t e g y $U^{(s)} \div p(t, w, s)$. Then there exists a solution U for the game of approach 1.1.

Assume the problem of evasion $w[t] \in \mathfrak{M}_2(t)$, $t_0 \leq t \leq \vartheta$, for the first player in the closed-loop m i n m a x differential game with complete information for system (3.3) is soluble by means of a c o u n t e r s t r a t e g y $U^{(s)} \div p(t, w, s)$. Then there exists a solution U for the game of evasion 1.2.

The basic element for the construction of solution strategies of the closed-loop approach evasion games for system (3.3) is t h e s t a b l e b r i d g e W. Here it suffices to define it as follows. A closed set $W = \{W(t), t_0 \leq t \leq \vartheta\}$ in the space $\{t, w\}$ constitutes a s t a b l e b r i d g e W that ends at $\mathfrak{M}_1$ at time ϑ (or evades $\mathfrak{M}_2$ until time ϑ) if:

(1) the inclusion $w \in W(\vartheta)$ yields $w \in \mathfrak{M}_1(\vartheta)$ $(W(t) \cap \mathfrak{M}_2(t) = \varnothing,\ t_0 \leq \leq t \leq \vartheta)$; (2) for any selected values of $t_* \in [t_0, \vartheta]$, $t^* \in (t_*, \vartheta]$ of vector s_* and of the element $w_* \in W(t_*)$ it is possible to indicate a solution $w(t, t_*, w_*)$ of equation $\dot{w} \in Y(t, w, s_*)$ that would satisfy either condition $w(t^*; t_*, w_*) \in \in W(t^*)$ or $w(\tau; t_*, w_*) \in \mathfrak{M}_1(\tau)$ for any $\tau \in [t_*, t^*]$ (condition: $w(t^*, t_*, w_*) \in \in W(t^*)$). Condition (2) may be formulated otherwise: (2) whatever are $t_* \in \in [t_0, \vartheta)$, s_*, $w_* \in W(t_*)$, $w_* \notin \mathfrak{M}_1(t_*)$ and the value $\alpha > 0$ there exists a number $\delta > 0$ such that the distance from set $\{x: x \in w_* + Y_\beta(t_*, w_*, s_*)\ (t - t_*)\}$ to set $W(t)$ for $t \in [t_*, t_* + \delta]$ would not exceed the value $\alpha(t - t_*)$. With the existence of such a bridge W that would also satisfy the inclusion $y[t_0] \in \in W(t_0)$ the initial problem 1.1 (problem 1.2) is solved by means of an e x t r e m a l strategy U^0 to this bridge W. The strategy is constructed as follows. Assume the values $\{w[\tau_i],\ y[\tau_i]\}$ have realized where $w[\tau_i] \in W(\tau_i)$. The program $\Pi[\tau_i, \tau_{i+1}) = \Pi^0[\tau_i, \tau_{i+1}]$ is then selected as the one that minimizes the value $\sigma(\tau_{i+1}, \tau_i, y[\tau_i], s[\tau_i], \Pi)$ (2.3) with $s[\tau_i] = y[\tau_i] - w[\tau_i]$. The value

$w(\tau_{i+1})$ is chosen as the solution of Eq. (3.3) that satisfies condition (2) in the definition of the stable bridge W (again for $s = s[\tau_i]$). The vector $y[\tau_{i+1}]$ is selected by minimizing the value $\| y[\tau_{i+1}] - w[\tau_{i+1}] \|$ under the inclusion

$$G[\tau_{i+1}] \subset G^*\big[\tau_{i+1}, y[\tau_{i+1}]\big] \tag{3.4}$$

with domains $G^*[\tau_{i+1}, y[\tau^i{}_{+1}]]$ varying among the variety $\{G^*[\tau_{i+1}, y[\tau_{i+1}]]; y[\tau_i], \Pi^0[\tau_i, \tau_{i+1})\}$. The initial vectors $y[t_0] = y[\tau_0]$ and $w[t_0] = w[\tau_0]$ are selected by minimizing the values $\| y[\tau_0] - w[\tau_0] \|$ under the restrictions

$$G[\tau_0] \subset G^*\big[\tau_0, y[\tau_0]\big] \tag{3.5}$$

and

$$w[\tau_0] \in W(\tau_0) \tag{3.6}$$

where the domains $G^*[\tau_0, y[\tau_0]]$ are varying over the initial variety $\{G^*[t_0, y[t_0]]\}$. (Conditions (3.4) and (3.5) which are now introduced for $y[t]$ are somewhat wider than the relations (2.1) and (2.2) that restricted $y[t]$ earlier. Precisely these restrictions are taken for selecting the values $y[t]$ in organizing the actual control process.) The strategy U^0 holds the motion $w[t]$ within $W(t)$ until the encounter with $\mathfrak{M}_1(t)$ (or until the moment ϑ) while keeping $y[t]$ and $w[t]$ close enough to each other (provided step $\delta(\Delta)$ is sufficiently small). The solution of problem 1.1 (problem 1.2) is thus ensured.

The control procedure described ensures the approach to $\mathfrak{M}$ until a certain moment of time ϑ selected in advance (the evasion from $\mathfrak{M}$ at $t_0 \leq \leq t \leq \vartheta$). However if apart from the approach or evasion it is desired to attain a best possible process quality evaluated by a certain cost criterion, it may then prove to be useful to correct the control procedure at the moment $t = \tau_{i+1}$ taking into consideration the given cost and passing at these moments to a new bridge W related to a better cost criterion. Consider e.g. the problem of ensuring an encounter in minimal time (the problem of evasion at maximal time). Denote with $\{W(t; \vartheta), \tau \leq t \leq \vartheta\}$ the stable bridge that ends at $\mathfrak{M}_1$ at moment ϑ (that evades $\mathfrak{M}_2$ until time ϑ). Suppose at initial moment $y[\tau_0] = w[\tau_0] \in \in W(\tau_0; \vartheta_0)$ where ϑ_0 is the least possible (the largest possible) value of ϑ that allows the inclusion $y[\tau_0] = w[\tau_0] \in W(\tau_0; \vartheta)$. Assume at time $t = \tau_i$ the values $y[\tau_i]$ and $w(\tau_i) \in W(\tau_i; \vartheta[\tau_i])$ have realized. Then after the choice of program $\Pi^0[\tau_i, \tau_{i+1})$ via procedure of above and after the realization of domain $G[\tau_{i+1}]$ at time $t = \tau_{i+1}$ one at first is to select a value $w^*[\tau_{i+1}]$ due to stability of bridge $\{W(t; \vartheta[\tau_i], \tau_i \leq t \leq \vartheta[\tau_i]\}$ from condition $w^*[\tau_{i+1}] \in W(\tau_{i+1}, \vartheta[\tau_i])$ or $w^*[\tau] \in \mathfrak{M}_1(\tau)$, $\tau \in [\tau_i, \tau_{i+1}]$ $(w^*[\tau_{i+1}] \in W(\tau_{i+1}, \vartheta[\tau_i]))$. Then one is to select the vector $y[\tau_{i+1}]$ for example from condition of minimum (or maximum)

for the value $\vartheta[\tau_{i+1}] \leq \vartheta[\tau_i]$ ($\vartheta[\tau_{i+1}] \geq \vartheta[\tau_i]$) under the restriction $\varrho(y[\tau_{i+1}], W(\tau_{i+1}, \vartheta[\tau_{i+1}])) \leq \gamma_i$ where $\gamma_i = \min_y \| y - w^*[\tau_{i+1}] \|$ under condition $G[\tau_{i+1}] \subset G^*[\tau_{i+1}, y]$. Here the symbol $\varrho(y, W)$ stands for the Euclidean distance of $y[\tau_{i+1}]$ to the section $W(\tau_{i+1}; \vartheta[\tau_{i+1}])$. The next step is to chose the value $w[\tau_{i+1}] \in W(\tau_{i+1}; \vartheta[\tau_{i+1}])$, $\| y[\tau_{i+1}] - w[\tau_{i+1}] \| = \gamma_i$. The described method of control ensures for any preassigned $\varepsilon > 0$ (for a certain $\varepsilon > 0$) an ε-approach by time $\vartheta^* = \min_i \vartheta[\tau_i] \leq \vartheta_0$ (an ε-evasion until time $\vartheta_* = \max_i \vartheta[\tau_i] - \varepsilon$) provided the step $\delta(\Delta)$ of subdivision $\Delta\{\tau_i\}$ would be sufficiently small.

4. Hence for the solution of problem (1.1) (problem 1.2) it suffices to know the ways of constructing the appropriate stable bridge. An effective procedure in this field for system (3.3) is given by the method of e x t r e m a l a i m i n g or the method of a p r i o r i s t a b l e i n t e g r a l m a n i f o l d s. The problems of this investigation being reduced to those of closed-loop differential games with complete information for system (3.3) are now possible to utilize these methods in an investigated form as given in book [1] (chapters VI–VIII). We shall therefore stop for details concerning only the linear equation (1.1)

$$\dot{x} = A(t)\, x + f(t, u, v) \tag{4.1}$$

when the given methods may be developed to the level of realizable algorithms with less effort. We start from e x t r e m a l a i m i n g method, considering sets $\mathfrak{M}_1(t)$ to be convex and compact. Here and in the following section we shall also select the domains G_* and G^* of conditions (1.11) and (2.1) to have the form $G^*[t, y] = y + G^*(t)$, $G_*[t] = x + G_*(t)$ where $\{G^*(t), t_0 \leq t \leq \vartheta\}$ and $\{G_*(t), t_0 \leq t \leq \vartheta\}$ are preassigned single-parametric varieties of closed and bounded domains. In the case of (4.1) equation (2.6) takes the form

$$\dot{w} \in A(t)\, w + Y_\beta(t, s) \tag{4.2}$$

where

$$Y_\beta(t, s) = \{x: s'x \geq \sigma^0(t, s),\ \| A(t)\, w + x \| \leq 2\beta\}.$$

The value $\sigma^0(t, s)$ is determined according to (2.3)–(2.5) by the equality

$$\sigma^0(t, s) = \min_{\Pi} \sigma(t, s, \Pi) \tag{4.3}$$

where

$$\sigma(t_*, s, \Pi) = \limsup_{t^* \to t_*} \frac{\sigma(t^*, t_*, s, \Pi)}{t^* - t_*} \tag{4.4}$$

and

$$\sigma(t^*, t_*, s, \Pi) = \max_{G_*[t^*]} \min_{y[t^*]} s'(y[t^*] - y[t_*]) - s'X(t^*, t_*)\, y[t_*] \tag{4.5}$$

where $X(t, t_*)$ is the fundamental matrix of solutions for the equation $\dot{x} = A(t)x$. In analogy with [1] (§ 41–§ 43) we take for Eq. (4.2) the values

$$\varepsilon(t, w, \tau) = \max_{||l||=1} \left[s'(t, \tau)\, w + \int_t^\tau \sigma^0\big(\zeta, s(\zeta, \tau)\big)\, d\zeta + \varrho_1(\tau, l) \right] \tag{4.6}$$

where the right-hand side (4.3) is nonnegative, $\varepsilon(t, w, \tau) = 0$ in other case and

$$\varepsilon_0(t, w) = \min_\tau \varepsilon(t, w, \tau) \qquad (t \leq \tau \leq \vartheta) \tag{4.7}$$

where $s(t, \tau)$ is the solution for the adjoint equation $\dot{s} = -A'(t)s$, $s(\tau, \tau) = l$ and $\varrho(\tau, l) = \min l'x$ with $(-x) \in \mathfrak{M}_1(\tau)$. A proof similar to [1] leads us to the following proposition.

Suppose $\mathfrak{L}(t, \tau, w)$ is the set of vectors $l^0 = s^0(\tau, \tau)$ for which a maximum (4.3) is attained. Assume the condition is fulfilled that for any choice of s and any pair $\{t, w\}$ that yield $\varepsilon_0(t, w) > 0$ it is possible to indicate at least one minimizing element τ_0 (4.7) such that the intersection of subspace $Y(t, s) = \{x: s'x \geq \sigma^0(t, s)\}$ with the intersection of all supspaces $X(t, l) = \{x: l'x \geq \sigma^0(t, l)\}$ $(l = l^0 \in \mathfrak{L}(t, \tau, w))$ is nonempty. Then the set $W = \{\{t, w\}: \varepsilon_0(t, w) = 0, t_0 \leq t \leq \vartheta\}$ forms a stable bridge that ends in $\mathfrak{M}_1$ at moment ϑ.

The given condition holds in particular if with $\varepsilon(t, w, \tau) > 0$ the maximum of (4.6) is attained at a unique vector $l = s^0(\tau, \tau)$.

5. We shall pass to discussion of the method of a priori stable integral manifolds and its application to Eq. (4.2). The following assertion is true.

Suppose $F(t)$ $(t_0 \leq t < \vartheta)$ is a variety of sets semicontinuous in t and such that for any choice of vector s and moment of time t the condition

$$F(t) \cap Y(t, s) \neq \emptyset \tag{5.1}$$

is fulfilled. Then the bundle $\{w(t; t_0, w_0)\}$ of all solutions $w(t, t_0, w_0)$ of the equation

$$\dot{w} \in A(t)\, w + F(t) \tag{5.2}$$

forms a stable bridge

$$W = \big\{\{t, w\}: w = w(t, t_0, w_0), t_0 \leq t \leq \vartheta\big\}. \tag{5.3}$$

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

If under the given condition the inclusion $w(\tau, t_0, w_0) \in \mathfrak{M}_1(\tau)$ taken for all $w(\tau, t_0, w_0)$ holds for at least one $\tau \in (t_0, \vartheta)$ then W will be a stable bridge that ends at $\mathfrak{M}_1$ by time ϑ. If for all $w(\tau, t_0, w_0)$ we have $w(t, t_0, w_0) \notin \mathfrak{M}_2(t)$ for all $t \in (t_0, \vartheta)$ then W will be a stable bridge that evades $\mathfrak{M}_2$ with $t_0 \leq t \leq \vartheta$.

Particularly condition (5.1) is bound to be fulfilled if for each $t \in [t_0, \vartheta]$ provided the intersection $Y(t) = \bigcap_s Y(t, s)$ be nonempty. Then in place of the bundle $\{w(t, t_0, w_0)$ we may select a certain single solution $w(t, t_0, w_0)$ of the equation

$$\dot{w} = A(t)\, w + Y(t) \tag{5.4}$$

for which the relation (5.2) attains the form of $\dot{w} = A(t)\, x + f(t)$ and which then constitutes a stable bridge W (5.3) in the form of a s t a b l e p a t h [1]. Under certain conditions the given construct gives a best solution in some sense. Namely, suppose the function $\sigma^0(t, s)$ (4.3) is concave in s. Consider the problem of selecting a minimal value of $\vartheta = \vartheta^0$ that would ensure the solubility of the problem of approach. According to the previous discussion it suffices for $Y(t)$ to be nonempty for each t that problem 1.1 is solved by the procedure of above at least for each $\vartheta = \vartheta^*$ where ϑ^* is the least value when at least one solution $w(t, t_0, w_0)$ of Eq. (5.4) satisfies the inclusion $w(\vartheta, t_0, w_0) \in \mathfrak{M}_1(\vartheta)$.

If now the first player will attempt to solve problem (1.1) on the basis of the estimating domains $G^*[\tau_i] + y[\tau_i]$ and the value $y[\tau_i]$ of (2.1) then it will turn out that there would exist no such procedures at all that would ensure the approach at time $\vartheta < \vartheta^*$. On the contrary, consider the problem of selecting the largest values $\vartheta = \vartheta_0$ that would ensure the solubility of problem (1.2) on the evasion from a certain fixed $\varkappa$-neighbourhood $\mathfrak{M}_2^{(\varkappa)} = \{\mathfrak{M}_2^{(\varkappa)}(t), t_0 \leq t \leq \vartheta\}$ of set $\mathfrak{M}_2$. According to the assertions of above it suffices for $Y(t)$ $(t_0 \leq t)$ to be nonempty that problem 1.2 would be solved by the given procedure for at least one value $\vartheta = \vartheta_*$ where ϑ_* is the largest of the values that yield the relation $w(t, t_0, w_0) \notin \mathfrak{M}_2^{(\varkappa)}(t)$ at $t_0 \leq t \leq \vartheta$ for at least one solution $w(t, t_0, w_0)$ of Eq. (5.4). It turns out that in solving the evasion problem 1.2 for $\mathfrak{M}_2^{(\varkappa)}$ on base of the domains $G^*[\tau_i] + y[\tau_i]$ and the values $y[\tau_i]$ (2.1) there exist no such procedure for the first player that would ensure the evasion at $t_0 \leq t \leq \vartheta$ where $\vartheta > \vartheta_*$. The proof of the assertions on the negative properties of ϑ^* and ϑ_* under concavity of function $\sigma^0(t, s)$ in s are proved in a way similar to [1] (pp. 207–226).

6. The previous sections give an idealized solution of problems 1.1 and 1.2 when the control is performed by program $\Pi[\tau_i, \tau_{i+1}]$. Actually it is control $u = u[t]$ that should be applied to the system (1.1). In Cases I, III and IV the

function $u[t]$ is determined in the semi-interval $[\tau_i, \tau_{i+1}]$ by condition (1) $u[t] = u_э(t)$ or (III, IV) $u[t] = u_э(t, v[t])$ where $u_э[t]$ or $u_э(t, v)$ are the functions that generate the elementary program $\Pi_э(\tau_i, \tau_{i+1})$ approximating $\Pi_э(\tau_i, \tau_{i+1})$. The result of the actual control action will yield a close approximation of the ideal solution if step $\delta(\Delta)$ is sufficiently small and the approximation of programs $\Pi[\tau_i, \tau_{i+1})$ by those of type $\Pi_э[\tau_i, \tau_{i+1}]$ is sufficiently accurate. In Case II the control $u[t]$ is formed as a piecewise constant function $u[t] = u_k^i$ $(\tau_k^i \leq t < \tau_{k+1}^i, \tau_0^i = \tau_i, \tau_{k(i)}^i = \tau_{i+1})$ where u_k^i are the results of probabilistic tests with distributions $\mu(du/\tau_k^i)$ where $\mu_i[du) = \mu[du/\tau_k^i)$ $(\tau_k^i \leq t < \tau_{k+1}^i)$ is the function measure that corresponds to the program $\Pi_э[\tau_i, \tau_{i+1})$ approximating $\Pi(\tau_i, \tau_{i+1})$. The result of the actual control will be arbitrary close to the ideal solution with probability arbitrary close to unity provided step $\delta(\Delta)$ is sufficiently small and the approximation of $\Pi[\tau_i, \tau_{i+1})$ by programs $\Pi_э[\tau_i, \tau_{i+1})$ is sufficiently accurate. In analogy with [1] (pp. 334–352) in Case II it is also assumed that within the minor time intervals $[\tau_k^i, \tau_{k+1}^i)$ the controls $u[t]$ and $v[t]$ are either stochastically independent or more or less weakly correlated.

References

1. *Krasovskii, N. N., Subbotin, A. I.*, Closed-loop Differential Games. Nauka, Moscow (1974).
2. *Krasovskii, N. N., Osipov, Yu. S.*, The Problem of Control under Incomplete Information. Izvestiya of the Acad. of Sciences of the USSR, Mechanics of a Rigid Body, **4** (1973).

Игровое управление при неполной фазовой информации

Н. Н. КРАСОВСКИЙ

(Свердловск)

В статье рассматривается задача об управлении в условиях конфликтующей или неопределенной помехи, осложненная неполнотой информации о реализующихся фазовых состояниях системы. Исходная проблема формализуется как задача сближения или задача уклонения в подходящей дифференциальной игре. При этом роль позиционного элемента, определяющего основу управления по принципу обратной связи, исполняет реализация информационной области, содержащая неизвестную более точно фазовую точку объекта. Предлагаются эффективные способы решения, которые могут быть реализованы с помощью ЭВМ.

Н. Н. Красовский
Институт математики и механики УНЦ АН СССР
СССР, Свердловск, К-66,
ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 5 (4), pp. 303—320 (1976)

A SPECIAL METHOD OF FAST FOURIER TRANSFORM TO DETERMINE THE LOW AND MIDDLE FREQUENCY PORTION OF THE FREQUENCY CHARACTERISTIC OF A CONTROL SYSTEM

A. FRIGYES

(Budapest)

(Received October 22, 1975)

There are numerous methods in control system analysis and synthesis where the information about the system dynamics is described by a time function, mostly by the weight function or the step response.

Various identification methods, deterministic as well as stochastic, yield the step response of the system.

Although there are some methods which carry out the synthesis directly in the time domain, the majority of the methods are based on the knowledge of the poles and zeros of the system transfer functions, or on their frequency characteristics. In the latter case the Fourier transform of the weight function is to be determined.

Various algorithms have been developed to evaluate numerically the Fourier transform of a time function. Some of them are called fast Fourier transform.

The characteristic point of these methods is the selection and grouping of the multiplications in a manner allowing the Fourier integral to be computed by a minimum number of operations. However, the use of these methods generally does not presume any constraints in respect of the time function to be transformed.

The method presented in this paper is restricted to time functions which occur most frequently as step responses (weight functions) of process control systems. In addition, the accuracy of the Fourier transform is required only in the frequency range decisive for the behaviour of the closed loop control. The simplicity and fastness of the method is due to these limitations.

The formulas derived by this method produce the real and imaginary part of the frequency characteristics of the system by adding together some samples of the step response $v(t)$, of its integral $V(t) = \int_0^t v(t)\,\mathrm{d}t$, and of its first and higher order derivatives $w(t) = \frac{\mathrm{d}v(t)}{\mathrm{d}t}$ $w'(t)$, etc.

1. General remarks

Let $w(t)$ be the weight function of a linear time-independent single-input single-output stable system. In this case the transfer function

$$W(s) = \int_0^\infty w(t)\, e^{-st}\,\mathrm{d}t \tag{1}$$

has only poles lying on the left side of the complex plane, and so there exists the Fourier transform of $w(t)$:

$$W(j\omega) = \int_0^\infty w(t)\, e^{-j\omega t}\, \mathrm{d}t\,. \tag{2}$$

If $w(s)$ has poles in the origin ($s = 0$) then Eq. (2) does not converge. Nevertheless we can apply the Fourier transform in this case, too. Namely, if the pole in the origin is of mth order (this means that the system contains an integrating element of mth order) then the Fourier transform of the mth derivative of the weight function exists, and is

$$(j\omega)^m\, W(j\omega) = \int_0^\infty w^{(m)}(t)\, e^{-j\omega t}\, \mathrm{d}t \tag{3}$$

where

$$w^{(m)}(t) = \frac{\mathrm{d}^m w(t)}{\mathrm{d}t^m}\,.$$

The method of fast Fourier transform to be demonstrated in this paper is related to those systems which have only poles in the left half plane and in the origin, so Eq. (3) may be applied.

Let $R(\omega)$ be the real and $Q(\omega)$ the imaginary part of the frequency function:

$$W(j\omega) = R(\omega) + jQ(\omega).$$

Applying the Euler relation, we get:

$$(j\omega)^m\, R(\omega) + j(j\omega)^m\, Q(\omega) = \int_0^\infty w^{(m)}(t)\, (\cos \omega t - j \sin \omega t)\, \mathrm{d}t\,. \tag{4}$$

Depending on the value of m, we may express $R(\omega)$ and $Q(\omega)$. If $m = 0$,

$$R(\omega) = \int_0^\infty w(t) \cos \omega t\, \mathrm{d}t\,, \tag{5}$$

$$Q(\omega) = -\int_0^\infty w(t) \sin \omega t\, \mathrm{d}t\,. \tag{6}$$

If $m = 1$,

$$R(\omega) = -\frac{1}{\omega} \int_0^\infty w'(t) \sin \omega t\, \mathrm{d}t\,, \tag{7}$$

$$Q(\omega) = -\frac{1}{\omega} \int_0^\infty w'(t) \cos \omega t\, \mathrm{d}t\,. \tag{8}$$

If $m = 2$,

$$R(\omega) = -\frac{1}{\omega^2} \int_0^\infty w''(t) \cos \omega t \, \mathrm{d}t , \tag{9}$$

$$Q(\omega) = \frac{1}{\omega^2} \int_0^\infty w''(t) \sin \omega t \, \mathrm{d}t . \tag{10}$$

(It is to be noted that if $m = 0$ then the expressions related to $m = 0$ are also applicable; if $m = 1$ then the expressions related to $m = 2$ are applicable, but not those related to $m = 0$ etc.)

Thus, to determine the frequency function from the weight function, it is necessary to evaluate the integral of those products which have the form $f(t)$ cos ωt or $f(t)$ sin ωt.

For a numerical evaluation it is necessary

(1) to substitute the integrand by partly analytical functions,
(2) to properly select the finite upper limit of the integration,
(3) by using correcting terms, to diminish the error caused by the finite upper limit of the integration.

The essence of our method is that not the time function $f(t)$, but the sin or cos functions are approximated by properly chosen straight lines, and the upper limit of the integration is extended to some integer numbers of the actual period.

2. The integral of the product of a continuous function $f(t)$ and a partly linear function $g(t)$

Let $g(t)$ be a partly continuous linear function in the interval $t_0 \leq t \leq t_n$, the breakpoints of which are at the points $t_1, t_2, \ldots$ (Fig. 1), and

$$g(t) = 0 \quad \text{if} \quad t \leq t_0 \text{ and } t \geq t_n . \tag{11}$$

Consider $f(t)$ as the second derivative of a function $h(t)$

$$f(t) = \frac{\mathrm{d}^2 h(t)}{\mathrm{d}t^2} . \tag{12}$$

One has to determine the integral

$$I = \int_{t_a}^{t_m} f(t) \, g(t) \, \mathrm{d}t . \tag{13}$$

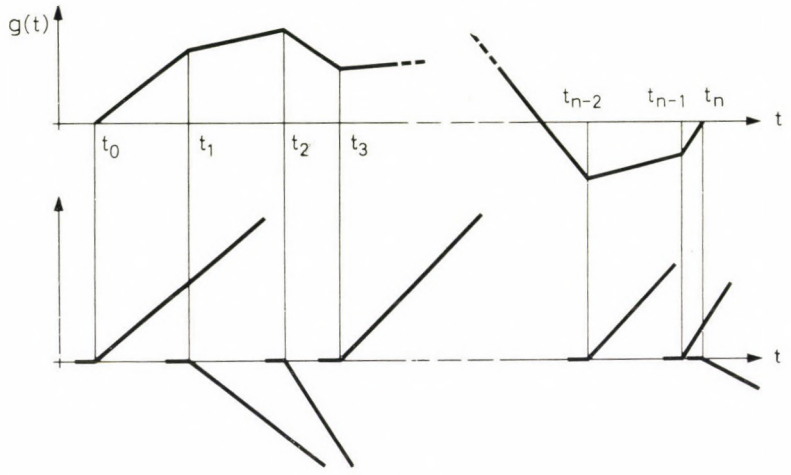


Fig. 1

According to Fig. 1 $g(t)$ can be considered as the sum of ramp functions entering at the breakpoints

$$t_0, t_1, \ldots, t_n .$$

Thus

$$g(t) = \sum_{i=0}^{n} \bar{1}(t - t_i)\, a_i(t - t_0) \tag{14}$$

where $\bar{1}\,(t-t_i)$ denotes the unit step response function entering at t_i and a_i is the change of the slope of $g(t)$ at the breakpoint t_i.

It is easy to prove that

$$\sum_{i=0}^{n} a_i = 0 , \tag{15}$$

$$\sum_{i=0}^{n} a_i t_i = 0 . \tag{16}$$

Let us namely apply Eq. (14) for an instant $t_p > t_n$, in which according to Eq. (11) $g(t) \equiv 0$.

$$0 \equiv g(t) = \sum_{i=0}^{n} a_i(t_p - t_i) = t_p \sum_{i=0}^{n} a_i - \sum_{i=0}^{n} a_i t_i . \tag{17}$$

(We may omit the notation of the unit step function $\bar{1}\,(t-t_i)$, because at the time t_p all steps have already entered.) Equation (17) can be satisfied for every $t_p > t_n$ only in the case if Eqs. (15) and (16) are true.

Let us now evaluate the integral I

$$I = \sum_{i=0}^{n} \int_{t_i}^{t_n} f(t)\, a_i(t - t_i)\, \mathrm{d}t\,. \tag{18}$$

According to Eq. (12) we substitute $f(t) = h''(t)$, and expand the ith member of the sum:

$$\int_{t_i}^{t_n} f(t)\, a_i(t - t_i)\, \mathrm{d}t = a_i \int_{t_i}^{t_n} h''(t)\, t\, \mathrm{d}t - a_i t_i \int_{t_i}^{t_n} h''(t)\, \mathrm{d}t\,.$$

The first member of the above equation will be integrated by part, while its second one simply:

$$\int_{t_i}^{t_n} f(t)\, a_i(t - t_i)\, \mathrm{d}t = a_i [h'(t) \cdot t]_{t_i}^{t_n} - a_i \int_{t_i}^{t_n} h'(t)\, \mathrm{d}t - [a_i t_i h'(t)]_{t_i}^{t_n} =$$

$$= a_i t_n h'(t_n) - a_i t_i h'(t) - a_i h(t_n) + a_i h(t_0) - a_i t_i h'(t_n) + a_i t_i h'(t_i) =$$

$$= a_i [t_n h'(t_n) - h(t_n)] - a_i t_i h'(t_n) + a_i h(t_i)\,.$$

Writing this expression in Eq. (18) and taking Eqs. (15) and (16) into consideration, we get

$$I = \sum_{i=0}^{n} a_i h(t_i)\,. \tag{19}$$

It is remarkable that all what we need to know to determine I are the changes of the slopes of $g(t)$ at the breakpoints, and the values of the second integral of $f(t)$ at the same points.

3. Approximation of the sin and cos function by straight lines

In order to approximate the sind and cos function in Eqs. (5)–(10) by straight lines we divide the period in eight equidistant intervals according to Fig. 2. Thus both the sin ωt and cos ωt functions are approximated by straight lines having a slope of 0 and $\pm\, c\omega$.

It is convenient to express the slope as a product of a constant c and then actual angular frequency ω, because the greatest slope of the periodic function (occurring at crossing the time axis) is related exactly to $c = 1$.

We shall approximate the periodical functions over a whole number of periods. This means that the approximating Fourier integral will be extended over a whole multiple of the actual period time, i.e. $k \dfrac{2\pi}{\omega}$.

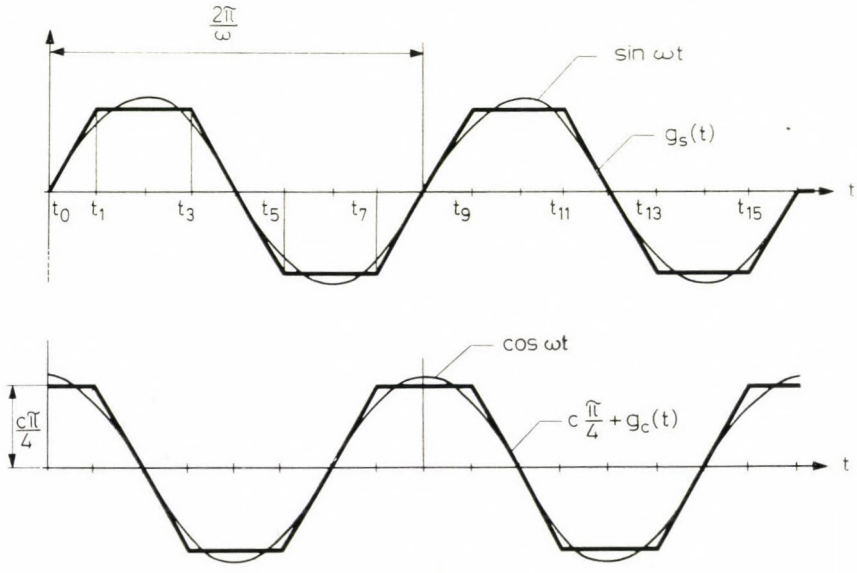


Fig. 2

Thus the approximation of the sin ωt function is

$$\sin \omega t \approx g_s(t) = \sum_{i=0}^{8k} \bar{1}(t - t_i)\, a_i\, t$$

where for even values of i — except for $i = 0$ and $i = 8k$ —

$$a_i = 0$$

$$a_0 = c\omega$$

$$a_{8k} = -c\omega.$$

If the value of i is an odd number then for

$$i = 8p + 1 \text{ and } i = 8p + 3 \quad (p = 0, 1, \ldots k-1)$$

$$a_i = -c\omega$$

and for

$$i = 8p + 5 \text{ and } i = 8p + 7$$

$$a_i = +\, c\omega.$$

It is to be noted that $g_s(t)$ fulfils the condition given in Eq. (11).

On the other hand the approximation of $\cos \omega t$ will be expressed as follows:

$$\cos \omega t \approx \frac{c\pi}{4} + g_c(t) = \frac{c\pi}{4} + \sum_{i=0}^{8k} \bar{1}(t - t_i)\, a_i\, t\,.$$

If the value of i is an even number then

$$a_i = 0.$$

If it is odd then

$$\text{for } i = 8p + 1 \text{ and } i = 8p + 7 \qquad (p = 0, 1, \ldots k-1)$$

$$a\text{i} = -c\omega$$

and for

$$i = 8p + 3 \text{ and } i = 8p + 5$$

$$a_i = + c\omega.$$

The value of c is determined so that the first harmonic of the Fourier series of the trapezoidal function should be equal to the sin (or cos) function approximated by it. This condition leads to a value

$$c = 1.13.$$

This value has been verified by tests made with various structures.

4. The approximating expressions of the Fourier transform

4.1. The method of the derivation of the expressions

The following method is used to derive the approximating expressions:

Starting from Eq. (4), by selecting the value of m we determine the expressions of $R(\omega)$ and $Q(\omega)$ according to Eqs. (5)–(10). The upper limit of the integrals will be $kT = k\, 2\pi/\omega$ instead of ∞. (Thus the upper limit depends on the actual frequency; however, this causes no difficulty, because ω is a constant with respect to the integral.) The sin and cos functions will be replaced by their trapezoidal approximations, according to Fig. 2. The trapezoidal function is extended, of course, also over k periods of the actual frequency. Both trapezoidal functions, $g_s(t)$ and $g_c(t)$, satisfy the condition given by Eq. (11).

Then Eq. (19) is applied to evaluate the integral corresponding to Eq. (13). The approximating transform obtained in this way will differ from the exact value of the Fourier transform. The reason of the difference is the straight line approximation of the sin and cos function on the one hand, and the finite upper limit of the integration on the other hand. We denote the real and

imaginary part of the approximating transform by $R_k(\omega)$ and $Q_k(\omega)$ respectively, where the index k relates to the fact that the integration is extended over k periods.

The approximating expression depends also on the choice of m in Eq. (4). Namely, Eqs. (5)–(10) are equivalent only in the case when no approximation is used. Therefore an upper index m will denote that the approximating expression is derived from a formula containing the mth derivative of the weight function. Accordingly, for instance, ${}^1R_2(\omega)$ denotes the approximation of the real part of the transform, if it is derived from Eq. (7) in which the integral is extended for two periods.

4.2 Expressions derived without intermediate integrating by part

Starting from Eqs. (5) and (6), we get the expressions $R_k(\omega)$ and $Q_k(\omega)$ The notations used in Eqs. (12) and (13) are now

$$f(t) = w(t)\,,$$

$$f(t) = \int_0^t \int_0^t w(t)\,\mathrm{d}t\,\mathrm{d}t = V(t)\,.$$

Thus $V(t)$ is the response of the system to a unit rump input.

After proper substitutions:

$${}^0R_k(\omega) = \int_0^{k\frac{2\pi}{\omega}} \left[w(t)\left(g_c(t) + c\pi/4\right)\right]\mathrm{d}t\,, \tag{21}$$

$${}^0Q_k(\omega) = -\int_0^{k\frac{2\pi}{\omega}} w(t)\,g_s(t)\,\mathrm{d}t. \tag{22}$$

Considering the solution given in Eq. (19) of the integrals of the type according to Eq. (13), and having the previously given values of a_i,

$${}^0R_k(\omega) = c\omega \sum_{p=0}^{k-1} (-V_{8p+1} + V_{8p+3} + V_{8p+5} + V_{8p+7}) + \frac{c\pi}{4} v_{8k}\,, \tag{23}$$

$${}^0Q_k(\omega) = c\omega \sum_{p=0}^{k-1} (V_{8p+1} + V_{8p+3} - V_{8p+5} - V_{8p-7}) + c\omega V_{8k}\,. \tag{24}$$

Because there is no danger of misunderstading in the above equations, the shorter notation V_i and v_i are used instead of $V(t_i)$ and $v(t_i)$, respectively.

Equations (7) and (8) are needed for the approximating formulas ${}^1R_k(\omega)$ and ${}^1Q_k(\omega)$,

$$ {}^1R_k(\omega) = -\frac{1}{\omega} \int_0^{k\frac{2\pi}{\omega}} w'(t)\, g_s(t)\, \mathrm{d}t\,, \tag{25} $$

$$ {}^1Q_k(\omega) = -\frac{1}{\omega} \int_0^{k\frac{2\pi}{\omega}} w'(t) \left[g_c(t) + \frac{c\pi}{4} \right] \mathrm{d}t\,. \tag{26} $$

The time functions used in Eqs. (12) and (13) are now

$$ f(t) = w'(t) \text{ and } h(t) = v(t)\,. $$

Similarly to the procedure applied previously

$$ {}^1R_k(\omega) = c \left[\sum_{p=0}^{k-1} (v_{8p+1} + v_{8p+3} - v_{8p+5} - v_{8p+7}) + v_{8k} \right], \tag{27} $$

$$ {}^1R_k(\omega) = c \sum_{p=0}^{k-1} (v_{8p+1} - v_{8p+3} - v_{8p+5} + v_{8p+7}) - \frac{c\pi}{4\omega} w_{8k}\,. \tag{28} $$

Remember that ${}^0R_k(\omega)$ and ${}^0Q_k(\omega)$ are expressed mainly by some instantaneous values of $V(t)$ (the integral of the step response of the system), and ${}^1R_k(\omega)$ and ${}^1Q_k(\omega)$ are expressed by the step responses themselves (except for one weight function value). Both expressions contain only those instantaneous values of the time function which are indexed by odd numbers (except for the last one). Of course, the distance of the discrete values depends on the actual frequency (Fig. 3).

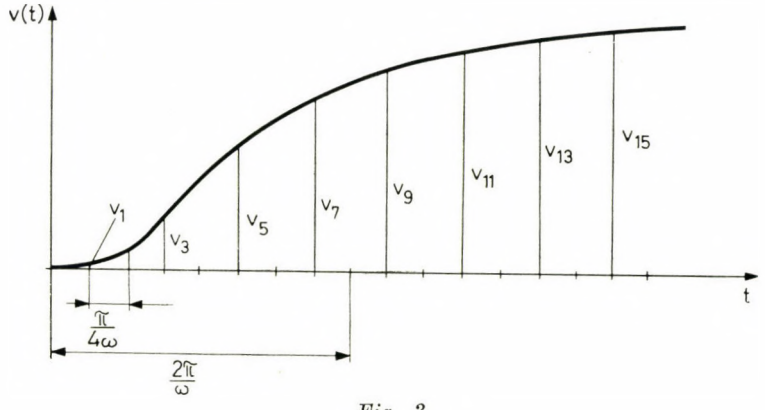


Fig. 3

It is possible to derive further expressions from Eqs. (9) and (10) containing the second derivative of the weight function. In this case the approximating transform will be expressed by the instantaneous values of the weight function. These expressions have less practical significance because the values of the weight function can be determined with considerable errors only.

4.3 Expressions derived by intermediate integrating by part

A somewhat different way to derive the approximating expressions is as follows:

In Eqs. (5)–(10) a finite upper limit of the integral will be introduced. Then, before approximating the sin and cos functions by straight lines, first an integration by part will be carried out in the way to diminish the order of the derivative of the weight function. Only after this procedure will the sin and cos function be approximated by trapezoidal function.

So let us start from Eqs. (7) and (8) containing the first derivative of the weight function. Introducing a finite upper limit of the integral, and then integrating by part, we get

$$-\frac{1}{\omega}\int_0^{k\frac{2\pi}{\omega}} w'(t)\sin\omega t\,\mathrm{d}t = -\frac{1}{\omega}\left[w(t)\sin\omega t\right]_0^{k\frac{2\pi}{\omega}} + \int_0^{k\frac{2\pi}{\omega}} w(t)\cos\omega t\,\mathrm{d}t, \tag{29}$$

$$-\frac{1}{\omega}\int_0^{k\frac{2\pi}{\omega}} w'(t)\cos\omega t\,\mathrm{d}t = -\frac{1}{\omega}\left[w(t)\cos\omega t\right]_0^{k\frac{2\pi}{\omega}} - \int_0^{k\frac{2\pi}{\omega}} w(t)\sin\omega t\,\mathrm{d}t\,. \tag{30}$$

In Eqs. (29) and (30) the value of the expressions in brackets in zero on the lower limit, and $\sin\omega k\,2\pi/\omega = 0$ and $\cos\omega k\,2\pi/\omega = 1$.

And now the sin and cos functions will be approximated by straight lines

$${}^1R_k^1(\omega) = \int_0^{k\frac{2\pi}{\omega}} w(t)\,[g_c(t) + c\pi/4]\,\mathrm{d}t\,, \tag{31}$$

$${}^1Q_k^1(\omega) = -\frac{1}{\omega}w\left(k\frac{2\pi}{\omega}\right) - \int_0^{k\frac{2\pi}{\omega}} w(t)\,g_s(t)\,\mathrm{d}t\,. \tag{32}$$

In the above expressions the right-hand upper index 1 indicates that *one* integration by part has been carried out during the derivation. The left-hand upper index 2 and the lower index have the same meaning as before.

A comparison of Eqs. (21) and (22) with Eqs. (31) and (32) leads to the relations

$$ {}^{1}R_k^1(\omega) = {}^{0}R_k(\omega)\,, \tag{33} $$

$$ {}^{1}Q_k^1(\omega) = -\frac{1}{\omega} w\left(k\frac{2\pi}{\omega}\right) + {}^{0}Q_k(\omega)\,. \tag{34} $$

But

$$ w\left(k\frac{2\pi}{\omega}\right) = w(kT) = w_{8k}\,. $$

This is the value of the weight function at the end point of the integration.

Substituting the expressions of Eqs. (23) and (24) in Eqs. (33) and (34), we get

$$ {}^{1}R_k^1(\omega) = c\omega \sum_{p=0}^{k-1} (-V_{8p+1} + V_{8p+3} + V_{8p+5} - V_{8p+7}) + \frac{c\pi}{4} v_{8k}\,, \tag{35} $$

$$ {}^{1}Q_k^1(\omega) = c\omega\left[\sum_{p=0}^{k-1} (V_{8p+1} + V_{8p+3} - V_{8p+5} - V_{8p+7}) + V_{8k}\right] - \frac{w_{8k}}{\omega}\,. \tag{36} $$

The same procedure will be used to get the expressions of ${}^{2}R_k^1(\omega)$ and ${}^{2}Q_k^1(\omega)$. In this case we have to start from Eqs. (9) amd (10), and then

$$ {}^{2}R_k^1(\omega) = -\frac{1}{\omega^2} w'\left(k\frac{2\pi}{\omega}\right) + {}^{1}R_k(\omega)\,, \tag{37} $$

$$ {}^{2}Q_k^1(\omega) = {}^{1}Q_k(\omega)\,. \tag{38} $$

Substituting the expressions of Eqs. (27) and (28) in Eqs. (37) and (38) respectively:

$$ {}^{2}R_k^1(\omega) = c\left[\sum_{p=0}^{k-1} (v_{8p+1} + v_{8p+3} - v_{8p+5} - v_{8p+7}) + v_{8k}\right] - \frac{w'_{8k}}{\omega^2}\,, \tag{39} $$

$$ {}^{2}Q_k^1(\omega) = c\sum_{p=0}^{k-1} (v_{8p+1} - v_{8p+3} - v_{8p+5} + v_{8p+7}) - \frac{c\pi}{4\omega} w_{8k}\,. \tag{40} $$

It is to be noted that by the intermediate integration by part not more than one member is added to the expressions derived without intermediate integration by part.

If more intermediate integrations by part are carried out then more terms will be added. The general equation will not be derived; however, in the next chapter the main properties of the different approximating formulas will be discussed.

5. The structure of the approximating Fourier transforms

The main part of the approximating Fourier transforms consists of some samples of the step response of the system ($v(t)$) or its integral $\int_0^t v(t)\,\mathrm{d}t = V(t)$. The actual period is to be divided in eight equidistant sampling points according to Fig. 3. In the main part only those samples of $v(t)$ or $V(t)$ occur which are indexed by odd numbers. The signs of these samples are repeated in four-term cycles, each of which is related to one period of the actual frequency. There are only two possible sequences of signs in one cycle:

$$\text{sequence } C: \quad + \; - \; - \; +$$

$$\text{and sequence } S: \quad + \; + \; - \; - \; .$$

If the main part of the approximating Fourier transform is expressed by the samples of $V(t)$ then sequence C occurs in the real component of the transform and sequence S in its imaginary component.

If the transform is expressed by the samples of $v(t)$ then sequence S occurs in the real and sequence C in the imaginary component.

The main part of the approximating transform is influenced neither by the number of the intermediate integrations by part nor by the choice of the equation pairs (5)–(10) from which the formulas are derived.

The samples of $V(t)$ are multiplied by $c\omega$, while those of $v(t)$ are by c. The approximating formulas have a complementary part, too, which is a series consisting of the value of functions $V(t)$, $v(t)$, $w(t)$ and their derivatives at the upper limit of the integration, i.e. at the $8k$th sampling point.

The expressions of the approximating transforms are given as follows:

main part ⟵	⟶ complementary part	
$R_{[0,k]}=c\omega(-V_1+V_3+V_5-V_7-V_9+-\ldots-V_{8k-1})$	$+\frac{c\pi}{4}v_{8k}-\frac{w_{8k}}{\omega^2}+\ldots$	(41)
$Q_{[0,k]}=c\omega(V_1+V_3-V-V+V+-\ldots\quad -V_{8k-1})$	$+c\omega V_{8k}-\frac{w_{8k}}{\omega}+\ldots$	(42)
$R_{[1,k]}=c(v_1+v_3-v-v+v+-\ldots\quad -v_{8k-1})$	$+\quad cv_{8k}-\frac{w'_{8k}}{\omega^2}+\ldots$	(43)
$Q_{[1,k]}=c(v_1-v_3-v_5+v_7+v_9+-\ldots\quad +v_{8k-1})$	$-\frac{c\pi}{4\omega}w_{8k}-\frac{w''_{8k}}{\omega^3}+\ldots$	(44)

The derivation of these equations is not presented in this paper, but their main parts and the first terms of the complementary parts can be recognized in Eqs. (23), (23), (27), (28), (35), (36), (39) and (40). The first index in square brackets is 0 or 1, indicating that the main part is expressed by the samples of $V(t)$ or $v(t)$, respectively. The second index expresses that the integration is carried out over k periods of the actual frequency.

6. Valuation of the method

To valuate the practical applicability of the method, numerical tests have been carried out. The step responses of systems of different structures have been calculated. All the systems are linear and stable, the number of the poles of their transfer function is between 2 and 6. Some of the tested transfer functions have finite zeros, and some of them contain a transportation lag.

In order to valuate the different approximating formulas, the Bode plots calculated from them have been compared with the exact Bode plot of the tested system.

There are three questions to be answered:

(1) Comparison of the formulas, the main part of which is expressed by the samples of $V(t)$ (Eqs. (41) and (42)), with those expressed by the samples of $v(t)$ (Eqs. (43) and (44)).

(2) For how many periods should the integration be extended, i.e. what should be the minimum value of k?

(3) How many terms of the complementary part are to be considered?

To distinguish the different approximating formulas being tested, we introduce a three-term index $[i, k, n]$. The first and second term have the same meaning as those used in the indices of R and Q in Eqs. (41)–(44). The third term n indicates the highest derivative of $v(t)$, which is still considered in the complementary part. It is to be noted that at most the second derivative of $v(t)$ can be measured or calculated from the measurements with acceptable error. Thus only $n \leq 2$ are of importance for practical applications.

In addition, the tests have shown that, with respect to those frequencies which are interesting in the control system synthesis, the integration is to be extended over at most two periods of the actual frequency.

From the real and imaginary parts $R_{[i,k,n]}(\omega)$ and $Q_{[i,k,n]}(\omega)$ respectively, of the approximating frequency function the amplitude and phase functions have been calculated:

$$a_{[i,k,n]} = \sqrt{R^2_{[i,k,n]} + Q^2_{[i,k,n]}}$$

$$\varphi_{[i,k,n]} = \operatorname{arc\ tan} \frac{Q_{[i,k,n]}}{R_{[i,k,n]}} .$$

Two illustrative examples of the numerous (approximately 50) systems, which have been tested, are demonstrated in Figs. 4 and 5.

The different approximations are marked by a triple index $[i, k, n]$, and the exact curve by p. According to the remarks mentioned above, the following values of i, k and n turned out to be of interest: $i = 0,1$; $k = 1, 2$; $n = 1, 2$.

The main viewpoint of the valuation of the formulas is: which of them approximates the exact transform in the frequency range decisive for the synthesis of control systems. This extends to those frequencies where the phase function exceeds the value 225°.

Another point of view is the accuracy which can be achieved in measuring or calculating the samples of the time functions used in the formulas. In this respect the formulas containing the samples of the step response $v(t)$ $(i = 1)$ are to be preferred to those which contain the samples of the integral of the step response $V(t)$ $(i = 0)$, i. e. measurements on the system usually produce the step response itself, and its numerical integration involves additional errors.

The tests have led to the following conclusions:

1. Large errors occur more often in the phase curve than in the amplitude curve. More exactly: if the amplitude error is large, then the phase error is also large, but this is not true inversely. Therefore, the goodness of the approximation can be judged more reliably by the phase error than by the amplitude error.

2. In the lower frequency range the approximations characterized by $k = 2$ and $n = 2$ are not better than those characterized by $k = 1$ and $n = 1$. Therefore, in this frequency range it is not worth using other formulas than the most simple ones indexed by [0, 1, 1] and [1, 1, 1]. These formulas give very good approximation in the frequency range where the value of the phase function is less than 135°. In general, the amplitude error is less than 1—2 dB, and the phase error less than 2°–3°.

3. In the middle frequency range (i.e. where the phase function is between —135° and —225°), we have to choose $k = 2$. So the formulas indexed by [0, 2, 1], [0, 2, 2], [1, 2, 1] and [1, 2, 2] are to be used. Among them [1, 2, 1] is to be preferred, because it is expressed by the samples of the step response, and its complementary part does not contain the second derivative of the step response, but only the first derivative (the weight function).

4. At higher frequencies (where the phase function is between —235° and —270°) only the formula indexed by [1, 2, 2] yields acceptable results.

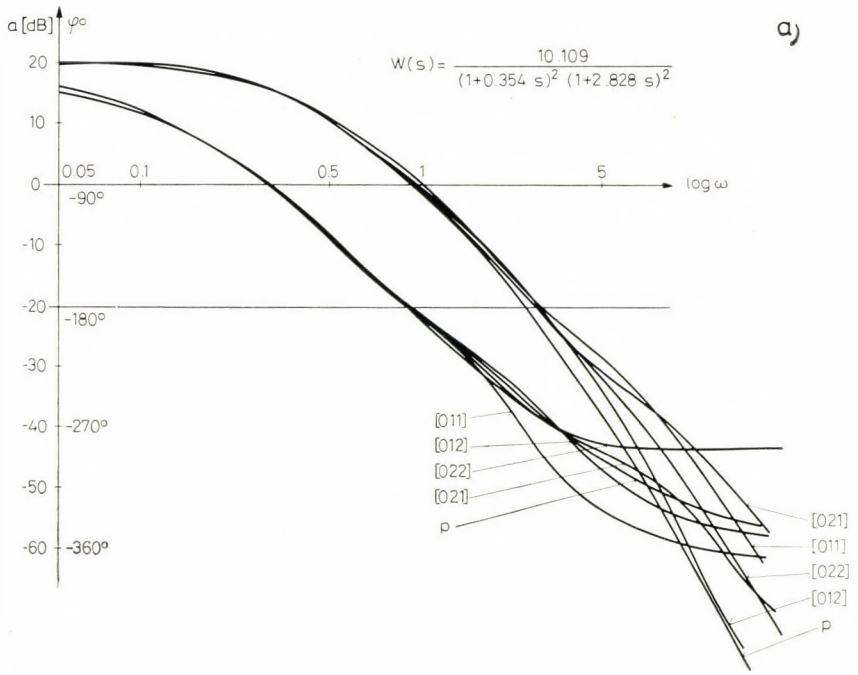

a)
a[dB]
φ°
W(s) = 10.109 / (1+0.354 s)² (1+2.828 s)²
20
10
0
-10
-20
-30
-40
-50
-60
0.05
0.1
0.5
1
5
log ω
-90°
-180°
-270°
-360°
[011]
[012]
[022]
[021]
p
[021]
[011]
[022]
[012]
p

Fig. 4a

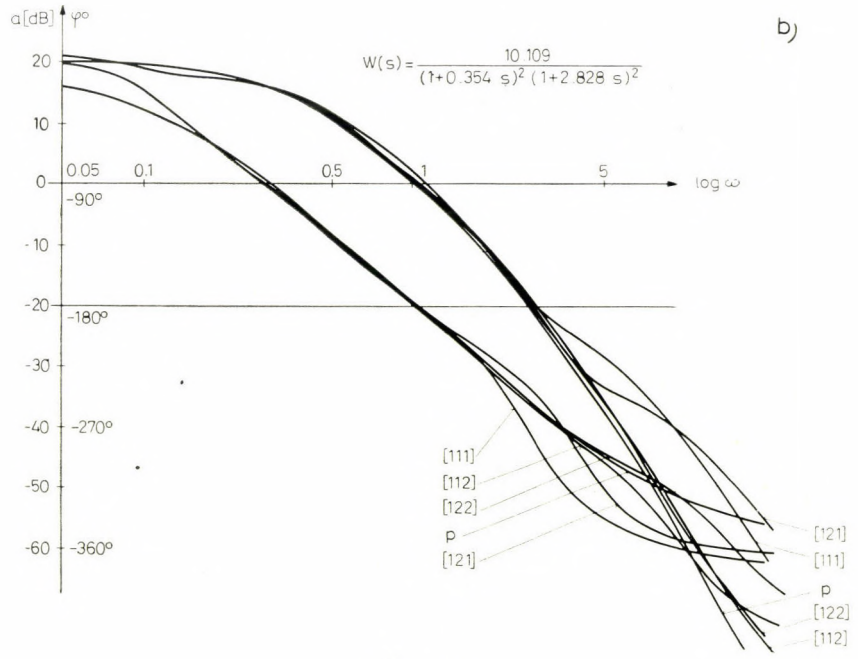

b)
a[dB]
φ°
W(s) = 10.109 / (1+0.354 s)² (1+2.828 s)²
20
10
0
-10
-20
-30
-40
-50
-60
0.05
0.1
0.5
1
5
log ω
-90°
-180°
-270°
-360°
[111]
[112]
[122]
p
[121]
[121]
[111]
p
[122]
[112]

Fig. 4b

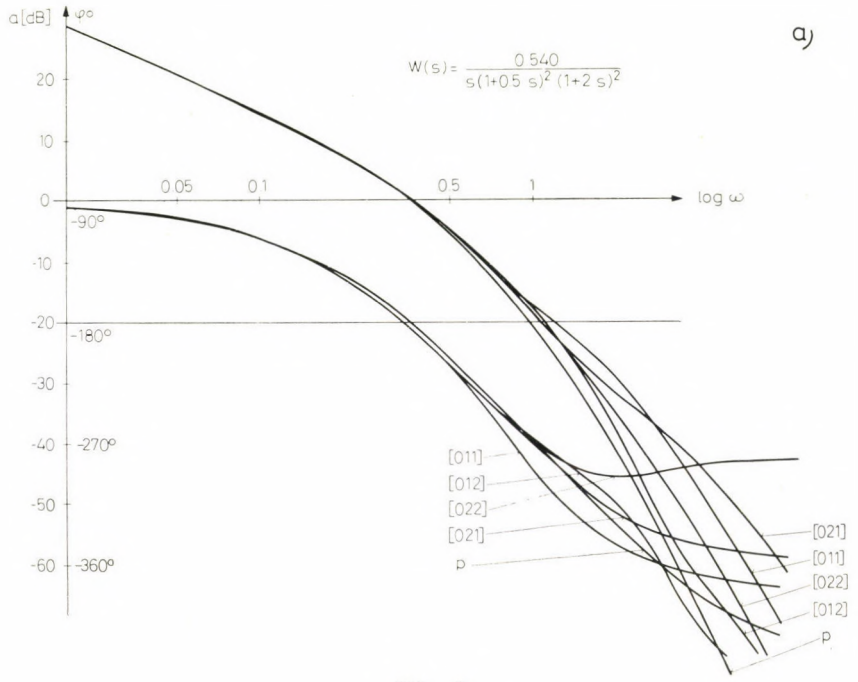

a)
a[dB]
φ°
W(s) = 0.540 / s(1+0.5 s)² (1+2 s)²
20
10
0
-10
-20
-30
-40
-50
-60
0.05
0.1
0.5
1
log ω
-90°
-180°
-270°
-360°
[011]
[012]
[022]
[021]
p
[021]
[011]
[022]
[012]
p

Fig. 5a

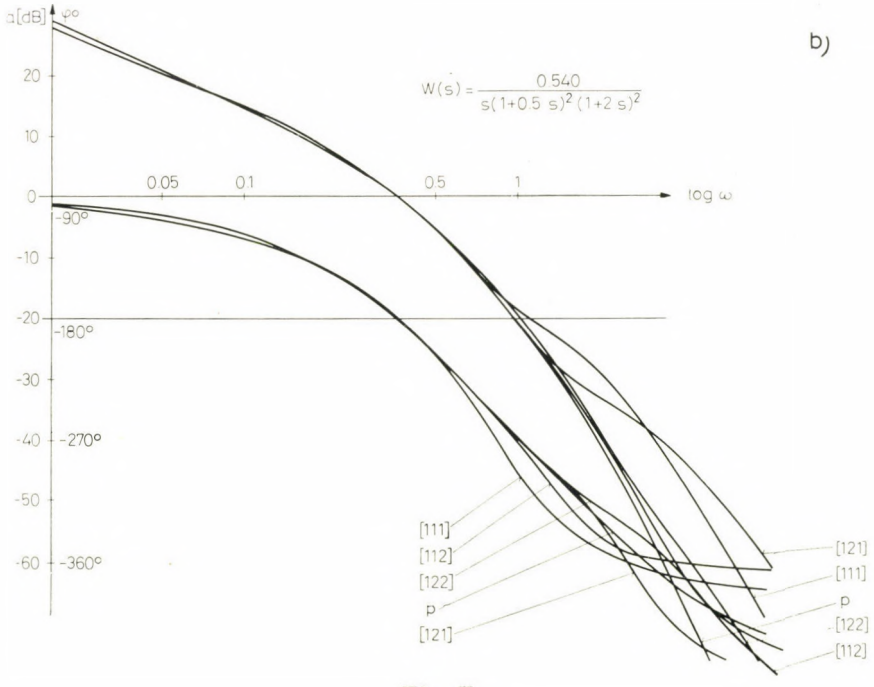

b)
a[dB]
φ°
W(s) = 0.540 / s(1+0.5 s)² (1+2 s)²
20
10
0
-10
-20
-30
-40
-50
-60
0.05
0.1
0.5
1
log ω
-90°
-180°
-270°
-360°
[111]
[112]
[122]
p
[121]
[121]
[111]
p
[122]
[112]

Fig. 5b

7. Conclusion

A fast Fourier transform to calculate the frequency characteristics of a system from its step response has been demonstrated. The method gives good results in solving numerous problems occurring in process control.

The main advantage of the method is the simplicity of the formulas used. The accuracy of the method is sufficient in the frequency range in which the phase shift of the frequency characteristics is less than 225°.

Just in this frequency range are the values of the frequency function decisive for the behaviour of the closed control loop. Therefore, the knowledge of the frequency function in this frequency range gives enough information for the synthesis of control systems. If the suitable formula is used, the error of the amplitude and phase does not generally exceed 2–3 dB and 3–4°, respectively.

Due to the simplicity of the formulas, the computing time needed to calculate the frequency function is very small. Therefore, the method can be applied in computer control procedures of on-line identification and automatic control parameter setting.

Acknowledgement. The author would like to express his thanks to Mr. L. Langer for his help and assistance.

Reference

1. *Frigyes, A., Langer, L.*, A new method for the fast computing of the frequency characteristic of control systems. Periodica Polytechnica **18**, 1 (1974).

Специальный метод ускоренного преобразования Фурье для определения областей низких и средних частот частотной характеристики системы автоматического управления

А. ФРИДЕШ

(Будапешт)

Для анализа и синтеза систем автоматического управления разработаны различные методы описания динамических свойств системы, основанные на применении временных функций, и, чаще всего, весовой функции и переходной функции.

В различных, как детерминированных, так и стохастических методах идентификации применяется переходная функция системы.

Хотя и существуют такие методы, с помощью которых синтез системы производится непосредственно во временной области, но в большинстве случаев они основаны на применении известных полюсов и нулей передаточной функции или частотной характеристики системы. В последнем случае необходимо определить преобразование Фурье весовой функции.

Разработаны различные алгоритмы расчета преобразования Фурье; некоторые из них носят название «ускоренного преобразования Фурье».

Характерной чертой таких методов является такой выбор и такая группировка операций умножения, с помощью которых расчет интеграла Фурье производится с применением минимального количества операций. В то же время обычно при использовании таких методов не предполагается никакое ограничение в отношении преобразуемой временной функции.

Предложенный в настоящей статье метод использует такие временные функции как переходные функции (или весовые функции) систем управления объектами. Далее, в отношении точности ставится условие лишь в такой области частот, в которой точность существенно действует на качество замкнутой системы автоматического управления.

Выведенными на основе предложенного метода формулами определяются действительная и мнимая часть частотной характеристики: переходная функция $v(t)$ путем определения некоторых проб ее интеграла $V(t) = \int_0^t v(t)\,dt$, первой и высших ее производных $w(t) = \frac{dv(t)}{dt}$, $w'(t)$ и т. д.

A. Frigyes
Department of Process Control
Technical University of Budapest
H-1111 Budapest, XI. Stoczek u. 2.

Problems of Control and Information Theory, Vol. 5 (4), pp. 321—327 (1976)

ОБ ОДНОМ ПОДХОДЕ К ЗАДАЧЕ ДЕКОМПОЗИЦИИ ВЕРОЯТНОСТНЫХ АВТОМАТОВ

А. Х. ГИОРГАДЗЕ

(Тбилиси)

(Поступила в редакцию 27 января 1975 г.)

Отличие предлагаемого подхода к задаче декомпозиции вероятностных автоматов заключается в том, что подавтомат, выделяемый из системы автоматов, в которую декомпозируется исходный автомат, может быть не изоморфным исходному автомату и представлять то же событие, что и исходный автомат, но с произвольной точкой сечения. Декомпозиция такого типа названа λ'-декомпозицией и рассмотрена для одного класса вероятностных автоматов. Определены условия декомпозиции обобщенных автоматов. Доказано, что если автомат λ'-декомпозируем, он декомпозируем с произвольным λ, $0 < \lambda < 1$.

Задача декомпозиции автомата обычно ставится следующим образом [1]: для заданного автомата A строится автомат B (при условии, если A удовлетворяет условиям соответствующих теорем по декомпозиции), который представляет собой систему автоматов, соединенных друг с другом определенным образом и меньших по числу состояний, чем автомат A и такой, что в нем можно выделить подавтомат B', изоморфный исходному автомату A. Если A является вероятностным автоматом (в. а.), представляющим событие E с точкой сечения λ, то ясно, что и B' будет представлять то же событие с той же точкой сечения λ.

В данной заметке ставится следующая задача: пусть задан в. а. A, представляющий событие E с точкой сечения λ и не удовлетворяющий обычным [1] условиям декомпозиции. Однако пусть существует система автоматов (автомат B^0) с заданными наперед ограничениями на связи между ними, число состояний которых меньше числа состояний автомата A, причем некоторый подавтомат B_0' автомата B_0 представляет это же событие E, но с точкой сечения λ' в общем случае не обязательно совпадающей с λ. Декомпозицию такого вида будем называть λ'-декомпозицией вероятностного автомата.

Определим некоторые условия λ'-декомпозиции.

Идея заключается в следующем: вводится понятие декомпозиции обобщенных автоматов (о. а.). Обобщенные автоматы вводятся и рассматриваются в [2]. Определяются необходимые и достаточные условия их декомпозиции. Далее, заданный в. а. преобразуется в некоторый обобщенный автомат, и если этот о. а. удовлетворяет условиям декомпозиции обобщенных автоматов,

определяется система обобщенных автоматов, в которую он декомпозируется. Затем каждый автомат в найденной системе заменяется вероятностным автоматом, так, чтобы в полученной системе в. а. можно было выделить некоторый подавтомат, представляющий то же событие, что и исходный в. а.

Определение 1. По аналогии с [2] о. а. A^0 есть пятерка $A^0 = (X, S, \{\mathbf{M}(x), g_0, f_0)$, где X — входной алфавит, $S = (s_1, s_2, \ldots, s_n)$ — множество состояний, $M(x)$ — матрица, связанная с буквой $x \in X$, g_0 — n-мерный вектор-строка (начальный вектор), f_0 — n-мерный вектор-строка (начальный вектор), f_0 — n-мерный вектор-столбец (вектор конечных состояний); компоненты $M(x)$, g_0 и f_0 — суть действительные числа.

Иногда под о. а. будем подразумевать четверку $(X, S, \{\mathbf{M}(x)\}, g_0)$.

Определение 2. Каскадное произведение обобщенных автоматов $A_1^0 = (X, S^1 = (s_1^1, \ldots, s_n^1), \{\mathbf{M}^1(x_i)\}, g_0^1 = (p_1, p_2, \ldots, p_n))$, $A_2^0 = (X \times Y, S^2 = (s_1^2, \ldots, s_m^2), \{\mathbf{M}^2(x_i, y_j)\}, g_0^2)$, где Y — множество букв $y_1, y_2, \ldots, y_n$, генерируемых автоматом A_1^0 при нахождении его в состояниях $s_1^1, s_2^1, \ldots, s_n^1$, соответственно, есть автомат $A_1^0 \otimes A_2^0 = (X, S^1 \times S^2, \{\mathbf{M}(x_i)\}, g_0)$, где $g_0 = (p_1 g_0^2, p_2 g_0^2, \ldots, p_n g_0^2)$, $\mathbf{M}(x_i) = \mathbf{M}^1(x_i) \otimes \{M^2(x_i, y_j)\}$ — каскадное произведение [3] матрицы $\mathbf{M}^1(x_i)$ и множества $\{\mathbf{M}^2(x_i, y_j)\}$, т. е. матрица вида

$$\mathbf{M}(x_i) = \begin{bmatrix} p_{11}\mathbf{M}^2(x_i, y_1) & \ldots & p_{1n}\mathbf{M}^2(x_i, y_1) \\ p_{21}\mathbf{M}^2(x_i, y_2) & \ldots & p_{2n}\mathbf{M}^2(x_i, y_2) \\ \ldots\ldots\ldots\ldots & & \\ p_{n1}\mathbf{M}^2(x_i, y_n) & \ldots & p_{nn}\mathbf{M}^2(x_i, y_n) \end{bmatrix}, \qquad (1)$$

где p_{ij} — элементы $\mathbf{M}^1(x_i)$. Каскадное произведение автоматов означает их последовательное соединение. Если для каждой буквы $x_i \in X$ $\mathbf{M}^2(x_i, y_j) = \mathbf{M}^2(x_i)$ для всех $y_j \in Y$, тогда произведение $A_1^0 A_2^0$ называется прямым произведением автоматов и означает их параллельное соединение.

Определение 3. Автомат A^0 допускает последовательную декомпозицию на автоматы A_1^0 и A_2^0, если в автомате $A_1^0 \otimes A_2^0$ выделяется подавтомат $A^{0'}$, изоморфный A^0 и $|A_1^0| < |A^0|$, $|A_2^0| < A^0|$ ($|A|$ — число состояний автомата A).

Теорема 1. Для того, чтобы о. а. A^0 допускал последовательную декомпозицию на автоматы A_1^0 и A_2^0 необходимо и достаточно существование двух разбиений π и τ множества S состояний автомата A^0 на непересекающиеся подмножества, $\pi = (\pi_1, \ldots, \pi_m)$, $\tau = (\tau_1, \ldots, \tau_k)$, таких что:

1. $\pi \cdot \tau = 0$

2. $m < |A^0|$, $k < |A^0|$, $|\pi_i| < |A^0|$, $|\tau_j| < |A^0|$, $i = \overline{1, m}$, $j = \overline{1, k}$ ($|\pi_i|$ — число элементов подмножества π_i разбиения π),

3. Для каждого π_a, произвольных π_b, π_c и всех $x \in X$ либо $\operatorname{sign} P^x_{i\tau_a \cap \pi_b} = \operatorname{sign} P^x_{i\tau_a \cap \pi_c}$ для всех $i \in \pi_a$ и всех $\tau_a \in \tau$ либо $\operatorname{sign} p^x_{i\tau_a \cap \pi_b} = -\operatorname{sign} p^x_{i\sigma_a \cap \pi}$ для всех $i \in \pi_a$ и всех $\tau_a \in \tau$.

4. Для каждого π_a, произвольных $i, j \in \pi_a$ и произвольных τ_b, τ_c (τ_b может совпадать с τ_c) и всех $x \in X$

$$\frac{P^x_{i\tau_b \cap \pi_1}}{P^x_{j\tau \cap \pi_1}} = \frac{P^x_{i\tau_b \cap \pi_2}}{P^x_{j\tau_c \cap \pi_2}} = \ldots ! = \frac{P^x_{i\tau_b \cap \pi_m}}{P^x_{j\tau_c \cap \pi_m}},$$

где P^x_{ij} — элементы матрицы $\mathbf{M}(x)$ автомата A^0.

Пусть условия теоремы выполнены. Рассмотрим π_i и π_j. Пусть $P^x_{i\tau_b \cap \pi_i} = \alpha^x_{ij} P^x_{i\tau_b \cap \pi_j}$, где α^x_{ij} — некоторое действительное число. Покажем, что для любых $\tau_k \in \tau$ $P^x_{i\tau_k \cap \pi_i} = \alpha^x_{ij} P^x_{i\tau_k \cap \pi_j}$. Действительно, это следует из равенства $P^x_{i\tau_b \cap \pi_i} / P^x_{i\tau_k \cap \pi_i} = P^x_{i\tau_b \cap \pi_j} / P^x_{i\tau_k \cap \pi_j}$.

Рассмотрим некоторое π_a. Используя условия 3 и 4 теоремы, можно доказать также, что $p^x_{i\tau_k \cap \pi_i} = \alpha^x_{ij} p^x_{i\tau_k \cap \pi_j}$ и для любого $i \in \pi_i$.

Подмножества π_i и π_j определяют соответствующие подматрицы в матрицах $\mathbf{M}(x)$ автомата A^0. Пусть π_d — максимальное подмножество по числу элементов в π. Подматрицы $\pi_i \pi_j$, $i, j = 1, 2, \ldots, m$ увеличим до размеров $|\pi_d| \times |\pi_d|$, дополняя их столбцами и строками. Их можно заполнить так, чтобы полученная матрица имела вид (1). При этом подматрица $\pi_i \pi_j$, умноженная на α^x_{ij}, будет совпадать поэлементно с подматрицей $\pi_i \pi_j$. Согласно определениям 2 и 3 числа α^x_{ij} определяют элементы матрицы $\mathbf{M}^1(x)$ автомата A^0_1, а подматрица $\pi_i \pi_j$ есть матрица $\mathbf{M}^2(x, y_i)$ автомата A^0_2. Необходимость условий теоремы. очевидна.

Пусть автомат $A^0 = (X, S = (s_1, \ldots, s_n), \{\mathbf{M}(x)\}, g_0, f_0)$, представляющий событие E с точкой сечения η, (декомпозируется на два параллельно соединенные автоматы $A^0_1 = (X, S^1 = (s^1_1, \ldots, s^1_m), \{\mathbf{M}_1(x)\}, g^1_0)$ и $A^0_2 = (X, S^2 = (s^2_1, \ldots, s^2_k), \{\mathbf{M}_2(x)\}, g^2_0)$, и при этом элементы матрицы $\mathbf{M}(x)$ неотрицательны.

Теорема 1. Существует система четырех в. а. A_1, A_2, A_0 и A_4, соединенных параллельно, представляющая то же событие, что и о. а. A^0, и при этом $|A_1| = |A_2| = m + 3$, $|A_3| = |A_4| = k + 3$.

Сначала докажем лемму о существовании автомата $A^{0\prime} = (X, S' = (s_1, \ldots, s'_{(m+3)(k+3)}), \{\mathbf{M}'(x)\}, g'_0, f'_0)$, представляющего то же событие, что и автомат A^0, и декомпозирующего на два параллельно соединенные автоматы

$$A^{0\prime}_1 = (X, S^{1\prime} = (s^{1\prime}_1, \ldots, s^{1\prime}_{m+3}), \{\mathbf{M}'_1(x)\}, g^{1\prime}_0)$$

и

$$A^{0\prime}_2 = (X, S^{2\prime} = (s^{2\prime}_1, \ldots, s^{2\prime}_{k+3}), \{\mathbf{M}'(x)\}, g^{2\prime}_0)$$

так, что матрицы $\mathbf{M}'(x)$, $\mathbf{M}'_1(x)$, $\mathbf{M}'_2(x)$ — стохастические.

Доказательство. Пусть $\overline{X}$ — множество всех слов x над алфавитом X и $\mathbf{M}_1(\bar{x})$ — соответствующая матрица, описывающая работу автомата A_1^0 при поступлении слова $\bar{x} \in \overline{X}$. Обозначим через $\delta(\bar{x})$ число большее, чем наибольшая из сумм строчек матрицы $\mathbf{M}_1(\bar{x})$ и введем числа $\delta_i(\bar{x})$ такие, чтобы матрица

$$\mathbf{M}_1'(\bar{x}) = \left[\begin{array}{c|ccc|cc} 1 & 0 & \ldots & 0 & 0 & 0 \\ \hline \delta_1(x) & & & & 0 & 0 \\ \vdots & & \mathbf{M}_1(\bar{x})/\delta(\bar{x}) & & \vdots & \vdots \\ \delta_m(\bar{x}) & & & & 0 & 0 \\ \hline 0 & 0 & \ldots & 0 & 1/\delta(\bar{x}) & 1-1/\delta(\bar{x}) \\ 0 & 0 & \ldots & 0 & 0 & 1 \end{array}\right]$$

была стохастической. Аналогично для автомата A_2^0 введем числа $\gamma(\bar{\ })$ и $\gamma_i(\bar{x})$ и стохастическую матрицу $\mathbf{M}_2'(\bar{x})$. Положим $\mathbf{M}'(\bar{x}) = \mathbf{M}_1'(\bar{x}) \otimes \mathbf{M}_2'(\bar{x})$ согласно (1). Кроме того, положим $g_0^{1'} = (0, g_0^1, \eta_1, 0)$ и $g_2^{0'} = (0, g_0^2, \eta_2, 0)$, где $\eta_1 \cdot \eta_2 = \eta$. Введем $g_0' = g_0^{1'} \cdot g_0^{2'}$ и $f_0' = (0, 0,0, 0, 0, f_0, 0, 0,0, 0, -1, 0, 0, 0, 0)^T$. Нетрудно проверить, что

$$g_0' \mathbf{M}'(\bar{x}) f_0' = \left(1/\delta(\bar{x})\, \gamma(\bar{x})\right) \cdot \left(g_0^1 g_0^2 \mathbf{M}(\bar{x}) f_0 - \eta\right),$$

что доказывает лемму.

Пусть теперь автомат $A^0 = \left(X, S = (s_1, \ldots, s_n), \{\mathbf{M}(x)\}, g_0, f_0\right)$, представляющий событие E с точкой сечения Θ такой, что все компоненты $g_0 = (p_1, p_2, \ldots, p_n)$ положительные и $\{\mathbf{M}(x)\}$ — множество стохастических матриц.

Введем автомат $A_c^0 = \left(X, S, \{\mathbf{M}(x)\}, p^{-1} g_0, f + (c, c, \ldots, c)^T\right)$, где $p = p_1 + \ldots + p_n$, а c — число, выбранное так, чтобы все компоненты вектора конечных состояний автомата A_c^0 были положительны. Тогда

$$p^{-1} g_0 \mathbf{M}(\bar{x}) \left(f + (c, c, \ldots, c)^T\right) = p^{-1} g_0 \mathbf{M}(\bar{x})\, f + c,$$

и автомат A_c^0 представляет событие E. Далее, компоненты вектора $f + (c, c, \ldots, c)^T$ обозначим через $d_1, d_2, \ldots, d_n$. Введем $d = d_1 + d_2 + \ldots + d_n$ и $d_i' = d_i/d$. Согласно [2] (теорема 16), существует вероятностный автомат

$$A = \left(X, S = (s_1, \ldots, s_n, s_{n+1}, \ldots, s_{n^n}), \{H(x)\}, g, F\right),$$

где $g = n^{-1} (p^{-1} g_0, \ldots, p^{-1} g_0)$, $p^{-1} g_0$ взято n раз,

$F = (s_1, s_{n+2}, s_{2_{n+3}}, \ldots, s_{n^n})$ и

$$\mathbf{H}(x) = \begin{bmatrix} d_1'\mathbf{M}(x) & \ldots & d_n'\mathbf{M}(x) \\ d_1'\mathbf{M}(x) & \ldots & d_n'\mathbf{M}(x) \\ \ldots\ldots\ldots\ldots & & \\ d_1'\mathbf{M}(x) & \ldots & d_n'\mathbf{M}(x) \end{bmatrix}.$$

Согласно обычным условиям декомпозиции в. а., автомат A декомпозируется на два параллельно работающие автоматы $A_1 = (X, S_1 = (s_1^1, \ldots, s_n^1), \{\mathbf{M}_1(x)\}, g_1)$ и $A_2 = (X, S_2 = (s_1^2, \ldots, s_n^2), \{\mathbf{M}_2(x)\}, g_2)$, где $g_1 = (1/n, \ldots, 1/n)$, $g_2 = p^{-1}g_0$, $\mathbf{M}_2(x) = \mathbf{M}(x)$ и

$$\mathbf{M}_1(x) = \begin{bmatrix} d_1 & \ldots & d_n' \\ d_1' & \ldots & d_n' \\ \ldots\ldots\ldots & & \\ d_1' & \ldots & d_n' \end{bmatrix}.$$

Пусть $A = (X, S = (s_1, \ldots, s_n), \{M(x)\}, g, f)$ — в. а., представляющий событие E с точкой сечения λ, не допускающий декомпозицию. Однако при этом пусть существуют матрицы $\mathbf{C}(x) = |c_{ij}^x|$, $c_{1k}^x = c_{2k}^x = \ldots = C_{nk}^x$, $\forall\, k = \overline{1, n}$ такие, что матрицы $\mathbf{M}(x) + \mathbf{C}(x)$ удовлетворяют условиям теоремы 1 о параллельной декомпозиции по разбиениям $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$ и τ. Рассмотрим матрицы $\mathbf{M}'(x)$, полученные из $\mathbf{M}(x)$ следующим образом. К каждой подматрице матрицы $\mathbf{M}(x)$, определяемой парой блоков π_i, π_j, припишем справа столбец с элементами $-1/m$ и снизу строку, равную первой строке этой подматрицы. Таким же способом изменим матрицы $\mathbf{C}(x)$, добавляя строку и столбец, но с элементами $1/m$. Обозначим полученные матрицы через $\mathbf{C}'(x)$. Введем сумму

$$\mathbf{M}'(x) + \mathbf{C}'(x) = \mathbf{M}''(x).$$

Нетрудно проверить, что для любого слова $\bar{x}$

$$\mathbf{M}''(\bar{x}) = \mathbf{M}'(\bar{x}) + \alpha^{l(\bar{x})-1} \cdot \mathbf{C}'(\bar{x}),$$

где $\alpha = \sum_j c_{ij}'$, $l(\bar{x})$ — длина слова $\bar{x}$. Введем матрицу

$$\mathbf{Q}(\bar{x}) = \begin{bmatrix} \mathbf{M}''(\bar{x}) & \mathbf{0} \\ \mathbf{0} & \alpha^{l(\bar{x})-1} \cdot \mathbf{C}'(\bar{x}) \end{bmatrix}$$

и рассмотрим автомат $A_t = (X, S_t, \{\mathbf{Q}(x)\}, g_t, f_t)$, где $g_t = (g, g)$, $f_t = (f, -f)$. Тогда

$$g_t\, \mathbf{Q}(\bar{x})\, f_t = g\, \mathbf{M}'(\bar{x}) f, \quad \forall\, \bar{x}.$$

Т. е. автомат A_t представляет то же событие, что и автомат $A_0 = (X, S_0\ \mathbf{M}'(x), g_0, f_0)$, подавтоматом которого является автомат A. Далее, легко видеть, что автомат A_t декомпозируется на два последовательно соединенные автоматы A_t^1 и A_t^2. Автомат A_t^1 описывается матрицей $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, а автомат A_t^2 — матрицами $\mathbf{M}''(x)$ и $\mathbf{C}'(x)$, и при этом A_t^2 декомпозируется на два автомата, работающие параллельно. Таким образом, получаем систему каскадно [3] соединенных о. а, в которой выделяется подавтомат, представляющий то же событие, что и исходный в. а.

Пример. В качестве примера рассмотрим автономный в. а., описываемый матрицей

$$\mathbf{M} = \begin{bmatrix} 0{,}2 & 0{,}1 & 0{,}2 & 0{,}4 & 0{,}1 \\ 0{,}2 & 0{,}2 & 0 & 0{,}4 & 0{,}2 \\ 0{,}2 & 0 & 0{,}4 & 0{,}4 & 0 \\ 0{,}2 & 0{,}1 & 0{,}2 & 0{,}4 & 0{,}1 \\ 0{,}2 & 0{,}2 & 0 & 0{,}4 & 0{,}2 \end{bmatrix}.$$

Легко проверить, что соответствующая цепь Маркова укрупнима только по разбиениям $\pi_1 = (\overline{1{,}4};\ \overline{2{,}5};\ \overline{3})$, $\pi_2 = (\overline{1{,}4};\ \overline{2};\ \overline{5};\ \overline{3})$, $\pi_3 = (\overline{1};\ \overline{4};\ \overline{2{,}5};\ \overline{3})$, из которых только два π_2 и π_0 удовлетворяют условию $\pi_2 \cdot \pi_0 = 0$, но при этом условию независимости [1] не удовлетворяют. Это означает, что рассматриваемый автомат не декомпозируется на два автомата, работающие параллельно. Однако, если к матрице $\mathbf{M}$ добавить матрицу $\mathbf{C}$, все строчки которой одинаковы и равны [0,2 0 0 0 0], то легко убедиться, что автономный в. а., описываемый матрицей $\mathbf{M} + \mathbf{C}$, декомпозируем на автоматы, описывамые матрицами

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \text{ и } \begin{bmatrix} 0{,}4 & 0{,}1 & 0{,}2 \\ 0{,}4 & 0{,}2 & 0 \\ 0{,}4 & 0 & 0{,}4 \end{bmatrix}.$$

Применяя к автомату A_t^2 результат теоремы 2, придем к декомпозиции исходного автомата A. Легко убедиться, что эта декомпозиция является λ'-декомпозицией. Пусть в. а. A представляет событие E с точкой сечения λ и допускает λ'-декомпозицию. Тогда справедлива следующая теорема.

Теорема 3. Автомат A допускает λ-декомпозицию.

Доказательство. Пусть B_0 — в. а., описывающий систему в. а. $(A_1, A_2, \ldots, A_m)$, удовлетворяющую условиям $|A_i| < |A|$ и ограничениям на связи между A_i, $i = \overline{1, m}$. Пусть B_0^1 — подавтомат в. а. B_0, представляющий E с точкой

сечения λ', и пусть $\{\mathbf{M}(x)\}$ — множество его матриц переходов. Согласно[2] существует в. а. (обозначим его через B_0^2), представляющий E с любой точкой сечения η, $0 < \eta < 1$, и при этом множество матриц переходов в. а. B_0^2 имеет вид

$$\left\{\begin{bmatrix} \mathbf{M}(x) & 0 \\ 0 & 1 \end{bmatrix}\right\}.$$

Нетрудно видеть, что в. а. B_0^2 допускает декомпозицию на два последовательно соединенные автоматы; первый — автономный в. а. с матрицей переходов $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, и второй — некоторый в. а. (обозначим его через B_0^3), множество матриц переходов которого есть $\{\{\mathbf{M}(x)\}, \mathrm{I}\}$, где I — единичная матрица той же размерности, что и $\mathbf{M}(x)$. Так же, как и B^2, в. а. B^3 допускает декомпозицию и это завершает доказательство.

Литература

1. *Bacon, G.*, The Decomposition of stochastic Automata, Information and Control, **7**, 3, 1964, pp. 320—339.
2. *Turakainen, P.*, On probabilistic automata and their generalizations, Ann. Acad. Sci. Fenn. **A1**, 429, 1968, pp. 10—52.
3. *Paz, A.*, Introduction to probabilistic automata. Academic Press, New York and London, 1971, ch. 2.

An approach to decomposition of probabilistic automata

A. KH. GIORGADZE

(Tbilisi)

An automaton B which is a system of smaller connection restricted automata is constructed so that some subautomaton of B should exist (which is in general not isomorphic to the original one) to represent the same event as the original automaton. The corresponding cut-points are allowed to be different. This type of decomposition is called λ'-decomposition and an example of such decomposition is given for a certain class of probabilistic automata (PA). A representative of the class is transformed into a generalized automaton (decomposition conditions had been introduced) which is decomposed into smaller generalized automata and then their reverse transformation into probabilistic ones is performed.

Finally it is proved that if an automaton is λ'-decomposable (λ' — some cut-point), then it can be decomposed with an arbitrary cut-point λ, $0 < \lambda < 1$.

А. Х. Гиоргадзе
Институт кибернетики АН ГрССР,
СССР, Тбилиси, 380086,
ул. С. Эули, 5

Problems of Control and Information Theory, Vol. 5 (4), pp. 329—358 (1976)

DECOUPLING AND POLE PLACEMENT IN LINEAR MULTIVARIABLE SYSTEMS: A DIRECT ALGEBRAIC APPROACH

P. RÓZSA*, N. K. SINHA

(Budapest) (Hamilton)

(Received October 6, 1975)

Starting with the controllable canonical form representation of a linear multivariable system, necessary and sufficient conditions are established for decoupling a linear multivariable system and placing its poles at arbitrary locations by using linear state variable feedback and a constant nonsingular transformation of the input. Since these conditions are based only on the matrix **C** of the triple **A**, **B**, **C** in the canonical representation, they are very easy to apply. An efficient algorithm for decoupling and pole placement is then presented.

Introduction

The problem of decoupling an interacting linear time-invariant multivariable system has been considered by many authors. For decoupling to be of practical value one must also be able place to all the poles of the transfer function of the decoupled system at suitable locations in order to have a satisfactory and stable response. Important contributions to the field have been made by Morgan and Rekasius [2], Falb and Wolovich [3], Gilbert [4] and Wonham and Morse [5]. A recent paper by Morse and Wonham [6] presents a detailed account on the status of noninteracting control. Falb and Wolovich [3] have developed a necessary and sufficient condition for decoupling by state-variable feedback and a constant transformation of the input. Gilbert [4] introduced a dynamic precompensator with state feedback and obtained a precompensator of minimal order. Wonham and Morse [5] have developed a geometric approach for minimal precompensation. Silverman [7] has proposed two algorithms for decoupling invertible systems using state feedback and precompensation.

Although the problem appears to have been "completely solved", there is still need for developing a straightforward method for being able to know when a multivariable system may be decoupled by applying state feedback and a constant transformation of the input as well as the poles of the closed loop system may be placed arbitrarily. In all cases where this is possible, it is generally more desirable and convenient to use such a "static" compensator instead of the "dynamic" precompensator proposed by Gilbert [4], Silverman [7], or Wonham and Morse.

* On leave at McMaster University for Oct. 1973 to Aug. 1974.

The object of this paper is to develop such a procedure. Starting with the controllable canonical form representation, necessary and sufficient conditions will be developed for decoupling and pole placement.

A direct algebraic approach will be used. Although the derivations will require considerable amount of matrix manipulations, the final results are very easy to apply. Some simple tests on the matrix **C** of the triple **A**, **B**, **C** in the canonical representation answer the question whether the decoupling and pole-placement is possible. These then lead to a straightforward procedure for determining the static precompensator and the feedback matrix for decoupling and desired pole placement. The procedure appears much simpler than that of Falb and Wolovich.

Statement of the problem

Given

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{1}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} \tag{2}$$

where $\mathbf{x} \in R_n$, $\mathbf{u} \in R_m$, $\mathbf{y} \in R_m$ and $m \leq n$. It is required to find the state-feedback law

$$\mathbf{u} = \mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{v}$$

where **G** is a nonsingular $m \times m$ constant matrix, such that the resulting closed-loop system transfer function matrix

$$\mathbf{D}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A} - \mathbf{B}\mathbf{F})^{-1}\mathbf{B}\mathbf{G} \tag{3}$$

is a diagonal matrix of proper rational functions of s.

For the sake of simplicity, it will be assumed that the matrices **A** and **B** are in the controllable canonical form, since any system can be transformed to this form following the procedure given by Luenberger [8], and more recently by Applevich [9].

The forms are shown below,

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \dots & \mathbf{A}_{1m} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \dots & \mathbf{A}_{2m} \\ \vdots & & & \\ \mathbf{A}_{m1} & \mathbf{A}_{m2} & \dots & \mathbf{A}_{mm} \end{bmatrix} \tag{4}$$

where $\mathbf{A}_{ij}$ is an $n_i \times n_j$ matrix $\left(\sum_{i=1}^{m} n_i = n\right)$ of the following form:

$$\mathbf{A}_{ij} = \delta_{ij} \cdot \mathbf{J}_{n_i} - e_{n_i} \cdot a^T(ij), \tag{5}$$

the superscript T representing transposition.

In the above expression, δ_{ij} is the Kronecker delta, $\mathbf{J}_{n_i}$ is a nilpotent matrix of index n_i as shown below

$$\mathbf{J}_{n_i} = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdot & \cdot & \cdots & 1 \\ 0 & 0 & & & & 0 \end{bmatrix}. \tag{6}$$

Also, $\mathbf{e}_{n_i}$ is the n_ith unit vector of dimension n_i, and

$$\mathbf{a}^T(ij) = [a_1(ij)\, a_2(ij) \dots a_n(ij)]. \tag{7}$$

For example

$$\mathbf{A}_{11} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & & 1 \\ -a_1(11) & -a_2(11) & & & -a_{n_1}(11) \end{bmatrix} \tag{8}$$

and

$$\mathbf{A}_{12} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 0 & 0 & & 0 \\ -a_1(12) & -a_2(12) & \dots & -a_{n_2}(12) \end{bmatrix}. \tag{9}$$

Also,

$$\mathbf{B} = \begin{bmatrix} \mathbf{e}_{n_1} & 0 & 0 & \dots & 0 \\ 0 & \mathbf{e}_{n_2} & 0 & \dots & 0 \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ 0 & 0 & 0 & & \mathbf{e}_{n_m} \end{bmatrix} = \langle \mathbf{e}_{n_i} \rangle. \tag{10}$$

The $m \times n$ matrix $\mathbf{C}$ can be partitioned accordingly, as shown below

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}^T(11) & \mathbf{c}^T(12) & \dots & \mathbf{c}^T(1m) \\ \mathbf{c}^T(21) & \mathbf{c}^T(22) & \dots & \mathbf{c}^T(2m) \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \mathbf{c}^T(m1) & \mathbf{c}^T(m2) & \dots & \mathbf{c}^T(mm) \end{bmatrix} = [\mathbf{c}^T(ij)] \tag{11}$$

where $\mathbf{c}^T(ij)$ is a row vector of dimension n_j.

The closed-loop transfer function matrix $\mathbf{D}(s)$, is desired to be a diagonal matrix, and will be represented as

$$\mathbf{D}(s) = \left\langle \frac{R_i(s)}{Q_i(s)} \right\rangle = \left\langle \frac{b_0(i) + b_1(i)\, s + \dots + b_{n_i-1}(i)\, s^{n_i-1}}{q_0(i) + q_1(i)\, s + \dots + b_{n_i-1}(i)\, s^{n_i-1} + s^{n_i}} \right\rangle \tag{12}$$

where $i = 1,2,\dots, m$.

If the poles of the closed-loop system are pre-assigned then all the coefficients $q_j(i)$ are known. The problem, therefore, is to find $\mathbf{F}$ and $\mathbf{G}$ so that the system may be decoupled (if at all possible) and also the coefficients $b_j(i)$ which will result.

Derivation of necessary conditions

As is well known, the closed-loop transfer function matrix is given by

$$\mathbf{D}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A} - \mathbf{BF})^{-1}\, \mathbf{BG}. \tag{13}$$

Since $\mathbf{G}$ and $\mathbf{D}(s)$ are nonsingular, the above equality may be converted to the following form making use of the inverse of a modified matrix,*

$$(\mathbf{G}\, \mathbf{D}^{-1}(s)\, \mathbf{C} + \mathbf{F})\, (sI - \mathbf{A})^{-1}\, \mathbf{B} = \mathbf{I}. \tag{14}$$

It follows that (noting that $\mathbf{B}^T\mathbf{B} = \mathbf{I}$),

$$\mathbf{G}\, \mathbf{D}^{-1}(s)\, \mathbf{C} + F = \mathbf{B}^T(s\mathbf{I} - \mathbf{A}) + \mathbf{GH}(s)\, \mathbf{Y}(s) \tag{15}$$

where $\mathbf{Y}(s)$ is a matrix of full rank satisfying the equation

$$\mathbf{Y}(s)(s\mathbf{I} - \mathbf{A})^{-1}\, \mathbf{B} = 0 \tag{16}$$

and $\mathbf{H}(s)$ is, for the time being, an arbitrary $m \times (n-m)$ matrix.

* The derivation of Eq. (14) is given in Appendix 1.

The problem is then to find $\mathbf{G}$ and $\mathbf{H}(s)$ in such a way that $\mathbf{F}$ is a constant matrix.

Define

$$\mathbf{P} = \mathbf{D}^{-1}(s)\,\mathbf{C} - \mathbf{G}^{-1}\mathbf{B}^T s - \mathbf{H}(s)\mathbf{Y}(s). \tag{17}$$

Then

$$\mathbf{F} = -\mathbf{GP} - \mathbf{B}^T\mathbf{A}. \tag{18}$$

It follows from Eq. (18) that $\mathbf{P}$ must also be a constant matrix. Also, it is easy to see that

$$-\mathbf{B}^T\mathbf{A} = [\mathbf{a}^T(ij)]. \tag{19}$$

In order to find the $(n-m)\times n$ matrix $\mathbf{Y}(s)$ satisfying Eq. (16), one may again use the inverse of a modified matrix.* Hence

$$\mathbf{Y}(s)\langle s\mathbf{I}_{n_i} - \mathrm{J}_{n_i}\rangle^{-1}\mathbf{B} = 0. \tag{20}$$

The structure of $\mathbf{Y}(s)$ is of the following form

$$\mathbf{Y}(s) = \langle \mathbf{Y}_i(s)\rangle = \Big\langle \underbrace{\begin{bmatrix} s & -1 & 0 & \cdots & \\ 0 & s & -1 & 0 & \\ \vdots & & \ddots & \ddots & \\ & & & \ddots & \ddots \\ 0 & \cdots & & s & -1 \end{bmatrix}}_{n_i} \Big\}\, n_i - 1 \Big\rangle \tag{21}$$

where $i = 1, 2, \ldots, m$.

Returning to Eq. (17), the matrix $\mathbf{P}$ may be partitioned in the same way as $\mathbf{C}$ and $\mathbf{B}^T$, i.e.,

$$\mathbf{P} = [\mathbf{p}^T(ij)]. \tag{22}$$

For a fixed i, we have

$$\mathbf{P}^T(ij) = \frac{Q_i(s)}{R_i(s)}\,\mathbf{c}^T(ij) - sg_{ij}^{(-1)}\,\mathbf{e}_{n_j}^T - \mathbf{h}^T(ij)\,\mathbf{Y}_j(s) \tag{23}$$

where $g_{ij}^{(-1)}$ are the elements of $\mathbf{G}^{-1}$ and $\mathbf{h}^T(ij)$ are rows of $\mathbf{H}$ partitioned in a way similar to $\mathbf{P}$ in Eq. (22) but the rows now have $n_j - 1$ elements.

* The derivation of Eq. (20) is described in detail in Appendix 2.

Since

$$\mathbf{h}^T(ij)\,\mathbf{Y}_j(s) + s g_{ij}^{(-1)}\,\mathbf{e}_{n_j}^T = [h_1(ij)\ldots h_{n_j-1}(ij)\,g_{ij}^{(-1)}] \begin{bmatrix} s & -1 & & & \\ & s & -1 & & \\ & & s & & \\ & & & \ddots & \\ & & & & s\ -1 \end{bmatrix} \tag{24}$$

we may write Eq. (23) as

$$\mathbf{P}^T(ij) = \frac{Q_i(s)}{R_i(s)}\,\mathbf{c}^T(ij) - [h_1(ij)\ \ldots\ h_{n_j-1}(ij)\ g_{ij}^{(-1)}]\ (s\mathbf{I}_{n_j} - \mathbf{J}_{n_j}). \tag{25}$$

Postmultiplying Eq. (25) by

$$(s\mathbf{I}_{n_j} - \mathbf{J}_{n_j})^{-1} = \frac{\mathbf{I}}{s}\mathbf{I}_{n_j} + \frac{1}{s^2}\mathbf{J}_{n_j} + \frac{1}{s^3}\mathbf{J}_n^2 + \ldots + \frac{1}{s^n j}\mathbf{J}_{nj}^{n_j-1} \tag{26}$$

which is an upper triangular matrix of Toeplitz type, and taking the last element in each term of (25), we obtain the following equation

$$p_1(ij) + p_2(ij)s + \ldots + p_{n_j}(ij)s^{n_j-\mathbf{1}} =$$

$$= \frac{Q_i(s)}{R_i(s)}\,(c_1(ij) + c_2(ij)\,s + \ldots + c_{n_j}(ij)\,s^{n_j-1}) - g_{ij}^{(-1)}\,s_{n_j}. \tag{27}$$

Multiplying by $R_i(s)$, we obtain an identity in s. Comparing the coefficients we get $n_i + n_j$ equations for $p_k(ij)$ and $c_k(ij)$, $k = 1,2,\ldots,n_j$, which can be written in the following matrix form

$$\begin{array}{c} \\ n_i + n_j \end{array} \overbrace{\begin{bmatrix} 1 & 0 & 0 & & \\ q_{n_i-1}(i) & 1 & 0 & & \\ q_{n_i-2}(i) & q_{n_i-1}(i) & 1 & & \\ \vdots & \vdots & & \ddots & \\ \vdots & \vdots & & & \ddots \\ \vdots & \vdots & & & & 1 \\ q_1(i) & q_2(i) & & & & q_{n_i-1}(i) \\ q_0(i) & q_2(i) & & & & q_{n_i-2}(i) \\ 0 & q_0(i) & & & & \vdots \\ \vdots & \vdots & & & & \vdots \\ \vdots & \vdots & & & & \vdots \\ & & & & & q_0(i) \end{bmatrix}}^{n_j} \overbrace{\begin{bmatrix} 0 & 0 & & \\ b_{n_i-1}(i) & 0 & & \\ b_{n_i-2}(i) & b_{n_i-1}(i) & & \\ \vdots & \vdots & \ddots & \\ \vdots & \vdots & & \ddots \\ \vdots & \vdots & & & 0 \\ b_1(i) & b_2(i) & & & b_{n_i-1}(i) \\ b_0(i) & b_1(i) & & & b_{n_i-2}(i) \\ 0 & b_0(i) & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ & & & & b_0(i) \end{bmatrix}}^{n_j} \times$$

$$\times \begin{bmatrix} c_{n_j}(ij) \\ c_{n_j-1}(ij) \\ \vdots \\ c_1(ij) \\ -p_{n_j}(ij) \\ -p_{n_j-1}(ij) \\ \vdots \\ -p_1(ij) \end{bmatrix} = \begin{bmatrix} b_{n_i-1}(i) \\ b_{n_i-2}(i) \\ \\ \\ b_0(i) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} g_{ij}^{(-1)} . \tag{28}$$

When writing this system of equations for all j, the first equations of these systems may be collected to give

$$[c_{n_1}(i1)\, c_{n_2}(i2) \ldots c_{n_m}(im)] = b_{n_i-1}(i)\, [g_{i1}^{(-1)}\, g_{i2}^{(-1)} \ldots g_{im}^{(-1)}] . \tag{29}$$

Now there are two possibilities: either all of the elements $c_{n_j}(ij)$, $j = 1, 2, \ldots, m$ are zero, or at least one of them is nonzero. For the latter case, $b_{n_i-1}(i)$ is different from zero, and, without any loss in generality, we may consider it equal to one, i.e.,

$$b_{n_i-1}(i) = 1 . \tag{30}$$

For the former case, we may introduce the term "the degree of simplicity".

Definition 1. If $c_{n_j-k}(ij) = 0$ for $j = 1,2, \ldots, m$, and $k = 0,1,2, \ldots, \alpha_i-1$, but at least one of the elements $c_{n_j-\alpha_i}(ij)$ is nonzero then α_i will be called the degree of simplicity corresponding to the ith row of the matrix **C**.

From Eq. (28) it is obvious that if the degree of simplicity is α_i then

$$b_{n_i-1}(i)\, b_{n_i-2}(i) = \ldots = b_{n_i-\alpha_i}(i) = 0 . \tag{31}$$

By putting

$$b_{n_i-\alpha_i-1}(i) = 1 \tag{32}$$

there is no loss in generality in this case. Consequently,

$$g_{ij}^{(-1)} = c_{n_j-\alpha_i}(ij) . \tag{33}$$

Hence, we get a necessary condition for the system to be decoupled, i.e., the matrix $[c_{n_j-\alpha_i}(ij)]$, $i, j = 1,2, \ldots, m$ must be nonsingular. In that

case

$$\mathbf{G} = [c_{n_j-\alpha_i}(ij)]^{-1}. \tag{34}$$

The result can be formulated as the following

Theorem 1. A necessary condition for a multivariable system to be decoupled by linear state feedback and constant linear transformation of the input is that the matrix $[c_{n_j-\alpha_i}(ij)]$ where $i, j = 1,2, \ldots, m$, is nonsingular, where **A**, **B** and **C** are in the controllable canonical form and α_i are the degrees of simplicity defined in Definition 1.

As Eq. (34) gives us the matrix **G**, the next step is to determine **P** from Eq. (28) in order to determine **F** according to Eq. (18). The difficulty of solving the system of Eq. (28) is that the coefficients $b_k(i)$ are also unknown. Therefore, we shall first eliminate $p_k(ij)$ from the equations in order to get further information about $b_k(i)$.

There are two cases to be distinguished, depending upon whether $n_j < n_i$ or $n_j \geq n_i$. First we will prove that in the former case a further necessary condition can be obtained for the system to be decoupled. This will be formulated as the following theorem.

Theorem 2. A necessary condition for a multivariable system to be decoupled by linear state feedback and constant linear transformation of the input is that

$$c_k(ij) = 0 \quad \text{if} \quad n_j < n_i \qquad \text{for} \quad k = 1,2, \ldots, n_j - \alpha_i. \tag{35}$$

Proof. If the degree of simplicity for the ith row of matrix **C** is equal to α_i and $n_j < n_i$ then Eq. (28) can be written as follows:

$$n_i - n_j + \alpha_i \left\{ \begin{array}{c} \overbrace{\hphantom{1 \quad 0 \quad \ldots}}^{n_j - \alpha_i} \quad \overbrace{\hphantom{1 \quad 0 \quad 0}}^{n_j + 1} \\ \left[\begin{array}{ccc|ccc} 1 & 0 & \ldots & 1 & 0 & 0 \\ q_{n_i-1}(i) & 1 & & b_{n_i-\alpha_i-2}(i) & 1 & \\ & \ddots & & & & \\ q_{n_i-2}(i) & q_{n_i-1}(i) & \ddots & b_{n_i-\alpha_i-3}(i) & b_{n_i-\alpha_i-2}(i) & \\ \vdots & \vdots & 1 & \vdots & b_{n_i-\alpha_i-3}(i) & \\ q_1(i) & q_2(i) & q_{n_i-1}(i) & b_1(i) & & \\ q_0(i) & q_1(i) & q_{n_i-2}(i) & b_0(i) & b_1(i) & 0 \\ 0 & q_0(i) & \cdot & \cdot & b_0(i) & 1 \\ \cdot & \cdot & \cdot & \cdot & \vdots & b_{n_i-\alpha_i-2}(i) \\ \cdot & \cdot & \cdot & \cdot & \cdot & \vdots \\ \cdot & \cdot & q_0(i) & 0 & 0 & b_0(i) \end{array}\right] \end{array} \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ c_{n_j-\alpha_i-1}(ij) \\ \vdots \\ c_1(ij) \\ -c_{n_j-\alpha_i}(ij) \\ -p_{n_j}(ij) \\ -p_{n_j-1}(ij) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ -p_1(ij) \end{bmatrix} = 0. \tag{36}$$

Here the vector on the right-hand side of Eq. (28) was subtracted from both sides and included as the $n_j - \alpha_i + 1$th column of the coefficient matrix.

Let us form a lower triangular matrix of the Toeplitz type of order $n_i + n_j - \alpha_i$ with the coefficients of the polynomial $R_i(s)$, as given below

$$\mathbf{R}(n_i + n_j - \alpha_i) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ b_{n_i-\alpha_{i-2}}(i) & 1 & & \vdots \\ b_{n_i-\alpha_{i-3}}(i) & b_{n_i-\alpha_{i-2}}(i) & & \vdots \\ \vdots & \vdots & & \vdots \\ b_0(i) & b_1(i) & & \vdots \\ & b_0(i) & & \vdots \\ 0 & 0 & & \vdots \\ 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & 1 & 0 \\ 0 & & b_{n_i-\alpha_{i-2}}(i) & 1 \end{bmatrix}. \tag{37}$$

Premultimultiplying Eq. (36) by the inverse of the matrix (37) and taking into account the commutativity of two (lower) triangular matrices of the Toeplitz type we have

$$\begin{bmatrix} \begin{bmatrix} 1 & 0 & & 0 \\ q_{n_i-1}(i) & 1 & & \vdots \\ q_{n_i-2}(i) & q_{n_i-1}(i) & \ddots & \vdots \\ \vdots & & & \vdots \\ q_1(i) & q_2(i) & & \vdots \\ q_0(i) & q_1(i) & & \vdots \\ 0 & q_0(i) & & \vdots \\ 0 & & & \vdots \\ \vdots & & 1 & 0 \\ 0 & 0 & q_{n_i-1}(i) & 1 \end{bmatrix} \begin{bmatrix} \\ \mathbf{R}^{-1(n_j-\alpha_i)}_{(n_i+n_j-\alpha_i)} \\ \\ \end{bmatrix} & \vdots & \mathbf{I}^{(n_j+1)}_{n_i+n_j-\alpha_i} \end{bmatrix} \begin{bmatrix} c_{n_i-\alpha_i}(ij) \\ c_{n_j-\alpha_{i-1}}(ij) \\ \vdots \\ c_1(ij) \\ -c_{n_j-\alpha_i}(ij) \\ -p_{n_j}(ij) \\ -p_{n_j-1}(ij) \\ \vdots \\ -p_1(ij) \end{bmatrix} = 0, \tag{38}$$

where $\mathbf{R}^{-1(n_j-\alpha_i)}_{(n_i+n_j\alpha_i)}$ denotes the first $n_j - \alpha_i$ columns of $\mathbf{R}^{-1}(n_i + n_j - \alpha_i)$ and $\mathbf{I}^{(n_j+1)}_{n_i+n_j-\alpha_i}$ denotes the first $n_j + 1$ columns of an identity matrix of order $n_i + n_j - \alpha_i$. Taking the

last $n_i - \alpha_i - 1$ equations from (38) we get an equation for $c_k(ij)$, $k = 1, 2, \ldots, n_j - \alpha_i$ as follows

$$\begin{bmatrix} q_{n_i-n_j-1}(i) & q_{n_i-n_j}(i) & \cdots & 1 & & 0 & \cdots \\ \vdots & \vdots & & \ddots & & & \vdots \\ q_0(i) & q_1(i) & & & \ddots & & \vdots \\ 0 & q_0(i) & & & & 0 & 0 \\ \vdots & \vdots & & & & 1 & 0 \\ 0 & 0 & & & q_{n_i-1}(i) & & 1 \end{bmatrix} \begin{bmatrix} \mathbf{R}^{-1(n_j-\alpha_i)}_{(n_i+n_j-\alpha_i)} \end{bmatrix} \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ c_{n_j-\alpha_i-1}(ij) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ c_1(ij) \end{bmatrix} = 0. \quad (39)$$

After some algebraic manipulations,* it can be shown that Eq. (39) leads to

$$\mathbf{T}\mathbf{R}^{-1}(n_j - \alpha_i) \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ \vdots \\ c_1(ij) \end{bmatrix} = 0, \quad (40)$$

where $\mathbf{T}$ is the submatrix formed from the last $n_j - \alpha_i - 1$ columns of the matrix $\tilde{\mathbf{T}} = \mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_4^{-1}\mathbf{W}_3$ obtained from the indicated partitioning of the following matrix:

$$\mathbf{W} = \left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{W}_2 \\ \hline \mathbf{W}_3 & \mathbf{W}_4 \end{array}\right] =$$

$$= \left[\begin{array}{cccc|cccccc} q_0(i) & q_1(i) & q_2(i) & \cdots & \cdots & q_{n_i-1}(i) & 1 & 0 & 0 & \\ 0 & q_0(i) & q_1(i) & \cdots & \cdots & q_{n_i-2}(i) & q_{n_i-1}(i) & 1 & 0 & \\ \vdots & \ddots & \ddots & & & & & & & \\ & & & q_0(i) & q_1(i) & & & & & 1 \\ \hline b_0(i) & b_1(i) & \cdots & b_{n_i-\alpha_i-2}(i) & 1 & 0 & 0 & \cdots & 0 & \\ 0 & b_0(i) & \cdots & b_{n_i-\alpha_i-3}(i) & b_{n_i-\alpha_i-2}(i) & 1 & 0 & \cdots & & \\ \vdots & \ddots & \ddots & & & \ddots & \ddots & & & \\ & & b_0(i) & & & & & & & \\ & & & & 0 & & & & & 1 \end{array}\right] \begin{matrix} \} n_i - \alpha_i - 1 \\ \\ \\ \\ \\ \\ \\ \} n_i \\ \\ \\ \\ \\ \\ \\ \end{matrix} \quad (41)$$

$$\underbrace{\qquad\qquad}_{n_i - \alpha_i - 1} \qquad \underbrace{\qquad\qquad}_{n_i}$$

* These are described in detail in Appendix 3.

It may be shown* that the matrix $\tilde{\mathbf{T}}$ is always nonsingular unless the rational functions $\frac{R_i(s)}{Q_i(s)}$ in $\mathbf{D}(s)$ are reducible, which may be ruled out if the system is to be both controllable and observable. Consequently, the columns of $\mathbf{T}$, being columns of the nonsingular matrix $\tilde{\mathbf{T}}$ are always linearly independent. Thus, Eq. (40) implies

$$\lfloor c_k(ij) = 0, \quad k = 1, 2, \ldots, n_j - \alpha_i. \tag{42}$$

This proves the theorem.

Corollary. From Eq. (38) it follows that $p_\mu(ij)$ $\mu = 1,2, \ldots n_i$ are homogenous linear functions of $c_k(ij)$, $k = 1,2, \ldots, n_j - \alpha_i$. Hence, they are all zeros if the condition of Theorem 2 is satisfied.

We shall now consider the second case, when $n_j \geq n_i$. Starting with an equation corresponding to (36), one may proceed as before and obtain the following equation corresponding to (39),

$$\begin{bmatrix} 0 & 0 & \ldots & 0 & q_0(i) & q_1(i) & \ldots & q_{n_i-1}(i) & 1 & & 0 & \ldots & 0 \\ & & & & 0 & q_0(i) & \ldots & & q_{n_i-1}(i) & 1 & & & \\ & & & & & \ddots & & & \ddots & \ddots & \ddots & & \\ & & & & & & \ddots & & & \ddots & \ddots & \ddots & \\ & & & & & & & q_0(i) & & & q_{n_i-1}(i) & 1 \end{bmatrix} \times$$

$$\times \; \mathbf{R}^{-1(n_j-\alpha_i)}_{(n_i+n_j-\alpha_i)} \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ c_{n_j-\alpha_i-1}(ij) \\ \vdots \\ c_1(ij) \end{bmatrix} = 0. \tag{43}$$

It can be shown** that Eq. (43) leads to

$$\tilde{\mathbf{T}}\, \tilde{\mathbf{R}}^{-1(n_i-\alpha_i-1)}_{(n_j-\alpha_i)} \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ \vdots \\ c_1(ij) \end{bmatrix} = 0, \tag{44}$$

where $\tilde{\mathbf{R}}^{-1(n_i-\alpha_i-1)}_{(n_j-\alpha_i)}$ is a rectangular matrix consisting of the last $n_i - \alpha_i - 1$ rows of $\mathbf{R}^{-1}(n_j - \alpha_i)$.

* The details are given in Appendix 4.
** This is discussed in Appendix 5.

Since $\tilde{\mathbf{T}}$ is always nonsingular, Eq. (44) implies that

$$\tilde{\mathbf{R}}^{-1(n_i-\alpha_i-1)}_{(n_j-\alpha_i)} \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ \cdot \\ \cdot \\ \cdot \\ c_1(ij) \end{bmatrix} = 0. \tag{45}$$

This equation gives necessary conditions if $n_j \geq n_i$. Equation (45) may now be used to determine the coefficients $b_k(i)$ since $c_k(ij)$ are known. For further investigation let us introduce the term "degree of deficiency".

Definition 2. If $c_k(ij) = 0$ for $j = 1,2, \ldots, m$ and $k = 1,2, \ldots, \beta_i$, but at least one of the elements $c_{\beta_i+1}(ij)$ is nonzero, then β_i will be called the degree of deficiency corresponding to the ith row of the matrix $\mathbf{C}$.

Since Eq. (45) holds for all j where $n_j \geq n_i$, in order to obtain the coefficients $b_k(i)$, let us consider the case where $j = i$ and the degree of deficiency is β_i. Equation (45) becomes

$$\tilde{\mathbf{R}}^{-1(n_i-\alpha_i-1)}_{(n_i-\alpha_i)} \begin{bmatrix} c_{n_i-\alpha_i}(ii) \\ c_{n_i-\alpha_i-1}(ii) \\ \cdot \\ \cdot \\ \cdot \\ c_{\beta_i+1}(ii) \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix} = 0. \tag{46}$$

The partitioned form of $\tilde{\mathbf{R}}^{-1(n_i-\alpha_i-1)}_{(n_i-\alpha_i)}$ is then

$$\tilde{\mathbf{R}}^{-1(n_i-\alpha_i-1)}_{(n_i-\alpha_i)} = \left[\begin{array}{c|c} -\mathbf{R}^{-1}(n_i-\alpha_i-1) \begin{bmatrix} b_{n_i-\alpha_i-2}(i) \\ \vdots \\ b_{\beta_i}(i) \\ \vdots \\ b_0(i) \end{bmatrix} & \mathbf{R}^{-1}(n_i-\alpha_i-1) \end{array}\right]. \tag{47}$$

Premultiplying by $\mathbf{R}^{-1}(n_i - \alpha_i - 1)$ and substituting in (46) we obtain

$$\begin{bmatrix} -b_{n_i-\alpha_i-2}(i) & 1 & & \\ \vdots & & \ddots & \\ -b_{\beta_i}(i) & & & 1 \\ \vdots & & & \\ -b_0(i) & & & \end{bmatrix} \begin{bmatrix} c_{n_i-\alpha_i}(ii) \\ \vdots \\ c_{\beta_i+1}(ii) \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0. \tag{48}$$

Hence

$$\begin{bmatrix} b_{n_i-\alpha_i-2}(i) \\ \vdots \\ b_{\beta_i}(i) \\ b_{\beta_i-1}(i) \\ \vdots \\ b_0(i) \end{bmatrix} c_{n_i-\alpha_i}(ii) = \begin{bmatrix} c_{n_i-\alpha_i-1}(ii) \\ \vdots \\ c_{\beta_i+1}(ii) \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{49}$$

Theorems 1 and 2 imply that

$$c_{n_i-\alpha_i}(i) \neq 0,$$

otherwise the matrix $[c_{n_j-\alpha_i}(ij)]$ becomes singular. Thus from Eq. (49) we have

$$b_0(i) = b_1(i) = \ldots = b_{\beta_i-1}(i) = 0 \tag{50}$$

and

$$b_k(i) = \frac{c_{k+1}(ii)}{c_{n_i-\alpha_i}(ii)} \quad \text{for} \quad k = \beta_i,\ \beta_i + 1, \ldots, n_i - \alpha_i - 2\,. \tag{51}$$

In particular, if $\beta_i = n_i - \alpha_i - 1$, then

$$b_k(i) = 0 \ \text{ for } \ k = 0,1,2, \ldots, n_i - \alpha_i - 2\,,$$

i.e.

$$\mathbf{R}(n_j - \alpha_i) = \mathbf{I}_{n_j-ii}$$

and Eq. (45) is trivially satisfied.

Since we already know the coefficients of Eq. (45), now we have to determine the relationship among the quantities $c_k(ij)$ (for $n_j \geq n_i$) so that the solution of Eq. (45) exists (third necessary condition of decoupling and pole placement).

Equation (45) can be written in the following partitioned form:

$$\begin{bmatrix} \mathbf{R}(n_j - n_i + 1) & 0 & 0 \\ \mathbf{B}_4 & \mathbf{R}(n_i - \alpha_i - \beta_i - 1) & 0 \\ 0 & \mathbf{B}_5 & \mathbf{R}(\beta_i) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{c}(n_j - n_i + 1) \\ \mathbf{c}(n_i - \alpha_i - \beta_i - 1) \\ 0 \end{bmatrix} = 0 \tag{52}$$

where

$$\mathbf{B}_4 = \begin{bmatrix} b_{2n_i - n_j - \alpha_i - 2}(i) & \cdots & b_{n_i - \alpha_i - 2}(i) \\ \vdots & & \vdots \\ b_{\beta_i}(i) & \ddots & \vdots \\ & \ddots & \vdots \\ & & b_{\beta_i}(i) \end{bmatrix}, \tag{53}$$

$$\mathbf{B}_5 = \begin{bmatrix} b_{\beta_i}(i) & \cdots & b_{n_i - \alpha_i - 2}(i) \\ \ddots & & \\ & b_{\beta_i}(i) \cdots & b_{n_i - \alpha_i - \beta_i - 1}(i) \end{bmatrix} \tag{54}$$

and

$$\mathbf{c}(n_j - n_i + 1) = \begin{bmatrix} c_{n_j - \alpha_i}(ij) \\ \vdots \\ c_{n_i - \alpha_i}(ij) \end{bmatrix}; \quad \mathbf{c}(n_i - \alpha_i - \beta_i - 1) = \begin{bmatrix} c_{n_i - \alpha_i - 1}(ij) \\ \vdots \\ c_{\beta_i + 1}(ij) \end{bmatrix}. \tag{55}$$

The partitioned form of the inverse in (52) has the following form:

$$\begin{bmatrix} \mathbf{R}^{-1}(n_j - n_i + 1) & & \\ -\mathbf{R}^{-1}(n_i - \alpha_i - \beta_i - 1)\,\mathbf{B}_4\,\mathbf{R}^{-1}(n_j - n_i + 1) & & \times \\ \mathbf{R}^{-1}(\beta_i)\,\mathbf{B}_5\,\mathbf{R}^{-1}(n_i - \alpha_i - \beta_i - 1)\,\mathbf{B}_4\,\mathbf{R}^{-1}(n_j - n_i + 1) & & \end{bmatrix}$$

$$\times \begin{bmatrix} 0 & 0 \\ \mathbf{R}^{-1}(n_i - \alpha_i - \beta_i - 1) & 0 \\ -\mathbf{R}^{-1}(\beta_i)\,\mathbf{B}_5\,\mathbf{R}^{-1}(n_i - \alpha_i - \beta_i - 1) & \mathbf{R}^{-1}(\beta_i) \end{bmatrix}; \tag{56}$$

Substituting into Eq. (52) and taking into consideration that the first two blocks of the third row can be obtained from the first two blocks of the second row by premultiplying them with $-\mathbf{R}^{-1}(\beta_i)\,\mathbf{B}_5$, Eq. (52) leads to the

following equation

$$[-\mathbf{R}^{-1}(n_i - \alpha_i - \beta_i - 1)\,\mathbf{B}_4\,\mathbf{R}^{-1}(n_j - n_i + 1)\;\; \mathbf{R}^{-1}(n_i - \alpha_i - \beta_i - 1)] \times \times \begin{bmatrix} \mathbf{c}(n_j - n_i + 1) \\ \mathbf{c}(n_i - \alpha_i - \beta_i - 1) \end{bmatrix} = 0\,. \tag{57}$$

Now, let us denote by s_k, $k = 1,2, \ldots, n_i - \alpha_i - \beta_i - 1$ the nonzero roots of the polynomial

$$R_i(s) \equiv s^{n_i-\alpha_i-1} + b_{n_i-\alpha_i-2}(i)\, s^{n_i-\alpha_i-2} + \ldots + b_{\beta_i}\, s^{\beta_i}\,. \tag{58}$$

The matrix $\mathbf{R}(n_j - \alpha_i - \beta_i)$ can be written as

$$\mathbf{R}(n_j - \alpha_i - \beta_i) = \prod_{k=1}^{n_i-\alpha_i-\beta_i-1} (\mathbf{I}_{n_j-\alpha_i-\beta_i} - s_k \mathbf{J}^T_{n_j-\alpha_i-\beta_i}) \tag{59}$$

and its inverse is obtained as

$$\mathbf{R}^{-1}(n_j - \alpha_i - \beta_i) = \prod_{k=1}^{n_i-\alpha_i-\beta_i-1} \sum_{\nu=0}^{n_j-\alpha_i-\beta_i-1} s_k^{\nu} (\mathbf{J}^T_{n_j-\alpha_i-\beta_i})^{\nu}\,. \tag{60}$$

Substituting this expression into Eq. (52) and taking into consideration that it leads to Eq. (57), further introducing the polynomials

$$L_{ij}(s) = \sum_{\mu=1}^{n_j-\alpha-\beta_i} c_{\beta_i+\mu}(ij)\, s^{\mu-1} \tag{61}$$

it is not difficult to see* that Eq. (57) implies

$$L_{ij}(s_k) = 0 \qquad \begin{matrix} k = 1, 2, \ldots, n_i - \alpha_i - \beta_i - 1 \\ j = 1, 2, \ldots, m. \end{matrix} \tag{62}$$

Hence, a necessary condition for decoupling is that $L_{ij}(s)$ have $\dfrac{L_{ii}(s)}{c_{n_i-\alpha_i}(ii)}$ as a factor for each given i and all j.

The polynomials $R_i(s)$ are obtained from (49) as

$$R_i(s) = \frac{L_{ii}(s)}{c_{n_i-\alpha_i}(ii)}\, s^{\beta_i}. \tag{63}$$

* This is described in detail in Appendix 6.

Hence, from Eq. (38), making use of Eq. (45) we obtain*

$$\begin{bmatrix} p_{n_j}(ij) \\ p_{n_j-1}(ij) \\ \vdots \\ p_1(ij) \end{bmatrix} = \tag{64}$$

$$= \begin{bmatrix} q_{n_i-1}(i) & 1 & 0 \cdots & 0 \\ q_{n_i-2}(i) & q_{n_i-1}(i) & 1 & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & & \ddots \\ \vdots & \vdots & q_{n_i-1}(i) & 1 \\ q_0(i) & q_1(i) & \cdots & q_{n_i-1}(i) \\ 0 & q_0(i) & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \vdots & & q_0(i) \end{bmatrix} \cdot \mathbf{R}^{-1}(n_j - n_i + 1) \cdot \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ c_{n_j-\alpha_i-1}(ij) \\ \vdots \\ c_{n_i-\alpha_i}(ij) \end{bmatrix}.$$

($n_j - n_i + 1$ columns)

Now we can formulate the main theorem on pole placement and decoupling.

Theorem 3. The system equations in the given canonical form can be decoupled and all poles arbitrarily placed if and only if

(i) the condition of Theorem 1 holds

(ii) the condition of Theorem 2 holds for elements of $\mathbf{C}$ for which $n_j < n_i$

* This is described in detail in Appendix 7.

(iii) the polynomials $L_{ij}(s)$ constructed from the elements of **C** which $n_j > n_i$, according to Eq. (61), have $L_{ii}(s)$ as a factor for each given i and all j.

Remark. The last condition is trivially satisfied for the case

$$\beta_i = n_i - \alpha_i - 1.$$

The sufficiency of these conditions follows from the fact that once the $R_i(s)$ are obtained from Eq. (63) and the $p_k(ij)$ are obtained from Eq. (64) then the matrix **F** is directly obtained from Eq. (18).

Description of the algorithm

A step-by-step procedure will now be discussed.

(1) It will be assumed that the matrices **A** and **B** are in the given canonical form. If not, Luenberger's procedure may be followed for the desired transformation.

(2) From the canonical form, we get the orders $n_1, n_2, \ldots, n_m$ directly, and also obtain the appropriate partitioning of **C**.

(3) Using this partitioning, the values of the degree of simplicity, α_i, and the degree of deficiency, β_i can be seen immediately for each i.

(4) The matrix $[c_{n_j-\alpha_i}(ij)]$ is now formed and checked for nonsingularity (first necessary condition). The inverse gives the matrix **G**.

(5) The elements $c_k(ij)$ should be zero for $k = 1,2, \ldots, n_j - \alpha_i$ if $n_j < n_i$ (second necessary condition).

(6) The polynomials $L_{ij}(s)$ are constructed from the elements of **C** for which $n_j \geq n_i$, according to Eq. (61). The divisibility, for each i of $L_{ij}(s)$ is then checked for all j (third necessary condition).

Evidently, this condition is trivially satisfied if, for particular i, $n_i = \alpha_i + \beta_i + 1$.

(7) The coefficients, $b_k(i)$, of the numerator polynomial $R_i(s)$ are now obtained from Eq. (63).

(8) Since all the coefficients $b_k(i)$ are known, Eq. (64) may now be used to determine the elements of the matrix **P**.

(9) Finally, Eq. (18) provides the feedback matrix **F** in a straightforward manner.

Example. The following example will illustrate the general procedure.

Let the triple **A**, **B**, **C** be given as

$$\mathbf{A} = \left[\begin{array}{ccc|cccc|cc} 0 & 1 & 0 & & & & & & \\ 0 & 0 & 0 & & & & & & \\ -6 & 5 & 2 & -1 & -2 & -3 & -4 & -5 & -6 \\ \hline & & & 0 & 1 & 0 & 0 & & \\ & 0 & & 0 & 0 & 1 & 0 & & \\ & & & 0 & 0 & 0 & 1 & & \\ & & & -6 & -1 & 7 & 1 & -7 & -8 \\ \hline & & & & & & & 0 & 1 \\ & 0 & & & 0 & & & -2 & 3 \end{array}\right]$$

$$\mathbf{B} = \left[\begin{array}{ccc} 0 & & \\ 0 & & \\ 1 & & \\ & 0 & \\ & 0 & \\ & 0 & \\ & 1 & \\ & & 0 \\ & & 1 \end{array}\right]; \quad \mathbf{C} = \left[\begin{array}{ccc|cccc|cc} 3 & 1 & 0 & 12 & 7 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 3 & 1 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 3 & 4 & 0 & 1 \end{array}\right].$$

It is easy to see that n_i, α_i, β_i are the numbers given below:

$$n_1 = 3,\ \alpha_1 = 1,\ \beta_1 = 0;\quad n_1 - \alpha_1 - \beta_1 - 1 = 1$$
$$n_2 = 4,\ \alpha_2 = 0,\ \beta_2 = 1;\quad n_2 - \alpha_2 - \beta_2 - 1 = 2$$
$$n_3 = 2,\ \alpha_3 = 0,\ \beta_3 = 1;\quad n_3 - \alpha_3 - \beta_3 - 1 = 0.$$

From the structure of the matrix **C** it is evident that the conditions of Theorem 2 are satisfied and since

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 2 & 4 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ -2 & -2 & 1 \end{bmatrix}$$

the condition of Theorem 1 is satisfied as well.

The polynomials $R_i(s)$ can be obtained in the following manner while verifying that the third condition of Theorem 3 is also satisfied.

$$L_{11}(s) = s + 3,\ L_{12}(s) = s^2 + 7s + 12 = (s + 3)(s + 4),\ L_{13}(s) = 0$$
$$R_1(s) = s^{\beta_1}(s + 3) = s + 3$$

$$L_{21}(s) = 0,\ L_{22}(s) = s^2 + 3s + 2,\ L_{23}(s) = 0$$
$$R_2(s) = s^{\beta_2}(s^2 + 3s + 2) = s^3 + 3s^2 + 2s.$$

For $i = 3$, since $n_3 = \alpha_3 + \beta_3 + 1$, $R_3(s) = s^{\beta_3} = s$.

The elements of the matrix **P** can be obtained according to Eq. (64) if the poles are pre-assigned. Let the polynomials $Q_i(s)$ be

$$Q_1(s) = (s+1)^2(s+2) = s^3 + 4s^2 + 5s + 2$$

$$Q_2(s) = (s^2 + s + 1)^2 = s^4 + 2s^3 + 3s^2 + 2s + 1$$

$$Q_3(s) = s^2 + s + 1.$$

Thus

$$\mathbf{p}(13) = \mathbf{p}(21) = \mathbf{p}(23) = 0,$$

further

$$\begin{bmatrix} p_3(11) \\ p_2(11) \\ p_1(11) \end{bmatrix} = \begin{bmatrix} q_2(1) \\ q_1(1) \\ q_0(1) \end{bmatrix} c_{n_1-\alpha_1}(11) = \begin{bmatrix} 4 \\ 5 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} p_4(12) \\ p_3(12) \\ p_2(12) \\ p_1(12) \end{bmatrix} = \begin{bmatrix} q_2(1) & 1 \\ q_1(1) & q_2(1) \\ q_0(1) & q_1(1) \\ 0 & q_0(1) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ b_0(1) & 1 \end{bmatrix}^{-1} \begin{bmatrix} c_{n_2-\alpha_1}(12) \\ c_{n_1-\alpha_1}(12) \end{bmatrix} =$$

$$= \begin{bmatrix} 4 & 1 \\ 5 & 4 \\ 2 & 5 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 7 \end{bmatrix} = \begin{bmatrix} 8 \\ 21 \\ 22 \\ 8 \end{bmatrix}$$

$$\begin{bmatrix} p_4(22) \\ p_3(22) \\ p_2(22) \\ p_1(22) \end{bmatrix} = \begin{bmatrix} q_3(2) \\ q_2(2) \\ q_1(2) \\ q_0(2) \end{bmatrix} c_{n_2-\alpha_2}(22) = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} p_3(31) \\ p_2(31) \\ p_1(31) \end{bmatrix} = \begin{bmatrix} q_1(3) & 1 \\ q_0(3) & q_1(3) \\ 0 & q_0(3) \end{bmatrix} \begin{bmatrix} c_{n_1-\alpha_3}(31) \\ c_{n_3-\alpha_3}(31) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} p_4(32) \\ p_3(32) \\ p_2(32) \\ p_1(32) \end{bmatrix} = \begin{bmatrix} q_1(3) & 1 & 0 \\ q_0(3) & q_1(3) & 1 \\ 0 & q_0(3) & q_1(3) \\ 0 & 0 & q_0(3) \end{bmatrix} \begin{bmatrix} c_{n_2-\alpha_3}(32) \\ c_{n_2-\alpha_3-1}(32) \\ c_{n_3-\alpha_3}(32) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 8 \\ 4 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} p_2(33) \\ p_1(33) \end{bmatrix} = \begin{bmatrix} q_1(3) \\ q_0(3) \end{bmatrix} c_{n_3-\alpha_3}(33) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Thus the matrix **P** is given as

$$\mathbf{P} = \left[\begin{array}{ccc|c|ccc|c|cc} 2 & 5 & 4 & 8 & 22 & 21 & 8 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 & 2 & 0 & 0 \\ 1 & 3 & 3 & 1 & 4 & 8 & 7 & 1 & 1 \end{array}\right]$$

and finally

$$\mathbf{F} = -\mathbf{GP} - \mathbf{B}^T\mathbf{A} = \left[\begin{array}{cccc|cccc|cc} 4 & -10 & -6 & & -6 & -18 & -15 & -2 & 5 & 6 \\ 0 & 0 & 0 & & 5 & -1 & -10 & -3 & 7 & 8 \\ 3 & 7 & 5 & & 17 & 44 & 40 & 13 & 1 & -4 \end{array}\right]$$

It can be verified that

$$\mathbf{D}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A} - \mathbf{BF})^{-1}\mathbf{BG} = \begin{bmatrix} \dfrac{s+3}{(s+3)^2(s+1)} & 0 & 0 \\ 0 & \dfrac{(s+1)(s+2)s}{(s^2+s+1)^2} & 0 \\ 0 & & \dfrac{s}{s^2+s+1} \end{bmatrix}.$$

Conclusions

Necessary and sufficient conditions have been established for decoupling and pole-placement in a linear multivariable system using state feedback and a constant transformation of the input. Although the derivations required considerable amount of matrix manipulations, the final results are very simple and easy to apply. Starting with the controllable canonical representation of the system, all the necessary information is obtained from the matrix **C**. This compares very favourably with the method of Falb and Wolovich.

It may be added the procedure discussed in this paper makes the numerical work required in the problem fairly simple. No ranks of matrices have to be determined. One general matrix of order m has to be inverted while determining the static precompensator; this occurs in other papers as well. All other matrices to be inverted are of the triangular and Toeplitz type, and are generally small so that they can be easily inverted.

The emphasis in this paper has been on finding the necessary and sufficient conditions for only those cases where all the poles of the closed-loop system can be placed arbitrarily, in addition to decoupling. Although this may appear as a limitation, one may argue that this is the case of most practical importance.

Acknowledgement. This work was supported by the cultural Exchange Programme between McMaster University and the Institute of Cultural Relations in Budapest, as well as by the National Research Council of Canada.

Appendix 1

The inverse of a modified matrix is given by

$$(s\mathbf{I} - \mathbf{A} - \mathbf{BF})^{-1} = (s\mathbf{I} - \mathbf{A})^{-1} + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\{\mathbf{I} - \mathbf{F}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\}^{-1}\mathbf{F}(s\mathbf{I} - \mathbf{A})^{-1} \quad (65)$$

which can be easily verified.

Substituting into (13) we obtain

$$\mathbf{D}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A} - \mathbf{BF})^{-1}\mathbf{BG} = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{BG} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}[\mathbf{I} - \mathbf{F}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}]^{-1}$$
$$\mathbf{F}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{BG}. \quad (66)$$

Factoring out $\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$ to the left and $\mathbf{G}$ to the right, respectively, we have

$$\mathbf{D}(s) = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\{\mathbf{I} + [\mathbf{I}-\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}]^{-1}\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\}\mathbf{G}. \tag{67}$$

Taking into consideration that $\mathbf{I} = [\mathbf{I}-\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}]^{-1}[\mathbf{I}-\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}]$, we get

$$\mathbf{D}(s) = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}[\mathbf{I}-\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}]^{-1}\mathbf{G}. \tag{68}$$

Postmultiplying by $\mathbf{G}^{-1}$ then premultiplying by $\mathbf{D}^{-1}(s)$ and $\mathbf{G}$, and finally postmultiplying by $\mathbf{I}-\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$ we obtain

$$\mathbf{I}-\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} = \mathbf{G}\mathbf{D}^{-1}(s)\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}. \tag{69}$$

Taking $\mathbf{F}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$ to the right hand side and factoring out $(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$ to the right, (14) is obtained.

Appendix 2

The matrix $\mathbf{A}$ can be expressed in the form

$$\mathbf{A} = \langle \mathbf{J}_{n_i} \rangle + \mathbf{B}\mathbf{B}^T\mathbf{A}. \tag{70}$$

Thus, for the inverse

$$(s\mathbf{I}-\mathbf{A})^{-1} = \{\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle - \mathbf{B}\mathbf{B}^T\mathbf{A}\}^{-1} \tag{71}$$

the formula for the modified inverse can be applied again:

$$(s\mathbf{I}-\mathbf{A})^{-1} = \langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1} + \langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}\mathbf{B}\{\mathbf{I}-\mathbf{B}^T\mathbf{A}\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}\mathbf{B}\}^{-1}\mathbf{B}^T\mathbf{A}$$
$$\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}. \tag{72}$$

Postmultiplying by $\mathbf{B}$ and factoring out $\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}\mathbf{B}$ to the left, we obtain, as in Appendix 1,

$$(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B} = \langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}\mathbf{B}\{\mathbf{I}-\mathbf{B}^T\mathbf{A}\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}\mathbf{B}\}^{-1}. \tag{73}$$

Substituting into (16) and postmultiplying by $\mathbf{I}-\mathbf{B}^T\mathbf{A}\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}\mathbf{B}$, we obtain Eq. (20). Since $\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle^{-1}\mathbf{B}$ is consisting of the last columns of each block column of $\langle s\mathbf{I}_{n_i} - \mathbf{J}_{u_i}\rangle^{-1}$, a complete set of linearly independent solutions of Eq. (20) are obtained from $\langle s\mathbf{I}_{n_i} - \mathbf{J}_{n_i}\rangle$ by omitting the last rows of each block row. Hence the structure of $\mathbf{Y}(s)$ is of the form shown in Eq. (21).

Appendix 3

In order to get Eq. (40), let the matrices in (39) be partitioned in the following way

$$\overbrace{\qquad\qquad}^{n_j-\alpha_i} \quad \overbrace{\qquad\qquad\qquad}^{n_i-\alpha_i-1} \quad \overbrace{\qquad}^{\alpha_i+1}$$

$$\left.\begin{matrix} \\ \\ n_i-\alpha_i-1 \\ \\ \\ \end{matrix}\right\{ \left[\begin{array}{ccc|ccc|ccc} q_{n_i-n_j-1} & \cdots & q_{n_j-\alpha_i-2} & q_{n_i-\alpha_i-1} & \cdots & 1 & & & \\ \vdots & & & \vdots & \iddots & & & & \\ \vdots & & & \vdots & & 1 & & & \\ q_0 & \iddots & & \vdots & & & 1 & \iddots & \\ & \iddots & & \vdots & & & \iddots & \iddots & \\ & & q_0 & q_1 & \cdots & q_{n_i-\alpha_i-1} & q_{n_i-\alpha_i-1} & \cdots & 1 \end{array}\right] =$$

$$= [\mathbf{Q}_1\ \mathbf{Q}_2\ \mathbf{Q}_3]. \tag{74}$$

$$\mathbf{R}(n_i+n_j-\alpha_i) = \begin{bmatrix} \mathbf{R}(n_j-\alpha_i) & 0 & 0 \\ \mathbf{B}_1 & \mathbf{R}(n_i-\alpha_i-1) & 0 \\ 0 & \mathbf{B}_2 & \mathbf{R}(\alpha_i+1) \end{bmatrix} \tag{75}$$

where

$$\mathbf{B}_1 = \begin{bmatrix} b_{n_i-n_j-1} & \cdots & b_{n_i-\alpha_i-2} \\ \vdots & & \vdots \\ b_0 & \ddots & \vdots \\ & \ddots & b_0 \end{bmatrix} \tag{76}$$

and

$$\mathbf{B}_2 = \begin{bmatrix} b_0 & \cdots & b_{n_i-\alpha_i-2} \\ & \ddots & \vdots \\ & & b_{n_i-2\alpha_i-2} \end{bmatrix}. \tag{77}$$

It is easy to verify that

$$\mathbf{R}^{-1}(n_i+n_l-\alpha_i) = \begin{bmatrix} \mathbf{R}^{-1}(n_j-\alpha_i) & & \\ -\mathbf{R}^{-1}(n_i-\alpha_i-1)\,\mathbf{B}_1\,\mathbf{R}^{-1}(n_j-\alpha_i) & \times & \\ \mathbf{R}^{-1}(\alpha_i+1)\,\mathbf{B}_2\,\mathbf{R}^{-1}(n_i-\alpha_i-1)\,\mathbf{B}_1\mathbf{R}^{-1}(n_j-\alpha_i) & & \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 \\ \times \quad \mathbf{R}^{-1}(n_i-\alpha_i-1) & 0 \\ -\mathbf{R}^{-1}(\alpha_i+1)\,\mathbf{B}_2\,\mathbf{R}^{-1}(n_i-\alpha_i-1) & \mathbf{R}^{-1}(\alpha_i+1) \end{bmatrix}. \tag{78}$$

Hence

$$\overset{(n_j-\alpha_i)}{\mathbf{R}^{-1}(n_i+n_j-\alpha_i)} = \begin{bmatrix} \mathbf{R}^{-1}(n_j-\alpha_i) \\ -\mathbf{R}^{-1}(n_i-\alpha_i-1)\,\mathbf{B}_1\mathbf{R}^{-1}(n_j-\alpha_i) \\ \mathbf{R}^{-1}(\alpha_i+1)\,\mathbf{B}_2\mathbf{R}^{-1}(n_i-\alpha_i-1)\,\mathbf{B}_1\mathbf{R}^{-1}(n_i-\alpha_i) \end{bmatrix}. \tag{79}$$

Multiplying the first two matrices in Eq. (39) and factoring out $\mathbf{R}^{-1}(n_j-\alpha_i)$ to the right, we obtain

$$\overset{(n_j-\alpha_i)}{[\mathbf{Q}_1\,\mathbf{Q}_2\,\mathbf{Q}_3]\,\mathbf{R}^{-1}(n_i+n_j-\alpha_i)} =$$

$$= \{\mathbf{Q}_1 - \mathbf{Q}_2\mathbf{R}^{-1}(n_i-\alpha_i-1)\,\mathbf{B}_1 + \mathbf{Q}_3\mathbf{R}^{-1}(\alpha_i+1)\,\mathbf{B}_2\mathbf{R}^{-1}(n_i-\alpha_i-1)\mathbf{B}_1\}\,\mathbf{R}^{-1}(n_i-\alpha_i)\,. \tag{80}$$

Introducing the notation

$$\mathbf{T} = \mathbf{Q}_1 - \mathbf{Q}_2\,\mathbf{R}^{-1}(n_i-\alpha_i-1)\mathbf{B}_1 + \mathbf{Q}_3\mathbf{R}^{-1}(\alpha_i+1)\,\mathbf{B}_2\mathbf{R}^{-1}(n_i-\alpha_i-1)\,\mathbf{B}_1 \tag{81}$$

it is easy to see that $\mathbf{T}$ can be written in the form

$$\mathbf{T} = \mathbf{Q}_1 - [\mathbf{Q}_2\,\mathbf{Q}_3]\begin{bmatrix} \mathbf{R}(n_i-\alpha_i-1) & 0 \\ \mathbf{B}_2 & \mathbf{R}(\alpha_i+1) \end{bmatrix}^{-1}\begin{bmatrix} \mathbf{B}_1 \\ 0 \end{bmatrix}. \tag{82}$$

Taking into consideration that $\mathbf{Q}_1$ and $\mathbf{B}_1$ consist of the last $n_j-\alpha_i$ columns of the following upper triangular matrices of Toeplitz type

$$\tilde{\mathbf{Q}}_1 = \begin{bmatrix} q_0 & q_1 & \cdots & q_{n_i-\alpha_i-2} \\ 0 & q_0 & \cdots & q_{n_i-\alpha_i-3} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & q_0 \end{bmatrix} \tag{83}$$

and

$$\widetilde{\mathbf{B}}_1 = \begin{bmatrix} b_0 & b_1 & \dots & b_{n_i-\alpha_i-2} \\ 0 & b_0 & \dots & b_{n_i-\alpha_i-3} \\ . & . & . & . \\ 0 & 0 & \dots & b_0 \end{bmatrix} \tag{84}$$

respectively, it is obvious that $\mathbf{T}$ is consisting of the last $n_j - \alpha_i$ columns of the matrix $\widetilde{\mathbf{T}}$:

$$\widetilde{\mathbf{T}} = \widetilde{\mathbf{Q}}_1 - [\mathbf{Q}_2\, \mathbf{Q}_3] \begin{bmatrix} \mathbf{R}(n_i - \alpha_i - 1) & 0 \\ \mathbf{B}_2 & \mathbf{R}(\alpha_i + 1) \end{bmatrix}^{-1} \begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ 0 \end{bmatrix}. \tag{85}$$

Now, constructing the block matrix $\mathbf{W}$ in Eq. (41):

$$\mathbf{W} = \left[\begin{array}{c|cc} \widetilde{\mathbf{Q}}_1 & \mathbf{Q}_2 & \mathbf{Q}_3 \\ \hline \widetilde{\mathbf{B}}_1 & \mathbf{R}(n_i - \alpha_i - 1) & 0 \\ 0 & \mathbf{B}_2 & \mathbf{R}(\alpha_i + 1) \end{array}\right] = \left[\begin{array}{c|c} \mathbf{W}_1 & \mathbf{W}_3 \\ \hline \mathbf{W}_3 & \mathbf{W}_4 \end{array}\right] \tag{86}$$

we see that really

$$\widetilde{\mathbf{T}} = \mathbf{W}_1 - \mathbf{W}_2 \mathbf{W}_4^{-1} \mathbf{W}_3 . \tag{87}$$

Appendix 4

Let s_k be the zeros of the polynomial

$$R_i(s) \equiv b_0(i) + b_1(i)\, s + \dots + b_{n_i-\alpha_i-2}(i)\, s^{n_i-\alpha_i-2} + s^{n_i-\alpha_i-2} \tag{88}$$

and let the matrix $\mathbf{W}$ in (41) be postmultiplied by the vectors

$$s_k = [1\ s_k\ s_k^2 \dots s_k^{2n_i-\alpha_i-2}]^T . \tag{89}$$

There are two cases to be distinguished.

(i) First let us assume that the polynomial $R_i(s)$ has $n_i - \alpha_i - 1$ distinct zeros. Let us form a matrix $\mathbf{S}$ of the vectors s_k and let $\mathbf{S}$ be partitioned into two blocks consisting of the first $n_i - \alpha_i - 1$ rows and of the last n_i rows respectively:

$$\mathbf{S} = [s_1\, s_2 \dots s_{n_i-\alpha_i-1}] = \underbrace{\begin{bmatrix} \mathbf{S}_1 \\ \hline \mathbf{S}_2 \end{bmatrix}}_{n_i-\alpha_i-1} \begin{matrix} \} n_i - \alpha_i - 1 \\ \} n_i \end{matrix} . \tag{90}$$

Taking into consideration that

$$[\widetilde{\mathbf{Q}}_1\, \mathbf{Q}_2\, \mathbf{Q}_3] \begin{bmatrix} 1 \\ s_k \\ s_k^2 \\ . \\ . \\ . \\ s_k^{2n_i-\alpha_i-} \end{bmatrix} = \begin{bmatrix} 1 \\ s_k \\ . \\ . \\ . \\ s_k^{n_i-\alpha_i-2} \end{bmatrix} Q_i(s_k) , \tag{91}$$

postmultiplying $\mathbf{W}$ by $\mathbf{S}$ we obtain

$$\begin{bmatrix} \mathbf{W}_1 & \mathbf{W}_2 \\ \mathbf{W}_3 & \mathbf{W}_4 \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \langle Q_i(s_k) \rangle \\ 0 \end{bmatrix} \tag{92}$$

where $\langle Q_i(s_k) \rangle$ is a diagonal matrix formed by the elements $Q_i(s_1), Q_i(s_2), \ldots, Q_i(s_{n_i - \alpha_i})$.

Eliminating $\mathbf{S}_2$ from this equation, we get

$$(\mathbf{W}_1 - \mathbf{W}_2 \mathbf{W}_4^{-1} \mathbf{W}_3)\, \mathbf{S}_1 = \mathbf{S}_1 \langle Q_i(s_k) \rangle \tag{93}$$

which shows that $Q_i(s_k)$, $k = 1,2, \ldots, n_i - \alpha_i - 1$ are the eigenvalues of

$$\tilde{\mathbf{T}} = \mathbf{W}_1 - \mathbf{W}_2 \mathbf{W}_4^{-1} \mathbf{W}_3 \tag{94}$$

and the columns of $\mathbf{S}_1$ are its corresponding eigenvectors. Since the elements of $\mathbf{D}(s)$ are necessarily the irreducible rational functions $R_i(s)/Q_i(s)$, the polynomials $R_i(s)$ and $Q_i(s)$ must not have any common zeros. Thus $Q_i(s_k) \neq 0$ for $k = 1,2, \ldots, n_i - \alpha_i - 1$ and therefore $\tilde{\mathbf{T}}$ is nonsingular.

(ii) If a particular zero s_k of the polynomial $R_i(s)$ is a multiple, say a λ_k-tuple one then postmultiplying the matrix $\mathbf{W}$ by the set of the following λ_k vectors

$$\begin{bmatrix} 1 \\ s_k \\ s_k^2 \\ s_k^3 \\ \cdot \\ \cdot \\ \cdot \\ s_k^{2n_i - \alpha_i - 2} \end{bmatrix}; \begin{bmatrix} 0 \\ 1 \\ 2s_k \\ 3s_k^2 \\ \cdot \\ \cdot \\ \cdot \\ (2n_i - \alpha_i - 2)\, s_k^{2n_i - \alpha_i - 3} \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \\ 2 \\ 3 \cdot 2 s_k \\ \cdot \\ \cdot \\ \cdot \\ (2n_i - \alpha_i - 2)\,(2n_i - \alpha_i - 3)\, s_k^{2n_i - \alpha_i - 4} \end{bmatrix}; \ldots$$

we obtain

$$\begin{bmatrix} & \\ & \\ & \\ \mathbf{W}_1 & \mathbf{W}_2 \\ & \\ \mathbf{W}_3 & \mathbf{W}_4 \\ & \\ & \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \ldots \\ s_k & 1 & 0 & \ldots \\ s_2^k & 2s_k & 2 & \ldots \\ s_k^3 & 3s_2^k & 3 \cdot 2 s_k & \ldots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix} = \begin{bmatrix} Q_i(s_k) & Q_i'(s_k) & Q_i''(s_k) & \ldots \\ s_k Q_i(s_k) & (sQ_i(s))_k' & (sQ_i(s))_k'' & \\ s_k^2 Q_i(s_k) & (s^2 Q_i(s))_k' & (s^2 Q_i(s))_k'' & \\ s_k^3 Q_i(s_k) & (s^3 Q_i(s))_k' & (s^3 Q_i(s))_k'' & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ R_i(s_k) & R_i'(s_k) & R_i''(s_k) & \\ s_k R_i(s_k) & (sR_i(s))_k' & (sR_i(s))_k'' & \\ s_k^2 R_i(s_k) & (s^2 R_i(s))_k' & (s^2 R_i(s))_k'' & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix} \tag{95}$$

where $Q_i'(s_k)$ is the derivation of $Q_i(s)$ with respect to s and evaluated at $s = s_k$.

Taking into consideration that s_k is a λ_k-tuple zero of $R_i(s)$, i.e.

$$R_i(s_k) = R_i'(s_k) = \ldots = R_i^{(\lambda_k-1)}(k_k) = 0\,, \tag{96}$$

the last n_i rows of the matrix on the right-hand side of Eq. (95) are zeros and so the last n_i rows of the matrix

$$\mathbf{S}_k = \begin{bmatrix} 1 & 0 & 0 & \ldots \\ s_k & 1 & 0 & \ldots \\ s^2 & 2s_k & 2 & \ldots \\ s^3 & 3s_k^2 & 3\cdot 2s_k & \ldots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{bmatrix} \tag{97}$$

can be eliminated. Furthermore, it is easy to see that

$$\begin{bmatrix} Q_i(s_k) & Q_i'(s_k) & Q_i''(s_k) & \ldots & Q_i^{(\lambda_k-1)}(s_k) \\ s_kQ_i(s_k) & (sQ_i(s))_k' & (sQ_i(s)))_k'' & \ldots & \\ s_k^2Q_i(s_k) & (s^2Q(_is))_k' & (s^2Q_i(s))_k'' & \ldots & \\ s_k^3Q_i(s_k) & (s^3Q_i(s))_k' & (s^3Q_i(s))_k'' & \ldots & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \\ \cdot & \cdot & \cdot & & \end{bmatrix} = \mathbf{S}_k \begin{bmatrix} Q_i(s_k) & Q_i'(s_k) & Q_i''(s_k) & \ldots \\ & Q_i(s_k) & 2Q_i'(s_k) & \ldots \\ & & Q_i(s_k) & \ldots \\ & & & \ddots \end{bmatrix} \tag{98}$$

hence we get

$$(\mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_4^{-1}\mathbf{W}_3)\,\mathbf{S}_k = \mathbf{S}_k \begin{bmatrix} Q_i(s_k) & Q_i'(s_k) & Q_i''(s_k) & \ldots & Q_i^{(\lambda_k-1)}(s_k) \\ & Q_i(s_k) & 2Q_i'(s_k) & \ldots & \\ & & Q_i(s_k) & \ldots & \\ & & & & \\ & & & & Q_i(s_k) \end{bmatrix} \tag{99}$$

and so the columns of $\mathbf{S}_k$ span a λ_k dimensional invariant subspace corresponding to the eigenvalue $Q_i(s_k)$ with multiplicity λ_k of $\widetilde{\mathbf{T}} = \mathbf{W}_1 - \mathbf{W}_2\mathbf{W}_4^{-1}\mathbf{W}_3$. Thus $\widetilde{\mathbf{T}}$ cannot be singular in the general case either.

Appendix 5

The same partitioning as shown in Appendix 3 now leads to

$$[\widehat{\mathbf{Q}}_1\,\mathbf{Q}_2\,\mathbf{Q}_3]\,\mathbf{R}^{-1(n_j-\alpha_i)}_{(n_i+n_j-\alpha_i)} =$$

$$= \left\{\widehat{\mathbf{Q}}_1 - [\mathbf{Q}_2\,\mathbf{Q}_3]\begin{bmatrix} \mathbf{R}(n_i-\alpha_i-1) & 0 \\ \mathbf{B}_2 & \mathbf{R}(\alpha_i+1) \end{bmatrix}^{-1}\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ 0 \end{bmatrix}\right\}\mathbf{R}^{-1}(n_j-\alpha_i) \tag{100}$$

where

$$\hat{\mathbf{Q}}_1 = \begin{bmatrix} \overbrace{0 \ldots 0}^{n_j - n_i + 1} & q_0(i) & q_1(i) \ldots q_{n_i - \alpha_i - 2}(i) \\ 0 \ldots 0 & 0 & q_0(i) \ldots q_{n_i - \alpha_i - 3}(i) \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \\ 0 \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot & q_0(i) \end{bmatrix} \tag{101}$$

and

$$\hat{\mathbf{B}}_1 = \begin{bmatrix} \overbrace{0 \ldots 0}^{n_j - n_i + 1} & b_0(i) & b_1(i) \ldots b_{n_i - \alpha_i - 2}(i) \\ 0 \ldots 0 & 0 & b_0(i) \ldots b_{n_i - \alpha_i - 3}(i) \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \\ 0 & & b_0(i) \end{bmatrix}. \tag{102}$$

Since the first $n_j - n_i + 1$ columns in both of the matrices $\hat{\mathbf{Q}}_1$ and $\hat{\mathbf{B}}_1$ are zero, the first $n_j - n_i + 1$ rows of $\mathbf{R}^{-1}(n_j - \boldsymbol{\alpha}_i)$ are to be deleted, further $\hat{\mathbf{Q}}_1$ and $\hat{\mathbf{B}}_1$ are to be replaced by $\tilde{\mathbf{Q}}_1$ and $\tilde{\mathbf{B}}_1$, respectively. Thus, making use of the denotation of Appendix 3:

$$\mathbf{T} = \mathbf{W}_1 - \mathbf{W}_2 \mathbf{W}_4^{-1} \mathbf{W}_3, \tag{103}$$

Eq. (44) is obtained.

Appendix 6

Substituting (60) into Eq. (57), it can be written in the following form:

$$\left[\begin{array}{ccc|cccc} \cdots \sum_{i \le j \le k} s_i s_j s_k & \sum_{i \le j} s_i s_j & \sum s_i & 1 & & & \\ \cdot \quad \cdot \quad \cdot & & & & & & \\ & \sum_{i \le j \le k} s_i s_j s_k & \sum_{i \le j} s_i s_j & \sum s_i & 1 & & \\ & & & & \cdot & & \\ & & & & & \cdot & \\ & & & & & & \cdot \\ & & & & & \sum_{i \le j} s_i s_j & \sum s_i \quad \cdot \quad 1 \end{array}\right] \left[\begin{array}{c} c_{n_j - \alpha_i}(ij) \\ \cdot \\ \cdot \\ \cdot \\ c_{n_i - \alpha_i}(ij) \\ \hline c_{n_i - \alpha_i - 1}(ij) \\ \cdot \\ \cdot \\ \cdot \\ c_{\beta_i + 1}(ij) \end{array}\right] = 0. \tag{104}$$

Subtracting each row multiplied by s_1 from the next one, then starting with the second row, subtracting each row multiplied by s_2 from the next one; finally the last but one row multiplied by $s_{n_i - \alpha_i - \beta_i - 2}$ and subtracting from the last one, thel atter yields

the condition

$$c_{\beta_i+1}(ij) + c_{\beta_i+2}(ij)\, s_{n_i-\alpha_i-\beta_i-1} + \ldots + c_{n_j-\alpha_i}(ij)\, s_{n_i-\alpha_i-\beta_i-1}^{n_j-\alpha_i-\beta_i-1} = 0. \tag{105}$$

Since the order of the factors s_k can arbitrarily chosen, we obtain all the conditions (62) in a similar way.

Appendix 7

The 2nd, 3rd, ..., n_j+1st equations of (38) immediately yield $p_{nj}(ij)$, $p_{n_j-1}(ij)$, ..., $p_1(ij)$ in the following form

$$\begin{bmatrix} p_{n_j}(ij) \\ p_{n_j-1}(ij) \\ \vdots \\ p_1(ij) \end{bmatrix} = \begin{bmatrix} q_{n_i-1}(i) & 1 & 0 & & \\ q_{n_i-2}(i) & q_{n_i-1}(i) & 1 & & \\ q_{n_i-3}(i) & q_{n_i-2}(i) & q_{n_i-1}(i) & \ddots & \\ \vdots & \vdots & \vdots & \ddots & \ddots \\ q_0(i) & q_1(i) & & & \\ 0 & q_0(i) & & & \\ & 0 & & & \\ \vdots & \vdots & & & \\ & & & & 1 \end{bmatrix} \mathbf{R}^{-(n_j-\alpha_i)}_{(n_i+n_j-\alpha_i)} \begin{bmatrix} c_{n_j-\alpha_i}(ij) \\ c_{n_j-\alpha_i-1}(ij) \\ \vdots \\ c_2(ij) \\ c_1(ij) \end{bmatrix}. \tag{106}$$

Taking into consideration that the partitioned form of $\mathbf{R}^{-1(n_j-\alpha_i)}_{(n_i+n_j-\alpha_j)}$ is

$$\mathbf{R}^{-1(n_j-\alpha_i)}_{(n_i+n_j-\alpha_i)} = \begin{bmatrix} \mathbf{R}^{-1}(n_j-\alpha_i) \\ -\mathbf{R}^{-1}(n_i)\,\mathbf{B}_3\,\mathbf{R}^{-1}(n_j-\alpha_i) \end{bmatrix} = \begin{bmatrix} \mathbf{I}(n_j-\alpha_i) \\ -\mathbf{R}^{-1}(n_i)\,\mathbf{B}_3 \end{bmatrix} \mathbf{R}^{-1}(n_j-\alpha_i) \tag{107}$$

where

$$\mathbf{B}_3 = \begin{bmatrix} \overbrace{0 \ldots 0}^{n_j-n_i+1} & \overbrace{b_0(i) \quad b_1(i) \quad \ldots \quad b_{n_i-\alpha_i-2}(i)}^{n_j-\alpha_i-1} \\ 0 \ldots 0 & 0 \quad b_0(i) \quad \ldots \quad b_{n_i-\alpha_i-3}(i) \\ \vdots & \vdots \quad \vdots \qquad\qquad \vdots \\ 0 & 0 \quad 0 \qquad\qquad b_0(i) \\ 0 & 0 \quad 0 \qquad\qquad 0 \\ \vdots & \vdots \quad \vdots \qquad\qquad \vdots \\ 0 & 0 \quad 0 \qquad\qquad 0 \end{bmatrix} \tag{108}$$

and $\mathbf{R}^{-1}(n_j - \alpha_i)$ can be partitioned into two blocks according to its rows as follows

$$\mathbf{R}^{-1}(n_j - \alpha_i) = \begin{bmatrix} \widetilde{\mathbf{R}}^{-1(n_j - n_i + 1)}_{(n_j - \alpha_i)} \\ \hline \widetilde{\mathbf{R}}^{-1(n_i - \alpha_i - 1)}_{(n_i - \alpha_i)} \end{bmatrix}, \tag{109}$$

we get

$$\begin{bmatrix} p_{n_j}(ij) \\ p_{n_j-1}(ij) \\ \vdots \\ p_1(ij) \end{bmatrix} = \begin{bmatrix} q_{n_i-1}(i) & 1 & & \\ q_{n_i-2}(i) & q_{n_i-1}(i) & & \\ \vdots & \vdots & \ddots & \\ q_0(i) & q_1(i) & & \\ & q_0(i) & & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I}(n_j - \alpha_i) \\ \\ -\mathbf{R}^{-1}(n_i)\mathbf{B}_3 \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{R}}^{-1(n_j - n_i + 1)}_{(n_j - \alpha_i)} \\ \\ \widetilde{\mathbf{R}}^{-1(n_i - \alpha_i - 1)}_{(n_j - \alpha_i)} \end{bmatrix} \times$$

$$\times \begin{bmatrix} c_{n_j - \alpha_i}(ij) \\ c_{n_j - \alpha_i - 1}(ij) \\ \vdots \\ c_1(ij) \end{bmatrix}, \tag{110}$$

Substituting Eq. (45), the product of the last two factors becomes

$$\begin{bmatrix} \mathbf{R}^{-1}(n_j - n_i + 1) \\ 0 \end{bmatrix} \begin{bmatrix} c_{n_j - \alpha_i}(ij) \\ c_{n_j - \alpha_i - 1}(ij) \\ \vdots \\ c_{n_i - \alpha_i}(ij) \end{bmatrix},$$

since the last $n_i - \alpha_i - 1$ columns of $\widetilde{\mathbf{R}}^{-1(n_j - n_i + 1)}_{(n_j - \alpha_i)}$ are zero.

Further, since the last $n_i - \alpha_i - 1$ columns of $\mathbf{I}(n_j - \alpha_i)$ and $\mathbf{B}_3$ are multiplied by zero, the product of the three last factors is

$$\begin{bmatrix} \mathbf{I}(n_j - n_i + 1) \\ 0 \end{bmatrix} \mathbf{R}^{-1}(n_j - n_i + 1) \begin{bmatrix} c_{n_j - \alpha_i}(ij) \\ c_{n_j - \alpha_i - 1}(ij) \\ \vdots \\ c_{n_i - \alpha_i}(ij) \end{bmatrix}$$

and finally the last $n_i - \alpha_i - 1$ columns of the first factor containing the elements $q_k(i)$ vanish as well. Hence we obtain Eq. (64).

References

1. *Morgan, B. S.*, "The synthesis of linear multivariable systems by state variable feedback", IEEE Trans. Automatic Control, **AC–9**, pp. 405–411, October 1964.
2. *Rekasius, Z. V.*, "Decoupling of multivariable systems by means of state variable feedback", Proc. 3rd Annual Allerton Conf. on Circuit and System Theory (Urbana, Illinois), pp. 439–447, October 1965.
3. *Falb, L. P., Wolovich, W. A.*, "Decoupling in the design and synthesis of multivariable control systems", IEEE Trans. Automatic Control, **AC–12**, pp. 651–659, December 1967.
4. *Gilbert, E. G.*, "The decoupling of multivariable systems by state feedback", SIAM Journal on Control, **7**, pp. 50–63, February 1969.
5. *Wonham, W. M., Morse, A. S.*, "Decoupling and pole assignment in linear multivariable systems: A geometric approach", SIAM Journal on Control, **8**, pp. 1–18, January 1970.
6. *Morse, A. S., Wonham, W. M.*, "Status of noninteracting control", IEEE Trans. Automatic Control, **AC–16**, pp. 568–581, December 1971.
7. *Silverman, S.*, "Decoupling with state feedback and precompensation", IEEE Trans. Automatic Control, **AC–15**, pp. 487–489, August 1970.
8. *Luenberger, D. G.*, "Canonical forms for linear multivariable systems", IEEE Trans. Automatic Control, **AC–12**, pp. 290–293, June 1967.
9. *Applevich, J. D.*, "Direct computation of canonical forms for linear systems by elementary matrix operations", IEEE Trans. Automatic Control, **AC–19**, pp. 124–126, Aprils, 1974.

Разбиение линейных, время-инвариантных многомерных систем

П. РОЖА, Н. К. СИНХА

(Будапешт, Гамильтон)

Проблематика разбиения линейных, время-инвариантных, многомерных систем управления рассматривалась в работах различных авторов. Для того, чтобы получить удовлетворительное и устойчивое состояние системы, необходимо найти возможность расположения полюсов передаточной функции системы, полученной после разбиения, в соответствующих точках. В области решения данной проблемы большой вклад внесли следующие авторы: Морган и Рекасиус [2], Фальб и Уолович [3], Жильбер [4] и Уонэм и Морзе [5]. Недавно в одной из опубликованных работ Морзе и Уонэм [6] дали широкий обзор о положении теории несвязанного управления. Фальб и Уолович [3] разработали необходимое и удовлетворительное условие разбиения путем применения установления обратной связи переменных состояния постоянного преобразования входа. Жильбер [4] ввел понятие предварительного компенсатора с обратной связью состояния и им был получен последующий компенсатор с минимальным порядком. Уонэм и Морзе [5] разработали геометрическое приближение предварительной компенсации с минимальным порядком. Сильверманом [7] были предложены два алгоритма для разбиения обратимых систем с использованием обратной связи состояния и предварительной компенсации.

Хотя проблема кажется почти полностью разрешенной, существует необходимость разработки некоторых непосредственных методов, с помощью которых определяется возможность разбиения многомерной системы управления путем образования обратной связи состояния и применения постоянного преобразования входа, с одной стороны, а с другой — возможность произвольного расположения полюсов замкнутой системы. В любом случае, как только имеется такая возможность, обычно желательно и более удобно применить некоторый «статический» компенсатор вместо «динамических» предварительных компенсаторов, предложенных Жильбером [4], Сильверманом [7] и Уонэмом и Морзе [5].

В настоящей статье разработан один из таких, методов. Разработано необходимое и удовлетворительное условие разбиения и расположения полюсов, исходя из управляемой формы канонического описания системы. Используется непосредственный алгебраический

метод. В то время как необходимо проводить довольно большое количество матричных операций при использовании разработанного метода, окончательные результаты совсем нетрудно применить. В связи с матрицей D тройственного канонического представления A, B, C даются некоторые простые тестовые ответы на вопрос: возможно ли осуществить разбиение и расположение полюсов. Из этого получится непосредственный метод определения статического предварительного компенсатора, обратной матрицы для разбиения и желательного расположения полюсов. Рекомендуемый метод кажется существенно более простым, чем метод, разработанный Фальбом и Уоловичем.

P. Rózsa
Department of Mathematics
Faculty of Civil Engineering
Technical University of Budapest
H-1111
Hungary

N. Sinha
Department of Electrical Engineering
McMaster University
Hamilton
Canada

Problems of Control and Information Theory, Vol. 5 (4), pp. 359–369 (1976)

CONNECTION BETWEEN THE WEIGHT DISTRIBUTIONS OF A GROUP CODE AND ITS DUAL

V. V. MASGRAS

(Bucharest)

(Received October 20, 1975)

The paper consists of three parts. The first is a general survey of the main connections between the weight distributions of a code and its dual when the code is over a alphabet which is a field, an Abelian group or an algebraic nonstructured alphabet. For all these, the well-known Mac Williams identity holds

$$A(z) = q^{k-n}(1 + (q-1)z)^n B\big((1-z)/(1+(q-1)z)\big) \tag{1}$$

where $A(z) = \sum_{i=0}^{n} A_i z^i$ and $B(t) = \sum_{i=0}^{n} B_i t^i$ are the weight enumerators of the group code and its dual.

In the second part, another form of (1) is given by the following

Theorem 1. If F is an Abelian group of order q and $C \subset F^n$ is a group code with k information symbols, then between the weight vector A of C and the weight vector B of C' (the dual of C) there exists the relation

$$\mathfrak{R} A^T = q^k \mathfrak{B} B^T, \tag{11}$$

where

$$\mathfrak{R} = \left[q^i \binom{j}{i} \right] \qquad (0 \le i \le n,\ 0 \le j \le n), \tag{12}$$

$$\mathfrak{B} = \left[(-1)^j \binom{n-j}{n-i} (q-1)^{i-j} \right] \qquad 0 \le i \le n,\ 0 \le j \le n. \tag{13}$$

(Superscript T denotes transposition).

The last part gives some identities. In the third part of the paper the nonbinary group code of length n over additive group Z_q of integers modulo q is defined as follows

$$C = \{c \mid c = (c_1, \ldots, c_n),\ \sum_{i \in I_k^j} c_i = 0 \pmod q,\ 1 \le j \le q-1,\ 1 \le k \le s\},$$

where s is the number of the digits of the radix-q representation of n. This representation is denoted by $(n)_q$ and I_k^j are the following sets

$$I_k^j = \{i \mid 1 \le i \le n,\ (i)_q = i_1 \ldots i_{k-1} j i_{k+1} \ldots i_s\},\ 1 \le j \le q-1,\ 1 \le k \le s.$$

Using (11) the weight distributions of two examples of the nonbinary group codes (for $n = 8$, $q = 3$ and for $n = 15$, $q = 4$) are determined.

Introduction

Let F be a finite alphabet with q symbols and let $V \subset F^n$ be a code of length n and k information symbols. If F is a finite field (q being a prime or a prime power and if V is a finite vector space of dimension k over F then the Mac Williams identity [11] holds

$$A(z) = q^{k-n}(1 + (q-1)z)^n B\big((1-z)/(1+(q-1)z)\big) \tag{1}$$

where $A(z) = \sum_{i=1}^{in} A_i z^i$ and $B(t) = \sum_{i=0}^{n} B_i t^i$ are the weight enumerators of the code V and of its dual V' defined as follows

$$V' = \{y \mid y \in F^n, \langle x, y\rangle = \forall, Vx \in V\} \tag{2}$$

where $\langle x, y\rangle$ denotes the inner product $\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$ if $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n) \in F^n$, the operations being from the field F.

However if F is not a field and if we suppose that F is only an Abelian group it is possible to define the concept of the dual of the code V (cf. Delsarte [3], [4]) in the case if V is a subgroup of F^n.

Let r be the period of the group $(F, +)$ and let Q_r be the cyclomotic field of the complex rth roots of unity. Let Φ be the group characters of $(F, +)$. A character $\varphi \in \Phi$ is a group homomorphism φ: $(F, +) \to (Q_r, .)$. The homomorphism $\varphi_0 \in \Phi$ defined by $\varphi_0(f) = 1$, $\forall f \in F$ is the principal character.

We have $|\Phi| = |F| = q$ and if we suppose $F = \{f_0, \ldots, f_\lambda\}$ with $\lambda = q-1$ and f_0 the null element of the group F and $\Phi = \{\varphi_0, \varphi_1, \ldots, \varphi_\lambda\}$ with principal character φ_0, it is always possible to choose indexing in such a way that $\varphi_i(f_j) = \varphi_i(f_j)$ $\forall i, j \in \{0,1, \ldots, n\}$.

With the above conventions and notations we can define the inner product of two elements $a, b \in V \subset F^n$, $a = (a_1, \ldots a_n)$, $b = (b_1, \ldots b_n)$ in the following way:

$$\langle a, b\rangle = \prod_{i=1}^{n} \varphi_{t_i}(a_i) \text{ with } f_{t_i} = b_i \tag{4}$$

or equivalently

$$\langle a, b\rangle = \prod_{i=1}^{n} \varphi_{t_i}(b_i) \text{ with } f_{t_i} = a_i. \tag{4'}$$

For any subgroup E of V the dual E' of E is defined by

$$E' = \{x \in V \mid \langle x, y\rangle = 1, \forall y \in E\}. \tag{5}$$

Taking into account the properties of the inner product (4) or (4′) it is not difficult to show that E' is itself a subgroup of V. If $E = V$, V' in (5) will be termed the dual of V.

In view of the above considerations it follows that the concept of the dual code of a linear code over a field may be extended to the group codes. Moreover, this concept can be extended to the codes over any alphabets, algebraic nonstructured, in the following way. Let F be an arbitrary finite alphabet and let C be a code of length n over F. We define $A = (A_0, A_1, \ldots, A_n)$ or $A(C) = (A_0(C), \ldots, A_n(C))$ the distance distributions of C, i.e.

$$A_i = A_i(C) = \frac{1}{|C|} |\{(a, b) \in C^2 \mid d_H(a, b) = i\}| \quad (0 \leq i \leq n), \tag{6}$$

where $d_H(a, b)$ is the Hamming distance between the code words a and b of C.

(6) implies that $A_i(C)$ is the mean number of pairs of code words at distance i $(0 \leq i \leq n)$.

Let C' be another code of length n over F and let $B = (B_0, \ldots, B_n)$ or $B(C) = (B_0(C), \ldots, B_n(C))$ be the distance distribution of C'. Then the following definition can be given.

Definition. The code C' is called dual of the code C if between the distance distributions of these codes there exists the relation:

$$B_k(C') = \frac{|1}{|C|} \sum_{i=0}^{n} P_k(i)\, A_i(C) \qquad (0 \leq k \leq n), \tag{7}$$

where $P_k(x)$ is the Krawtchouk polynomial ([5], [12]) of degree k defined by

$$P_k(x) = \sum_{j=0}^{k} (-1)^j \lambda^{k-j} \binom{x}{j} \binom{n-x}{k-j} \quad (0 \leq k \leq n). \tag{8}$$

In a matrix form (7) becomes

$$B = \frac{1}{|C|} A \mathcal{S} \tag{9}$$

where $\mathcal{S}$ is the Krawtchouk matrix defined by

$$\mathcal{S} = [P_k(i)] \quad (0 \leq i \leq n,\ 0 \leq k \leq n). \tag{10}$$

The transform (7) or (9) will be called the Mac Williams transform.

If a code C is a group code then the dual of C (cf. (5)) is the dual also in the sense of the above definition (Delsarte [3]).

For a linear code C of length n over a finite field F with q elements all the forms (1), (5) and (9) of the connection between the weights of a code C and the weight of its dual code C' are equivalent.

This connection also holds if the code C is only an Abelian group code.

Main results

Another form of the connection between the weight distributions of a group code and its dual code is given by the following

Theorem 1. If F is an Abelian group of order q and $C \subset F_n$ is a group code with k information symbols then between the weight vector $A = (A_0, \ldots, A_n)$ of C and the weight vector $B = (B_0, \ldots, B_n)$ of C', the dual of C, the following relation holds

$$\mathcal{A} A^T = q^k \mathcal{B} B^T \tag{11}$$

where

$$\mathcal{A} = \left[q^i \binom{j}{i} \right] \qquad (0 \leq i \leq n,\ 0 \leq j \leq n), \tag{12}$$

$$\mathcal{B} = \left[(-1)^j - \binom{n-j}{n-i} (q-1)^{i-j} \right] \qquad (0 \leq i \leq n,\ 0 \leq j \leq n) \tag{13}$$

with $\binom{u}{v} = 0$ if $u < v$ or if $v < 0$, and x^T is the transpose of the vector x.

Proof. Since (1), (5) and (9) are equivalent for group codes we start with (1) in the form

$$q^{n-k} \sum_{j=0}^{n} A_j z^j = \sum_{j=0}^{n} B_j (1-z)^j (1+\lambda z)^{n-j}$$

which being an identity by z we take the ith derivative of each side with respect to z.

Denoting $(j)_i = j(j-1)\ldots(j-i+1)$ we have

$$q^{n-k} \sum_{j=0}^{n} (j)_i A_j z^{j-i} = \sum_{j=0}^{n} B_j \sum_{\mu=0}^{i} \binom{i}{\mu} \left((1-z)^j\right)^{i-\mu} \left((1+\lambda z)^{n-j}\right)^{(\mu)} =$$

$$= \sum_{j} \sum_{\mu=0}^{i} (-1)^{i-\mu} \lambda^\mu (j)_{i-\mu} (n-j)_\mu \binom{i}{\mu} (1-z)^{i+\mu-i} (1+\lambda z)^{n-j-\mu} B_j.$$

If we set $z = 1$, the only nonzero terms are those for which $j + \mu - i = 0$. Because $1 + \lambda = q$, we obtain

$$q^{n-k} \sum_{j=i}^{n} (j)_i A_j = q^{n-i} \sum_{j+\mu=i} (-1)^{i-\mu} \lambda^\mu (j)_{i-\mu} (n-j)_\mu \binom{i}{\mu} B_j.$$

Dividing by $i!q^{n-k-i}$ and renouncing at the summation index μ, we have

$$q^i \sum_{j=i}^{n} \binom{j}{i} A_j = q^k \sum_{j=0}^{i} (-1)^j \lambda^{i-j} \binom{n-j}{n-i} B_j \quad (0 \leq i \leq n). \tag{14}$$

This relation represents (11) with notations (12) and (13). Q.E.D.

Remarks. 1. Another form of (13) is

$$\mathfrak{B} = \left[(-1)^j \binom{i}{j} v_i v_j^{-1} \right] \quad (0 \leq i \leq n,\ 0 \leq j \leq n) \tag{15}$$

where $v_e = \binom{n}{e} (q-1)^e$.

2. The matrix $\mathfrak{A}$ from (12) is an upper triangular matrix and $\mathfrak{B}$ from (13) is a lower triangular matrix so that in the ith relation from (14) enter the last $n - i + 1$ components of the weight vector of code C and the first $i + 1$ components of the weight vector of the dual code C' of C. Then for any i there are $n + 2$ interconnected components.

3. The two matrices $\mathfrak{A}$ and $\mathfrak{B}$ are nonsingular because

$$\det \mathfrak{A} = q^{n(n+1)/2}, \tag{16}$$

$$\det \mathfrak{B} = 1. \tag{17}$$

Corollary. Between the Krawtchouk matrix $\mathfrak{F}$ defined by (10) and the matrices $\mathfrak{A}$ and $\mathfrak{B}$ defined by (12) and (13) there exists the relation

$$\mathfrak{F}^T = \mathfrak{B}^{-1} A. \tag{18}$$

Proof. Taking into account that for a group code C with k information symbols we have $|C| = q^k$, (9) becomes

$$\mathfrak{F}^T A^T = q^k B^T. \tag{19}$$

Multiplying (19) by $\mathfrak{B}$ and considering (11) we obtain

$$\mathfrak{B}\mathfrak{F}^T A^T = q^k \mathfrak{B} B^T = \mathfrak{A} A^T$$

and since $\mathfrak{B}$ is invertible and A arbitrary, we obtain (18).

Relation (18) represents a decomposition of the Krawtchouk matrix in a product of two triangular matrices, so that an equation

$$\mathfrak{F} x^T = y^T$$

with the unknown vector x is easily solvable using this relation.

Let C be a group code, $A = (A_0, \ldots, A_n)$ be the weight distribution of C and $B = (B_0, \ldots, B_n)$ the weight distribution of its dual C'.

Let d denote the least integer k, $1 \leq k \leq n$, such hat $A_k \neq 0$, then d is the minimum distance of the code C. Let d' denote the least integer k, $1 \leq k \leq n$ such that $B_k \neq 0$, then d' is the minimum distance of code C' (dual of C) or equivalently (see Delsarte [5]) d' is the dual distance of the code C.

Theorem 2. If F is an Abelian group of order q, $C \subset F^n$ is a group code with k information symbols, d' is the dual distance and $A = (A_0, \ldots, A_n)$ the weight distribution, then the following identities hold

$$\sum_{j=i}^{n} \binom{j}{i} A_j = q^{k-i}(q-1)^i \binom{n}{i} \quad (0 \leq i < d'). \tag{20}$$

Proof. For $0 \leq i \leq d'$ we have $B_0 = 1$, $B_1 = B_2 = \ldots = B_{d'-1} = 0$ and $B_{d'} \neq 0$. Using identities (14) and this one, the only nonzero term in the right hand side of (14) is for $j = 0$. Dividing by q^i results (20). **Q.E.D.**

Remarks. 1. Substituting C by its dual C' in Theorem 2 and taking into account that the dual of the dual of a code is the code itself, $(C')' = C$, $(d')' = d$ and $k' = n-k$, we obtain the following identities

$$\sum_{j+i}^{n} \binom{j}{i} B_j = q^{n-k-i}(q-1)^i \binom{n}{i} \quad (0 \leq i < d). \tag{21}$$

2. For $i = 0$, (20) and (21) represent the well-known formulas for weight summation

$$\sum_{j=0}^{n} A_j = q^k \text{ and } \sum_{j=0}^{n} B_j = q^{n-k} \tag{22}$$

and for $d' > 1$, if $i = 1$, (20) becomes

$$\sum_{j=0}^{n} jA_j = q^{k-1}(q-1)\, n \tag{23}$$

which has been proved by many authors (such as Assmus and Mattson [1], Delsarte [4], Pless [10], Levy [7] and others).

3. If C is a projective group code (i.e. $d' > 2$), for $i = 2$, taking (23) into account, relation (20) becomes

$$\sum_{j=1}^{n} j^2 A_j = q^{k-2}\, m(m+1) \tag{24}$$

with $m = n(q-1)$, which has been proved by Pless [10] and Delsarte [4].

Application to the nonbinary group codes

In the following we shall apply Theorem 1 to the nonbinary group codes for two examples.

Let $F = Z_q$ be the additive group of integers modulo q. This is an Abelian group. Let us take $n > 2q-1$ and $n \neq jq^s$ for any integers j and s.

We denote by $(n)_q$ the radix-q representation of n. We suppose that this representation has s digits $(n)_q = n_1 n_2 \ldots n_s$ with $n_1 \neq 0$.

Let I_k^j be the following sets

$$\mathbf{I}_k^j = \{i \mid 1 \leq i \leq n,\ (i)_q = i_1 \ldots i_{k-1} j i_{k+1} \ldots i_s\}\ (1 \leq j \leq q-1,\ 1 \leq k \leq s). \tag{25}$$

Definition. The set $C \subset Z_q^n$ is the nonbinary group code [8], so that $c = (c_1, \ldots, c_n) \in C$ is a code word if and only if:

$$\sum_{i \in I_k^j} c_i \equiv 0 \pmod{q} \qquad (1 \leq j \leq q-1,\ 1 \leq k \leq s). \tag{26}$$

If we let $r = |\{I_k^j \mid I_k^j \neq \emptyset,\ 1 \leq j \leq q-1,\ 1 \leq k \leq s\}|$ then the group code C has r check symbols and $m = n-r$ information symbols. In a former paper [8] it has been proved that these codes are single error correcting codes over the alphabet $F = Z_q$ for any q. For $q = 2$ and $n = q^s - 1$, C is a binary Hamming code [2], [6]. For $q > 2$, C represents another generalization of the Hamming code ([8]) for the nonbinary case. This justifies the above definition. In [9] it has been proved that the nonbinary group codes may be regarded as a class of binoid single error correcting codes, defined in [13].

An unsolved problem for nonbinary group codes (in fact, for many other classes of codes) is the determination of the weight distributions.

Example 1. Let us consider the nonbinary group codes defined in (25) for $n = 8$, $q = 3$, over the additive group $F = Z_3 = \{0,1,2\}$ of integers modulo 3.

Let f be the function $f: Z_3^4 \to Z_3^8$ which maps the information word $a = (a_1, a_2, a_3, a_4)$ into the code word $c = (c_1, \ldots, c_8)$, $c \in Z_3^8$ (after (26)) i.e. the components c_i satisfy the following congruences:

$$\begin{aligned} c_1 + c_4 + c_7 &\equiv 0 \pmod 3 \\ c_2 + c_5 + c_8 &\equiv 0 \pmod 3 \\ c_3 + c_4 + c_5 &\equiv 0 \pmod 3 \\ c_6 + c_7 + c_8 &\equiv 0 \pmod 3. \end{aligned} \tag{27}$$

Of course four components of c will be the information symbols a_1, a_2, a_3, a_4 and the other four will be the check symbols. These check symbols will be chosen so that the congruences (27) will be satisfied.

The map f is an injective homomorphism of Abelian groups. Indeed if for $a \neq a'$, $a, a' \in Z_3^4$ $f(a) = c \neq c' = f(a')$ and if $a' = (a_1', a_2', a_3', a_4')$ then $f(a + a') = c + c' = f(a) + f(a')$ because if $c = (c_1, \ldots, c_8) = f(a)$ and $c' = (c_1 + c_1', \ldots, c_8 + c_8') = f(a + a')$ satisfies (27). On the other hand Im $f = C \cong Z_3^4$ and if we consider the dual code C' of C, $C' = Z_3^8/C$ we have the isomorphisms

$$C' = Z_3^8/C \simeq Z_3^8/Z_3^4 \simeq Z_3^4 \simeq C. \tag{28}$$

This means the code C is an autodual code. Thus, the two weight distributions are equal

$$A_i = B_i \qquad (0 \leq i \leq 8). \tag{29}$$

The matrices $\mathcal{A}$ and $\mathcal{B}$ from (12) and (13) become

$$\mathcal{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 3 & 6 & 9 & 12 & 15 & 18 & 21 & 24 \\ 0 & 0 & 9 & 27 & 54 & 90 & 135 & 189 & 252 \\ 0 & 0 & 0 & 27 & 108 & 270 & 540 & 945 & 1512 \\ 0 & 0 & 0 & 0 & 81 & 405 & 1215 & 2835 & 5670 \\ 0 & 0 & 0 & 0 & 0 & 243 & 1458 & 5103 & 13608 \\ 0 & 0 & 0 & 0 & 0 & 0 & 729 & 5103 & 20412 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2187 & 17496 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6561 \end{pmatrix}, \tag{30}$$

$$\mathcal{B} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 16 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 112 & -14 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 448 & -84 & 12 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1120 & -280 & 60 & -10 & 1 & 0 & 0 & 0 & 0 \\ 1792 & -560 & 160 & -40 & 8 & -1 & 0 & 0 & 0 \\ 1792 & -672 & 240 & -80 & 24 & -6 & 1 & 0 & 0 \\ 1024 & -448 & 192 & -90 & 32 & -12 & 4 & -1 & 0 \\ 256 & -128 & 64 & -32 & 16 & -8 & 4 & -2 & 1 \end{pmatrix}. \tag{31}$$

Considering (29) in which $A_0 = B_0 = 1$, (30) and (31) relation (11) will be transformed into the following system of linear equations

$$
\begin{cases}
A_1+ & A_2+ & A_3+ & A_4+ & A_5+ & A_6+ & A_7+ & A_8= & 80 \\
28A_1+ & 2A_2+ & 3A_3+ & 4A_4+ & 5A_5+ & 6A_6+ & 7A_7+ & 8A_8= & 432 \\
126A_1- & 8A_2+ & 3A_3+ & 6A_4+ & 10A_5+ & 15A_6+ & 21A_7+ & 28A_8= & 1008 \\
252A_1- & 36A_2+ & 4A_3+ & 4A_4+ & 10A_5+ & 20A_6+ & 35A_7+ & 56A_8= & 1344 \\
280A_1- & 60A_2+ & 40A_3+ & & 5A_5+ & 15A_6+ & 35A_7+ & 70A_8= & 1120 . \\
560A_1- & 160A_2+ & 40A_3- & 8A_4+ & 4A_5+ & 18A_6+ & 63A_7+ & 168A_8= & 1792 \\
672A_1- & 240A_2+ & 80A_3- & 24A_4+ & 6A_5+ & 8A_6+ & 64A_7+ & 252A_8= & 1792 \\
448A_1- & 192A_2+ & 90A_3- & 32A_4+ & 12A_5- & 4A_6+ & 28A_7+ & 216A_8= & 1024 \\
128A_1- & 64A_2+ & 32A_3- & 16A_4+ & 8A_5- & 4A_6+ & 2A_7+ & 80A_8= & 256
\end{cases}
\tag{32}
$$

From any eight equations of (32), (for example from the first eight) we obtain the solution

$$A_0 = 1, A_1 = A_2 = 0, A_3 = 8, A_4 = 10, A = 16, A = 36, A = 8, A_8 = 2 \tag{33}$$

which also verifies to 9th equation of (32).

Of course, the determination (33) of the weights of the nonbinary group code C for $n = 8, q = 3$ can also be obtained applying the Mac Williams transform (7) or (9) using the Krawtchouk matrix.

Example 2. Let us consider the nonbinary group code C of length $n = 15$ over the additive group $F = Z_4 = \{0,1,2,3\}$ of integers modulo 4, as it is defined in (26).

The following weight distribution of C has been obtained by a computer

i	0	1	2	3	4	5	6	7	8
A_i	1	0	0	27	81	216	1116	3231	9171
		9	10	11	12	13	14	15	
		24456	42336	63081	57531	39744	16740	4413	

$$\sum_{i=0}^{15} A_i = 462144 = 4^9.$$

Using (11) and verifying the results with (9), the following weight distribution of the dual code C' of C has been determined

i	0	1	2	3	4	5	6	7	8
B_i	1	0	0	0	18	0	30	54	153
		9	10	11	12	13	14	15	
		252	720	720	1404	468	234	42	

$$\sum_{i=0}^{15} B_i = 4096 = 4^6.$$

References

1. *Assmus, E. F., Mattson, H. F.*, "Error Correcting Codes — An Axiomatic Approach", Information and Control **6** (1963) 315–330.
2. *Berlekamp, E. R.*, Algebraic Coding Theory. McGraw Hill, New York 1968.
3. *Delsarte, P.*, "Bounds for Unrestricted Codes by Linear Programming", Philips Res. Repts. **27** (1972) 272–289.
4. *Delsarte, P.*, "Weights of Linear Codes and Strongly Regular Normed Spaces", Discrete Mathematics **3** (1972) 47–64.
5. *Delsarte, P.*, "Four Fundamental Parameters of a Code and Their Combinatorial Significance", Information and Control **23** (1973) 407–438
6. *Hamming, R. W.*, "Error Detecting and Error Correcting Codes", Bell System Technical Journal **29** (1950) 147–160.
7. *Levy, J. E.*, "A Weight Distribution Bound for Linear Codes", IEEE Transactions on Information Theory **IT–14**, 3 (1968).
8. *Masgras, V. V.*, "Coduri grupale nebinare", Studii si Cercetari matematice **24**, 6 (1972) 855–863.
9. *Masgras, V. V.*, "A Class of Binoid Single-Error-Correcting Codes", IEEE Transactions on Information Theory **IT–21** (1975) 105–106.
10. *Pless, V.*, "Power Moment Identities on Weight Distribution in Error Correcting Codes", Information and Control **6** (1963) 147–152.
11. *Mac Williams, F. J.*, "A Theorem on the Distribution of Weights in a Systematic Code", Bell System Technical Journal **42** (1963) 74–94.
12. *Szegő, G.*, "Orthogonal polynomials", American Mathematical Society Colloquium Publications **XXIII**, 1959, New York.
13. *Samoylenko, S. I.*, "Binoid Error-Correcting Codes", IEEE Transactions on Information Theory **IT–19**, 1 (1973) 95–101.

Связь между некоторым групповым кодом и весовыми распределениями его дуального кода

В. В. МАСГРАС

(Бухарест)

Статья составлена из трех частей. В первой части рассматривается вопрос о главных связях между кодом и весовыми распределениями его дуальных кодов тогда, когда код содержит весь алфавит, являющийся некоторым полем, т. е. алфавитом без алгебраической структуры или группой типа Абеля. Для всех кодов такого типа действительно хорошо известное тождество Мака Вильямса:

$$A(z) = q^{k-n}(1 + (q-1)z)^n \, B\big((1-z)/(1 + (q-1)z)\big), \tag{1}$$

где $A(z) = \sum_{i=0}^{n} A_i z^i$ и $B(t) = \sum_{i=0}^{n} B_i t^i$ являются весовыми счетчиками кода и его дуального кода.

Во второй части уравнение (1) определяется в другом виде с использованием следующей теоремы:

Теорема 1: Если F является группой типа Абеля порядка q и $C \subset F^n$ — групповым кодом, состоящим из k информационных символов, то имеет место соотношение

$$\mathfrak{R}\, A^T = q^k \mathfrak{B}\, B^T \tag{2}$$

между весовым вектором A группового вектора C и весовым вектором B группового кода C (здесь C' является дуальным кодом кода C), где

$$\mathfrak{A} = \left[q^i \binom{j}{i}\right] \qquad (0 \leq i \leq n,\ 0 \leq j \leq n), \tag{3}$$

$$\mathfrak{B} = \left[(-1)^j \binom{n-j}{n-i} (q-1)^{i-j}\right] \quad (0 \leq i \leq n,\ 0 \leq j \leq n) \tag{4}$$

(индексом T обозначено транспонирование).

В последней части определены некоторые тождества. Нелинейный групповой код длины n над аддитивной группой z_q целых чисел q modulo определяется в следующем виде:

$$C = \{c \mid c = (c_1, \ldots, c_n),\ \sum_{i \in I_k^j} c_i = 0(\bmod q),\quad 1 \leq j \leq q-1,\ 1 \leq k \leq s\},$$

где знаком s обозначается число радиксных знаков q представления n. Данное представление обозначается символом $(n)_q$, а символами I_k^j обозначаются следующие множества:

$$I_k^j = \{i \mid 1 \leq i \leq n,\ (i)_q = i_1 \ldots i_{k-1}\, j\, i_{k+1} \ldots i_s\} \quad 1 \leq j \leq q-1,\quad 1 \leq k \leq s\,.$$

Для представления весовых распределений нелинейных групповых кодов (при $n = 8$, $q = 3$ и $n = 15$, $q = 4$) приведены два примера.

V. V. Masgras
Department of Mathematics III
Polytechnic Institute
Splaiul Independentei Nr. 313
Sector 7, Bucharest
Rumania

Problems of Control and Information Theory, Vol. 5 (4), pp. 371–382 (1976)

ON THE MINIMIZATION OF CLASSIFICATION ERROR PROBABILITY IN STATISTICAL PATTERN RECOGNITION

J. FRITZ, L. GYÖRFI

(Budapest)

(Received October 20, 1975)

A two class pattern recognition algorithm is introduced for learning the optimal separating function within a given finite dimensional linear space of functions. It is proved that the algorithm converges with probability one to the separating function minimizing the probability of misclassification over the given class. This stochastic gradient process is based on asymptotically unbiased estimates of the gradient vector of error probability. This method is proposed to improve classification rules obtained by other methods.

1. Introduction

In statistical pattern recognition, where the patterns are coming from overlapping classes, the basic situation can be described only in probabilistic terms, namely in that of alternative hypothesis testing. In such cases the recognition problem is usually reduced to the iterative learning of the Bayesian decision rule: the learning process itself is based on the available training sequence.

We suppose that there are only two pattern classes H_+ and H_-, the recognition of which has to be based on the observed value of the r-dimensional feature vector $\xi \in E_r$. Such a recognition rule can be given by a real valued decision function $D(x)$ on E_r: we adopt H_+ or H_-, according as $D(\xi) \geq 0$ or $D(\xi) < 0$. The conditional density functions of ξ given H_+ or H_- will be denoted by $p_+(x)$ and $p_-(x)$ respectively, and $p_+ = P(H_+)$, $p_- = P(H_-) = 1 - p_+$ are the a priori probabilities of H^+ and H_-. Then the density function of ξ will be

$$p(x) = p_+ p_+(x) + p_- p_-(x),$$

while

$$P(H_+ \mid x) = P(H_+ \mid \xi = x) = \frac{p_+ p_+(x)}{p(x)}, \quad P(H_- \mid x) = P(H_- \mid \xi = x) = \frac{p_- p_-(x)}{p(x)}$$

are the a posteriori probabilities. Then, given an arbitrary decision function $D(x)$ on E_r (for which the set $D(x) = 0$ has 0 measure), the probability of decision error will be

$$P_e(D) = p_+ \int_{D<0} p_+(x)\,dx + p_- \int_{D>0} p_-(x)\,dx =$$

$$= \int_{E_r} p_+\, p_+(x)\,\frac{1 - \operatorname{sign} D(x)}{2}\,dx + \int_{E_r} p_-\, p_-(x)\,\frac{1 + \operatorname{sign} D(x)}{2}\,dx =$$

$$= \frac{1}{2} - \frac{1}{2}\int_{E_r} D_B(x) \operatorname{sign} D(x)\, p(x)\,dx\,, \tag{1}$$

where $D_B(x) = P(H_+ \mid x) - P(H_- \mid x)$ is the Bayesian decision function.

It is well known that the Bayesian decision is optimal in the sense that $P_e(D) \geq P_e(D_B)$ for any D. Unfortunately, in pattern recognition problems the a posteriori probabilities are not known. Instead, we are given a training sequence of independent pairs (ξ_1, ϱ_1), (ξ_2, ϱ_2), . . . , where $\xi_1, \xi_2 \ldots$ is a sequence of randomly selected points from E_r with common probability density $p(x)$, and $P(\varrho_n = 1 \mid \xi_n = x) = P(H_+ \mid x)$ $P(\varrho_n = -1 \mid \xi_n = x) = P(H_- \mid x)$; that is $\varrho_n = +1$ or -1 according as H_+ or H_- was true for ξ_n. Observe that this implies

$$E(\varrho_n \mid \xi_n = x) = D_B(x) \tag{2}$$

which makes it possible to estimate $D_B(x)$ from the training samples. Let us remark that — besides independence — (2) is the only essential assumption on the sequence (ξ_n, ϱ_n).

In the last few years several methods have been developed for learning the Bayesian decision rule. Among these some nonparametric estimates, as the nearest neighbor rule (see Cover and Hart [1]) and the Wagner–Wolverton algorithm (see [2] and [3]) have very nice asymptotic properties concerning the probability of misclassification, however, these nonparametric methods require extremely large amount of memory if realized on a computer. Therefore we restrict ourselves to decision functions in a parametrized form:

$$D = D(x, c) = \sum_{i=1}^{k} c^{(i)} \varphi_i(x) \tag{3}$$

where $c = (c^{(1)}, c^{(2)} \ldots c^{(k)}) \in E_k$ and $\varphi_1(x)$, $\varphi_2(x) \ldots \varphi_k(x)$ are given linearly independent interpolation functions on E_r, and we suppose that the subsets of E_r, where $\varphi_i(x) = 0$ or $D(x, c) = 0$, are of zero Lebesgue measure for each $i = 1,2 \ldots k$ and $c \neq 0 \in E_k$.

For this, the iterative learning procedures discussed in papers [4]–[10] approximate $D_B(x)$ by a sequence of decision functions $D_n(x) = D(x, c_n)$ of type (3). Unfortunately, apart from the estimations in [10] and [9], almost nothing is known on the behaviour of error probabilities $P_e(D_n)$ in these cases, since the algorithms are constructed to minimize loss functions, essentially different from P_e. This weakness of the known methods has been pointed out by Wassel and Sklansky [14] who have proposed a method to approximate the minimum error decision function of type $D(x, c) = x - c$, when $r = 1$, $k = 1$. The aim of this paper is to attack the problem in the general case. We introduce a stochastic gradient algorithm to minimize the probability of misclassification within the class of decision functions, described in (3), and convergence properties of the algorithms are investigated.

2. Results

Let us consider the error probability function

$$J(c) = P_e(D) = \frac{1}{2} - \frac{1}{2} \int_{E_r} D_B(x) \operatorname{sign} D(x, c)\, p(x)\, dx \tag{4}$$

defined on E_k by (1) and (3); we wish to find the minimum of $J(c)$ by means of a gradient algorithm. It is easy to show that if $p(x)$ is continuous, then J is differentiable under some mild conditions, at least for $c \neq 0$. The gradient vector of J at $c \neq 0 \in E_k$ will be denoted by $U(c)$, we suppose that U exists for $c \neq 0$.

Observe that $J(c)$ is a homogeneous function of order zero, that is $J(\alpha c) = J(c)$ and $\alpha U(\alpha c) = U(c)$ if $\alpha > 0$; thus we shall avoid some difficulties by minimizing J on the unit sphere $S_k = \{c;\ c \in E_k \mid\mid c \mid\mid = 1\}$; $\mid\mid \cdot \mid\mid$ and $(.\,,.)$ denote the Euclidean norm and scalar product in E_k, respectively. This, however, needs a slight modification of the usual stochastic gradient method as follows. Let $c_0 \in S_k$ be arbitrary, and define the sequence $c_1, c_2, \ldots c_n, \ldots$ by

$$b_{n+1} = c_n - \gamma_n w_{n+1}, \quad c_{n+1} = \frac{b_{n+1}}{\mid\mid b_{n+1} \mid\mid}, \tag{5}$$

$$w_{n+1} = Z_{n+1} - c_n(c_n, Z_{n+1}), \tag{6}$$

where $\gamma_n \geq 0$ is a given sequence of real numbers, and Z_{n+1} is a suitable estimator of $U(c_n)_j$; that is $Z_1, Z_2, \ldots$ is a sequence of k-dimensional random vectors constructed on the basis of the training sequence.

This modification is motivated by the observation that

$$\frac{\partial}{\partial \alpha} J(\alpha c) = \big(U(c), c\big) = 0\,,$$

that is $U(c)$ is orthogonal to c if $c \neq 0$. In accordance with this observation, w_{n+1} is the orthogonal projection of Z_{n+1} on the hyperplane, orthogonal to c_n, that is $(c_n, w_{n+1}) = 0$ and $(U(c_n), w_{n+1}) = (U(c_n), Z_{n+1})$. Observe that

$$|| b_{n+1} ||^2 = || c_n ||^2 + \gamma_n^2 || w_{n+1} ||^2 - 2\gamma_n (c_n, w_{n+1}) = 1 + \gamma_n^2 || w_{n+1} ||^2 \geq 1, \quad (7)$$

that is algorithm (5)–(6) is defined for each n.

Of course, the crucial step is to define the estimates Z_{n+1} in such a way that the algorithm (5) should converge, since no unbiased estimate is known for $U(c)$ in general. What we propose is a smoothing version of the Kiefer–Wolfovitz scheme; this technique of singular integrals has been introduced in probability theory by Bochner [11], Rosenblatt [12] and Parzen [13].

Let the absolutely integrable real function $g(t)$ satisfy the conditions:

$$\int_{-\infty}^{\infty} g(t)\, dt = 0, \quad \int_{-\infty}^{\infty} t\, g(t)\, dt = 1, \quad \int_{-\infty}^{\infty} t^2 \mid g(t) \mid dt < +\infty, \quad (8)$$

and set

$$h(t) = \int_{-t}^{\infty} g(u)\, du. \quad (9)$$

In typical cases $h(t)$ is a symmetric density function, for instance $h(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. With these notations we define

$$Z_{n+1}^{(i)} = -\varrho_{n+1} \operatorname{sign} \varphi_i(\xi_{n+1})\, \lambda_n h\left(\lambda_n \frac{D(\xi_{n+1} c_n)}{\varphi_i(\xi_{n+1})}\right), \quad (10)$$

where $\lambda_n > 0$ tends to infinity and $Z_{n+1}^{(i)}$ denotes the ith coordinate of Z_{n+1}, $i = 1,2, \ldots, k$.

It is clear that

$$|| Z_{n+1} || \leq \lambda_n K, \quad || w_{+n1} || \leq \lambda_n K_1, \quad (11)$$

where $K = \int_{-\infty}^{+\infty} | g(t) |\, dt$, and we shall show that

$$|| E(Z_{n+1} \mid c_n) - U(c_n) || \leq \lambda_n^{-1} L, \; L < \infty\,, \quad (12)$$

provided that $U(c)$ satisfies a Lipschitz condition.

If the constants γ_n and λ_n are specified as

$$\sum_{n=0}^{\infty} \gamma_n = +\infty, \quad \sum_{n=0}^{\infty} \gamma_n^2 \lambda_n^2 < \infty, \quad \sum_{n=0}^{\infty} \gamma_n \lambda_n^{-1} < +\infty, \tag{13}$$

then the following theorem holds.

Theorem. Suppose that $J(c)$ is differentiable for $c \neq 0$, and its gradient vector satisfies the uniform Lipschitz condition

$$\| U(x) - U(y) \| \leq L_1 \| x - y \| \text{ if } \| y \| = 1, \| x \| \geq \frac{1}{2}. \tag{14}$$

Then also (11) and (12) hold for the algorithm defined in (5), (6), (8)–(10) and (13), and

$$\lim_{n \to \infty} U(c_n) = 0$$

with probability one.

The proof of this theorem will be given in the Appendix; similar methods have been used in [5], [9], [18], [19]. Since $\| c_n \| = 1$, $\lim_{n \to \infty} U(c_n) = 0$ means that c_n converges to the set of stationary points of $J(c)$. Therefore, even the global minimum can be located, at least if the algorithm is started at several initial states. In Section 4 a simple condition is given, under which J has only one strict local minimum.

3. Proof of (12)

In this section we deduce the estimation method, the basic ideas are expressed by the Lemma.

Lemma. Let $0 \leq f(t) \leq 1$ be a differentiable function on E_1, the derivative of which satisfies

$$| f'(0) - f'(t) | \leq L_0 | t | \text{ if } | t | \leq \frac{1}{2}. \tag{15}$$

If $g(t)$ is specified as in (8) and $\lambda > 0$ then

$$| f'(0) - \lambda^2 \int_{-\infty}^{\infty} f(t)\, g(\lambda t)\, dt | \leq \lambda^{-1} \big(L_0 + 4 + 2 | f'(0) |\big) \int_{-\infty}^{\infty} u^2 | g(u) |\, du. \tag{16}$$

Proof. In view of (8) we have, by a change of variables, that

$$f'(0) = \lambda^2 \int_{-\infty}^{\infty} \big(f(0) + t f'(0)\big)\, g(\lambda t)\, dt,$$

thus

$$|f'(0) - \lambda^2 \int_{-\infty}^{\infty} f(t)\, g(\lambda t)\, dt| \leq \lambda^2 \int_{-\infty}^{\infty} |f(t) - f(0) - tf'(0)|\, |g(\lambda t)|\, dt,$$

however, from (15)

$$|f(t) - f(0) - tf'(0)| \leq L_0 t^2 \quad \text{if} \quad |t| \leq \frac{1}{2},$$

therefore, as $|f(t) - f(0)| \leq 1$, it follows that

$$|f'(0) - \lambda^2 \int_{-\infty}^{\infty} f(t)\, g(\lambda t)\, dt| \leq \lambda^2 L_0 \int_{|t|\leq \frac{1}{2}} t^2 |g(\lambda t)|\, dt +$$

$$+ \lambda^2 \int_{|t|>\frac{1}{2}} |g(\lambda t)|\, dt + \lambda^2 |f'(0)| \int_{|t|>\frac{1}{2}} |t|\, |g(\lambda t)|\, dt =$$

$$= \lambda^{-1} L_0 \int_{|u|\leq \frac{\lambda}{2}} u^2 |g(u)|\, du + \lambda \int_{|u|\geq \frac{\lambda}{2}} |g(u)|\, du + |f'(0)| \int_{|u|>\frac{\lambda}{2}} |u|\, |g(u)|\, du \leq$$

$$\leq \lambda^{-1} L_0 \int_{-\infty}^{\infty} u^2 |g(u)|\, du + \lambda^{-1} 4 \int_{-\infty}^{\infty} u\, |g(u)|\, du + \lambda^{-1} 2\, |f'(0)| \int_{-\infty}^{\infty} u^2 |g(u)|\, du,$$

as $\frac{4u^2}{\lambda^2} \geq 1$, $\frac{2|u|}{\lambda} \geq 1$ if $|u| \geq \frac{\lambda}{2}$, and they are nonnegative otherwise; thus the Lemma is proved.

To prove (12) set

$$J^{(i)}(c, t) = \frac{1}{2} - \frac{1}{2} \int_{E_r} D_B(x) \operatorname{sign} \big(D(x, c) + t\varphi_i(x)\big)\, p(x)\, dx \tag{17}$$

and observe that the ith coordinate $U^{(i)}(c)$ of $U(c)$ is just

$$U^{(i)}(c) = \frac{\partial}{\partial t} J^{(i)}(c, t) \quad \text{at} \quad t = 0\,. \tag{18}$$

However, in view of the Lemma, we have for the vector $U_\lambda(c)$ with coordinates

$$U_\lambda^{(i)}(c) = \lambda^2 \int_{-\infty}^{\infty} J^{(i)}(c, t)\, g(\lambda t)\, dt \tag{19}$$

the estimation

$$\| U(c) - U_\lambda(c) \| \leq \lambda^{-1} L \quad \text{if} \quad c \in S_k, \tag{20}$$

where, for example

$$L = K\left(L_1 + 4 + 2 \max_{y \in S_k} \| U(y) \|\right) \int_{-\infty}^{\infty} u^2 \mid g(u) \mid du < + \infty. \tag{21}$$

Taking into account (14), we can apply the Lemma for the functions $J^{(i)}(c, t)$ if $c \in S_k$, and we obtain (20)–(21).

As the last step, we prove that

$$U_\lambda^{(i)}(c) = - \lambda \int_{E_r} D_B(x) \operatorname{sign} \varphi_i(x)\, h\left(\lambda \frac{D(x, c)}{\varphi_i(x)}\right) p(x)\, dx\,, \tag{22}$$

whence by (2)

$$U_\lambda^{(i)}(c_n) = E(Z_{n+1} \mid c_n)\,, \tag{23}$$

which, in view of (20), implies (12).

From (8)

$$U_\lambda^{(i)}(c) = \lambda^2 \int_{-\infty}^{\infty} J^{(i)}(c, t)\, g(\lambda t)\, dt = \lambda^2 \int_{-\infty}^{\infty} J^{(i)}\left(c, \frac{u}{\lambda}\right) g(u)\, du =$$

$$= -\frac{1}{2} \lambda \int_{E_r} \int_{-\infty}^{\infty} D(x) \operatorname{sign}\left(D(x, c) + \frac{u}{\lambda} \varphi_i(x)\right) g(u)\, p(x)\, du\, dx\,, \tag{24}$$

further, if $\varphi_i(x) \neq 0$, we can write that

$$\int_{-\infty}^{\infty} \operatorname{sign}\left(D(x, c) + \frac{u}{\lambda} \varphi_i(x)\right) g(u)\, du =$$

$$= \int_{-\infty}^{\infty} \operatorname{sign} \varphi_i(x) \operatorname{sign}\left(\lambda \frac{D(x, c)}{\varphi_i(x)} + u\right) g(u)\, du =$$

$$= \operatorname{sign} \varphi_i(x) \left(\int_{-a}^{\infty} g(u)\, du - \int_{-\infty}^{-a} g(u)\, du \right) = \operatorname{sign} \varphi_i(x)\, 2h(a),$$

where $a = \lambda \dfrac{D(x, c)}{\varphi_i(x)}$; but $\varphi_i(x) \neq 0$ almost everywhere in E_r, as we supposed in (3); thus, substituting into (24), we obtain (22). We have proved that the assumptions of the Theorem imply the validity of (12), therefore the Theorem applies for our algorithm (5) and yields that $\lim_{n \to \infty} U_n(c_n) = 0$ with probability one. The remaining part of the proof will be given in the Appendix.

4. On the stationary points of J

The following simple result gives insight into the structure of stationary points of J, which may be very important in practical applications.

Proposition. If there exists a $c^* \in S_k$ with $J(c^*) = P_e(D_B)$, then

$$J(c) - J(\lambda c + \mu c^*) = \frac{1}{2} \int_A |D_B(x)| \, p(x) \, dx \geq 0 \tag{25}$$

for each $c \in E_k$, $\lambda > 0$, $\mu > 0$, where

$$A = \{x; \operatorname{sign} D(x, c) \neq \operatorname{sign} D(x, c^*), \lambda \mid D(x, c) \mid \leq \mu \mid D(x, c^*) \mid\}$$

further

$$(U(c), c - c^*) \geq 0 \text{ for each } c \neq 0. \tag{26}$$

Proof. The optimality of c^* immediately implies that $\operatorname{sign} D(x, c^*) = \operatorname{sign} D_B(x)$ a.s. on the set, where $|D_B(x)| \, p(x) > 0$, therefore

$$J(c) - J(\lambda c + \mu c^*) =$$

$$= \frac{1}{2} \int_{E_r} D_B(x) \left[\left(\operatorname{sign} D(x) \, \lambda c + \mu c^* \right) - \left(\operatorname{sign} D(x, c) \right) \right] p(x) \, dx =$$

$$= \frac{1}{2} \int_{E_r} |D_B(x)| \left[\operatorname{sign} \left(D(x, c^*) \cdot D(x, \lambda c + \mu c^*) \right) - \right.$$

$$\left. - \operatorname{sign} \left[D(x, c^*) \cdot D(x, c) \right] \right] p(x) \, dx = \int_A |D_B(x)| \, p(x) \, dx$$

as $D(x, \lambda c + \mu c^*) = \lambda D(x, c) + \mu D(x, c^*)$, thus (25) is proved. This means, however, that $J\left(c^* + t(c - c^*)\right)$ is an increasing function of $t > 0$, which implies (26), too.

Remarks. Under the assumptions of the Proposition we know that c^* is the only strict local minimum of J, if any. This is true, for example, if $p(x) > 0$ on E_r, and $D_B(x) \neq 0$ a.s. Of course, then $-c^*$ will be the only strict local maximum.

Appendix

Proof of the Theorem. From (13) we have

$$J(x) \leq J(y) + (U(y), x - y) + L_1 \| x - y \|$$

if $\| y \| = 1$, $\| x \| \geq \frac{1}{2}$ thus, taking into account (7), (11) and the relation $(U(c_n), Z_{n-1}) = (U(c_n), w_{n+1})$, we obtain that

$$J(c_{n+1}) = J(b_{n+1}) \leq J(c_n) - \gamma_n(U(c_n), Z_{n+1}) + \gamma_n^2 L_1 K^2 \lambda_n^2 = J(c_n) + L_1 K^2 \gamma_n^2 \lambda_n^2 -$$
$$- \gamma_n \| U(c_n) \|^2 - \gamma_n\big(U(c_n), Z_{n+1} - U(c_n)\big).$$

However

$$| E\big(U(c_n), Z_{n+1} - U(c_n) \,|\, c_n\big)| = | \big(U(c_n), E(Z_{n+1} \,|\, c_n) - U(c_n)\big)| \leq$$
$$\leq \| U(c_n) \| \cdot \| E(Z_{n+1} \,|\, c_n) - U(c_n) \| \leq K_2 L \lambda_n^{-1},$$

as $\max_{y \in S_k} \| U(y) \| = K_2 < + \infty$, and (12) holds.

Thus

$$E\big(J(c_{n+1})\big) \leq E\big(J(c_n)\big) + \gamma_n^2 \lambda^2 L_1 K^2 + \gamma_n \lambda_n^{-1} L K_2 - \gamma_n E \| U(c_n) \|^2,$$

whence, upon summing over n,

$$E\left(\sum_{n+0}^{\infty} \gamma_n \| U(c_n) \|^2\right) < + \infty \tag{27}$$

follows immediately; that is

$$\sum_{n=0}^{\infty} \gamma_n \| U(c_n) \|^2 < + \infty \text{ a.s.}$$

From (7) we see that $\lim_{n \to \infty} \| b_n \|^2 = 1$ a.s., therefore we have also

$$\sum_{n=0}^{\infty} \gamma_n \| U(b_n) \|^2 < + \infty \text{ a.s.} \tag{28}$$

as $U(c_n) = \| b_n \| U(b_n)$. It suffices to prove that $\lim_{n \to \infty} \| U(b_n) \| = 0$ a.s. which, of course, implies $\lim_{n \to \infty} U(c_n) = 0$ a.s.

Observe, that (28) is possible only if

$$\liminf_{n \to \infty} \| U(b_n) \| = 0 \text{ a.s.} \tag{29}$$

since $\sum_{n=0}^{\infty} \gamma_n = + \infty$; utilizing the continuity of U we deduce the statement from here. Clearly, as $1 \leq \| b_n \|^2 \leq K^2 \sum_{K=1}^{n} \gamma_k^2 \lambda_k^2 < + \infty$ from (7) for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\| U(b_{m+1}) - U(b_n) \| \leq \varepsilon \quad \text{if} \quad \| b_{m+1} - b_n \| < \delta. \tag{30}$$

From the algorithm

$$b_{n+1} = \frac{b_n}{\| b_n \|} - \gamma_n w_{n+1}$$

we have

$$b_{n+1} - b_n = b_n (\| b_n \|^{-1} - 1) - \gamma_n w_{n+1},$$

whence by (7) and (11), for each $m > n$ we obtain

$$\| b_{m+1} - b_n \| \leq \sum_{i=n}^{m} \| b_i \| \, | \, \| b_i \|^{-1} - 1 \, | + \| \sum_{i=n}^{m} \gamma_i w_{i+1} \| \leq$$

$$\leq \sum_{i=n}^{m} (\| b_i \|^2 - 1) + \| \sum_{i=n}^{m} \gamma_i w_{i+1} \| \leq K^2 \sum_{i=n}^{m} \gamma_{i+1}^2 \lambda_{i+1}^2 + \| \sum_{i=n}^{m} \gamma_i w_{i+1} \|, \quad (31)$$

where the first sum converges as m tends to infinity. On the other hand, we show that with the notation

$$I_n = \begin{matrix} 1 \text{ if } \| U(b_n) \| \geq \varepsilon \\ 0 \text{ if } \| U(b_n) \| < \varepsilon \end{matrix} \quad (32)$$

also the sequence

$$\sum_{n=1}^{m} \gamma_n I_n w_{n+1} \quad (33)$$

converges with probability one in E_r. Therefore, we can choose an integer N for almost every realization of the processes in such a way that

$$\left\| \sum_{i=n+1}^{m} \gamma_i I_i w_{i+1} \right\| < \frac{\delta}{3}, \qquad \gamma_n \| w_{n+1} \| \leq \gamma_n \lambda_n < \frac{\delta}{3}, \quad (34)$$

further

$$K^2 \sum_{i=n}^{m} \gamma_{i+1}^2 \lambda_{i+1}^2 < \frac{\delta}{3} \quad (35)$$

hold for each $m > n \geq N$. We know, however, that there exists an $n_0 = N$ with probability one, for which $\| U(b_{n_0}) \| \leq \varepsilon$; suppose that $\| U(b_{m+1}) \| \geq \varepsilon$ for an $m \geq n_0$, and let $n \leq m$ be the greatest integer with $\| U(b_n) \| \leq \varepsilon$; then from (31), (34), (35) we have

$$\| b_{m+1} - b_n \| < \frac{\delta}{3} + \frac{\delta}{3} + \frac{\delta}{3},$$

that is, in view of (30)

$$\| U(b_{m+1}) \| \leq \| U(b_n) \| + \| U(b_{m+1}) - U(b_n) \| < 2\varepsilon$$

which proves the statement of the Theorem, provided that the sum in (33) really converges. To show this, we apply a simple consequence of Proposition IV. 6.1. of Neveu [20].

Lemma 2. Let $\mathcal{A}_n \subset \mathcal{A}$ be an increasing sequence of σ-algebras in a probability space $(\Omega, \mathcal{A}, P)$, and suppose that the k-dimensional random vectors

X_n are measurable on $\mathfrak{A}$ for $n = 1,2, \ldots$ Then, if

$$\sum_{n=1}^{\infty} \| E\,(X_{n+1} | \mathfrak{A})_n \| < +\infty \quad \text{a.s.} \tag{36}$$

and

$$\sum_{n=1}^{\infty} E \,\| X_{n+1} \|^2 < +\infty\,, \tag{37}$$

then $\sum_{n=1}^{m} X_{n+1}$ converges in E_k with probability one.

Let $\mathfrak{A}$ denote the minimal σ-algebra on which $W_1, \ldots W_n$ are measurable, and set $X_{n+1} = \gamma_n I_n W_n$; then (37) follows directly from (11) and (13). Further, as I_n and c_n are measurable on $\mathfrak{A}_n$, we have

$$\begin{aligned} &\| E(\gamma_n I_n w_{n+1} \,|\, \mathfrak{A}_n) \| = \gamma_n I_n \,\| E(w_{n+1} \,|\, c_n) \| = \gamma_n I_n \,\| E(Z_{n+1} \,|\, c_n) - \\ &- c_n\big(E(Z_{n+1} \,|\, c_n), c_n\big) \| \leq \gamma_n I_n \,\| E(Z_{n+1} \,|\, c_n) \| \leq \gamma_n I_n \,\| U(c_n) \| + \\ &+ \gamma_n I \,\| E(Z_{n+1} \,|\, c_n) - U(c_n) \| \leq \gamma_n I_n \max_{y \in S_k} \| U(y) \| + \gamma_n \lambda_n^{-1} L\,, \end{aligned}$$

from the defitnition of I_n and from (12), that is

$$\sum_{n=1}^{\infty} \| E(X_{n+1} \,|\, \mathfrak{A}_n) \| \leq \max_{y \in S_k} \| U(y) \| \sum_{n=1}^{\infty} \gamma_n I_n + L \sum_{n=0}^{\infty} \gamma_n \lambda_n^{-1}\,.$$

The convergence of the first sum on the right hand side follows immediately from (28), while that of the second one is postulated in (13), therefore Lemma 2 yields (33), and the Theorem is completely proved.

References

1. *Cover, T., Hart, P.* (1967), The nearest neighbor pattern classification. IEEE Trans. on Information Theory **13**, 21–27.
2. *Wolverton, C. T., Wagner, T. J.* (1969), Asymptotically optimal discriminant functions for pattern classification. IEEE Trans. on Information Theory **15**, 258–265.
3. *Rejtő, L., Révész, P.* (1972), Density estimation and pattern classification. Problems of Control and Information Theory.
4. *Specht, D. F.* (1967), Generation of polynomial discriminant functions for pattern recognition. IEEE Trans. on Electronic Computers **15**, 308–319.
5. *Aizerman, M. A., Braverman, E. M., Rosonoer, L. I.* (1970), The Method of Potential Function in the Theory of Machine Learning. Nauka, Moscow (in Russian).
6. *Kashyap, R. L.* (1970), Algorithms for pattern classification. Adaptive, Learning and Pattern Recognition Systems, Mendel-Fu (Ed.) Academic Press, New York.
7. *Csibi, S.* (1971), Simple and compound processes in iterative machine learning. CISM Summer Course, Udine, Italy.
8. *Csibi, S.* (1972), Stability and complexity of learning algorithms. CISM Summer Course, Udine, Italy.

9. *Fritz, J.*, On a pattern classification algorithm of R. L. Kashyap. Problems of Control and Information Theory. v. 2, No. 1, 1973, pp. 81—93.
10. *Györfi, L.* (1972), Estimation of probability density and optimal cecision function in RKHS. Proceedings of European Meeting of Statisticians, Budapest.
11. *Bochner, S.* (1955), Harmonic Analysis and the Theory of Probability. Univ. of Calif. Press, Berkeley.
12. *Rosenblatt, M.* (1971), Curve estimates. Ann. Math. Statist. **42,** 1815–1842.
13. *Parzen, E.* (1962), On estimation of probability density function and mode. Ann. Math. Statist. **33,** 1065–1076.
14. *Wassel, G. N., Sklansky, J.* (1972), Training a one dimensional classifier to minimize the probability of error. IEEE Trans. on Systems, Man, and Cybernetics **2,** 533–541.
15. *Györfi, L.* (1974), On the estimation of asymptotic error probability. IEEE Trans. on Information Theory **IT-20,** 277–278.
16. *Venter, I. H.* (1966), On Dvoretzky stochastic approximation theorems. Ann. Math. Statist. **37,** 1534–1544.
17. *Venter, I. H.* (1967), On convergence of the Kiefer—Wolfowitz approximation procedures. Ann. Math. Statist. **38,** 1031–1036.
18. *Fritz, J.* (1973), Stochastic approximation for finding local maxima of probability densities. Studia Sci. Math. Hungar. (to appear).
19. *Nevelson, M. B., Khasminskii, R. Z.* (1972), Stochastic Approximation and Iterative kstimation Methods. Nauka, Moscow (in Russian).
20. *Neveu, J.* (1964), The Mathematical Foundations of Probability Theory. Masson, Paris.

О минимизации вероятности ошибки классификации в области статистического распознавания

Й. ФРИЦ, Л. ДЕРФИ

(Будапешт)

Рассматривается алгоритм распознавания образов по двум классам при обучении оптимальной разделяющей функции в заданном пространстве линейных функций ограниченной размерности. Доказывается сходимость с вероятностью 1 предложенного алгоритма к разделяющей функции, минимизирующей вероятность ошибочной классификации над данным классом образов. Предложенный стохастический градиентный метод основан на асимптотически неискаженной оценке вектора градиента вероятности ошибки. Методику рекомендуют применить для улучшения правил классификации, полученных путем применения других методов.

J. Fritz
Mathematical Research Institute
Hungarian Academy of Sciences
H-1053 Budapest, V. Reáltanoda u. 13—15.

L. Györfi
HEI, Technical University of Budapest
H-1111 Budapest, XI. Stoczek u. 2.
Hungary

MTA Könyvtára
988/19 77

BOOK REVIEW

R. Lauber: **Prozessautomisierung I (Process Control I)** (Berlin–Heidelberg–New York: Springer-Verlag, 1976, 263 pp, in German.). Reviewed by L. Lakatos, Department for Automation, Technical University, Budapest, Hungary.

Processing the data of an industrial process, gaining information about the operation of a process, controlling a process — these mean, that a new range of engineering has been developed in connection with industrial processes, the process control. The topics of process control are not only the elements and the programming methods of process control computers, but also the methods of using them for controlling industrial processes. In this book the computer is considered as an aid from the viewpoint of a control engineer. Though the process control computers spread around the whole world, e.g. nowadays the decentralized structures built from microprocessors are used more and more widely, the systematical study of this new branch of science does not keep level with the development of the installed process control systems. Thus most of the specialists dealing with process control can rely only on their experiences and on papers dealing with rather specific process control systems. The book presented here meets a long-felt need, it is a school-book. Its purpose is to give a deep basis and good practical methods.

The book is the first volume of a school-book consisting of two volumes. The first volume presents the structure and the operation of process control systems and the second volume, which will be published later, will present the methods arising in using the computers for supervisory control, sequential control, direct digital control, optimization, etc. The selection of the material of the book, the structure of the chapters were determined by the fact, that control engineers will read the book, not producting engineers or system programmers. Thus the chapters on instrumentation and programming are written only on a lower level, which is reasonable from the viewpoint of the user.

This first volume consists of three parts, each of them consists of four chapters.

In part A (Introduction to process control) the used technical terms are explained and the structure of process control systems is shown briefly. The chapters of this part are the following:

1. From simple open-loop and closed-loop control until the process control
2. The structure of process control systems
3. Elements of process control systems
4. Planning of process control systems

In part B (The instrumentation and hardware of process control systems) the structural questions are detailed.

The chapters of this part:

5. The central unit of a process control computer
6. Systems of process peripherical devices
7. Units for the input and output of process signals
8. Noiseless signal transfer between the process and the computer.

Part C (Programming of process control systems) is a very valuable part of this book. The author presents the basic methods of real time programming systematically

and clearly, thus the engineer can learn it easily. The operation of the real time operating systems is presented not only generally, by abstract words, but considering an example. A simplified example, a mini operating system is also shown, and it is extended step by step. This is an example-like, yet systematical presentation of real time operating systems, which is unique in literature. The chapters of part C:

9. Fundamentals of real time programming
10. Real time operating systems
11. Real time programming languages for process control computers (e.g. PEARL)
12. Programming systems for process control.

The Appendix of the book contains the fundamentals of the real time programming language PEARL, which is used as an example in chapter 11.

The readers of the book are supposed to be familiar with the fundamentals of computers and data processing. It would be worth translating this book into English.

Printed in Hungary

AKADÉMIAI NYOMDA, BUDAPEST

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

AN APPROACH TO DECOMPOSITION OF PROBABILISTIC AUTOMATA

A. KH. GIORGADZE

(Tbilisi)

The approach to probabilistic automata decomposition proposed here, is dinstinguished by the fact that a subautomaton separated from a set of automata into which the original automaton is decomposed may be not isomorphic to the original automaton but represent the same event as the original one with arbitrary point of cut. Decomposition of this kind, termed λ'-decomposition was considered for a class of probabilistic automata. For generalized automata, decomposition conditions are established. A λ'-decomposable automaton is shown to be decomposable with arbitrary λ, $0 < \lambda < 1$.

Usually, the problem of automaton decomposition is posed as follows [1]: for a given automaton A (provided that it meets appropriate decomposition theorems) an automaton B is constructed which is a system of automata interconnected after a certain pattern and having less states than automaton A and which enables separation of subautomaton B' isomorphic to automaton A. If A is a probabilistic automaton (PA) representing event E with cut-point λ, B' will obviously represent the same event with the same cut-point λ.

The present work considers the following problem: let PA A be given representing event E with cut-point λ and not satisfying the usual [1] decomposition conditions. Let, however, there be a system of automata (automaton B^0) with preassigned constraints on interconnections between automata (whose number of states is less than that of automaton A) such that some subautomaton B^0 of automaton B_0 represents the same event E but with cut-point λ' not necessarily coinciding with λ in the general case. This kind of decomposition will be referred to as λ'-decomposition of probabilistic automaton.

Define some of the conditions for λ'-decomposition.

The idea is as follows: a notion of generalized automaton (GA) decomposition is introduced. Generalized automata were introduced and discussed in [2]. Necessary and sufficient decomposition conditions are established followed by transformation of the given PA into a generalized one. If this GA satisfies the generalized automata decomposition conditions, a system of generalized automata is determined into which GA is decomposable. Next, probabilistic automata are substituted for each automaton in the system so

that in the resulting system of PA's a subautomaton could be identified representing the same event as the original PA.

Definition 1. By analogy with [2], GA A^0 is a quintuple $A^0 = (X, S, \{\mathbf{M}(x)\}, g_0, f_0)$ where X is input alphabet, $S = (s_1, s_2, \ldots, s_n)$ is the state set, $\mathbf{M}(x)$ is a matrix related to letter $x \in X$, g_0 is n-dimensional row vector (original vetor), f_0 is n-dimensional column vector (vector of finite states); components of $\mathbf{M}(x)$, g_0 and f_0 being real numbers.

Sometimes GA will be understood as a quadruple $(X, S \{\mathbf{M}(x)\}, g_0)$.

Definition 2. Cascade product of generalized automata $A_1^0 = (X, S^1 = (s^1, \ldots, s_n^1), \{\mathbf{M}^1(x_i)\}, g_0^i = (p_1, p_2, \ldots, p_n)$ and $A_2^0 = (X \times Y, S^2 = (s_1^2, \ldots, s_m^2), \{\mathbf{M}^2(x_i, y_i)\}, g_0^2$ where Y is the set of letters $y_1, y_2, \ldots, y_n$ generated by A_1^0 when it is in states $\varrho_1^1, \varrho_2^1, \ldots, \varrho_n^1$, respectively, is automaton

$$A_1^0 \oplus A_2^0 = (X, S^2 \times S^2, \{\mathbf{M}(x_i)\}, g_0)$$

where $g_0 = (p_1 g_0^2, p_2 g_0^2, \ldots, p_n g_0^2)$, $\mathbf{M}(x_i) = \mathbf{M}^1(x_i) \times \{\mathbf{M}^2(x_i, y_j)\}$ is a cascade product [3] of matrix $\mathbf{M}^1(x_i)$ and set $\{\mathbf{M}^2(x_i, y_j)\}$, i.e. a matrix of the following form:

$$\mathbf{M}(x_i) = \begin{bmatrix} p_{11}\mathbf{M}^2(x_i, y_1) & \ldots & p_{1n}\mathbf{M}^2(x_i, y_1) \\ p_{21}\mathbf{M}^2(x_i, y_2) & \ldots & p_{2n}\mathbf{M}^2(x_i, y_2) \\ \ldots & \ldots & \ldots \\ p_{n1}\mathbf{M}^2(x_i, y_n) & \ldots & p_{nn}\mathbf{M}^2(x_i\, y_n), \end{bmatrix} \tag{1}$$

where p_{ij} are elements of $\mathbf{M}^1(x_i)$. Cascade product of automata means that they are connected serially. If $\mathbf{M}^2(x_i, y_j) = \mathbf{M}^2(x_i)$ for each letter $x \in X$ and all $y_j \in Y$, product $A_1^0 A_2^0$ is called direct product of automata and means that they are connected in parallel.

Definition 3. Automaton A^0 allows serial decomposition into automata A_1^0 and A_2^0 if subautomaton $A^{0\prime}$ may be isolated in automaton $A_1^0 \oplus A_2^0$, isomorphic to A^0, and *if* $|A_1^0| < |A^0|$, $|A_2^0| < |A^0|$ ($|A|$ is the number of states of automaton A).

Theorem 1. For GA A^0 to allow serial decomposition into automata A_1^0 and A_2^0, it is necessary and sufficient that there be two decompositions of automaton A state set S into disjoint subsets $\pi = (\pi_1, \ldots, \pi_m)$ and $\tau(\tau_1, \ldots \tau_k)$ such that

1. $\pi - \tau = 0$;

2. $m < |A^0|$, $k < |A^0|$, $|\pi_i| < |A^0|$, $|\tau_j| < |A^0|$, $i = \overline{1, m}$, $j = \overline{1, k}$

$j = 1, x$ ($|\pi_i|$ is the number of elements in subset π_i of decomposition π);

3. for each π_a and arbitrary π_b, π_c and all $x \in X$ either sign $p^x_{i\tau_a \cap \pi_b} =$ $=$sign $p^x_{i\tau_a \cup \pi_b}$ for all $i \in \pi_a$ and all $\tau_a \in \tau$, or sign $p^x_{i\tau_a \cap \pi_b} =$ sign $p^x_{i\tau_a \cap \pi_c}$ for all $i \in \pi_a$ and all $\tau_a \in \tau$;

4. for all π_a, arbitrary $i, j \in \pi_a$ and arbitrary τ_b, τ_c (τ_b may coincide with τ_c) and all $x \in X$

$$\frac{p^x_{i\tau_b \cap \pi_1}}{p^x_{j\tau_c \cap \pi_1}} = \frac{p^x_{i\tau_b \cap \pi_2}}{p^x_{j\tau_c \cap \pi_2}} = \ldots = \frac{p^x_{i\tau_b \cap \pi_m}}{p^x_{i\tau_c \cap \pi_m}}$$

where p_{ij} are elements of matrix $\mathbf{M}(x)$ of automaton A^0.

Let conditions of the theorem be met. Consider π_i and π_j. Let $p^x_{i\tau_b \cap \pi_j} =$ $= \alpha^x_{ij} p^x_{i\tau_b \cap \pi_i}$ where α^x_{ij} is some real number. Demonstrate that $p^x_{i\tau_k \cap \pi_i} =$ $= \alpha^x_{ij} p^x_{i\tau_k \cap \pi_j}$ for any $\tau_k \in \tau$. In fact, this follows from equality $p^x_{i\tau_b \cap \pi} / p^x_{i\tau_k \cap \pi_i} =$ $p^x_{i\tau_k \cap \pi_j} / p^x_{i\tau_b \cap \pi_j}$ Consieer some π_a. On the basis of conditions 3 and 4 of the theorem, one can prove that also $p^x_{i\tau_k \cup \pi_i} = \alpha^x_{ij} p^x_{i\tau_k \cup \pi_j}$ for any $i \in \pi_i$.

Subsets π_i and π_j define corresponding submatrices in matrices $\mathbf{M}(x)$ of automaton A^0. Let π_d be maximal (with respect to number of elements) subset in π. Expand submatrices $\pi_i\pi_j$, $i, j = 1,2, \ldots, m$, up to $|\pi_d| \times |\pi_d|$ by adding to them columns and rows. They may be filled so that the resulting matrix has the form of (1). In doing so, submatrix $\pi_i\pi_j$ multiplied by α^x_{ij} will elementwise coincide with submatrix $\pi_i\pi_j$. According to definitions 2 and 3, numbers α^x_{ij} define elements of matrix $\mathbf{M}^1(x)$ of automaton A^0_1, and submatrix $\pi_i\pi_j$ is matrix $\mathbf{M}^2(x, y_i)$ of automaton A^0_2. Necessity of the theorem conditions is, thus, obvious.

Let automaton $A^0 = (X, S = (\varrho_1, \ldots, \varrho_n), \{\mathbf{M}(x)\}, g_0, f_0)$ representing event E with the cut-point η be decomposable into two parallel automata $A^0_1 = (X, S^1 = (\varrho^1_1, \ldots, \varrho^1_m), \mathbf{M}_1(x), g^1_0)$ and $A^0_2 = (X, S^2 = (\varrho^2_1, \ldots, \varrho^2_k), \{\mathbf{M}_2(x)\}, g^2_0)$, elements of matrix $\mathbf{M}(x)$ being non-negative.

Theorem 2. There is a system of four parallel GA's A_1, A_2, A_3 and A_4 representing the same event as GA A^0, $|A_1| = |A_2| = m + 3$, $|A_3| =$ $= |A_4| = k + 3$.

Prove, first, a lemma about existence of automaton

$$A^{01} = (X, S = (\varrho'_1, \ldots, \varrho'_{(m+3)(k+3)}), \{\mathbf{M}'(x)\}, g'_0, f'_0)$$

representing the same event as automaton A^0 and decomposable into two parallel automata

$$A^{0'}_1 = (X, S^{1'} = (\varrho^{1'}_1, \ldots, \varrho^{1'}_{m+3}), \{\mathbf{M}'_2(x)\}, g^{1'}_0)$$

and

$$A^{0'}_2 = (X, S^{2'} = (\varrho^{2'}_1, \ldots, \varrho^{2'}_{k+3}), \{\mathbf{M}'_2(x)\}, g^{2'}_0)$$

so that matrices $\mathbf{M}'(x)$, $\mathbf{M}'_1(x)$ and $\mathbf{M}'_2(x)$ are stochastic.

Proof. Let $\overline{X}$ be a set of all words $\bar{x}$ over alphabet X, and $\mathbf{M}_1(\bar{x})$ be a corresponding matrix describing operation of A_1^0 when word $\bar{x} \in \overline{X}$ occurs. Denote by $\delta(\bar{x})$ a number greater than the greatest sum of two rows of matrix $\mathbf{M}_1(\bar{x})$ and introduce numbers $\delta_i(\bar{x})$ such that matrix

$$\mathbf{M}'(x) = \left[\begin{array}{c|ccc|cc} 1 & 0 & \dots & 0 & 0 & 0 \\ \hline \delta_1(\bar{x}) & & & & 0 & 0 \\ \vdots & & \mathbf{M}_1(x)/\delta(\bar{x}) & & \vdots & \vdots \\ \delta_n(\bar{x}) & & & & 0 & 0 \\ \hline 0 & 0 & \dots & 0 & 1/\delta(\bar{x}), & 1-1/\delta/(\bar{x}) \\ 0 & 0 & \dots & 0 & 0 & 1 \end{array}\right]$$

be stochastic. For automaton A_2^0 introduce similarly numbers $\gamma(\bar{x})$ and $\gamma_i(\bar{x})$ and stochastic matrix $\mathbf{M}_2'(x)$. Take that $\mathbf{M}'(\bar{x}) = \mathbf{M}_1'(\bar{x}) \times \mathbf{M}_2'(\bar{x})$ according to (1). Assume, moreover, that $g_0^{1'} = (0, g_0^1, \eta_1, 0)$ and $g_0^{2'} = (0, g_0^2, \eta_2, 0)$ where $\eta_1 \cdot \eta_2 = \eta$. Introduce $g_0' = g_0^{1'}, g_0^{2'}$ and $f_0' = (0, 0, 0, 0, 0, f_0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0)^T$. It may be readily verified that

$$g_0' M'(\bar{x}) f_0' = \left(1/\delta(\bar{x})\, \gamma(\bar{x})\right) \cdot (g_0^1 g_0^2 \mathbf{M}(\bar{x}) f_0 - \eta),$$

and this completes the proof of the lemma.

Let now $A^0 = (X, S = (\varrho_1, \dots, \varrho_n), \{\mathbf{M}(x)\}, g_0, f_0)$ be an automaton representing event E with cut-point $\mathbf{G}$ such that all components of $g_0 = (p_1, p_2, \dots, p_n)$ are positive and $\{\mathbf{M}(x)\}$ is a set of stochastic matrices.

Introduce automaton $A_c^0 = (X, S, \{\mathbf{M}(x)\}, p^{-1}g_0, f + (c, c, \dots, c)^T)$ where $p = p_1 + \dots p_n$ and c is a number chosen so that all components of the finite state vector be positive. Then

$$p^{-1} g_0 \mathbf{M}(\bar{x}) \left(f + (c, c, \dots, c)^T\right) = p^{-1} g_0 \mathbf{M}(\bar{x}) f + c$$

and automaton A_c^0 represents event E. Further, denote components of vector $f + (c, c, \dots, c)^T$ through $d_1, d_2, \dots d_n$. Introduce $d = d_1 + d_2 + \dots + d_n$ and $d_i' = d_i/d$. According to Theorem 16 of [2], there is a probabilistic automaton

$$A = \left(X, S = (\varrho_1, \varrho_2, \dots \varrho_n, \varrho_{n+1}, \dots, \varrho_{n^2}), \{H(x)\}, g, F\right)$$

where $g = n^{-1}(p^{-1}g_0, \dots, p^{-1}g_0)$, $p^{-1}g_0$ is taken n times, $F = (S_1, S_{n+2}, S_{2n+3}, \dots, S_{n^2})$, and

$$\mathbf{H}(x) = \begin{bmatrix} d_1' \mathbf{M}(x) & \dots & d_n' \mathbf{M}(x) \\ d_1' \mathbf{M}(x) & \dots & d_n' \mathbf{M}(x) \\ \dots\dots\dots\dots\dots & & \\ d_1' \mathbf{M}(x) & \dots & d_n' \mathbf{M}(x) \end{bmatrix}.$$

According to usual decomposability conditions of PA's, automaton A may be decomposed into two parallel automata

$$A_1 = \left(X, S_1 = (\varrho'_1, \ldots, \varrho'_n), \{\mathbf{M}_1(x)\}, g_1\right)$$

and

$$A_2 = (X, S_2 = (\varrho^2_1, \ldots, \varrho^2_n), \{\mathbf{M}_2(x)\}, g_2$$

where $g_1 = (1/n, \ldots 1/n)$, $g_2 = p^{-1}g_0$, $\mathbf{M}_2(x) = \mathbf{M}(x)$ and

$$\mathbf{M}_1(x) = \begin{bmatrix} d'_1 & \ldots & d'_n \\ d'_1 & \ldots & d'_n \\ \ldots & \ldots & \ldots \\ d'_1 & \ldots & d'_n \end{bmatrix}.$$

Let $A = (X, S = (\varrho_1, \ldots, \varrho_n), \{\mathbf{M}(x)\}, g, f)$ be a PA representing event E with cut-point λ and not allowing decomposition. Let, however, matrices $\mathbf{C}(x) = |C^x_{ij}|$, $C^x_{1k} = C^x_{2k} = \ldots = C^x_{nk}$, $\forall_k = \overline{1, n}$ exist such that matrices $\mathbf{M}(x) + \mathbf{C}(x)$ satisfy conditions of Theorem 1 about parallel decomposition into $\pi = (\pi_1, \pi_2, \ldots, \pi_m)$ and τ. Consider matrices $\mathbf{M}'(x)$ obtained from $\mathbf{M}(x)$ in the following manner. To each submatrix of matrix $\mathbf{M}(x)$ defined by a pair of blocks π_i, π_j from the right add a column with elements $-1/m$ and from below a row equal to the first row of the submatrix. In the same manner transform matrices $\mathbf{C}(x)$ by adding a row and a column but with elements $1/m$. Denote resulting matrices by $\mathbf{C}'(x)$. Introduce a sum

$$\mathbf{M}'(x) + \mathbf{C}'(x) = \mathbf{M}''(x).$$

It is easy to verify that for any word $\bar{x}$

$$\mathbf{M}''(x) = \mathbf{M}'(x) + \alpha^{l(\bar{x})-1}, \ \cdot\mathbf{C}'(\bar{x})$$

where $\alpha = \sum_j c'_{ij}$, and $l(\bar{x})$ is the length of word $\bar{x}$. Introduce matrix

$$\mathbf{Q}(\bar{x}) = \begin{bmatrix} \mathbf{M}''(x) & \mathbf{0} \\ \mathbf{0} & \alpha^{l(\bar{x})-1}\mathbf{C}'(\bar{x}) \end{bmatrix}$$

and consider automaton $A_t = (X, S_t, \{\mathbf{Q}(x)\}, g_t, f_t)$ where $g_t = (g, g)$, $f_t = (f, -f)$. Then

$$g_t\,\mathbf{Q}(\bar{x})\,f_t = g\,\mathbf{M}'(\bar{x})\,f, \ \forall\bar{x}$$

i.e. automaton A_t represents the same event as automaton $A_0 = (X, S_0, \mathbf{M}'(x), g_0, f_0)$. Automaton A is subautomaton of the latter. Further, it is easy to see that automaton A_t is decomposable into two serially connected automata

A_t^1 and A_t^2. Automaton A_t^1 is described by matrix $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$, and automaton A_t^2 by matrices $\mathbf{M}''(x)$ and $\mathbf{C}'(x)$, automaton A_t^2 being decomposable into two parallel automata. Thus a system of cascaded generalized automata [3] results where a subautomaton is isolated representing the same event as the original PA.

Illustrate the above-said by an autonomous PA described by matrix

$$\mathbf{M} = \begin{bmatrix} 0.2 & 0.1 & 0.2 & 0.4 & 0.1 \\ 0.2 & 0.2 & 0 & 0.4 & 0.2 \\ 0.2 & 0 & 0.4 & 0.4 & 0 \\ 0.2 & 0.1 & 0.2 & 0.4 & 0.1 \\ 0.2 & 0.2 & 0 & 0.4 & 0.2 \end{bmatrix}.$$

It is easy to check that the corresponding Markov chain may be enlarged only with respect to decompositions $\pi_1 = (\overline{1,4};\ \overline{2,5};\ 3)$, $\pi_2 = (\overline{1,4};\ \overline{2,5};\ \overline{3})$ and $\pi_3 = (\overline{1,4};\ 2,5;\ \overline{3})$ of which only π_2 and π_3 satisfy the condition that $\pi_2 \cdot \pi_3 = 0$ but do not simultaneously satisfy the independence condition of [1]. This means that the automaton under study cannot be decomposed into two parallel automata. However, if a matrix $\mathbf{C}$ whose rows are similar and equal to [0.2 0 0 0 0 0] is added to matrix $\mathbf{M}$, autonomous PA described by matrix $\mathbf{M} + \mathbf{C}$ is decomposable into automata described by matrices

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0.4 & 01. & 0.2 \\ 0.4 & 0.2 & 0 \\ 0.4 & 0 & 0.4 \end{bmatrix}.$$

By applying results of Theorem 2 to automaton A_t^2, obtain decomposition of the original automaton A. It is easy to see that this decomposition is λ'-decomposition. Let PA A represent event E with cut-point λ and allow λ'-decomposition. Then the following theorem holds.

Theorem 3. Automaton A allows λ decomposition.

Proof. Let B_0 be a PA describing a system of PA's $(A_1, A_2, \ldots, A_m)$ meeting conditions $|A_i| < |A|$ and constraints on interconnections between A_i, $i = \overline{1, m}$. Let B_0^1 be a subautomaton of PA B_0 representing E with cut-point λ^1 and let $\{\mathbf{M}(x)\}$ be the set of its transition matrices. According to [2], there is a PA (denote it through B_0^2) representing E with any cut-point η, $0 < \eta < 1$, and the set of transition matrices of PA B_0^2 is as follows:

$$\left\{\begin{bmatrix} \mathbf{M}(x) & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}\right\}.$$

It is easy to see that PA B_0^2 allows decomposition into two serially connected automata, the first one being an autonomous PA with a transition matrix $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and the second one being some PA (denote it through B_0^3) whose set of transition matrices is $\{\{\mathbf{M}(x)\}, \mathbf{I}\}$ where $\mathbf{I}$ is identity matrix of the same dimensionality as $\mathbf{M}(x)$. Like B_0^2, PA B_0^3 allows decomposition. This completes the proof.

References

1. *Bacon, G.*, The Decomposition of stochastic Automata, Information and Control, **7**, 3, 1964, pp. 320–339.
2. *Turakainen, P.*, On probabilistic automata and their generalizations, Ann. Sci. Fenn. **A1**, 429, 1968, pp. 10—52.
3. *Paz, A.*, Introduction to probabilistic automata. Academic Press, New York and London, 1971, ch. 2.

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

H–1111 Budapest XI., Goldmann György tér 3, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4–5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to eturn by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которых должен содержать рисунки, таблицы и литературу) направляются

Е. Д. Теряеву — Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский проспект, 14, Москва В-71 СССР

или

Я. Кочишу — Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета
H–1111 Будапешт XI площадь Голдманн Д., 3,
Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4–5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру .

Статье должна предшествовать аннотация объемом 50–100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме — на русском языке, если статья написана на английском, и наоборот – должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем резюме не должен превышать 10–15% объема статьи.

Авторам высылается корректура статьи, которую они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Krasovskii, N. N.: Game-theoretic control under incomplete phase-state information (*Красовский Н. Н.* Игровое управление при неполной фазовой информации) 291

Frigyes, A.: A special method of fast Fourier transform to determine the low and middle frequency portion of the frequency characteristic of a control system (*Фридеш А.* Специальный метод ускоренного преобразования Фурье для определения областей низких и средних частот частотной характеристики системы автоматического управления) 303

Гиоргадзе А. Х. Об одном подходе к задаче декомпозиции вероятностных автоматов (*Giorgadze, A. Kh.:* An approach to decomposition of probabilistic automata) 321

Rózsa, P., Sinha, N. K.: Decoupling and pole placement in linear multivariable systems: A direct algebraic approach (*Рожа П., Синха Н. К.* Разбиение линейных время-инвариантных многомерных систем) 329

Masgras, V. V.: Connection between the weight distributions of a group code and its dual (*Масграс В. В.* Связь между некоторым групповым кодом и весовыми распределениями его дуального кода) 359

Fritz, J., Györfi, L.: On the minimization of classification error probability in statistical pattern recognition (*Фриц Й., Дерфи Л.* О минимизации вероятности ошибки классификации в области статистического распознавания) 371

Book review 383

316.920

VOL. 5 * NUMBER 5–6
TOM ТОМ НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1976

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

is an international quarterly sponsored jointly by the Presidium of the Academy of Sciences of the USSR and of the Hungarian Academy of Sciences. It offers publicity for original papers and short communications of the following topics:

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

— theory and methods of adaptation, learning, identification and pattern recognition

— methods of information processing; application of digital computers in control and communication systems

— new physical principles in constructing technical devices for automation, control and information processing

The six issues published per year make up a volume of some 480 pp.

While this quarterly is mainly a publication forum of the research results achieved within the USSR and Hungary, also papers of international interest from other countries are welcome.

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

H-1389 Budapest 62, P.O.B. 149, Hungary

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

H-1363 Budapest P.O.B. 24, Hungary

Академия наук СССР и Академия наук ВНР совместно издают журнал

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал публикует материалы о новых результатах научных исследований, разработок и краткие сообщения о достижениях в следующих областях науки:

— общая теория процессов управления и теория систем
— теория автоматов
— теория информации
— исследование операций; теория сложных систем
— теория управления экономическими системами; модели систем
— теория и методы адаптации, обучения, идентификации и распознавания образов

— методы обработки информации; применение вычислительных машин в системах управления и передачи информации

— новые физические принципы создания технических средств автоматики, управления и передачи информации

Журнал быхогит шестъ раз в год общим объёмом 480 печатных страниц.

Журнал публикует статьи о научных достижениях обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

H-1389 Будапешт 62, п. о. 149, Венгрия

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии наук Венгрии

H-1363 Будапешт п. о. 24, Венгрия

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)
F. CSÁKI (Budapest)

DEPUTY EDITORS

V. S. PUGACHEV (Moscow)
V. I. SIFOROV (Moscow)
S. CSIBI (Budapest)

COORDINATING EDITORS

M. A. GAVRILOV (Moscow)
L. KALMÁR (Szeged)

S. V. EMELYANOV (Moscow)
I. CSISZÁR (Budapest)

N. N. KRASOVSKII (Sverdlovsk)
A. PRÉKOPA (Budapest)

E. P. POPOV (Moscow)
L. VARGA (Budapest)

E. D. TERYAEV (Moscow)
J. KOCSIS (Budapest)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)
Ф. ЧАКИ (Будапешт)

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Ш. ЧИБИ (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

М. А. ГАВРИЛОВ (Москва)
Л. КАЛМАР (Сегед)

С. В. ЕМЕЛЬЯНОВ (Москва)
И. ЧИСАР (Будапешт)

Н. Н. КРАСОВСКИЙ (Свердловск)
А. ПРЕКОПА (Будапешт)

Е. П. ПОПОВ (Москва)
Л. ВАРГА (Будапешт)

Е. Д. ТЕРЯЕВ (Москва)
Я. КОЧИШ (Будапешт)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES

BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. (5, 6) pp. 385—399 (1976)

A MULTI-CRITERIA APPROACH TO APPLIED *R* AND *D* PLANNING: THE CASE OF QUALITATIVE CRITERIA

S. V. EMELYANOV, O. I. LARICHEV

(Moscow)

(Received March 20, 1975)

The paper considers development of a method for planning applied scientific research and development (*R* and *D*) allowing for both qualitative and quantitative estimates of *R* and *D* projects. Major stages of an heuristic algorithm are described that was developed for *R* and *D* plan generation.

1. Portfolio optimization problem

There is a familiar formulation of applied research and development planning: N, R and D projects are given, the ith project ($i = 1,2,3, \ldots N$) being characterized by a certain value v_i (income, usefulness, cost effectiveness, etc.). Projects require appropriate resources. Constraints $R^v(t)$ are imposed on resources of the vth type (manpower, money, material) necessary for implementation of R and D projects. Several versions of an R and D project are possible differing in volume and type of required resources, and in project duration (all versions of one project are assumed to lead to the same final result). It is required to select such a set of R and D projects (with due allowance for the resource constraints and the project versions) that would have the maximal effectiveness index. It should be noted here that, most commonly, the sum of values of R and D projects included into the plan is regarded as such an index.

This problem was termed "portfolio optimization" problem [1, 2, 3]. There is a bulky literature dealing with various approaches to the problem in terms of mathematical programming.

All the known "portfolio optimization" methods estimate each project through the totality of its characteristics (quality estimation criteria).

These criteria may be grouped into two categories. The first one involves criteria characterizing the expected resource expenditures and economical effectiveness of a project. Various criteria of this group are used within the framework of mathematical models of cost and economical effectiveness. For applied R and D, these models are rather beyond question. The very existence of approved techniques for calculation of project cost and economical

effectiveness makes it possible to regard the criteria of this group as objective in that both their composition and methods for estimate determination are independent of the governing body of some planning organ.

Along with these purely economical criteria, in the actual environment planning organs are using in increasing frequency other criteria reflecting the scientific and technical policy, and the specific features of a particular organization. Among such criteria there are, for example, estimates of the working quality of potential personnel, chances of success, correspondence of the expected results to the world level.

The necessity to involve into planning noneconomical criteria is due to the complexity of the environment, the impossibility of quantitative estimation of the effects of some R and D projects, and to directions of the superior levels. As a rule, the criteria have qualitative estimate scales, with each estimate being coached into words. Correspondence of the expected results to the world standards may be estimated, for instance, by the following scale:

a) expected R and D results are superior to the world standard;
b) expected R and D results correspond to the world standard; and
c) expected R and D results are inferior to the world standard.

Criteria of the second group are subjective in that both their composition and estimate scales are established by a particular planning body and cannot be accepted universally. Therefore, a model integrating criteria of both groups into a single project quality estimate cannot be universal.

The known methods of portfolio optimization may be grouped with respect to criteria used for project estimation and to models used for general evaluation of projects.

Methods evaluating project profitability [4, 5] use, as a rule, criteria of the first group only. A monetary index characterizing economical effect of project implementation is assigned to each project. Criteria of the second group cannot be integrated into the project profitability evaluation methods because of the impossibility of establishing their monetary equivalents. This is an evident demerit of this group of methods. Practically, when making decisions, the planning organs allow for many of the second-group criteria, although not systematically. Results obtained through project economical effectiveness models are corrected, not always consistently. Project profitability models cannot make the planning body sure that its orientation in science and technology was sufficiently taken into consideration during project selection.

Another popular method is that of scoring models of project profitability [6, 7]. This approach involves criteria of both groups using scales of scores; continuous scales being substituted by discrete ones for the first-group criteria. The major problem arising with application of scoring models is that of generating a model relating estimates with respect to various criteria and the general value of a project. How to establish proportion between criteria? How to compare the importance of the criteria of both groups? As it was demonstrated by critical surveys of existing methods of R and D planning [8, 9, 10], these questions have not yet found a satisfactory answer. Larichev [8] points to this fact as to one of the causes of the poor mathematical practicability of the R and D planning models. In our opinion, the very approach to the development of scoring models without representatives of the planning body is unsatisfactory.

Planning practice today requires methods for selection and evaluation of R and D projects that allow for the criteria of both (objective and subjective) groups, integration of the criterion estimates being done on the basis of the planning body (decision makers) policy. The present paper suggests a method [11] for the solution of "portfolio optimization" problems of some classes characterized below.

2. Peculiarities of the problem under consideration

The particular version of the portfolio optimization problem discussed below has the following features. The R and D projects presented for consideration to the planning organ are oriented to a particular product, many of them being closer to developments rather then to researches. Each R and D project is independent of other projects. There is an approved technique for calculation of economical effectiveness of R and D projects that involves evaluation of both consumption of various resources, and incomes which would be obtained with the attainment of the aims.

Along with economical effectiveness, the planning body uses for decision making noneconomical criteria as well. Such criteria, for example, may involve social effects of R and D, extent of application of the expected R and D results, correspondence of the results to the world level, etc.

Projects presented to the planning body may be in one or more versions. Such versions may include: a) in-house implementation of the R and D projects; b) technological cooperation; c) purchase and use of some results (patents, technology).

Resource consumption for each R and D project is small as compared with the resources at the planning organ's disposal. Moneys in national,

COMECON or hard currencies disposed by the planning organ are regarded as resources. It is assumed that with money one can buy any other resources, delivery time may be allowed for in the total duration of a given project version. The total allowable implementation time is limited also by decisions of the planning organ.

Resources of the planning organ are insufficient for implementation of all R and D projects. The management of the planning organ, i.e. decision maker(s) (DM), poses the problem of selecting from many projects presented to consideration the most preferable ones which are to be included into the financial plan. Those projects are regarded as the most preferable which comply with the DM sceintific and technical policy and result in the greatest possible total economical effect. The State Institute-Factory Combines of the People's Republic of Bulgaria feature the above peculiarities of R and D planning. The method presented below was developed with this end in view.

3. The class of projects of special importance

The noneconomical criteria manifest themselves most prominently when DM includes R and D projects into the plan on the basis of these criteria only. In this case, such projects, naturally, have the best estimates for all or some of the second-group criteria.

The idea of the proposed approach is to separate primarily those R and D projects where the second-group criteria are stronger than those of the economical nature.

In many practical cases it is important to establish those combinations of noneconomical criterion estimates which dictate that the R and D project be included into the plan. Such estimate combinations enable classification of all projects into two groups: especially important projects (K_1) which are necessarily included into the plan (one of the possible implementation versions), and ordinary projects (K_2) whose implementation depends on the expected economical effectiveness.

By condition, R and D projects having the highest estimates for all the qualitative criteria belong to K_1, and those having the lowest estimates belong to K_2.

The boundary between the two classes may be established by finding out preferences of the planning organ DM because they reflect DM's experience and scientific and technical policy.

Thus, the first stage of the solution is to separate the class of the most important projects. Researchers working in collaboration with employees of the planning organ work out, on the basis of the former decisions and the

environment, a list of noneconomical criteria involving both criteria which were actually taken into consideration in the previous decisions, and those which should be taken into consideration from the viewpoint of the DM scientific and technical policy. For each criterion a discrete scale with several qualitative verbal estimates is generated.

Formulation of the verbal estimates for criterion scales reflects the scientific and technical policy of DM, their desire to have certain qualities in R and D projects. In addition, the qualitative estimates play another very important role. The point is that the diversity, complexity and heterogeneity of R and D projects make their estimation very difficult for DM. This leads inevitably to inviting experts making a qualified and unbiased study and estimation of the R and D projects, and suggesting possible implementation versions. In this connection, the qualitative scales are a communication language enabling DM to tell the experts which degrees of quality should be discerned in the projects under consideration.

Now one meets with a problem from the decision making theory. It should be noted that the general problems of multi-criteria decision making were treated in monographs [20, 15, 21].

Let there be n criteria having a corresponding discrete scale with a_i estimates ($i = 1, 2, \ldots, n$) each. The total number of all possible combinations of criterion estimates is

$$A = \prod_{i=1}^{n} a_i. \tag{3.1}$$

It is required to break down these combinations into two classes of ordinary and especially important projects on the basis of information given by DM.

Solution of this problem may result in a decision rule representable as a set of two-criteria estimate combination tables, the number of the tables being equal to the number of combinations of other $(n - 2)$ criteria. Figure 1 illustrates such tables for three criteria.

Now pass to an algorithm for solving the above problem.

Any procedure for getting information from DM should be based upon psychological and psychometric data about possibility of getting transitive and repeatable information. Moreover, these procedures should envision the possibility of verifying the information given by DM.

Presently, one may regard the hypothesis sufficiently confirmed that the major cause of the preference DM intransitivity in the selection problem is due to numerous attributes of the objects being compared and to their multi-criteria estimates [14, 15]. Using this hypothesis, one can formulate two auxiliary hypotheses:

1. With only minor violations of transitivity, DM may compare objects differing in estimates with respect to two criteria, all other estimates being fixed.

This hypothesis was checked in [18, 19], their results indicate to its plausibility.

2. For a small (2—3) number of estimates along the generalized criterion scale (in this case, K_1 and K_2), DM may make stable (good repeatibility for repeated inquiries) and consistent (rare violations of transitivity)

	B_1	B_2	B_3
$Б_1$	K_1	K_1	K_1
$Б_2$	K_1	K_1	K_2
$Б_3$	K_1	K_1	K_2

A_1

	B_1	B_2	B_3
$Б_1$	K_1	K_1	K_1
$Б_2$	K_1	K_2	K_2
$Б_3$	K_2	K_2	K_2

A_2

	B_1	B_2	B_3
$Б_1$	K_1	K_2	K_2
$Б_2$	K_2	K_2	K_2
$Б_3$	K_2	K_2	K_2

A_3

Fig. 1

Criterion A

Importance of expected results for the R and D contractor's standing in the environment

A_1 — very great, A_2 — great, A_3 — insignificant

Criterion Б

Work done by the contractor in anticipation.

$Б_1$ — the contractor has already done a major part of the work required by the given R and D. The rest does not involve principal difficulties.

$Б_2$ — R and D involve a number of difficult problems. There are ideas for their solution, and areas of studies have been defined.

$Б_3$ — R and D require study of new and insufficiently explored problems. There are no ideas for their solution.

Criterion B

Correspondence of the anticipated result to the world standard

B_1 — anticipated R and D results are superior to the world standard;

B_2 — anticipated results of R and D correspond to the world standard;

B_3 — anticipated results of R and D are inferior to the world standard.

estimates with respect to the generalized criterion scale for any combination of estimates of two criteria under the assumption that with respect to other $(n-2)$ criteria estimates are the best.

This hypothesis was studied in [13] and also may be regarded plausible.

According to these hypotheses, information required for the solution of the problem above may be obtained from DM by means of Table T for all combinations of two criteria under the supposition that other $(n-2)$ criteria have the best estimates (see, for example, Fig. 2). For all pairs of criteria, DM puts K_1 or K_2 into each entry of the tables. Redundant information may be used for DM consistency checking.

	A_1	A_2	A_3
B_1	K_1	K_1	K_1
B_2	K_1	K_2	K_2
B_3	K_2	K_2	K_2

Fig. 2

Criterion A

Anticipated level of R and D

The anticipated results are
A_1 – superior to the world standard;
A_2 – corresponding to the today's world standard;
A_3 – inferior to the world standard.

Criterion B

Social effect of R and D implementation

B_1 – the R and D project has a direct and very great bearing upon the improvement of the life standards of wide sections of the population;

B_2 – the R and D project contributes directly to the improvement of life standards of wide sections of the population;

B_3 – implementation of the R and D project has no direct bearing upon the living conditions of wide sections of the population.

DM should meticulously analyse contradictions in the two-riteria estimate combination tables in order to correct its scientific and technical policy.

It may be readily seen that in the particular case where all estimates have binary estimate scales and where reduction of estimates for any criterion pair results in the class K_2, the information obtained from DM through tables T is sufficient for generation of the decision rule. Sometimes one can reduce the situation to this particular case by uniting similar rows and columns in tables T (for instance, one can pass to binary estimates of criteria$A$ and B in Fig. 2).

In the general case, the information in tables T is insufficient for decision rule generation. These tables are coordinate planes in the n-dimensional criterion space, and the boundary between the classes K_1 and K_2 may be restored through its projections on these planes.

Note, first of all, that some entires of the decision rule table are filled according to the evident rule: reduction of estimates with respect to some criterion does not result in increase of the class of quality.

To fill the balance of the decision rule tables, DM compares pairwise objects differing in estimates with respect to two criteria, quality class (K_1 or K_2) of one of the objects being known. Denote by $\rightarrow$ qualitative superiority of one object over another. Let O_1 and O_2 be compared objects. It is easy to see that the following relations hold:

$$\begin{aligned} &\text{if } O_1 \in K_2 \quad \text{and} \quad O_1 \rightarrow O_2, \quad O_2 \in K_2 \\ &\text{if } O_1 \in K_1 \quad \text{and} \quad O_2 \rightarrow O_1, \quad O_2 \in K_1. \end{aligned} \tag{3.2}$$

The entries left undetermined after application of (3.2) may be classified by putting the following question to DM: "Does reduction of quality with respect to one/two criteria move the object from K_1 to K_2?"

Note, that in practice the problem of the boundary between the classes K_1 and K_2 is essentially simplified by the fact that the number of estimate combinations classified as K_1 is small.

If several DM groups participate in the development of the planning body policy, filling and discussion of such tables is a convenient means of establishing a joint scientific and technical policy.

4. Formalized model

Separation of two classes of R and D projects allows to pass to another stage in the solution of our problem. A special difficulty involved in generation of a plan is due to the fact that no project implementation version can be selected independently of the whole package of proposals. Each project version requires at each stage certain resources of each kind. Consequently, under limited resources, inclusion of some versions of a project into the plan cannot but say upon the possibility of including other projects into the plan if they require resources of the same kind. Therefore, versions of each project should be considered jointly when forming a plan.

Let q_{ij} be the jth version of implementation of the ith project belonging to class K_1, and let p_{ij} be the jth implementation version of the ith project belonging to class K_2.

Impose the following constraint on the duration of projects:

$$T_{ij} \leq T_0 \tag{4.1}$$

where T_{ij} is implementation term for the jth version of the ith project, and T_0 is a constraint imposed by DM.

The problem under consideration may be formulated as follows:

$$\underset{x_{ij},y_{ij}}{\text{Max}}\Big(\sum_{i,j} D_i x_{ij} + \sum_{i,j} D_i y_{ij}\Big) = \sum_i D_i + \underset{x_{i,j},y_{i,j}}{\text{Max}} \sum_{i,j} y_{i,j} \tag{4.2}$$

$$\sum_j x_{i,j} = 1 \tag{4.3}$$

$$\sum_j y_{ij} \leq 1 \tag{4.4}$$

$$\sum_{i,j} x_{ij} R_{ij}^{(\nu t_k)} + \sum_{i,j} y_{ij} R_{ij}^{(\nu t_k)} \leq R^{(\nu t_k)} \tag{4.5}$$

where D_i is the economical effectiveness of the ith project ("income" of the planning body);

$R_{ij}^{(\nu t_k)}$ is consumption of the ν-th resource at t_kth planning period necessary to implement the jth version of the ith project;

$R^{(\nu t_k)}$ are constraints on the value of ν-th resource at tth planning stage.

$$x_{ij} = \begin{cases} 1 \text{ if the project } q_{ij} \text{ is included into the plan} \\ 0 \text{ if the project } q_{ij} \text{ is not included into the plan} \end{cases}$$

$$y_{ij} = \begin{cases} 1 \text{ if the project } p_{ij} \text{ is included into the plan} \\ 0 \text{ if the project } p_{ij} \text{ is not included into the plan.} \end{cases}$$

Note that Condition (4.3) necessarily requires that some versions of the R and D projects of class K_1 be included into the plan.

5. Solution method

The above problem (4.2)–(4.5) may be regarded as a version of the "multi-dimensional knapsack problem" [17] well known from the literature. Hence, the methods of integer programming [17] may be applied to this problem. It should be noted that significant computational difficulties may arise with application of precise methods for finding extremum owing to the great number of variables and constraints in (4.2)–(4.5).

Allowance for the particular features of the problem under consideration enables development of a special heuristic algorithm giving, as it has been verified in many cases, quite satisfactory results.

Introduce some notations.

Let W_{ij} be the economical effect index for implementation of the jth version of the ith project:

$$W_{ij} = \frac{D_i}{R_{ij}} \tag{5.1}$$

where R_{ij} is the total consumption of resources required for the jth version of the ith project.

R_{ij} may be computed as follows: denote by α_v deficiency index of the v-th resource. These indices may be found through the following conditions

$$\alpha_v = \frac{R^v_{\text{req}}}{R^v_{\text{pl}}}; \qquad \sum_v \alpha_v = 1 \tag{5.2}$$

where R^v_{pl} is the value of the v-th resource at the disposal of the planning organ;

R^v_{req} is the value of the v-th resource required for implementation of all versions of all R and D projects.

R_{ij} is computed as

$$R_{ij} = \sum_v \alpha_v \sum_{t_k} R_{ij}^{(vt_k)} \frac{1}{(1+k)^{t_k}} \tag{5.3}$$

where $R_{ij}^{(vt_k)}$ is the consumption of the v-th resource at the t_kth financing stage required for implementation of the jth version of the ith project; and

k is the discount coefficient assigned by the planning organ.

The heuristic algorithm below [11] is based upon the idea of successive filling of "many knapsacks". Owing to the peculiarities of the problem under consideration, all the obligatory items, i.e. class K_1 project versions, are first packed. Further, projects from the class K_2 are added in accordance to their index. When constraints are reached, some project versions are excluded in a prescribed order. The algorithm does not require selection of solution versions and is, therefore, very simple. Now we shall give its detailed description.

The algorithm [11] for the solution of problems (4.2)–(4.5) has some stages.

1. All the implementation versions of all class K_1 projects are ordered with respect to W_{ij}. The same is done separately with the class K_2 projects.

2. A hypothetical plan

$$P_h = \{q_{ij}\} \tag{5.4}$$

is formed involving all the versions of the class K_1 projects ordered with respect to W_{ij}.

Plan P_h is verified in terms of constraints (4.1) and (4.5). If temporal constraints are violated for all the versions of some q_{ij}, DM is informed in order that they assign another T_0 or revise the boundary between projects of classes K_1 and K_2.

If at some planning stage resource constraints (4.5) are violated, those projects q_{ij} are excluded from P_h that require this resource at this stage. (Condition (4.3) should be necessarily observed requiring that, at least, one version of each q_i is left in P_h.) If such an exclusion does not satisfy constraints (4.5), DM is informed in order that they move the boundary between classes K_1 and K_2.

3. Let plan P_h be formed without violations of the constraints (4.1) and (4.5). All the versions which do not satisfy (3) are excluded from the sequence of versions p_{ij} ordered with respect to W_{ij}. Versions p_{ij} are introduced into the plan successively beginning from those with greatest W_{ij} down until some of the constraints from (4.5) are violated.

4. Let introduction of the next version p_{ij} violate the financial constraint (4.5) for some kind v of resources at the t_kth planning stage. Then, beginning from the versions of q_{ij} with lowest W_{ij}, those are successively excluded from P_h that require resource v at the given planning stage. In doing so, one checks each time whether condition (4.3) is observed, i.e. whether some versions of each q_{ij} are left. After each exclusion of version q_{ij}, the next-by-value version p_{ij} is introduced into the plan.

5. Let only one implementation of each q_i be left and let addition of the next version p_{ij} violate the financial constraint (4.5) for some resource at the planning stage t_k. Those p_{ij} are separated from the previously included ones that were included into the plan in several versions. Those versions p_{ij} are successively eliminated that require at planning stage t_k greater resource v as compared with other versions of the same project (elimination begins with versions with lower W_{ij}). In doing so, the following condition is observed: some versions of projects p_i previously included into the plan should be left there. As versions p_{ij} are excluded, new ones are added into the plan from a sequence ordered with respect to W_{ij}.

6. Let all p_{ij} and q_{ij} be included into the plan in one of their implementation versions, and let addition of the next p_{ij} violate the constraint (4.5) for the v-th resource at the planning stage t_k. Then, from a list of ordered with respect to W_{ij} versions of p_{ij} which were not previously included into the plan, those versions are excluded which require resource v at the planning stage t_k. The balance of versions is included into the plan according to the procedure above.

7. The plan is regarded as formed if addition of new versions p_{ij} becomes impossible because resources are required at those planning stages where financial constraints have been reached.

For one planning interval and one kind of resource, this algorithm resembles the well-known method for approximate solution of the classical knapsack problem [12]. This method involves successive packing of objects according to the importance-to-weight ratio. The algorithm presented in [12] gives the optimal solution for objects of the same weight. Under a similar condition and for one planning interval, the above algorithm gives the optimal solution as well.

Indeed, let there be one planning period, and let expenses on all the projects coincide ($R_{ij} = R^*$). Let the above algorithm result in plan P_c, constraint (4.5) on resource v preventing introduction of new p_{ij}'s into the plan. Then plan P_c is optimal with respect to the criterion (4.2), provided the constraints (4.5) and (4.3) are observed.

By condition, all the projects require the same amount of resource. According to the above algorithm, resource v is required only by those q_{ij} which cannot be otherwise realized. Hence, changes in the plan P_c may be done by substituting p_{ij} which are not included in the plan for those which are included.

According to the above algorithm, this results in decrease of the functional (4.2).

The practical problems for which this algorithm is intended feature comparable project expenses that are small as compared with the total resource. In this connection, one can assume that the heuristic algorithm would enable fairly satisfactory solutions.

Results of practical applications of the heuristic algorithm were analysed as follows:

Each version of each project was represented by a point in the space of the following dimensions: execution time, expenses, various resources required for implementation, and income. Further, a dimensionality reduction method of major components [16] was used to represent projects on a plane where they were grouped by closeness of their characteristics. A similar representation was obtained for projects included into the plan through the above algorithm. The visual representation enabled DM to analyse both proposals and projects. Detailed study of the characteristics of the groups of excluded projects has demonstrated that the scientific and technical policy of DM was not violated. Analysis, carried out for several practical applications, allows to estimate the proposed algorithm as fairly satisfactory.

6. Plan sensitivity analysis

Constraints on various resources are prescribed to DM by the superior authority. The values of constraints, however, are not rigidly fixed. It is reasonable to expect that the superior organ would agree with a comparably small increase of some resources if it may result in significant increase of the criterion (4.2) value.

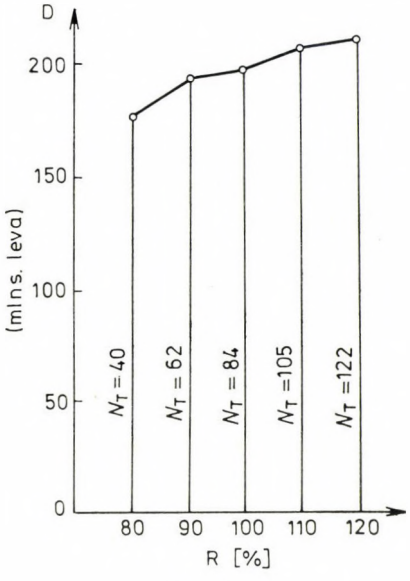


Fig. 3

The final verification of the chosen plan is done by the analysis of sensitivity of the extremum reached with respect to the criterion (4.2) to variations of constraints on each resource. Analysis is done by DM by means of the graphical relations between the total expenditures on a given kind of resource and the total income.

The method above was used for planning applied R and D for the State Institute-Factory Combines of the People's Republic of Bulgaria. Figure 3 illustrates analysis of R and D plan sensitivity to constraint variations (consideration was given to the constraints on national, COMECON and hard currencies, and also on the project implementation time). Variation of the total economic effect D with the total expenditures ($T_0 = 3$ years, N_T is the number of themes in the plan) is shown in Fig. 3.

It follows from Fig. 3 that 10% increase of expenditures results in appreciable increase of income, while reduction of grants by the same 10% does not practically effect the income.

Thus, expenditures should be either reduced or increased by 10% with respect to this characteristic value.

As Fig. 3 demonstrates, further increase of expenditures would be less effective.

7. Conclusions

The governing bodies of R and D organizations come across the following problems:

1. determination of the total R and D budget;
2. selection of particular R and D projects within the budget framework.

In our opinion, these two problems are inseparable, and possible expenditures on R and D projects should be largely dictated by their characteristics. The method presented here enables selection of R and D projects (one or another implementation version) and comparison of economical effects attained with various expenditures on R and D.

*

Acknowledgement

The authors are indebted to N. N. Moiseev, Corresponding Member of the Academy of Sciences of the USSR, for his valuable advice and criticism.

References

1. *Markowitz, H. M.* Portfolio selection: Efficient Diversification of Investment, New York, Wiley, 1959.
2. *Francis, J. C., Archer, S. H.* "Portfolio Analysis", Englewood Cliffs, New Jersey, Prentice-Hall, 1971.
3. *Gear, A. E., Lockett, A. G., Pearson, A. W.* "Analysis of some portfolio selection models for R and D", IEEE Transactions on Engng Manag., vol. EM—18, pp. 66—77, 1971.
4. *Dean, B. V., Sengupta, S. S.* "Research budgeting and project evaluation", IRE Trans. on Engng. Manag., vol. EM—9, No. 4, 1962.
5. *Atkinson, A. C., Bobis, A. H.* "A mathematical basis for the selection of R and D", IEEE Trans. on Engng. Manag., vol. EM—16, No. 1, 1969.
6. *Moore, J. R., Baker, N. R.* "Computational analysis of scoring models for R and D project selection", Manag. Science, vol. 16, No. 4, 1969.
7. *Seiler, R. E.* "Improving the effectiveness of R and D", New York, Mc Graw Hill, 1965.
8. *Clark, T. E.* "Decision making in technologically based organizations: a literature survey of present practice", IEEE Trans. on Engng Manag., vol. EM—21, No. 1, 1974.
9. *Larichev, O. I.* Organizational Systems and Strategic Planning Methods for Scientific Research and Development (Review), Automation and Remote Control, vol. 35, No. 5, Part 2.
10. *Baker, N. R., Pound, W. H.*, "R and D project selection: where we stand", IEEE Trans. on E. M., vol. EM—11, No. 4, 1964.

11. *Larichev O. I.*, et al., A Method For Generation of *R* and *D* Topical Plan, in "Long-Term Planning of Scientific *R* and *D*", Nauka, Moscow, 1974 (in Russian).
12. *Dantzig, G. B.*, Linear Programming and Extensions, Princeton Univ. Press, 1963.
12. *Larichev, O. I., Glotov, V. A.*, Organizational System and a Method for Evaluation of the Quality of Scientific and Technical Papers, in "Long-Term Planning of Scientific *R* and *D*", Nauka, Moscow, 1974 (in Russian).
14. *Mirkin, B. G.*, The Group Selection Problem, Nauka, Moscow, 1974 (in Russian).
15. *Marschak, J.*, "Decision making: economic aspects", International Encyclopedia of the Social Sciences, vol. 4, Collier and MacMillan, 1968.
16. *Terekhina, A. Yu.*, Methods of Multidimensional Data Scaling and Visuolization (Survey), Automation and Remote Control, vol. 34, No. 7, Part 1.
17. *Korbut, A. A., Finkel'stein, Yu. Yu.*, Discrete Programming, Nauka, Moscow, 1969 (in Russian).
18. *Larichev O. I., et al.*, A Method for Generation of a Classification of Applied *R* and *D*, in "Long-Term Planning of Scientific *R* and *D*", Nauka, Moscow, 1974 (in Russian).
19. *Adams, E., Fogot, R.*, "A model of riskless choice", Behavioral Science, vol. 4, No. 1, 1959.
20. *Germeier, Yu. B.*, Introduction into the Operations Research, Nauka, Moscow, 1974 (in Russian).
21. *Fishburn, P.*, "Utility Theory for Decision Making", New York, Wiley, 1970.

Многокритериальный подход к проблеме планирования прикладных научных исследований и разработок при качественных критериях

С. В. ЕМЕЛЬЯНОВ, О. И. ЛАРИЧЕВ

(Москва)

В статье рассматривается проблема выбора из совокупности проектов научных исследований и разработок (НИР) наилучших проектов при ограничениях, наложенных на различные виды ресурсов. Каждый из проектов может быть выполнен в одном из нескольких вариантов, отличающихся количеством и видом потребляемых ресурсов и длительностью выполнения проекта. Предполагается, что проекты независимы. Качество каждого из проектов оценивается по совокупности критериев, часть которых имеет экономический характер (расходы различных ресурсов, ожидаемый экономический эффект), а другая часть отражает научно-техническую политику планового органа (например, соответствие директивным указаниям, социальный эффект от выполнения проекта и т. д.). Критерии второй группы имеют качественный характер. Основная проблема состоит в разработке метода планирования, позволяющего учесть как качественные, так и количественные критерии. В качестве критерия оптимальности плана НИР рассматривается максимум суммарного экономического эффекта от выполнения проектов при соблюдении научно-технической политики планового органа.

Основные идеи предлагаемого подхода заключаются в следующем.

На основе предпочтений планового органа определяются такие сочетания оценок критериев неэкономического характера, при которых проект относится к классу особо важных проектов, которые обязательно должны быть включены в план. Разработан эвристический алгоритм для ЭВМ формирования субоптимального плана НИР, обеспечивающий выбор совокупности проектов НИР в определенных вариантах их выполнения, причем в план входят все особо важные проекты. Дан пример анализа чувствительности выбранного плана к изменению ряда параметров.

С. В. Емельянов, О. И. Ларичев
Институт проблем управления
СССР, Москва В-485,
Профсоюзная ул., 81

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 5 (5, 6), pp. 401—419 (1976)

К ТЕОРИИ РЕЛЕЙНО-ЛИНЕЙНЫХ ЗАКОНОВ УПРАВЛЕНИЯ

К. А. ПУПКОВ, А. И. МОРОЗ

(Москва)

(Поступила в редакцию 18 июня 1975 г.)

К настоящему времени достаточно полно разработана теория стабилизации линейных динамических систем [1]. Однако реальные системы содержат нелинейности, вследствие чего выводы линейной теории не всегда оказываются состоятельными. Типичной нелинейностью такого рода является насыщение в исполнительном органе системы управления. Формально в этом случае при решении задачи синтеза закона управления необходимо учитывать ограничения на величину управляющего воздействия. Это приводит к существенному усложнению задачи. Общим методом её решения является динамическое программирование. Однако имеющийся к настоящему моменту опыт указывает на исключительную трудность решения уравнения Беллмана в этом случае. «Проклятие размерности» остаётся непреодолимым.

В данной статье разрабатывается теория специального вида релейно-линейных законов (релионов) управления, которые в фиксированной окрестности начала координат пространства состояний являются линейными, а вне этой окрестности — релейными, определяемыми кусочно-линейными поверхностями переключения, которые аппроксимируют поверхности переключения оптимального по времени закона управления, так что при синтезе этого закона можно учесть ограничения на управление. Введение линейной зоны приводит к улучшению качества переходных процессов в системе, поскольку при этом исключаются в окрестности начала координат автоколебательные или скользящие режимы, которые всегда возникают в чисто релейных системах. Релион при больших отклонениях состояния от начала координат обладает достоинствами релейного оптимального по времени закона управления, одеспечивая максимальным по интенсивности управлением перевод системы в окрестность начала координат за время, близкое к минимально возможному, и при этом замкнутая система имеет максимально возможную при наличии ограничений на регулирующее воздействие область устойчивости. В линейной зоне он обеспечивает гладкое затухание переходных процессов. Включая только релейные и линейные операции, релион легко реализуется с помощью стандартных средств автоматики.

1. Постановка задачи

Рассмотрим процесс, описываемый уравнением

$$\frac{dx}{dt} = \mathrm{A}(t)\,x + \sum_{i=1}^{m} b_i(t)\,u_i + f(t)\,, \tag{1.1}$$

где x — n-мерный вектор, $\mathrm{A}(t)$ — матрица, $b_i(t)$ — n-мерный вектор, $f(t)$ — известная суммируемая вектор-функция, u_i — управление. Область допустимых значений управления U задаётся условием

$$|u_i| \leq 1, \qquad i = 1, 2, \ldots, m. \tag{1.2}$$

Допустимыми будем считать управления $u(\tau) = \mathrm{col}(u_1(\tau), u_2(\tau), \ldots, u_m(\tau))$, которые измеримы и удовлетворяют ограничению

$$\operatorname{vrai}\max_{[t_0, t_1]} \max_{1 \le i \le m} \{|u_i(\tau)|\} \le 1. \tag{1.3}$$

Пусть $\Phi(t, \tau)$ — фундаментальная матрица системы (1.1) и $\mathrm{B}(t)$ — матрица, i-м столбцом которой является $b_i(t)$. Система (1.1) называется неособенной на интервале $[t_0, t_1]$, если для любого ненулевого вектора l вектор-функция $l^*\Phi(t_1, \tau)\,\mathrm{B}(\tau)$ обращается на $[t_0, t_1]$ в нуль только на множестве нулевой меры Лебега. Система (1.1) называется неособенной на $[t_0, t_1)$ по управлению u_α, если система $\dfrac{dx}{dt} = \mathrm{A}(t)x + b_\alpha(t)u_\alpha$ является неособенной. Система (1.1) называется t_0-дифференциально управляемой, если она является неособенной на (t_0, ∞). Очевидно, такая система является вполне управляемой [2] на любом интервале $[\tau_0, \tau_1] \subset [t_0, \infty)$.

Обозначим через $G(t_0, t_0 + r)$ множество точек $\hat{x}$, для каждой из которых существует на интервале $[t_0, t_0 + r]$ допустимое управление $u(t)$, такое, что решение $x(t)$ уравнения (1.1), соответствующее $u = u(t)$, удовлетворяет граничным условиям $x(t_0) = \hat{x}$ и $x(t_0 + r) = 0$. Областью t_0-управляемости системы (1.1) будем называть множество $G(t_0) = \bigcup_{r=0}^{\infty} G(t_0, t_0 + r)$.

Релион будет строиться в виде аппроксимации оптимального по времени закона управления, поэтому вначале рассмотрим следующую задачу. Определить для каждого $t \in [a_0, a_1)$ и каждого $x \in G(t)$ вектор-функцию $F^0(t, x)$, удовлетворяющую ограничению $|F_i^0(t, x)| \le 1$, $i = 1, 2, \ldots, m$, и такую, что для любых $t_0 \in (a_0, a_1)$ и $\hat{x} \in G(t_0)$ существует $T^0(\hat{x})$ и на интервале $[t_0, t_0 + T^0(\hat{x})]$ решение $\hat{x}(t)$ уравнения

$$\frac{dx}{dt} = \mathrm{A}(t)\,x + \sum_{i=1}^{m} b_i(t)\,F_i^0(t, x) + f(t), \tag{1.4}$$

которое удовлетворяет граничным условиям $\hat{x}(t_0) = \hat{x}$ и $\hat{x}(t_0 + T^0(\hat{x})) = 0$, причем для любого допустимого управления $u(t)$, переводящего на соответствующем ему интервале $[t_0, t_0 + T]$ точку $\hat{x}$ в 0, выполняется $T \ge T^0(\hat{x})$.

2. Синтез оптимального по времени закона управления

Можно показать, что если система (1.1) является t_0 дифференциально управляемой по каждой компоненте u_i управления u, то последнее из условия принципа максимума Понтрягина определяется однозначно. В этом случае можно таким же образом, как и в [3] (стр. 127), показать, что оптимальное по времени программное управление $u_0(t)$ является единственным. В рассматриваемой задаче для любой пары (t_0, x^0) начальных условий, где $x^0 \in G(t_0)$,

существует оптимальное по времени управление [2]. Если же, кроме того, выполняется условие, что значения функции $f(t)$ при почти всех t принадлежат внутренности множества $B(t)U$, то принцип максимума является достаточным условием оптимальности по времени.

В силу существования и единственности оптимальных траекторий $x^0(\tau)$, которые заканчиваются в начале координат, $F^0(t, x)$ определяется следующим образом. Фиксируем $\hat{x} \in G(t)$ и пусть $u^0(\tau)$ — оптимальное по времени управление, такое, что соответствующая ему оптимальная траектория $\hat{x}^0(\tau)$ удовлетворяет условиям $\hat{x}^0(t) = \hat{x}$ и $\hat{x}^0(T) = 0$ при некотором $T > 0$. Тогда зададим $F^0(t, \hat{x}) = u^0(t)$. На основании вышеизложенного такое задание в каждой паре (t, x) будет однозначным, если соответствующее оптимальное управление $u^0(\tau)$ в этот момент t определено однозначно. Это в принципе решает задачу синтеза закона управления. Однако такое решение является абстрактно математическим и пока совершенно не ясно, как оно может быть использовано для целей оптимального управления, когда требуется по заданной в момент t точке x указать соответствующее ей значение $F^0(t, x)$. Поэтому приступим к более детальному анализу закона $F^0(t, x)$.

В силу принципа максимума каждая компонента $u_i^0(\tau)$ оптимального программного управления при каждом τ будет принимать значение $u_i = 1$ либо $u_i = -1$, либо в этой точке будет происходить изменение знака управления. Поэтому, если в момент $\tau = t$ оптимальная траектория проходит через точку x, то соответствующее этой точке значение оптимального управления $F^0(t, x)$ будет равно либо одному из граничных значений, либо в точке x в момент t при движении по оптимальной траектории будет происходить изменение знака управления. Следовательно, для фиксированного t оптимальная синтезирующая функция $F^0(t, x)$ является кусочно-постоянной по x, причем относительно каждой компоненты u_i множество $G(t)$ разбивается на два непересекающихся подмножества, в одном из которых $G_i^+(t)$ в момент t выполняется $F_i^0(t, x) = 1$, и в другом $G_i^-(t)$ выполняется $F_i^0(t, x) = -1$. Исследуем множество тех точек, по которому $G_i^+(t)$ соприкасается с $G_i^-(t)$.

Рассмотрим сопряженную к (1.1) систему

$$\frac{d\psi}{dt} = -A^*(t)\psi. \tag{2.1}$$

Будем говорить, что x принадлежит поверхности переключения $\Omega_i(t)$ управления u_i, соответствующей моменту t, если можно указать моменты t_0 и t_1, где $t_0 \leq t < t_1$, и на $[t_0, t_1]$ нетривиальное решение $\psi(\tau)$ уравнения (2.1), такое, что соответствующее ему в силу принципа максимума управление $u^0(\tau)$ порождает на $[t_0, t_1]$ оптимальную траекторию $x^0(\tau)$ уравнения (1.1), удовлетворяющую условиям $x^0(t) = x$, $x^0(t_1) = 0$, причем при $\tau = t$ выполняется $\psi^*(t)b_i(t) = 0$.

Это определение выделяет такие точки в $G(t)$, в которых при движении вдоль оптимальной траектории i-я компонента u_i вектора управления u меняет знак, поскольку из принципа максимума следует, что $u_i^0(\tau) = \operatorname{sign} \psi^*(\tau) b_i(\tau)$, $\tau \in [t_0, t_1]$. Заметим, что в точках множества $\Omega_i(t)$ задание функции $F_i^0(t, x)$ в соответствии с вышесформулированным правилом не будет однозначным, поскольку u_i в момент t из принципа максимума не определяется однозначно. Ниже будет показано, что если рассматривать решение оптимальной замкнутой системы (1.4) в обобщенном смысле А. Ф. Филиппова [4], то задание $F_i^0(t, x)$ в точках $\Omega_i(t)$ может быть произвольным. Вместе с тем $F_i^0(t, x)$ можно определить в $\Omega_i(t)$ так, что решение уравнения (1.4) будет существовать в классическом смысле. Для этого достаточно $F_i^0(t, x)$ задать так, чтобы в точках переключения решение было продолжаемо по оптимальной траектории, которая получается из принципа максимума, т. е. задать $F_i^0(t, x) = u_i^0(t + 0)$, где $x = x^0(t; t_0, u^0(\cdot))$ и $x^0(\tau; t_0, u^0(\cdot))$ — оптимальная траектория, соответствующая управлению $u^0(.)$.

В [5] найдено, что поверхность переключения $\Omega_\alpha(t)$ для компоненты u_α описывается параметрическими уравнениями

$$x = -\int_t^{t+r} \Phi(t, \varphi) \sum_{i=1}^{m} b_i(\varphi) \operatorname{sign} w_\alpha^*(t)\, \Phi(t, \varphi)\, b_i(\varphi)\, d\varphi - \int_t^{t+r} \Phi(t, \varphi) f(\varphi)\, d\varphi, \quad (2.2)$$

где вектор-параметр $w_\alpha(t)$ принимает такие значения, что $w_\alpha^*(t) b_\alpha(t) = 0$ и $||w_\alpha(t)|| = 1$, и параметр r меняется в интервале $[0, \infty)$. Поверхность $\Omega_\alpha(t)$ обладает двумя важными свойствами [5]. Во-первых, она однозначна в направлении вектора $b_\alpha(t)$, т. е. для любой точки $x \in G(t)$ существует одна и только одна точка $\omega_\alpha \in \Omega_\alpha(t)$, такая, что выполняется разложение $x = \omega_\alpha + \lambda_\alpha b_\alpha(t)$. Другими словами, из этого уравнения ω_α и число λ_α определяются моментом t и точкой x однозначно. Тем самым определяется функция $\lambda_\alpha(t, x)$. Во-вторых, поверхность $\Omega_\alpha(t)$ гомеоморфна $(n-1)$-мерной плоскости. Из этих свойств следует, что она подобно гиперплоскости разделяет область t-управляемости системы на две изолированные части, и что Лебегова мера множества $\Omega_\alpha(t)$ равна нулю.

Остаётся установить знак закона $F_\alpha^0(t, x)$ по определённую сторону $\Omega_\alpha(t)$. Как указывалось, для произвольной точки $x \in G(t) \setminus \Omega_\alpha(t)$ выполняется разложение

$$x = \omega_\alpha^0(t, x) + \lambda_\alpha^0(t, x)\, b_\alpha(t), \quad (2.3)$$

причём $\lambda_\alpha^0(t, x) \neq 0$. В [7] показано, что тогда

$$F_\alpha^0(t, x) = -\operatorname{sign} \lambda_\alpha^0(t, x). \quad (2.4)$$

Полученную формулу можно рассматривать как абстрактный синтез оптимального по времени закона. Уравнения (2.2) и (2.3) можно использовать для нахождения по заданным t и x значения параметра $\lambda_\alpha^0(t, x)$.

3. Кусочно-линейная аппроксимация поверхностей переключения

Формула (2.2), хотя и полностью описывает поверхности $\Omega_i(t)$, $i = 1, 2, \ldots, m$, ещё не пригодна для использования в реальных процессах управления. Получить аналитическое выражение функций $\lambda_\alpha^0(t, x)$, $\alpha = 1, \ldots, m$, из (2.2). и (2.3) для сколько-нибудь общих систем не удаётся, поэтому единственный путь использования решения, определяемого поверхностями $\Omega_i(t)$, состоит в каком-либо способе их аппроксимации. В качестве класса аппроксимаций можно рассмотреть кусочно-линейные поверхности $\Omega_i^A(t)$, нульмерные грани которых лежат на $\Omega_i(t)$. Для построения такой поверхности достаточно вычислять по формуле (2.2) точки поверхности $\Omega_i(t)$. Достоинством такой аппроксимации является то, что алгоритм вычисления закона управления в текущей точке x будет при этом состоять только из вычисления линейных функций и релейных операций от значений, принимаемых этими функциями. Чтобы избежать несущественных осложнений в процессе аппроксимации, далее будем считать $f(t) \equiv 0$. Тогда $\Omega_\alpha(t)$ будут симметричны относительно начала координат.

Аппроксимацию поверхности $\Omega_\alpha(t)$ будем осуществлять итеративно, добавляя некоторую гиперплоскость к уже построенной кусочно-линейной поверхности. В качестве первого приближения $\Omega_\alpha^1(t)$ будем использовать гиперплоскость, которая натянута на $n-1$ точек поверхности $\Omega_\alpha(t)$ и проходит через начало координат. Для этого фиксируем число $r_1 > 0$, $n-1$ векторов $w_{\alpha 1}(t), \ldots, w_{\alpha n-1}(t)$ и воспользуемся формулой (2.2) для вычисления $n-1$ точек $\omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)$ на $\Omega_\alpha(t)$. Заметим, что r_1 имеет смысл минимального времени для перевода $\omega_{\alpha i}(t)$ в 0, и при рассмотрении конкретной задачи можно обосновать его выбор. При выборе $w_{\alpha i}(t)$, $i = 1, \ldots, n-1$ можно столкнуться со следующим обстоятельством. В результате вычислений на ЦВМ по формуле (2.2) для различных $w_{\alpha i}(t)$ может получаться одна и та же точка $\hat{\omega}_\alpha(t)$ (или ей симметричная $-\hat{\omega}_\alpha(t)$. Это объясняется следующим.

Согласно определению, $\hat{\omega}_\alpha(t)$ является точкой границы выпуклого множества $G(t, t+r)$, в которой гиперплоскость, ортогональная вектору $w_\alpha(t)$, будет опорной к $G(t, t+r)$, [3]. В случае многомерных систем $G(t, t+r)$ обычно оказывается сильно вытянутым в одном направлении и сплюснутым в других. Это обстоятельство является причиной неэффективности методов типа Нойштадта-Итона [3] при численном решении задачи оптимального по времени программного управления. Очевидно, что в такой ситуации произвольное задание векторов $w_\alpha(t)$, как правило, будет приводить к нахождению либо одной и той же точки $\hat{\omega}_\alpha(t)$, либо близких к ней точек в удалённой части множества $G(t, t+r)$ (при численном решении на ЦВМ они практически совпадают). Для получения существенно различных линейно независимых $\omega_{\alpha 1}(t)$,

. , ., $\omega_{\alpha n-1}(t)$ можно использовать следующие соображения. Поскольку $G(t, t+r)$ является выпуклым телом, то если вектор $w_{\alpha k}(t)$ на k-м шаге выбирать так, чтобы соответствующая ему опорная гиперплоскость оказалась параллельной $(k-1)$-мерной плоскости Π_{k-1}, построенной на точках $\omega_{\alpha 1}(t)$, . . ., $\omega_{\alpha k-1}(t)$, полученных на предыдущих этапах, то $\omega_{\alpha k}(t)$ будет максимально удалённой от Π_{k-1} в направлении $w_{\alpha k}(t)$ точкой поверхности $\Omega_\alpha(t) \cap G(t, t+r)$, т. е. точки $\omega_{\alpha 1}(t)$, . . ., $\omega_{\alpha k}(t)$ будут линейно независимы. Из сказанного следует, что $w_{\alpha k}(t)$ нужно выбирать так, чтобы выполнялось $w^*_{\alpha k}(t)\,\omega_{\alpha i}(t) = 0$, $i = 1$, . . ., $k-1$. Найденные таким способом $n-1$ точек $\omega_{\alpha i}(t)$ будут определять гиперплоскость $x^*\,[\omega_{\alpha 1}(t) \times \omega_{\alpha 2}(t) \times \ldots \times \omega_{\alpha n-1}(t)] = 0$, где $[\omega_{\alpha 1}(t) \times \ldots \times \omega_{\alpha n-1}(t)]$ — векторное произведение. Левую часть этого соотношения удобно записывать в виде $[x, \omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)]$ и рассматривать как определитель матрицы, столбцами которой являются $x, \omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)$. Тогда его запишем как

$$[x, \omega_{\alpha 1}(t), \omega_{\alpha 2}(t), \ldots, \omega_{\alpha n-1}(t)] = 0. \tag{3.1}$$

Учитывая критерий $F^0_\alpha(t, b_\alpha(t)) = -1$, получим первое приближение закона $F^0_\alpha(t, x)$

$$u_\alpha = F^1_\alpha(t, x) = \operatorname{sign} p^*_{\alpha 1}(t)\, x\,, \tag{3.2}$$

где $p^*_{\alpha 1}(t)\, x = -\,[x, \omega_{\alpha 1}(t), \omega_{\alpha 2}(t), \ldots, \omega_{\alpha n-1}(t)] \operatorname{sign} [b_\alpha(t), \omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)]$.

В силу симметрии относительно начала координат поверхности $\Omega_\alpha(t)$ точки $-\omega_{\alpha 1}(t)$, . . ., $-\omega_{\alpha n-1}(t)$ также будут принадлежать $\Omega_\alpha(t)$. Обозначим полученное множество $2(n-1)$ точек на $\Omega_\alpha(t)$ через M^1_α. Будем называть гиперплоскость $w^*x + d = 0$ отделяющей для некоторого множества M, если для любой точки $z \in M$ выполняется $w^*z + d \geq 0$.

Для построения следующей аппроксимирующей плоскости необходимо фиксировать набор из $n-1$ точек $\hat\omega_{\alpha 1}(t)$, . . ., $\hat\omega_{\alpha n-1}(t)$ множества M^1_α, которые не лежат в одной $(n-3)$-мерной плоскости, т. е. точки $\hat\omega_{\alpha i}(t) - \omega_{\alpha 1}(t)$, $i = 2$, . . ., $n-1$ будут линейно независимы, и таких что гиперплоскость

$$[x - \hat\omega_{\alpha 1}(t), \hat\omega_{\alpha 2}(t) - \hat\omega_{\alpha 1}(t), \ldots, \hat\omega_{\alpha n-1}(t) - \hat\omega_{\alpha 1}(t), b_\alpha(t)] \cdot$$
$$\cdot \operatorname{sign} [-\hat\omega_{\alpha 1}(t), \hat\omega_{\alpha 2}(t), \ldots, \hat\omega_{\alpha n-1}(t), b_\alpha(t)] = w^*_{\alpha 0}(t)\, x + d^0_{\alpha 0}(t) = 0 \tag{3.3}$$

была бы отделяющей для M^1_α. Последнее означает, что кусок многогранной поверхности, которая аппроксимирует $\Omega_\alpha(t)$ в окрестности начала координат и представляет собой выпуклую оболочку точек множества M^1_α, лежит по одну сторону гиперплоскости (3.3). Это позволяет строить следующую грань аппроксимирующей поверхности по другую сторону гиперплоскости (3.3) путём приклейки её к $(n-2)$-мерной грани, натянутой на точки $\hat\omega_{\alpha 1}(t)$, . . ., $\hat\omega_{\alpha n-1}(t)$. Для этого достаточно найти по формуле (2.2) одну точку $\omega_{\alpha n}(t)$,

такую что

$$w^*_{\alpha 0}(t)\,\omega_{\alpha n}(t) + d^0_{\alpha 0}(t) < 0\,, \tag{3.4}$$

и натянуть гиперплоскость на точки $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t), \omega_{\alpha n}(t)$. Для получения точки $\omega_{\alpha n}(t)$, такой что выполняется (3.4) можно фиксировать $r = r_2 \geq r_1$, взять $w_\alpha(t) = w_{\alpha 0}(t)$ и вычислить $\omega_{\alpha n}(t)$ по формуле (2.2). Точки $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t), \omega_{\alpha n}(t)$ в силу строгой выпуклости $G(t, t+r)$ не будут лежать в одной $(n-2)$-мерной плоскости, поэтому они определяют гиперплоскость

$$[x - \omega_{\alpha n}(t), \hat{\omega}_{\alpha 1}(t) - \omega_{\alpha n}(t), \ldots, \hat{\omega}_{\alpha n-1}(t) - \omega_{\alpha n}(t)] \cdot$$
$$\cdot \operatorname{sign}[-\omega_{\alpha n}(t), \hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)] = p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t) = 0\,. \tag{3.5}$$

Обозначим через $G^+_{\alpha 1}$ множество $\{x\colon F^1_{\alpha 1}(t, x) = 1\}$ и $G^-_{\alpha 1} = \{x\colon F^1_{\alpha 1}(t, x) = -1\}$. Предположим вначале, что полученная точка $\omega_{\alpha n}(t)$ принадлежит $G^+_{\alpha 1}$. Тогда грань, натянутая на точки $\omega_{\alpha n}(t), \hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)$, будет принадлежать $G^+_{\alpha 1}$, и закон $F^1_\alpha(t, x)$ следует уточнять в $G^+_{\alpha 1}$. Заметим, что в силу того, что точки множества M^1_α принадлежат гиперплоскости (3.1) и гиперплоскость (3.3) является для него отделяющей, любая гиперплоскость, проходящая через точки $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)$, в том числе и (3.5), будет отделяющей для M^1_α. Тогда (3.5) делит $G^+_{\alpha 1}$ на две части, в одной из которых, задаваемой условием $p^*_{\alpha 2}(t)x + d_{\alpha 2}(t) \geq 0$, закон $F^2_\alpha(t, x)$ определяется уже построенной аппроксимацией $F^1_\alpha(t, x)$, а в другой, определяемой условием $p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t) < 0$, в соответствии с (2.4) уточняется в виде $F^2_\alpha(t, x) = -1$.

Таким образом, добавление гиперплоскости (3.5) приводит в точках $x \in G^*_{\alpha 1}$ к закону

$$F^2_\alpha(t, x) = \operatorname{sign}\left(p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\right). \tag{3.6}$$

Заметим, что в силу симметрии $\Omega_\alpha(t)$ гиперплоскость $-\,p^*_{\alpha 2}(t)x + d_{\alpha 2}(t) = 0$ будет служить в качестве аппроксимации в $G^-_{\alpha 1}$, и тогда закон $F^1_\alpha(t, x)$ уточняется в виде

$$F^2_\alpha(t, x) = -\operatorname{sign}\left(-\,p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\right),\ x \in G^-_{\alpha 1}\,. \tag{3.7}$$

Соотношения (3.6) и (3.7) можно записать в виде одной формулы

$$F^2_\alpha(t, x) = \begin{cases} \operatorname{sign}\left(p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\right) & \text{при} \quad F^1_\alpha(t, x) = 1\,, \\ \operatorname{sign}\left(p^*_{\alpha 2}(t)\,x - d_{\alpha 2}(t)\right) & \text{при} \quad F^1_\alpha(t, x) = -1\,, \end{cases}$$

или в виде

$$F^2_\alpha(t, x) = \operatorname{sign}\left(p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\,F^1_\alpha(t, x)\right). \tag{3.8}$$

Таким образом, добавление плоскости (3.5) привело к закону (3.8).

Рассмотрим теперь случай, когда $\omega_{\alpha n}(t)$ принадлежит $G^-_{\alpha 1}$. В силу симметрии $\Omega_\alpha(t)$ точка $\hat{\omega}_{\alpha n} = -\,\omega_{\alpha n}$ также будет ей принадлежать, но уже будет выполняться $\hat{\omega}_{\alpha n} \in G^+_{\alpha 1}$. Однако заметим, что для точки $\hat{\omega}_{\alpha n}(t)$ не будет выпол-

няться (3.4), так как из $d^0_{\alpha 0}(t) > 0$ следует $w^*_{\alpha 0}(t)\hat{\omega}_{\alpha 0}(t) + d^0_{\alpha 0}(t) > 0$. Тогда, рассмотрев симметричную грань, натянутую на точки $-\hat{\omega}_{\alpha 1}(t), \ldots, -\hat{\omega}_{\alpha n-1}(t)$, окажемся в условиях, при которых была построена аппроксимация (3.8) . В данном случае будем иметь

$$p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t) = [x + \omega_{\alpha n}(t),\ -\hat{\omega}_{\alpha 1}(t) + \omega_{\alpha n}(t), \ldots,\ -\hat{\omega}_{n-1}(t) + \omega_{\alpha n}(t)] \cdot$$
$$\cdot\ \operatorname{sign}[\omega_{\alpha n}(t),\ -\hat{\omega}(t), \ldots,\ -\hat{\omega}_{\alpha n-1}(t)]\,. \qquad (3.9)$$

В дальнейшем уточнение закона управления при добавлении следующих аппроксимирующих гиперплоскостей осуществляется итеративно согласно процедуре, которая была описана выше. Пусть после k шагов получен закон $F^k_\alpha(t, x)$. При его построении было вычислено $2(n-1) + 2(k-1)$ точек поверхности $\Omega_\alpha(t)$, множество которых обозначим через M^k_α. Для построения $F^{k+1}_\alpha(t, x)$ фиксируем набор из $n-1$ точек $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)$, принадлежащих M^k_α, которые не лежат в одной $(n-3)$-мерной плоскости и таких, что плоскость

$$[x - \hat{\omega}_{\alpha 1}(t), \hat{\omega}_{\alpha 2}(t) - \hat{\omega}_{\alpha 1}(t), \ldots,\ \hat{\omega}_{\alpha n-1}(t) - \hat{\omega}_{\alpha 1}(t), b_\alpha(t)] \cdot$$
$$\cdot\ \operatorname{sign}[-\hat{\omega}_{\alpha 1}(t), \hat{\omega}_{\alpha 2}(t), \ldots, \hat{\omega}_{\alpha n-1}(t), b_\alpha(t)] = w^*_{\alpha k-1}(t)\,x + d^0_{\alpha k-1}(t) = 0 \qquad (3.10)$$

будет отделяющей для M^k_α. Это всегда можно сделать. Фиксируем число $r_{k+1} \geq r_k$ и вычислим по формуле (2.2) точку $\omega_{\alpha n}(t)$, соответствующую этому r_{k+1} и вектору $w_{\alpha k-1}(t)$. Если $\omega_{\alpha n}(t)$ принадлежит $G^+_{\alpha k}$, т. е. $F^k_\alpha(t, \omega_{\alpha n}) = 1$, то вычислим $(k+1)$-ю гиперплоскость $p^*_{\alpha k+1}(t)\,x + d_{\alpha k+1}(t) = 0$, которая аппроксимирует $\Omega_\alpha(t)$, по формуле (3.5). Если выполняется $F^{k+1}_\alpha(t, \omega_{\alpha n}) = -1$, то по формуле (3.9). Если на данном этапе гиперплоскость $p^*_{\alpha k+1}(t)\,x + d_{\alpha k+1}(t) = 0$ является отделяющей для M^k_α, то, повторив приведённые выше выкладки, иайдём, что $F^{k+1}_\alpha(t, x)$ определяется по формуле

$$F^{k+1}_\alpha(t, x) = \operatorname{sign}\left(p^*_{\alpha k+0}(t)\,x + d_{\alpha k+1}(t)\,F^k_\alpha(t, x)\right). \qquad (3.11)$$

Если при всех $i = 3, 4, \ldots, k+1$ гиперплоскости $p^*_{\alpha i}(t)\,x + d_{\alpha i}(t) = 0$ являются отделяющими для множеств M^{i-1}_α, то, используя рекуррентное соотношение (3.11), можно $F^{k+1}_\alpha(t, x)$ получить в виде

$$F^{k+1}_\alpha(t, x) = \operatorname{sign}\left(p^*_{\alpha k+1}(t)\,x + \right.$$
$$+ d_{\alpha k+1}(t)\operatorname{sign}\left(p^*_{\alpha k}(t)\,x + d_{\alpha k}(t)\operatorname{sign}\left(p^*_{\alpha k-1}(t)\,x + \ldots + d_{\alpha 3}(t)\operatorname{sign}\cdot\right.\right.$$
$$\left.\left.\cdot\left(p^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\operatorname{sign} p^*_{\alpha 1}(t)\,x\right)\right)\ldots\right)\Big). \qquad (3.12)$$

В прикладных задачах обычно оказывается достаточно формулы (3.12), так как при аппроксимации ограничиваются малым числом плоскостей и их можно выбрать так, чтобы они были отделяющими. В случае систем с одним управлением, когда аппроксимация проводится в области пространства состояний, где оптимальные процессы являются неколебательными [6], аппроксимирующие гиперплоскости всегда будут отделяющими.

4. Синтез релейно-линейного закона

В реальных процессах управления использование релейных законов часто оказывается нецелесообразным. Недостатки релейного закона начинают сказываться в окрестности начала координат. Но здесь его можно заменить линейным. Изложенный выше метод построения аппроксимаций оптимального по времени закона позволяет строить законы, которые при больших отклонениях состояния от нуля имеют релейную структуру, а при малых — линейную. При решении этой задачи необходимо, во-первых, построить критерий перехода с одной структуры на другую, и во-вторых, синтезировать линейный закон.

Обозначим через $G_{\alpha 0}(t)$ множество в X, где управление определяется гиперплоскостью (3.6). Из (3.11) следует, что если в некоторой точке x выполняется

$$\operatorname{sign}\left(p^*_{\alpha i}(t)\,x + d_{\alpha i}(t)\right) = \operatorname{sign}\left(-\,p^*_{\alpha i}(t) + d_{\alpha i}(t)\right) = 1, \quad i = 2, 3, \ldots, k+1, \quad (4.1)$$

то $$F^{k+1}_{\alpha}(t, x) = F^{k}_{\alpha}(t, x) = \ldots = F^{1}_{\alpha}(t, x) = \operatorname{sign} p^*_{\alpha 1}(t)\,x,$$

т. е. закон определяется только гиперплоскостью, аппроксимирующей $\Omega_\alpha(t)$ в окрестности начала координат. Множество тех x, которые удовлетворяют (4.1), определяется системой неравенств $p^*_{\alpha i}(t)\,x + d_{\alpha i}(t) \geq 0$, $-\,p^*_{\alpha i}(t)\,x + d_{\alpha i}(t) \geq 0$, $i = 2, \ldots, k+1$, или $-\,d_{\alpha i}(t) \leq p^*_{\alpha i}(t)x \leq d_{\alpha i}(t)$, $i = 2, \ldots, k+1$. Следовательно, $G_{\alpha 0}(t)$ — выпуклое многогранное множество. Условия (4.1) могут служить критерием перехода с линейной структуры на релейную и наоборот.

Итак, в $G_{\alpha 0}(t)$ закон $F^{k+1}_{\alpha}(t, x)$ определяется гиперплоскостью $p^*_{\alpha 1}(t)\,x = 0$ в качестве переключающей. Заметим, что в окрестности точки $x = 0$ вектор скорости $A(\tau)\,x + \sum_{i=1}^{m} b_i(\tau)\ \operatorname{sign}\ p^*_{i1}(\tau)x$ траектории $x(\tau)$ определяется членом $\sum_{i=1}^{m} b_i(\tau)\ \operatorname{sign}\ p^*_{i1}(\tau)x$, а в точках плоскости переключения $p^*_{i1}(t)x = 0$ скорость $dx(\tau)/d\tau$ будет определяться вектором, являющимся взвешенным средним векторов $-\,b_i(\tau)$ и $b_i(\tau)$, т. е. в точках плоскости $p^*_{i1}(t)x = 0$ она будет близкой нулю. Поэтому в качестве линейного закона целесообразно использовать закон $u_i = \operatorname{sat} a\ k_i p^*_{i1}(t)x$, где k_i — положительная константа и $\operatorname{sat} a = 1$ при $a \geq 1$, $\operatorname{sat} a = a$ при $|a| \leq 1$ и $\operatorname{sat} a = -1$ при $a \leq -1$, который внутренне связан с исходным релейным законом и представляет собой лишь ослабление интенсивности управляющих воздействий в окрестности начала координат. Число k_i определяет крутизну линейной характеристики и ширину линейной зоны в направлении вектора $p^*_{i1}(t)$. При $k_i = \infty$ получаем $\operatorname{sat} k_i p^*_{i1}(t)x = \operatorname{sign} p^*_{i1}(t)\,x$ — релейный закон.

Для получения формулы релиона введём следующие переменные

$$\xi_{\alpha i}(t,x)=\frac{1}{2}\left(1+\operatorname{sign}\left(p^*_{\alpha i}(t)\,x+d_{\alpha i}(t)\right)\right),$$
$$\eta_{\alpha i}(t,x)=\frac{1}{2}\left(1+\operatorname{sign}\left(-p^*_{\alpha i}(t)\,x+d_{\alpha i}(t)\right)\right), \tag{4.1a}$$

$i=2,3,\ldots,k+1$. Тогда x принадлежит $G_{\alpha 0}(t)$ в том и только том случае, если $\xi_{\alpha i}(t,x)=\eta_{\alpha i}(t,x)=1$ для всех $i=2,3,\ldots,k+1$, т. е. если выполняется $\prod_{i=2}^{k+1}\xi_{\alpha i}(t,x)\,\eta_{\alpha i}(t,x)=1$, и релион можно записать в виде

$$Rl^{k+1}_{\alpha}(t,x)=\operatorname{sat}\left(k_\alpha\,p^*_{\alpha 1}(t)\,x\prod_{i=2}^{k+1}\xi_{\alpha i}\,\eta_{\alpha i}+\left(1-\prod_{i=2}^{k+1}\xi_{\alpha i}\,\eta_{\alpha i}\right)F^{k+1}_{\alpha}(t,x)\right). \tag{4.2}$$

5. Синтез релейной аппроксимации в общем случае

Выше при построении закона $F^{i+1}_{\alpha}(t,x)$ предполагалось, что аппроксимирующая гиперплоскость $p^*_{\alpha i+1}(t)x+d_{\alpha i+1}(t)=0$ является отделяющей для M^i_α. Сейчас рассмотрим случай, когда это условие не выполняется.

Фиксируем число $r_{i+1}\geq r_i$ и $n-1$ точек $\omega_{\alpha 1}(t),\ldots,\omega_{\alpha n-1}(t)$ множества M^i_α, которые не лежат в одной $(n-3)$-мерной плоскости и таких, что гиперплоскость

$$[x-\omega_{\alpha 1}(t),\omega_{\alpha 2}(t)-\omega_{\alpha 1}(t),\ldots,\omega_{\alpha n-1}(t)-\omega_{\alpha 1}(t),b_\alpha(t)]\cdot$$
$$\cdot\operatorname{sign}\left[-\omega_{\alpha 1}(t),\omega_{\alpha 2}(t),\ldots,\omega_{\alpha n-1}(t),\,b_\alpha(t)\right]=w^*_{\alpha i}(t)\,x+d^0_{\alpha i}(t)=0 \tag{5.1}$$

будет отделяющей для M^i_α. В силу однозначности $\Omega_\alpha(t)$ в направлении вектора $b_\alpha(t)$ это всегда можно сделать. Пусть $\omega_{\alpha n}(t)\in\Omega_\alpha(t)$ найдена по формуле (2.2) при $r=r_{i+1}$ и $w_\alpha(t)=w_{\alpha i}(t)$. Если выполняется $F^i_\alpha(t,\omega_{\alpha n})=1$, то рассмотрим гиперплоскость

$$[x-\omega_{\alpha n}(t),\omega_{\alpha 1}(t)-\omega_{\alpha n}(t),\ldots,\omega_{\alpha n-1}(t)-\omega_{\alpha n}(t)]\cdot$$
$$\cdot\operatorname{sign}\left[-\omega_{\alpha n}(t),\omega_{\alpha 1}(t),\ldots,\omega_{\alpha n-1}(t)\right]=p^*_{\alpha i+1}(t)\,x+d_{\alpha i+1}(t)=0\,. \tag{5.2}$$

Если же выполняется $F^i_\alpha(t,\omega_{\alpha n}(t))=-1$, то рассмотрим гиперплоскость

$$[x+\omega_{\alpha n}(t),-\omega_{\alpha 1}(t)+\omega_{\alpha n}(t),\ldots,-\omega_{\alpha n-1}(t)+\omega_{\alpha n}(t)]\cdot$$
$$\cdot\operatorname{sign}\left[\omega_{\alpha n}(t),-\omega_{\alpha 1}(t),\ldots,-\omega_{\alpha n-1}(t)\right]=p^*_{\alpha i+1}(t)x+d_{\alpha i+1}(t)=0\,. \tag{5.3}$$

Пусть гиперплоскость (2) не является отделяющей для M_α^i. Разобьём пространство X на три множества: (1) множество, для точек которого выполняется

$$|w_{\alpha i}^*(t)\,x| \leq d_{\alpha i}^0(t)\,, \tag{5.4}$$

(2) множество тех x, для которых выполняется неравенство

$$w_{\alpha i}^*(t)\,x + d_{\alpha i}^0(t) < 0\,, \tag{5.5}$$

(3) множество, определяемое условием

$$w_{\alpha i}^*(t)\,x - d_{\alpha i}^0(t) > 0\,. \tag{5.6}$$

По построению гиперплоскость (5.1) является отделяющей для M_α^i, поэтому все точки M_α^i принадлежат (5.4), т. е. в качестве аппроксимации в (5.4) будет закон, определяемый кусочно-линейной поверхностью переключения, натянутой на точки множества M_α^i, т. е. закон $F_\alpha^i(t, x)$. Далее, точка $\omega_{\alpha n}(t)$ принадлежит множеству (5.5). Учитывая соотношение $F_\alpha^{i+1}(t, b_\alpha(t)) = -1$, получим, что в множестве (5.5) аппроксимация, определяемая гиперплоскостью (5.3), будет иметь вид $F_\alpha^{i+1}(t, x) = \operatorname{sign}\,(p_{\alpha i+1}^*(t)x + d_{\alpha i+1}(t))\varkappa$, где $\varkappa = -\operatorname{sign} p_{\alpha i+1}^*(t)\,b_\alpha(t)$. В силу симметрии в точках множества (5.6) в качестве аппроксимации будет закон управления $F_\alpha^{i+1}(t, x) = \operatorname{sign}\,(p_{\alpha i+1}^*(t)\,x - d_{\alpha i+1})\varkappa$.

Таким образом, определена $(i+1)$-я аппроксимация во всех точках пространства X, которую можно записать в виде формулы

$$F_\alpha^{i+1}(t,x) = \begin{cases} F_\alpha^i(t,x), & \text{если} \quad |w_{\alpha i}^*(t)\,x| \leq d_{\alpha i}^0(t)\,, \\ \operatorname{sign}\left(p_{\alpha i+1}^*(t)\,x + d_{\alpha i+1}(t)\right)\varkappa, & \text{если} \quad w_{\alpha i}^*(t)\,x < -d_{\alpha i}^0(t)\,, \\ \operatorname{sign}\left(p_{\alpha i+1}^*(t)\,x - d_{\alpha i+1}(t)\right)\varkappa, & \text{если} \quad w_{\alpha i}^*(t)\,x > d_{\alpha i}^0(t)\,. \end{cases} \tag{5.7}$$

Этим итеративный процесс построения кусочно-линейных аппроксимаций закона $F^0(t, x)$ для произвольных систем полностью определён.

6. Синтез релейно-линейного закона в общем случае

Пусть $G_{\alpha 0}(t)$ — множество в X, где закон $F_\alpha^k(t, x)$ определяется гиперплоскостью переключения $p_{\alpha 1}^*(t)x = 0$. Как и выше, модифицируем в $G_{\alpha 0}(t)$ релейный закон $u_\alpha = \operatorname{sign} p_{\alpha 1}^*(t)x$ законом $u_\alpha = \operatorname{sat} k_\alpha p_{\alpha 1}^*(t)x$. В общем случае условие принадлежности текущего состояния x области $G_{\alpha 0}(t)$ будет определяться не только гиперплоскостями аппроксимаций $p_{\alpha i}^*(t)x + d_{\alpha i}(t) = 0$, но, в том случае, когда они не являются отделяющими для множеств M_α^{i-1}, также и гиперплоскостями $w_{\alpha i}^*(t)\,x + d_{\alpha i}^0(t) = 0$. Пусть для некоторого i гиперплоскость $p_{\alpha i}^*(t)\,x + d_{\alpha i}(t) = 0$ не является отделяющей. Тогда, очевидно, для этого i

следует функции $\xi_{\alpha i}(t, x)$ и $\eta_{\alpha i}(t, x)$ определить как

$$\xi_{\alpha i}(t, x) = \frac{1}{2}\left(1 + \operatorname{sign}(w^*_{\alpha i}(t)\, x + d^0_{\alpha i}(t))\right)$$

и

$$\eta_{\alpha i}(t, x) = \frac{1}{2}\left(1 + \operatorname{sign}(-\, w^*_{\alpha i}(t)\, x + d^0_{\alpha i}(t))\right).$$

Тогда релейно-линейный закон опять получается в виде (4.2), где $F^i_\alpha(t, x)$ при каждом $i = 2, 3, \ldots, k+1$ определяется либо по формуле (3.12), либо по формуле (5.7).

7. Анализ релейной аппроксимации

При использовании приближенного закона управления естественно возникает вопрос о динамическом качестве получающейся в результате замкнутой системы. В частности, представляет интерес вопрос о близости субоптимальной и оптимальной траекторий. Последняя проблема в принципе была решена А. Ф. Филипповым в 1960 году [4] при разработке теории дифференциальных уравнений с разрывной правой частью.

Необходимость использования понятия обобщенного решения дифференциального уравнения обусловлена следующим. Как было указано выше, оптимальный закон $F^0(t, x)$ можно так доопределить на поверхностях $\Omega_1(t), \ldots, \Omega_m(t)$, что замкнутая система (1.4) будет для каждого начального условия обладать классическим абсолютно непрерывным решением $x(\tau)$; это решение будет оптимальным и единственным. Однако при использовании приближенных законов $F^{k+1}(t, x)$ с кусочно-линейными поверхностями переключения такое доопределение, вообще говоря, не будет возможным, т. е. замкнутая система не будет обладать решениями в классическом понимании, и здесь возникает необходимость расширения понятия решения на такие ситуации. Только после этого окажется возможным сравнение оптимального и субоптимального решений.

Обозначим через $B^n_\gamma(x, \delta)$ n-мерный шар в смысле нормы γ в X радиуса δ с центром в точке x. Лебегову меру множества θ обозначим через $\mu\theta$. Пусть $f(t, x)$ — функция, определённая и измеримая на некотором множестве $[t_0, t_1] \times Q$, значения которой принадлежат X. Тогда под $f(t_0, P)$ понимается множество, являющееся образом в X множества $(t \times P)$ при отображении f. Вектор-функция $x(t)$, $t \in (t_0, t_1)$ называется обобщенным решением уравнения

$$\frac{dx}{dt} = f(t, x), \tag{7.1}$$

если она абсолютно непрерывна и при почти всех t выполняется

$$\frac{dx(t)}{dt} \in \bigcap_{\delta>0} \bigcap_{\mu\theta=0} \overline{\operatorname{conv}} f(t, \mathrm{B}_\gamma^n(x(t), \delta) \setminus \theta). \tag{7.2}$$

Заметим, что решение в обобщенном смысле не зависит от значений функции $f(t, x)$ на множестве нулевой меры в X. Поэтому если рассматривать решения системы (1.4) в указанном смысле, то они не зависят от значений функции $F^0(t, x)$ в точках поверхностей $\Omega_1(t)$, . . ., $\Omega_m(t)$, т. е. задание $F^0(t, x)$ на поверхностях переключения может быть произвольным. Однако априори нельзя гарантировать, что решения Филиппова системы (1.4) будут определять оптимальные траектории, так как последние существенно зависят от значений закона управления на поверхностях переключения. Как показал Бруновски [8], системы, у которых обобщенные решения не являются оптимальными, существуют, и они не являются чем-то экстраординарным. Поэтому далее ограничимся рассмотрением таких систем вида (1.1), которые обладают свойством, состоящим в том, что обобщенные решения уравнения (1.4) совпадают с классическими оптимальными, когда $F^0(t, x)$ задаётся на $\Omega_1(t)$, . . ., $\Omega_m(t)$ таким образом, что последние существуют. В этом случае решение Филиппова уравнения (1.4), соответствующее произвольному начальному условию $(\hat{t}_0, x)$, где $x \in G(\hat{t}_0)$, является единственным.

Обозначим в соответствии с обозначениями работы [4] правую часть дифференциального уравнения (1.1) при $u = F^0(t, x)$ через $f(t, x)$ и при $u = F^k(t, x)$ через $f_k(t, x)$. Пусть $\varrho_c(x(\cdot))$ — норма в пространстве C непрерывных функций, т. е. $\varrho_c(x(\cdot)) = \max\limits_{t_0 \le t \le t_1} \gamma(x(t))$. Дифференциальные уравнения

$$\frac{dx}{dt} = f(t, x), \tag{7.3}$$

$$\frac{dx}{dt} = f_k(t, x), \tag{7.4}$$

очевидным образом удовлетворяют условиям теорем 3 и 11 Филиппова (здесь $g_k(t, x) \equiv 0$), поэтому вторую часть следствия 4 теоремы 11 можно сформулировать в виде следующего утверждения.

Теорема 1. Для заданной точки x^0 и заданного числа $\varepsilon > 0$ существует $d(x^0, \varepsilon) > 0$, такое, что если выполняется,

$$\Omega_\alpha(t) \subset \bigcup_{x_0 \in \Omega_\alpha^k(t)} \mathrm{B}_\gamma^n(x, d(x^0, \varepsilon)), \quad \alpha = 1, 2, \ldots, m, \tag{7.5}$$

т. е. $\Omega_\alpha(t)$ принадлежит d-окрестности поверхности $\Omega_\alpha^k(t)$, то выполняется $\varrho_c(x(\cdot\,; t_0, x^0) - x^\Delta(\cdot\,; t_0, x^0)) < \varepsilon$, где $x^\Delta(t; t_0, x^0)$ — решение уравнения (7.4), $x^\Delta(t_0; t_0, x^0) = x^0$.

Это утверждение можно несколько усилить. Пусть D — компактное множество в X. Для заданного $\varepsilon > 0$ по теореме 11 [4] для каждой точки x^0 существует $d(x^0, \varepsilon) > 0$, такое, что если $x^0 + \Delta x$ принадлежит $B_\gamma^n(x^0, d(x^0, \varepsilon))$ и выполнено (7.5), то $\varrho_c(x(\cdot\,; t_0, x^0) - x^\Delta(\cdot\,; t_0, x^0 + \Delta x)) < \frac{\varepsilon}{2}$. В силу непрерывной зависимости решения уравнения (7.3) от начального условия существует $\delta(x^0, \varepsilon) > 0$, такое, что $\varrho_c(x(\cdot\,; t_0, x^0 + \Delta x) - x(\cdot\,; t_0, x^0)) < \frac{\varepsilon}{2}$, как только $x^0 + \Delta x \in B_\gamma^n(x^0, \delta(x^0, \varepsilon))$. Если $\delta(x^0, \varepsilon) > d(x^0, \varepsilon)$, то при $x^0 + \Delta x \in B_\gamma^n(x^0, d(x^0, \varepsilon))$ и выполнении условия (7.5) получим $\varrho_c(x(\cdot\,; t_0, x^0 + \Delta x) - x^\Delta(\cdot\,; t_0, x^0 + \Delta x)) = \varrho_c(x(\cdot\,; t_0, x^0 + \Delta x) - x(\cdot\,; t_0, x^0) + x(\cdot\,; t_0, x^0) - x^\Delta(\cdot\,; t_0, x^0 + \Delta x)) \leq \varrho_c(x(\cdot\,; t_0, x^0 + \Delta x) - x(\cdot\,; t_0, x^0)) + \varrho_c(x(\cdot\,; t_0, x^0) - x(\cdot\,; t_0, x^0 + \Delta x)) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$.

Таким образом, множество D покрывается семейством $\{B_\gamma^n(x^0, d(x^0, \varepsilon)) \mid \forall x^0 \in D\}$, так что если $\hat{x} \in B_\gamma^n(x^0, d(x^0, \varepsilon))$, то $\varrho_c(x(\cdot\,; t_0, \hat{x}) - x^\Delta(\cdot\,; t_0, x)) < \varepsilon$. Поскольку D — компакт, из этого покрытия можно выделить конечное покрытие $\{B_\gamma^n(x_i^0, d(x_i^0, \varepsilon)) \mid i = 1, 2, \ldots, N\}$. Тогда существует $d(\varepsilon) = \min\limits_{1 \leq i \leq N} \{d(x_i^0, \varepsilon)\} > 0$ и если $\Omega_\alpha(t) \subset \bigcup\limits_{x \in \Omega_\alpha^k(t)} B_\gamma^n(x, d(\varepsilon))$, то $\varrho_c(x(\cdot\,; t_0, x^0) - x^\Delta(\cdot\,; t_0, x^0)) < \varepsilon$ для любого начального условия $x^0 \in D$. В итоге получаем следующий результат.

Теорема 2. Для любого $\varepsilon > 0$ существует $d(\varepsilon) > 0$, такое, что если выполняется условие (7.5) при $d(x^0, \varepsilon) = d(\varepsilon)$, то $\varrho_c(x(\cdot\,; t_0, x^0) - x^\Delta(\cdot\,; t_0, x^0)) < \varepsilon$ при любом $x^0 \in D$.

Заметим, что условие принадлежности $\Omega_\alpha(t)$ d-окрестности $\Omega_\alpha^k(t)$ в этой теореме можно заменить условием, что $\Omega_\alpha^k(t)$ принадлежит d-окрестности $\Omega_\alpha(t)$. Этим вопрос о качестве субоптимального закона в принципе решается. В каждом конкретном случае проверка условий этой теоремы потребует специального исследования. Вместе с тем имеется один частный случай систем, когда близость поверхностей $\Omega(t)$ и $\Omega^k(t)$ можно непосредственно оценить. Более того, в этом случае обобщенное решение заведомо является оптимальным.

8. Стационарные системы с одним управлением

Рассмотрим систему вида (1.1), где A — постоянная матрица, все собственные значения которой действительны, $f(t) \equiv 0$ и $m = 1$. Для этого случая в [7] был построен закон, который в $G(O, T)$, где T фиксировано, аппроксими-

рует закон $F^0(x)$ и имеет вид

$$F^k(x) = \operatorname{sign}(\varphi_k(x)\operatorname{sign}\chi_k(x) + \operatorname{sign}(\varphi_{k-1}(x)\operatorname{sign}\chi_{k-1}(x) + \ldots + \\ + \ldots + \operatorname{sign}(\varphi_n(x)\operatorname{sign}\chi_n(x) + \operatorname{sign}\varphi_{n-1}(x))\ldots)),$$

где $k \geq n$, $\varphi_j(x) = -\max\limits_{i\in N(j)}\{|p_{ij}^* x|\}$, $\chi_j(x) = \min\limits_{i\in N(j)}\{p_{ij}^* x + 1\}$,

$$N(j) = \{i \mid C_{j-1}^{n-1} + 1 \leq i \leq C_j^{n-1}\}, \quad j = n,\ n+1, \ldots, k;$$

C_j^n — число сочетаний из j по n;

$$p_{ij}^* x = \frac{[x, v_{q_1} v_{q_2}, \ldots, v_{q_{n-1}}]}{\left[\sum\limits_{\alpha=1}^{n-1}(-1)^{\alpha-1}\sum\limits_{i=q_{\alpha-1}+1}^{q_\alpha} v_i, v_{q1}, v_{p2}, \ldots, v_{q_{n-1}}\right]},$$

$v_i = g(t_i) - g(t_{i-1})$, $g(t) = \int\limits_0^t e^{-\tau A} b\,d\tau$, $i = 1, 2, \ldots, k$; t_i — заданные числа, такие, что $0 = t_0 < t_1 < t_2 < \ldots < t_{k-1} < t_k = T$, параметры $q_1, q_2, \ldots, q_{n-1}$ пробегают все целочисленные значения, причем $1 \leq q_1 < q_2 < \ldots < q_{n-2} < q_{n-1} = j$ и

$$\varphi_{n-1}(x) = -\frac{[x, v_1, v_2, \ldots, v_{n-1}]}{[v_1 v_2, v_3, \ldots, v_n]}.$$

Этот закон управления может быть также представлен в виде (3.12). Кусочно-линейная поверхность переключения Ω^A этого закона является объединением $(n-1)$-мерных параллелотопов, гомеоморфна оптимальной поверхности Ω и все нульмерные грани Ω^A принадлежат Ω. Такая структура Ω^A легко позволяет оценить её близость к Ω. Как показано в [7], Ω можно представить в виде объединения непересекающихся множеств, каждое из которых соответствует фиксированному набору $\{q_1, q_2, \ldots, q_{n-1}\}$ и параметрически описывается как

$$\omega = \sum_{\alpha=1}^{n-2}(-1)^{\alpha+1} g(\zeta_\alpha t_{q_\alpha - 1} + (1-\zeta_\alpha) t_{q_\alpha}) + \\ + (-1)^n g(\zeta_{n-1} t_{q_{n-1}-1} + (1-\zeta_{n-1}) t_{q_{n-1}}), \tag{8.1}$$

где параметры $\zeta_1, \ldots, \zeta_{n-1}$ изменяются в интервалах $0 \leq \zeta_i \leq 1$. Множество, соответствующее набору $\{q_1, q_2, \ldots, q_{n-1}\}$, аппроксимируется параллелотопом вида

$$\omega^A = \sum_{\alpha=1}^{n-2}(-1)^{\alpha+1}\{2g(t_{q_\alpha}) - 2\bar{\zeta}_\alpha v_{q_\alpha}\} + (-1)^n\{g(t_{q_{n-1}}) - \bar{\zeta}_{n-1} v_{q_{n-1}}\}, \tag{8.2}$$

где $0 \leq \bar{\zeta}_i \leq 1$, $i = 1, 2, \ldots, n-1$. Если показать, что для любой точки поверхности Ω можно указать такую точку поверхности Ω^A, что расстояние между ними будет меньше заданного d, то это и будет означать, что Ω при-

надлежит d-окрестности поверхности Ω^{Δ}. Но каждая точка поверхности Ω принадлежит одному из множеств (8.1). Поэтому рассмотрим произвольное множество из (8.1) и произвольную точку ω в нём, которая характеризуется некоторыми числами $q_1, q_2, \ldots, q_{n-1}$ и $\zeta_1, \zeta_2, \ldots, \zeta_{n-1}$. Рассмотрим точку $\omega^{\Delta} \in \Omega^{\Delta}$, которая определяется следующими значениями параметров: $\bar{\zeta}_1 = \zeta_1, \bar{\zeta}_2 = \zeta_2, \ldots, \bar{\zeta}_{n-1} = \zeta_{n-1}$ и оценим разность $\gamma(\omega - \omega^{\Delta})$. Имеем

$$\gamma(\omega - \omega^{\Delta}) = \gamma\left\{2\sum_{\alpha=1}^{n-2}(-1)^{\alpha+1}\{g(\zeta_\alpha t_{q_{\alpha-1}} + (1-\zeta_\alpha)t_{q_\alpha}) - g(t_{q_\alpha}) + \zeta_\alpha v_{q_\alpha}\} + \right.$$

$$\left. + (-1)^n\{g(\zeta_{n-1}t_{q_{n-1}-1} + (1-\zeta_{n-1})t_{q_{n-1}} - g(t_{q_{n-1}}) + \zeta_{n-1}v_{q_{n-1}}\}\right) \leq$$

$$\leq 2\sum_{\alpha=1}^{n-2}\gamma\left(-\int_{\zeta_\alpha t_{q_{\alpha-1}} + (1-\zeta_\alpha)t_{q_\alpha}}^{t_{q_\alpha}} e^{-\tau A}b\,d\tau + \zeta_\alpha\int_{t_{q_{\alpha-1}}}^{t_{q_\alpha}} e^{-\tau A}b\,d\tau\right) +$$

$$+\gamma\left(-\int_{\zeta_{n-1}t_{q_{n-1}-1}}^{t_{q_{n-1}}} e^{-\tau A}b\,d\tau + \zeta_{n-1}\int_{t_{q_{n-1}-1}}^{t_{q_{n-1}}} e^{-\tau A}b\,d\tau\right) =$$

$$= 2\sum_{\alpha=1}^{n-2}\gamma\left(\zeta_\alpha(t_{q_\alpha} - t_{q_\alpha-1})(e^{-\theta_{q_\alpha}A} - e^{-\hat{\theta}_{q_\alpha}A})b\right) +$$

$$+\gamma\left(\zeta_{n-1}(t_{q_{n-1}} - t_{q_{n-1}-1})(e^{-\theta_{q_{n-1}}A} - e^{-\hat{\theta}_{q_{n-1}}A})b\right),$$

где числа θ_{q_α} и $\hat{\theta}_{q_\alpha}$ удовлетворяют ограничениям $t_{q_\alpha-1} < \hat{\theta}_{q_\alpha} < t_{q_\alpha}$, $\zeta_\alpha t_{q_\alpha-1} + (1-\zeta_\alpha)t_{q_\alpha} < \theta_{q_\alpha} < t_{q_\alpha}$. Отсюда следует, что выполняется оценка $|\theta_{q_\alpha} - \hat{\theta}_{q_\alpha}| < t_{q_\alpha} - t_{q_\alpha-1}$. Учитывая, что $0 \leq \theta_{q_\alpha} \leq T$ при всех α, будем иметь неравенство

$$\gamma\left((e^{-\theta_{q_\alpha}A} - e^{-\hat{\theta}_{q_\alpha}A})b\right) \leq \gamma\left(e^{-\theta_{q_\alpha}A}(\hat{\theta}_{q_\alpha} - \theta_{q_\alpha})Ab\right) + o(\theta_{q_\alpha} - \hat{\theta}_{q_\alpha}) \leq$$

$$\leq \gamma(e^{-\theta_{q_\alpha}A}Ab)\,|\hat{\theta}_{q_\alpha} - \theta_{q_\alpha}| + o(\theta_{q_\alpha} - \hat{\theta}_{q_\alpha}) \leq K_1|\hat{\theta}_{q_\alpha} - \theta_{q_\alpha}| + o(\theta_{q_\alpha} - \hat{\theta}_{q_\alpha}),$$

где K_1 — константа, не зависящая от q_α. В результате получаем

$$\gamma(\omega - \omega^{\Delta}) \leq 2\sum_{\alpha=1}^{n-2}\zeta_\alpha(t_{q_\alpha} - t_{q_{\alpha-1}})K_1|\hat{\theta}_{q_\alpha} - \theta_{q_\alpha}| +$$

$$+\zeta_{n-1}K_2(t_{q_{n-1}} - t_{q_{n-1}-1})|\hat{\theta}_{q_{n-1}} - \theta_{q_{n-1}}| +$$

$$+o\Big(\max_{1\leq\alpha\leq n-1}|\theta_{q_\alpha} - \hat{\theta}_{q_\alpha}|\Big)\max_{1\leq\alpha\leq n-1}(t_{q_\alpha} - t_{q_\alpha-1}) < \hat{K}\sum_{\alpha=1}^{n-1}(t_{q_\alpha} - t_{q_\alpha-1})^2 +$$

$$+o\Big(\max_{1\leq\alpha\leq n-1}(t_{q_\alpha} - t_{q_\alpha-1})^2\Big) < K\max_{1\leq\alpha\leq n-1}(t_{q_\alpha} - t_{q_\alpha-1})^2 +$$

$$+o\Big(\max_{1\leq\alpha\leq n-1}(t_{q_\alpha} - t_{q_\alpha-1})^2\Big).$$

Таким образом, если требуется, чтобы выполнялось $\gamma(\omega - \omega^{\Delta}) < d$ для любой точки $\omega \in \Omega$, достаточно так выбрать k и $t_1, t_2, \ldots, t_k$, чтобы удовлетворялось неравенство $\max\limits_{1 \le i \le k} (t_i - t_{i-1})^2 < \dfrac{d}{K}$. По матрице A и числу T легко получить оценку K и по требуемой точности аппроксимации d задать числа $t_1, t_2, \ldots, t_k$.

Теорема 2 в рассматриваемом случае гарантирует, что для любого заданного $\varepsilon > 0$ можно подобрать такой набор параметров $t_1, t_2, \ldots, t_k$, где $k = k(\varepsilon)$, что для любого процесса $x^{\Delta}(\tau)$ системы при $u = F^k(x)$, начинающегося в $G(0, T)$, и оптимального процесса $x(\tau)$, $x^{\Delta}(0) = x(0)$, будет выполняться $\max\limits_{0 \le \tau \le T} \gamma(x(\tau) - x^{\Delta}(\tau)) < \varepsilon$. Из этого соотношения получаем, что если ε таково, что $B_{\gamma}^{n}(0, \varepsilon) \subseteq G(0, T)$, то для всех $\tau > 0$ выполняется $x^{\Delta}(\tau) \in B_{\gamma}^{n}(G(0, T), \varepsilon)$, т. е. аппроксимирующий закон будет обеспечивать устойчивость систем (ассимптотическая устойчивость здесь не гарантируется). Пусть $\hat{T}$ — минимальное из тех T, для которых $B_{\gamma}^{n}(0, \varepsilon) \subseteq G(0, T)$. Тогда $x^{\Delta}(\tau) \in B_{\gamma}^{n}(G(0, \hat{T}), \varepsilon)$ при всех τ. Заметим, что для заданного динамического качества, т. е. требования, чтобы выполнялось $x^{\Delta}(\tau) \in B_{\gamma}^{n}(0, \delta)$ можно выбрать ε такое, что $B_{\gamma}^{n}(G(0, \hat{T}), \varepsilon) \subseteq B_{\gamma}^{n}(0, \delta)$.

Таким образом, метод кусочно-линейных аппроксимаций оптимальной поверхности переключения, разработанный в [7], позволяет строить субоптимальные законы управления, которые обеспечивают заданную точность решений.

Замечание. Изложенные в этом разделе результаты без изменения переносятся на случай произвольных линейных систем с одним управлением, когда аппроксимация строится в таком $G(t_0, t_0 + T)$, что на $[t_0, t_0 + T]$ система является неколебательной, т. е. оптимальное по времени управление имеет не более $n - 1$ перемен знака. В этом случае таким же образом, как было показано в [7] для стационарных систем, можно показать, что функция $D(t; z_1, z_2, \ldots, z_{n-1}) = [\Phi(t, z_1)\, b(z_1), \Phi(t, z_2)\, b(z_2), \ldots, \Phi(t, z_n)\, b(z_n)]$ будет знакопостоянной в области $t \le z_1 \le z_2 \le \ldots \le z_n \le t + T$, откуда следует возможность построения параллелотопической аппроксимации оптимальной поверхности переключения. Векторы v_i в этом случае определяются как $v_i(t) = \int\limits_{t+t_{i-1}}^{t+t_i} \Phi(t, \zeta)\, b(\zeta)\, d\zeta$, $i = 1, 2, \ldots, k$, где $0 = t_0 < t_1 < t_2 < \ldots < t_k \le T$.

9. Анализ релиона

Теория Филиппова позволяет решать вопросы качества не только релейных аппроксимаций оптимального по времени закона, но также аппроксимаций в виде релиона. Следствие 2 из теоремы 11 [4], устанавливает, что если $f^{\Delta}(t, x)$ является средним значением (с любым неотрицательным весом) функции

$f(t, x)$ по области $B_\gamma^n(x, \Delta)$, то при $\Delta \to 0$ решения уравнения $\frac{dx}{dt} = f^\Delta(t, x)$ равномерно сходятся к соответствующим решениям уравнения $\frac{dx}{dt} = f(t, x)$.

Линейная зона всегда содержит как точки x, в которых $F_i^0(t, x) = 1$, так и точки x, в которых $F_i^0(t, x) = -1$. Вместе с тем при аппроксимации релионом линейный закон с насыщением представляет собой в каждой точке x среднее с неотрицательным весом этих граничных значений. Поэтому, выбрав достаточно малой линейную зону, на основании вышеизложенного можно обеспечить требуемую близость траекторий к оптимальной по времени траектории. По аналогии с релейным случаем можно показать, что путём выбора Δ можно построить такую линейную зону, что линейная часть релиона будет обеспечивать устойчивость системы.

Литература

1. *Красовский Н. Н.* Теория оптимальных управляемых систем. Механика в СССР за 50 лет, т. 1, «Общая и прикладная механика», Наука, М., 1968.
2. *Красовский Н. Н.* Теория управления движением. Наука, М., 1968.
3. *Болтянский В. Г.* Математические методы оптимального управления. Наука, М., 1969.
4. *Филиппов А. Ф.* Дифференциальные уравнения с разрывной правой частью. Математический сборник, **51,** 1, 1960.
5. *Мороз А. И.* К проблеме синтеза оптимального по времени управления. Автоматика и телемеханика, **1**, 1970.
6. *Тонков Е. Л.* Неосцилляция и число переключений в линейной нестационарной системе, оптимальной по быстродействию. Дифференциальные уравнения, **12,** 1973.
7. *Мороз А. И.* Элементы теории оптимальных систем. Издательство Московского института электронного машиностроения, М., 1974.
8. *Brunovský, P.*, The closed-loop time-optimal control. I. Optimality. "SIAM J. Control", **12**, *4*, 1974.

Contribution to the theory of relay-linear control laws

K. A. PUPKOV, A. I. MOROZ

(Moscow)

The problem of synthesizing of control law in the form of system state function is described. System (1.1) is considered where x is an n-dimensional vector of the state space X, $A(t)$ is an $(n \times n)$ matrix, $b_i(t)$ is a vector, $i = 1, 2, \ldots, m$, $f(t)$ is a known vector-function, u_i is a control, $i = 1, 2, \ldots, m$. Controls $u_1(\tau)$, $u_2(\tau)$, $\ldots$, $u_m(\tau)$ are considered to be admissible if they are measurable and if at all i they satisfy the restriction (1.2).

The set $G(t_0) \subset X$, each point of which for the finite time can be transmitted by a admissible control to the origin, if the process being at the moment t_0, is called the domain of t_0-controllability.

The time optimal control law for system (1.1) is a function $F^0(t, x)$ which satisfied the following requirement:

(1.1) $|F_i^0(t, x)| \leq 1$, $i = 1, 2, \ldots, m$, is valid for all (t, x),

(1.2) a moment $T^0(x^0)$ for any (t_0, x^0), $x^0 \in G(t_0)$, exists, and in the interval $[t_0, t_0 + T^0(x^0)]$ the solution $\hat{x}(t)$ of Eq. (1.4) exists and satisfies the boundary conditions $\hat{x}(t_0) = x^0$ and $\hat{x}(T^0(x^0)) = 0$, (1.3) for any admissible control $u(t)$, which transmits x^0 to 0 in corresponding interval $[t_0, T]$ the condition $T \geq T^0(x^0)$ is met.

The problem of syntesizing a relay-linear control which is linear in a fixed neighbourhood of the origin of space X and which is relay out of this neighbourhood is considered, this synthesizing being determined by a piecewise linear switching surface, which approximates switching surface of time optimal control law.

It is shown that for each $\alpha = 1, 2, \ldots, m$ the switching surface $\Omega_\alpha(t)$ of the law $F^0_\alpha(t, x)$ is described by parametric equation (2.2) where $\omega_\alpha \in \Omega_\alpha(t) \subset G(t)$, the parameter r changes in the interval $[0, \infty)$, an n-vector $w_\alpha(t)$ runs all directions in the hyperplane $b^*(t)\, x = 0$, i.e. $b^*(t)\, w_\alpha(t) = 0$ for all $w_\alpha(t)$. The problem of approximating the surfaces $\Omega_\alpha(t)$ by piecewise surfaces $\Omega^\Delta(t)$ and synthesizing the algorithm of control calculations defined by surfaces $\Omega^\Delta(t)$ are considered. Constructing $\Omega^\Delta(t)$ is carried out iteratively by using the points $\omega_\alpha(t)$ obtained by formulae (1.3). Let $M_\alpha(i)$ be a set of points of $\Omega_\alpha(t)$ which were obtained at the i-th step of constructing $\Omega^\Delta_\alpha(t)$. which were obtained at the i-th step of constructing $\Omega^\Delta(t)$. For determining $(i+1)$-th hyperplane used for constructing $\Omega^\Delta_\alpha(t) i + 1)$ from $\Omega^\Delta(t,i)$ $n-1$ points of this set is fixed and the new point $\omega_{\alpha n}(t) \in \Omega_\alpha(t)$ is calculated.

The hyperplane $w^* x + d = 0$ is called separating for some set M if $w^* z + d \geq 0$ is satisfied for any $z \in M$.

It is found that if the hyperplane (3.5) where square brackets denote a vector product, is separating for $M_\alpha(i)$ the law $F^{i+1}_\alpha(t, x)$ is determined by $F_\alpha(t, x)$ and the hyperplane (4) by formulae (3.11). It the above mentioned condition is satisfied for all $i = 1, 2, \ldots, k+1$ we obtain from (3.11)

$$u_\alpha = F^{k+1}_\alpha(t, x) = \operatorname{sign}\left(p^*_{\alpha k+1}(t)\, x + d_{\alpha k+1}(t) \operatorname{sign}\left(p^*_{\alpha k}(t)\, x + d_{\alpha k}(t) \operatorname{sign}\left(p^*_{\alpha k-1}(t)\, x + \right.\right.\right.$$

$$\left.\left.\left. + \ldots + \operatorname{sign} p^*_{\alpha 1}(t)\, x) \ldots\right)\right)\right).$$

While synthesizing relay-linear laws the control $F^{k+1}_\alpha(t, x)$ was taken as a base. This control in the domain where switching is determined by the hyperplane $p^*_1(t)x = 0$ of surface $\Omega^{k+1}_\alpha(t)$ that is in the domain defined by the condition $\prod_{i=2}^{k+1} \xi_{\alpha i}(t, x)\, \eta_{\alpha i}(t, x) = 1$, where (4.1a) is modified by the control $u_\alpha = \operatorname{sat} k_\alpha p^*_1(t)x$ that results in (4.2). In corresponding to A. F. Filippov's theorem relay-linear law (4.2) will provide desired closure approximative trajectory of system (1.1) to optimal one in metric continuous functions space as sufficient closure approximation of optimal law is hold.

К. А. Пупков, А. И. Мороз
Институт электронного машиностроения, каф. кибернетики
СССР, Москва Ж-28, Б. Вузовский пер, д. 3/12

Problems of Control and Information Theory, Vol. 5 (5, 6), pp. 421—436 (1976)

РАСПОЗНАВАНИЕ СИТУАЦИЙ В ЗАДАЧЕ О ПРЕСЛЕДОВАНИИ

Д. А. ПОСПЕЛОВ, Н. Н. ЧЕРЕЗОВА, Д. И. ШАПИРО

(Москва)

(Поступила в редакцию 7 апреля 1975 г.)

В работе рассматривается задача о преследовании. Процесс преследования описан в виде последовательности ситуаций. Предлагается способ распознавания данных ситуаций на основе специального набора логических функций и решающих таблиц.

В данной работе предлагается алгоритм распознавания ситуаций, возникающих в процессе преследования двух объектов одним. Объекты передвигаются по плоскости, обладают скоростью, направлением, кроме того, преследующий объект обладает некоторым запасом ресурса.

В ходе работы алгоритма вырабатывается решение о достижимости или недостижимости каждого из преследуемых объектов в анализируемой ситуации.

Для простоты изложения будем называть преследующий объект собакой, а преследуемые объекты — зайцами, хотя вместо них могут выступать любые другие объекты, обладающие перечисленными выше параметрами

1. Постановка задачи

Задача о преследовании рассматривается в следующей постановке: по плоскости передвигается собака и два зайца. Собака должна поймать одного или двух зайцев, если это возможно. В каждый фиксированный момент времени собака должна принять решение, исходя из анализа сложившейся ситуации.

Возможны следующие варианты решений: оба зайца недостижимы; достижим первый заяц (второй недостижим) и наоборот; достижимы оба зайца. Если достижимы оба зайца, собака должна решить, за каким зайцем бежать.

В каждый момент принятия решения анализируется следующая ситуация:

$$S_c^n(\varrho_c^n,\ v_c^n,\ \nu_c^n,\ m_c^n,\ v_{з_1}^n,\ \nu_{з_1}^n,\ m_{з_1}^n,\ v_{з_1}^n,\ v_{з_2}^n,\ m_{з_1}^n), \tag{1.1}$$

где:

n — номер ситуации (или номер момента принятия решения);
S^n — ситуация с номером n;
ϱ_c^n — ресурс собаки в момент n;
(с течением времени убывает, возрастает прямо пропорционально времени)
v_c^n — скорость собаки в момент n;
ν_c^n — направление движения собаки в момент n;
m_c^n — место нахождения собаки в момент n; место нахождения собаки на плоскости характеризуется двумя координатами (x_c^n, y_c^n);
$v_{з_1}^n$ — скорость первого зайца в момент n;
$\nu_{з_1}^n$ — направление движения первого зайца в момент n;
$m_{з_1}^n$ — место нахождения первого зайца в момент n; место нахождения первого зайца на плоскости характеризуется двумя координатами $(x_{з_1}, y_{з_1})$;
$v_{з_2}^n$ — скорость второго зайца в момент n;
$\nu_{з_2}^n$ — направление движения второго зайца в момент n;
$m_{з_2}^n$ — место нахождения второго зайца в момент n; место нахождения второго зайца на плоскости характеризуется двумя координатами $(x_{з_2}^n, y_{з_2}^n)$.

В случае представления ситуации в виде набора характеристических параметров (1.2) процедура выбора решения осуществляется следующим образом.

Конкретная ситуация S разбивается на две ситуации S^{n_1} и S^{n_2}, каждая из которых содержит набор параметров, описывающих положение собаки относительно каждого зайца

$$S^{n_1}(\varrho_c^{n_1}, v_c^{n_1}, \nu_c^{n_1}, m_c^{n_1}, v_{з_1}^{n_1}, \nu_{з_1}^{n_1}, m_{з_1}^{n_1}),$$
$$S^{n_2}(\varrho_c^{n_2}, v_c^{n_2}, \nu_c^{n_2}, m_c^{n_2}, v_{з_2}^{n_2}, \nu_{з_2}^{n_2}, m_{з_2}^{n_2}). \qquad (1.2)$$

Такое разбиение возможно, т. к. собака последовательно оценивает положение каждого зайца в зависимости от своих возможностей, и следовательно, последовательно принимает решение относительно каждого зайца. Предполагается, что зайцы передвигаются независимо относительно друг друга.

Основываясь на характеристических параметрах ситуаций, регулярным методом решается задача о достижимости первого зайца (соответственно второго зайца).

2. Регулярный метод решения задачи

В частности, может быть предложен геометрический способ решения данной задачи. Пусть в некоторый момент времени фиксируется ситуация S^n.

Соединим местоположение собаки (x_c^n, y_c^n) с местоположением первого зайца $(x_з^n, y_з^n)$ — отрезок AB и с местоположением $(x_{з_2}^n, y_{з_2}^n)$ — (отрезок AM). Проведем прямую $CD \perp AB$ и прямую $KL \perp AM$. Прямая CD делит плоскость, по которой передвигаются собака и первый заяц, на две полуплоскости A_1 и B_1; прямая KL делит плоскость, по которой передвигаются собака и второй заяц, на полуплоскости A_2 и B_2. Соответственно разделяется множество возможных решений о достижимости и недостижимости первого (второго) зайца в зависимости от конкретн ыхпараметров ситуаций S^{n_1} и S^{n_2}. Если направление движения первого (второго) зайца лежит в полуплоскости A_1 (A_2) и $v_{з_1} \geq v_c (v_{з_2} \geq v_c)$, то заяц недостижим. Если при сохранении первого условия $v_c > v_{з_1} (v_c > v_{з_2})$, то необходимо рассчитать точку возможной встречи собаки и зайца — R, а значит оценить расстояние AR (направление, по которому должна бежать собака). Заяц будет достижим, если у собаки хватит ресурса преодолеть расстояние AR. (Ресурс собаки убывает (возрастает) прямо пропорционально времени. Если $\varrho_c^n - \Delta\varrho_c t_c^n \geq 0$, то заяц достижим, где $\Delta\varrho$ — величина уменьшения ресурса в единицу времени. t_c — время движения собаки.

Если направление движения первого (второго) зайца лежит в полуплоскости B_1 (B_2), то первый (второй) заяц достижим, если $v_c \geq v_{з_1}$ $(v_c > v_{з_2})$ и если собаке хватит ресурса, чтобы преодолеть кратчайшее расстояние AS до возможной точки встречи собаки и зайца. Кроме того, первый (второй) заяц может быть достижим, если $v_c \leq v_{з_1}$ $(v_c \leq v_{з_2})$. В этом случае важно, чтобы собака успела преодолеть расстояние AS за время, равное или более короткое, чем время движения зайца, и чтобы у собаки хватило ресурса.

Если по ситуации $S^{n_1}(S^{n_2})$ принимается решение «заяц достижим», то параметры данной ситуации запоминаются и записываются в таблицу с названием «заяц достижим», представляющую собой матрицу, строки которой — наборы характеристических параметров каждой ситуации,

1-й столбец — столбец ресурсов собаки;
2-й столбец — столбец скоростей собаки;
3-й столбец — столбец направлений собаки;
4-й столбец — столбец местонахождения собаки;
5-й столбец — столбец скоростей зайца;
6-й столбец — столбец направлений зайца;
7-й столбец — столбец местонахождения зайца.

Аналогично формируется таблица «Заяц недостижим».

3. Анализ и использование таблиц «Заяц достижим», «Заяц недостижим»

Процесс формирования данных таблиц представляет собой процесс обучения ЦВМ, параметры данных таблиц используются для анализа конкретных ситуаций, их распознавания и принятия решений по каждой из них. Анализ и использование таблиц проводится следующим образом.

Пусть таблица «Заяц недостижим» содержит r строк — ситуаций, а таблица «Заяц достижим» — k строк — ситуаций. Таким образом, таблица «Заяц недостижим» представляет собой матрицу размером $(r\times 7)$, а таблица «Заяц достижим» — матрицу размером $(k\times 7)$.

Первую таблицу обозначим A_{r_7} — матрица недостижимости, вторую — A_{k_7} — матрица достижимости.

A_{r_7} преобразуется следующим образом с целью ее дальнейшего использования в качестве материала обучения в процессе распознавания конкретно возникшей ситуации:

Строки — ситуации в данной матрице располагаются в порядке убывания величины ресурса

$$\varrho_c^1 \leq \varrho_c^2 \leq \ldots \leq \varrho_c^r, \tag{3.1}$$

где 1, 2, . . ., r — номера строк — ситуаций, входящих в A_{r_7}. A_{k_7} преобразуется следующим образом: строки — ситуации, входящие в нее, располагаются в порядке возрастания первого характеристического параметра — ресурса собаки, т. е.

$$\varrho_c^1 \geq \varrho_c^2 \geq \ldots, \geq \varrho_c^k, \tag{3.2}$$

где 1, 2, . . ., k — номера ситуаций, входящих в A_{k_7}.

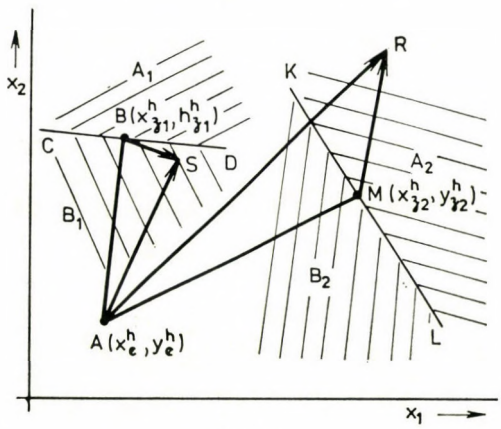


Рис. 1

Пусть в процессе преследования зафиксирована ситуация S^n, которую необходимо проанализировать и принять по ней некоторое решение.

Обработка вновь возникшей конкретной ситуации S^n начинается с анализа ее первого характеристического параметра — ресурса собаки и отыскания места данной ситуации в A_r:

$$\varrho_c^{e-1} \leq \varrho_c^n \leq \varrho_c^e, \tag{3.3}$$

где $1 \leq e-1, e \leq r$ — номера строк. Таким образом, выделяется усеченная матрица A_{r_r}, включающая строки с $l+1, \ldots, r$ номерами.

Строки-ситуации усеченной матрицы содержат параметрыресурса собаки большие, чем параметр ресурса собаки в конкретной ситуации S^n.

Таким образом, если считать, что остальные параметры, входящие в строки-ситуации зафиксированы, а изменяется только ресурс, то вывод о недостижимости зайца, верный для всех ситуаций усеченной матрицы, будет верен и для конкретной ситуации S^n. Кроме того, считается, что параметры, входящие в конкретную строку-ситуацию, обуславливают друг друга. Каждой величине ресурса соответствует определенная скорость собаки, скорость зайца и т. д. Связь эта выражается в том, что все эти параметры попали в одну ситуацию и объясняется разумным поведением собаки и зайца.

Таким образом, если ресурс собаки в конкретной ситуации S^n является некоторым обобщением величин ресурсов, входящих в строки ситуации усеченной матрицы (в том смысле, что вывод о недостижимости зайца, верный для строк-ситуаций усеченной матрицы, верен и для ситуации S^n, если изменяется только ресурс), то необходимо сделать некоторое обобщение скоростей собаки, зайца и других параметров усеченной матрицы, чтобы сравнить данные обобщенные параметры с соответствующими параметрами конкретной ситуации и установить, насколько они схожи между собой. Сходство ситуаций определяется одинаковыми выводами, которые можно сделать по ним в смысле достижимости и недостижимости зайца. Анализ и обобщение параметров усеченной матрицы проводится следующим образом:

$$v_c^e = \min(v_c^e, \ldots, v_c^r); \tag{3.4}$$

$$\bar{v}_c^e = \min(v_c^e, \ldots, v_c^r); \tag{3.5}$$

$$v_з^e = \min(v_з^e, \ldots, v_з^r); \tag{3.6}$$

$$\bar{v}_з^e = \max(v_з^e, \ldots, v_з^r); \tag{3.7}$$

$$f^e = \min(f^e, \ldots, f^r); \tag{3.8}$$

$$\bar{f}^e = \max(f^e, \ldots, f^r), \text{ где}$$

$$f^e = \sqrt{(x_c^e - x_з^e)^2 + (y_c^e - y_з^e)^2};$$

$$\nu^e = \{\nu_з^e, \ldots, \nu_з^r\}. \tag{3.9}$$

Сходство между конкретной ситуацией и обобщениями, проведенными на материале усеченной матрицы, устанавливается при помощи последовательного просчета логических функций $M_1, M_2, \ldots, M_8$. Данные функции описываются ниже. Аналогично используется A_{k_7}, в которой строки-ситуации расположены в порядке возрастания первого характеристического параметра — ресурса собаки.

При возникновении конкретной ситуации S^n, в которой $\varrho_c^{d-1} \geq \varrho_c^n \geq \varrho_c^d$, где

$$1 \leq d - 1, \quad d \leq k \tag{3.10}$$

номера строк матрицы достижимости. Выделяется усеченная матрица $A^1_{k_7}$, включающая строки с $d, d+1, \ldots, k$ номерами и проводится обобщение параметров данной матрицы по столбцам:

$$v_c^d = \min(v_c^d, \ldots, v_c^k); \tag{3.11}$$

$$\bar{v}_c^d = \max(v_c^d, \ldots, v_c^k); \tag{3.12}$$

$$v_з^d = \min(v_з^d, \ldots, v_з^k); \tag{3.13}$$

$$\bar{v}_з^d = \max(v_з^d, \ldots, v_з^k); \tag{3.14}$$

$$f^d = \min(f^d, \ldots, f^k); \tag{3.15}$$

$$\bar{f}^d = \max(f^a, \ldots, f^k), \text{ где}$$

$$f^d = \sqrt{(x_c^d - x_з^d)^2 + (y_c^d - y_з^d)^2};$$

$$\nu_з^d = \{\nu_з^d, \ldots, \nu_з^k\}. \tag{3.16}$$

Данный материал позволяет сформировать некоторую обобщенную ситуацию, которая сравнивается с конкретной ситуацией. Сходство между обобщенной и конкретной ситуациями устанавливается при помощи логических функций $F_1, \ldots, F_8$, которые описываются ниже.

4. Описание процесса распознавания ситуаций

Пусть возникла конкретная ситуация S^n. Она разбивается на две ситуации S^{n_1} и S^{n_2}. Каждая из ситуаций S^{n_1} и S^{n_2} выражает отношение собаки соответственно к первому и второму зайцу.

Сначала решается вопрос о недостижимости первого зайца. Ситуация S^{n_1} анализируется по таблице «Заяц недостижим»: находится соответствующее сечение в таблице «Заяц недостижим» и проводится обобщение параметров ситуаций, входящих в усеченную матрицу по формулам (3.4)—(3.9). Далее сравниваются параметры конкретной ситуации S^{n_1} с параметрами, получен-

ными в результате обобщения. Для этого последовательно вычисляются значения некоторых логических функций ($M_1, \ldots, M_8$), построенных на параметрах конкретной ситуации и материале обобщения с использованием некоторых логических операторов [1].

Если значение хотя бы одной из вычисляемых логических функций равно 1, то принимается решение о недостижимости первого зайца и начинается анализ ситуации S^{n_2} по таблице «Заяц недостижим». Если значения всех вычисляемых логических функций равны 0, то решается вопрос о достижимости первого зайца. В этом случае ситуация S^{n_1} анализируется по таблице «Заяц достижим».

На параметрах ситуации S^{n_1} и материале обобщения усеченной матрицы «Заяц достижим», полученном по формулам (3.11)—(3.16), строятся логические функции $F_1, \ldots, F_8$ с использованием логических операторов [1]. Если значение хотя бы одной из вычисляемых логических функций $F_1, \ldots, F_8$ равно 1, то принимается решение о достижимости первого зайца и начинается анализ ситуации S^{n_2} по таблице «Заяц недостижим».

Ситуация S^{n_2} анализируется по таблице «Заяц недостижим» по той же схеме, что и ситуация S^{n_1}. Для выработки решения о недостижимости применяются те же логические функции $M_1, \ldots, M_8$. Если данный анализ не дает положительных результатов (т. е. ни одна из функций $M_1, \ldots, M_8$ не принимает значение 1, то ситуация S^{n_2} анализируется по таблице «Заяц достижим», как и ситуация S^{n_1}. Для выработки решения о достижимости второго зайца используются функции $F_1, \ldots, F_8$.

Может представиться случай, что хотя бы одна из составляющих конкретной ситуации $S^n - S^{n_1}$ или S^{n_2} не распознается по таблицам «Заяц достижим» и «Заяц недостижим». Это значит, что такой вариант представляется в том случае, если сравнение параметров конкретной ситуации S^{h_1} и материала обобщения усеченной матрицы «Заяц недостижим» при помощи логических функций $M_1, \ldots, M_8$ не дает положительных результатов, т. е. все последовательно просчитанные логические функции $M_1, \ldots, M_8$ принимают значение равное 0.

Предположим, что последующая проверка ситуации S^{n_1} на достижимость также не дает положительных результатов, т. е. все логические функции $F_1, \ldots, F_8$ принимают значение равное 0. Поэтому решение по ситуации S^{n_1} принять невозможно, а значит и в целом о ситуации, например S^n, ничего сказать нельзя.

В этом случае ситуация S^n оценивается каким-либо регулярным методом, в результате чего принимается решение о достижимости и недостижимости первого и второго зайца.

Результаты решения оформляются в виде записи параметров ситуации в соответствующие таблицы «Заяц достижим» или «Заяц недостижим». Таким образом, происходит дообучение таблиц.

5. Алгоритм распознавания ситуаций

Итак, для анализа конкретных ситуаций по матрицам A_{r_7} и $A_{к_7}$ вводятся следующие логические операторы:

$$L_1(Y)=\begin{cases}1, & \text{если} \quad Y \le к \\ 0, & \text{если} \quad Y > к, \quad к \text{ — задано;}\end{cases} \tag{5.1}$$

$$L_2(Y)=\begin{cases}1, & \text{если} \quad Y \subset \Omega \\ 0, & \text{если} \quad Y \not\subset \Omega, \quad \Omega \text{ — задано;}\end{cases} \tag{5.2}$$

$$L_3(Y)=\begin{cases}1, & \text{если} \quad P \le Y \le к \\ 0, & \text{если} \quad Y > к \vee Y < P, \quad к, P \text{ — заданы;}\end{cases} \tag{5.3}$$

$$L_4(Y)=\begin{cases}1, & \text{если} \quad Y \ge Z; \\ 0, & \text{если} \quad Y < Z, \quad Z \text{ — задано.}\end{cases} \tag{5.4}$$

Рассмотрим функции $M_1, \ldots, M_8$, которые используются для сравнения параметров конкретной ситуации, например с обобщенными параметрами усеченной матрицы «Заяц недостижим».

$$\mathrm{M}_1 [L_1(\varrho_c^{n_1} \le \varrho_c^e) \wedge L_1(v_c^{n_1} \subseteq v^e) \wedge$$
$$\wedge L_2(\nu_{з_1}^{n_1} \subset \nu_з^e) \wedge L_4(v_{з_1}^{n_1} \ge \bar{v}_з^e) \wedge L_4(f^{n_1} \ge f_1^e). \tag{5.5}$$

Если $\mathrm{M}_1 = 1$, то заяц № 1 недостижим. Причина недостижимости такова: у собаки слишком мало ресурса, очень мала скорость, а у зайца слишком большая скорость, и он далеко.

В этом случае переходим к анализу S^{n_1} — второго зайца — по таблице недостижимости. Если $\mathrm{M}_1 = 0$, то считаем следующую функцию:

$$\mathrm{M}_2 [L_1(\varrho_c^{n_1} \le \varrho_c^e) \wedge L_з(v_c^e \le v_c^{n_1} \le \bar{v}_c^e) \wedge L_2(\nu_{з_1}^{n_1} \subset \nu\,) \wedge$$
$$\wedge L_4(v_{з_1}^{n_1} \ge \bar{v}^e) \wedge L_4(f_1^{n_1} \ge \bar{f}^e)]. \tag{5.6}$$

Если $\mathrm{M}_2 = 1$, то заяц № 1 недостижим, т. к. у собаки мало ресурса, а у зайца слишком большая скорость, и он далеко. Далее переходим к анализу S^{n_1} второго зайца по таблице недостижимости.

Если $\mathrm{M}_2 = 0$, то считаем следующую функцию для первого зайца:

$$\mathrm{M}_3 [L_1(\sigma_c^{n_1} \le \varrho_c^e) \wedge L_1(v^{n_1} \le v_c^e) \wedge L_2(\nu_{з_1}^{n_1} \subset \nu_з^e) \wedge L_3(v_з^e \le v_{з_1}^{n_1} \le \bar{v}_з^e) \wedge$$
$$\wedge L_4(f^{n_1} \ge \bar{f}^e)] \to \tag{5.7}$$

Если $M_3 = 1$, то заяц № 1 недостижим, т. к. у собаки слишком маленькая скорость, а заяц очень далеко.

Переходим к анализу S^{n_2} второго зайца по таблице недостижимости. Если $M_3 = 0$, то считаем следующую функцию для второго зайца:

$$M_4 [L_1(\varrho_{3_1}^{n_1} \leq \varrho_c^e) \wedge L_1(v_c^{n_1} \leq v_c^e) \wedge L_2(\nu_{3_1}^{n_1} \subset \nu_c^e) \wedge$$
$$\wedge L_4(v_{3_1}^{h_1} \geq \bar{v}_3^e) \wedge L_3(f^e \leq f^{n_1} \leq \bar{f}^e); \quad (5.8)$$

Если $M_4 = 1$, то заяц № 1 недостижим. У собаки мало ресурса, мала скорость, а у зайца очень большая скорость. Анализируем S^{n_2} второго зайца по таблице недостижимости.

Если $M_4 = 0$, то считаем M_5 для первого зайца:

$$M_5 [L_1(\varrho_c^{n_1} \leq \varrho_c^e) \wedge L_1(v_c^{n_1} \leq v_c^e) \wedge L_2(\nu_3^{n_1} \subset \nu_3^e) \wedge$$
$$\wedge L_3(v_3^e \leq v_3^{n_1} \leq \bar{v}_3^e) \wedge L_3(f^e \leq f^{n_1} \leq \bar{f}^e). \quad (5.9)$$

Если $M_5 = 1$, то заяц № 1 недостижим, т. к. у собаки мало ресурса и мала скорость. Далее анализируем S^{n_2} второго зайца по таблице недостижимости. Если $M_5 = 0$, то считаем M_6 для первого зайца:

$$M_6 [L_1(\varrho_c^{n_1} \leq \varrho_c^e) \wedge L_3(v_c^e \leq v_c^{n_1} \leq \bar{v}_c^e) \wedge$$
$$\wedge L_2(\nu^{n_1} \subset \nu_3^l) \wedge L_4(v^{n_1} \geq \bar{v}_3^e) \wedge L_3(f^e \leq f^{n_1} \leq \bar{f}^e)]. \quad (5.10)$$

Если $M_6 = 1$, то заяц № 1 недостижим, т. к. у собаки мало ресурса, а заяц имеет очень большую скорость. Далее анализируем S^{n_2} второго зайца по таблице недостижимости.

Если $M_6 = 0$, то считаем M_7 для первого зайца:

$$M_7 [(L_1(\varrho_c^{n_1} \leq \varrho_c^e) \wedge L_3(v_c^e \leq v_c^{n_1} \leq \bar{v}_c^e) \wedge$$
$$\wedge L_2(\nu_{3_1}^{n_1} \subset \nu_3^e) \wedge L^3(v_3^e \leq v_{3_1}^{n_1} \leq \bar{v}_3^e) \wedge L_4(f^{n_1} \geq \bar{f}^e); \quad (5.11)$$

Если $M_7 = 1$, то первый заяц недостижим, т. к. у собаки мало ресурса, и заяц очень далеко.

Переходим к анализу подситуации S^{n_2} по таблице недостижимости. Если $M_7 = 0$, то считаем M_8 для первого зайца:

$$M_8 [L_1(\varrho_c^{n_1} \leq \varrho_c^e) \wedge L_3(v_c^e \leq v_c^{n_1} \leq \bar{v}_c^e) \wedge L_2(\nu_{3_1}^{n_1} \subset \nu_3) \wedge$$
$$\wedge L_3(v_3^e \leq v_3^{n_1} \leq \bar{v}_3^e) \wedge L_3(f^e \leq f^{n_1} \leq \bar{f}^e); \quad (5.12)$$

Если $M_8 = 1$, то первый заяц недостижим, т. к. у собаки мало ресурса. Переходим к анализу подситуации S^{n_2} по таблице недостижимости.

Если $M_8 = 0$, то считаем, что ситуация S^{n_1}, описывающая положение собаки относительно первого зайца, не распознается по таблице недостижимости.

Возможно, что первый заяц достижим. Анализируем ситуацию S^{n_1} по соответствующей ей усеченной матрице таблицы достижимости. С помощью следующих логических функций:

$$F_1 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_4(v_c^{n_1} \geq \overline{v}_c^e) \wedge L_2(\nu_{3_1}^{n_1} \subset \nu_3^e) \wedge \\ \wedge L_1(v_{3_1}^{n_1} \leq v_3^d) \wedge L_1(f^{n_1} \leq f^d)]. \tag{5.13}$$

Если $F_1 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса, очень большая скорость, а у зайца маленькая скорость, и он близко.

Переходим к анализу S^{n_2} по таблице недостижимости.

Если $F_1 = 0$, то анализируем S^{n_1} при помощи функции F_2:

$$F_2 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_3(v_c^d \leq v_c^{n_1} \leq \overline{v}_c^d) \wedge \\ \wedge L_2(\nu_3^{n_1} \subset \nu_3^d) \wedge L_1(v_3^{n_1} \leq \overline{v}_3^d) \wedge L_1(f^{n_1} \leq f^d); \tag{5.14}$$

Если $F_2 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса, а у зайца очень маленькая скорость, и он близко. Анализируем S^{n_2} по таблице недостижимости.

Если $F_2 = 0$, то анализируем S^{n_1} при помощи функции F_3:

$$F_3 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_4(v_c^{n_1} \geq \overline{v}_c^d) \wedge L_2(\nu_{3_1}^{n_1} \subset \nu_3^d) \wedge \\ \wedge L_3(v_3^d \leq v_{3_1}^{n_1} \leq \overline{v}_3^d) \wedge L_1(f^{n_1} \leq f^d). \tag{5.15}$$

Если $F_3 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса, очень большая скорость, и заяц очень близко. Далее, переходим к анализу второго зайца по таблице недостижимости. Если $F_3 = 0$, то считаем для первого зайца:

$$F_4 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_4(v_c^{n_1} \geq \overline{v}_c^d) \wedge L_2(\nu_{3_1}^{n_1} \subset \nu_3^d) \wedge \\ \wedge L_1(v_{3_1}^{n_1} \leq v_3^d) \wedge L_3(f^d \leq f^{n_1} \leq f^d)]; \tag{5.16}$$

Если $F_4 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса, очень большая скорость, а у зайца очень маленькая скорость. Далее переходим к анализу второго зайца по таблице недостижимости.

Если $F_4 = 0$, то считаем F_5 для первого зайца:

$$F_5 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_3(v_c^d \leq v_c^{n_1} \leq \overline{v}_c^d) \wedge \\ \wedge L_2(\nu_{3_1}^{n_1} \subset \nu_3^d) \wedge L_3(v_3^d \leq v_{3_1}^{n_1} \leq \overline{v}_3^d) \wedge L^1(f^{n_1} \leq f^d). \tag{5.17}$$

Если $F_5 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса, и заяц очень близко.

Переходим к анализу второго зайца по таблице недостижимости.

Если $F_5 = 0$, то считаем F_6 для первого зайца:

$$F_6 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_3(\underline{v}_c^d \leq v_c^{n_1} \leq \bar{v}_c^d) \wedge$$
$$\wedge L_2(\nu_3^{n_1} \subset \nu_3^d) \wedge L_1(v_3^{n_1} \leq \underline{v}_3^d) \wedge L_3(\underline{f}^d \leq f^{n_1} \leq \bar{f}^d)]. \quad (5.18)$$

Если $F_6 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса, а у зайца мала скорость.

Далее переходим к анализу второго зайца по таблице недостижимости.

Если $F_6 = 0$, то считаем следующую логическую функцию для первого зайца:

$$F_7 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_4(v_c^{n_1} \geq \bar{v}_c^d) \wedge L_2(\nu_3^{n_1} \subset \nu_3^d) \wedge$$
$$\wedge L_3(\underline{v}_3^d \leq v_3^{n_1} \leq \bar{v}_3^d) \wedge L_3(\underline{f}^d \leq f^{n_1} \leq \bar{f}^d)]. \quad (5.19)$$

Если $F_7 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса и очень большая скорость.

Анализируем второго зайца по таблице недостижимости.

Если $F_7 = 0$, то считаем следующую логическую функцию F_8 для первого зайца:

$$F_8 [L_4(\varrho_c^{n_1} \geq \varrho_c^d) \wedge L_3(\underline{v}_c^d \leq v_c^{n_1} \leq \bar{v}_c^d) \wedge L_2(\nu_3^{n_1} \subset \nu_3^d) \wedge L_3(\underline{v}_3 \leq v_3^{n_1} \leq \bar{v}_3^d) \wedge$$
$$\wedge L_4(v_c^{n_1} > v_3^{n_2}) \wedge L_3(\underline{f}^d \leq f^{n_1} \leq \bar{f}^d)]. \quad (5.20)$$

Если $F_8 = 1$, то первый заяц достижим, т. к. у собаки достаточно ресурса. Далее анализируем позицию второго зайца по таблице недостижимости.

Если $F_5 = 0$, то считаем, что ситуация S^{n_1} не распознается по таблице достижимости.

Таким образом, считаем, что если хотя бы одна из составляющих ситуации S^n, S^{n_1} или S^{n_2} не распознается по таблицам достижимости и недостижимости при помощи перечисленных логических функций, то ситуация S^n не может быть узнака.

В данном случае вопрос о достижимости или недостижимости зайца решается регулярным методом. Результат решения записывается в соответствующую таблицу. Таким образом, идет процесс дообучения таблиц.

Кроме того, при переходе от одной ситуации к другой в моделируемом процессе преследования и при распознавании данных ситуаций при помощи функций недостижимости M_i ($i = 1, \ldots, 8$) и функций достижимости F_j ($j = 1, \ldots, 8$) представляется возможным дать качественную оценку процесса преследования в смысле ухудшения процесса — улучшения процесса преследования.*

* Оценка качественных параметров «хуже-лучше» может осуществляться на основании теории расплывчатых множеств Заде [3], что описано, например, в [4].

Понятия «хуже-лучше» формируются на основе смыслового и количественного анализа параметров конкретной ситуации, подлежащей распознаванию при помощи перечисленных логических функций.

Так при распознавании ситуаций по функциям недостижимости (M_i) намечается ухудшение состояния процесса преследования с точки зрения достижимости, если $M_k \rightarrow M_e$ и $e > k$. При распознавании ситуаций по функциями достижимости (F_j) намечается улучшение состояния процесса преследования с точки зрения достижимости, если $F_m \rightarrow F_n$ и $m > n$. Переход от любой функции M к любой функции F означает улучшение состояния процесса преследования.

6. Пример

На основе приведенного алгоритма распознавания ситуаций, возникающих в процессе преследования, была составлена программа для ЭВМ БЭСМ-6 на языке Алгол-60. Результаты работы программы выводились на графопостроитель.

Программа производила распознавание последовательно вводимых ситуаций при помощи указанного набора логических функций и введенных в память машины решающих таблиц. Ниже приводится контрольный пример, просчитанный на ЭВМ по распознаванию 14 ситуаций при помощи решающих таблиц, состоящих из 6 ситуаций.

Таблица «заяц недостижим»:

ϱ v_c ν_c x_c y_c $v_з$ $\nu_з$ $x_з$ $y_з$

(100; 11; 0,2000; 5; 7; 14; 0,5236; 10; 10;)
(140; 15; 0,4000; 3; 3; 20; 1,0647; 13; 6;)
(170; 5; 0,3665; 13; 12; 20; 1,0472; 21; 14)

Таблица «заяц достижим»:

(180; 20; 0,5236; 3; 3; 10; 1,0472; 11; 5;)
(130; 80; 0,5236; 3; 4; 10; 2,6180; 8; 9;)
(100; 20; 0,4014; 4; 9; 5; 2,0944; 10; 10;)

III) Процесс погони описывается последовательной сменой следующих ситуаций:

1) (135; 10; 0,4000; 2; 1; 24; 1,0472; 15; 2; 28; 1,0647; 25; 4); Заяц 1 недостижим и заяц 2 недостижим. Распознавание по M_2;

2) (130; 30; 0,4014; 2; 1; 6; 2,0944; 6; 6; 7; 2,0944; 7; 6); Заяц 1 достижим и заяц 2 достижим. Распознавание по F_8;

3) (180; 60; 0,5236; 2; 1; 9; 1,0472; 6; 7; 9; 1,0472; 5; 8); Оба зайца достижимы. Распознавание по F_8;

4) (170; 60; 0,5236; 3; 2; 9; 2,6180; 4; 6; 9; 2,6180; 5; 8); Оба зайца достижимы. Распознавание по F_5;

5) (160; 90; 0,5236; 3; 4; 4; 2,6180; 6; 10; 9; 2,6180; 8; 9); Заяц 1 достижим. Распознавание по F_4. Заяц 2 достижим. Распознавание по F_5;

6) (150; 90; 0,4014; 4; 9; 6; 2,6180; 2; 4; 9; 2,6180; 9; 4); Заяц 1 достижим. Распознавание по F_3. Заяц 2 достижим. Распознавание по F_4;

7) (140; 70; 0,5236; 5; 1; 4; 2,0944; 7; 6; 4; 2,0944; 2; 7); Заяц 1 достижим. Распознавание по F_2. Заяц 2 достижим. Распознавание по F_3;

8) (130; 30; 0,4014; 2; 1; 6; 2,0944; 6; 6; 7; 2,0944; 7; 6); Оба зайца достижимы. Распознавание по F_8;

9) (110; 30; 0,4014; 4; 9; 3; 2,0944; 8; 10; 5; 2,0944; 9; 10); Оба зайца достижимы. Распознавание по F_1;

10) (105; 30; 0,4014; 4; 9; 3; 2,0944; 4; 9; 5; 2,0944; 9; 10); Первый заяц пойман. Заяц 2 достижим. Распознавание по F_1;

11) (108; 25; 0,4014; 5; 8; 0; 0; 5; 8; 5; 2,0944; 8; 6); Заяц I пойман. Заяц 2 достижим. Распознавание по F_1;

12) (105; 25; 0,4014; 3; 8; 0; 0; 3; 8; 5; 2,0944; 5; 4); Заяц 1 пойман. Заяц 2 достижим. Распознавание по F_1;

13) (102; 23; 0,4014; 2; 2; 0; 0; 2; 3; 5; 2,0944; 1; 8); Заяц 1 пойман. Заяц 2 достижим. Распознавание по F_1;

14) (100; 23; 0,4014; 2; 1; 0; 0; 2; 1; 5; 2,0944; 2; 1); Заяц 1 пойман. Заяц 2 пойман.

Равенство 0 периметров направления и скорости зайца, а также совпадение координат зайца и собаки означает факт поимки зайца. Приводятся графики снятые с графикостоителя и иллюстрирующие процесс погони, состояний из 14 анализируемых ситуаций.

I) Качественная оценка процесса погони за первым зайцем и за вторым зайцем (рис. 2);

II) График изменения координат собаки и зайцев в процессе преследования. Отмеченные точки (вершины) — координаты собаки и зайцев (рис. 3);

C — координаты собаки;

$з_1$ — координаты зайца 1; $з_2$ — координаты зайца 2;

Цифры при вершинах — моменты времени или номера анализируемых ситуаций.

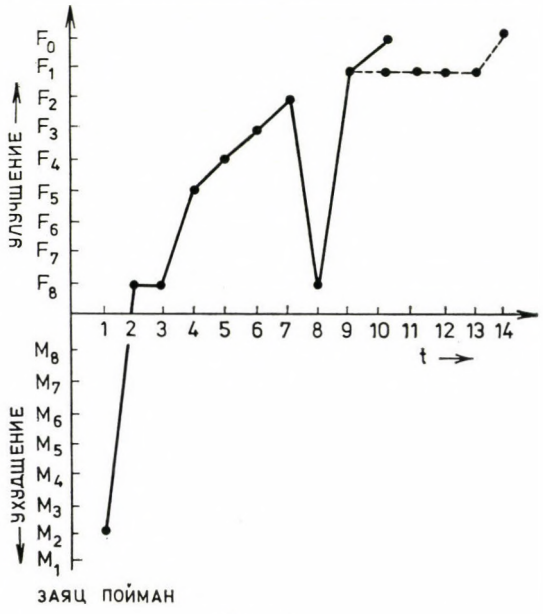

F0
F1
F2
F3
F4
F5
F6
F7
F8
УЛУЧШЕНИЕ
1 2 3 4 5 6 7 8 9 10 11 12 13 14
t
M8
M7
M6
M5
M4
M3
M2
M1
УХУДШЕНИЕ
ЗАЯЦ ПОЙМАН

Рис. 2

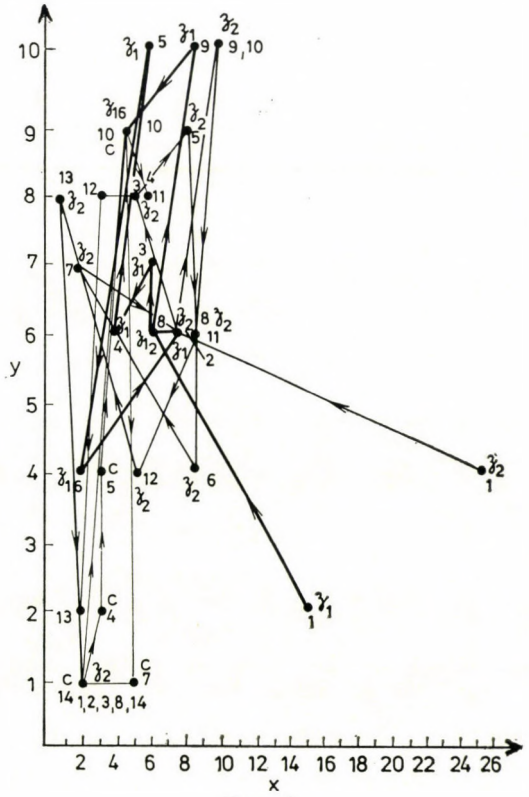

y
x
2 4 6 8 10 12 14 16 18 20 22 24 26
1 2 3 4 5 6 7 8 9 10
9,10
C
1,2,3,8,14

Рис. 3

В настоящее время ведутся работы по составлению рабочих программ для получения решающих таблиц на основе одного из регулярных методов, а также программ дообучения, данных таблиц, позволяющих произвести стыковку распознающей программы с программой получения решающих таблиц.

В результате процесса узнавания ситуаций и дообучения таблиц может наступить такой момент, когда большинство ситуаций будет распознаваться по таблицам без дальнейшего их дообучения. Определение min размерности таблиц для такого случая — важный результат решения данной задачи.

Литература

1. *Бонгард М. М.* Проблема узнавания., М., Наука, 1967.
2. *Поспелов Д. А., Пушкин В. Н.* Мышление и автоматы. М., «Советское радио», 1972.
3. *Zadeh Z.* Fuzzy sets. Information and control. IV 8, 1965, pp. 338–352.
4. *Шапиро Д. И.* Математические методы в проблеме принятия решения (Препринт), М., Совет по кибернетике АН СССР, 1974.

Situation recognition in the pursuit problem

D. A. POSPELOV, N. N. CHEREZOVA, D. I. SHAPIRO

(Moscow)

This paper deals with the problem of pursuit.

The problem of pursuit is considered in the following situation: a dog and two hares are moving on the surface. The dog should catch either one or the two hares, if possible.

The process of pursuit is described as a finite sequence of situations. Each situation is described by a set of parameters of a dog and two hares such as: the dog's resources, the dog's speed, the dog's direction, the dog's coordinates, the speed of the first hare and the second one, the directions of the first hare and the second one, the coordinates of the first hare and the second one.

At very fixed moment the dog should make a decision on the basis of the analysis of the current situation.

The paper suggests using the method of recognizing the given situations on the basis of the specific set of logical functions and decision tables of two kinds: "The hare is unattainable" and "The hare is Attainable".

At first each of the concrete situations is analyzed according to the table "The hare is unattainable". Then the process of generalization of the situation parameters, being the elements of the table "The hare is attainable", takes place. We should compare then the parameters of the concrete situation with the parameters which are the result of the generalization.

For this it is necessary to make consecutive calculation of values of several logical functions constructed on the parameters of the concrete situation and on the basis of the generalization.

Even if the value of only one of the calculated logical functions is equal to 1, the decision of unattainability of the hare is made. If the values of all the calculated logical functions are equal to 0, we make a decision of attainability of the hare.

In this case the concrete situation is analyzed according to the table "The hare is attainable" through the analogous scheme.

If the concrete situation is not recognized according to the tables "The hare is attainable" and "The hare is unattainable" with the help of the given method, it is evaluated by some regular methods (this paper suggests using the geometric method of evaluation).

The result of the decision is written into one of the tables "The hare is attainable" or "The hare is unattainable".

Thus, a process of teaching decision tables occurs.

Д. А. Поспелов, Н. Н. Черезова, Д. И. Шапиро
Вычислительный центр АН СССР
СССР, Москва, В-333,
ул. Вавилова, д. 40

Problems of Control and Information Theory, Vol. 5 (5–6), pp. 437–448 (1976)

К ВОПРОСУ О МЕТОДЕ ДИНАМИКИ МОМЕНТОВ

В. Г. РОМАНОВ

(Москва)

(Поступила в редакцию 5 мая 1975 г.)

Совершенствуется метод динамики моментов, предназначенный для исследования систем, включающих в себя разнотипные элементы, которые могут находиться в дискретном множестве состояний и изменять свое состояние под действием пуассоновских потоков.

1. Введение

Исследование больших систем и создание автоматических систем управления сопряжено с необходимостью моделирования процессов, протекающих в этих системах, например моделирования процессов сбора, накопления и обработки информации, функционирования отдельных подсистем и системы в целом при различных вариантах управлений.

В настоящей статье рассматриваются системы, включающие в себя разнотипные элементы, каждый из которых может находиться в одном из дискретного множества состояний и изменять свое состояние под действием пуассоновских потоков.

Исследование таких систем при небольших количествах элементов и их состояний может быть проведено с помощью уравнений для вероятностей состояний системы. При больших количествах элементов в системе широко применяется метод динамики средних. Кроме того, в этом случае может быть применен предложенный в [1], [2] метод динамики моментов, а когда интенсивности перехода элементов являются несложными (в основном линейными) функциями чисел элементов в различных состояниях, уравнения динамики моментов могут быть получены с помощью производящих функций [3].

Однако, во-первых, каждый из этих методов применим к ограниченному подклассу систем, а во-вторых, в большинстве случаев для составления уравнений динамики моментов требуется решать задачу каждый раз заново.

В статье получены дифференциальные уравнения для моментов в достаточной степени произвольной системы, которые сравнительно просто конкретизируются для исследуемой системы.

2. Дифференциальные уравнения для моментов

Рассмотрим некоторую систему, которая в каждый момент времени $t \geq 0$ может находиться в одном из состояний, образующих счетное множество A. Вероятность нахождения системы в момент $t \geq 0$ в состоянии $x \in A$ будем обозначать $p(x)$.

С течением времени система может переходить из одного состояния в другое. Это явление будем характеризовать случайным процессом $\{X(t), t \geq 0\}$, полагая, что $X(t) = x$, если в момент t система находится в состоянии x.

Будем предполагать процесс, протекающий в системе, марковским, т. е. для любых x, $y(\sigma) \in A$ при $0 \leq \sigma \leq \tau$ и $\tau < t$ считать выполняющимся равенство

$$P\{X(t) = x \,|\, X(\sigma) = y(\sigma),\ 0 \leq \sigma \leq \tau\} = P\{X(t) = x \,|\, X(\tau) = y(\tau)\}. \quad (2.1)$$

Более того, при всех $t \geq 0$ и x, $y \in A$ будем предполагать существование предела

$$\lim_{\Delta t \to 0} \frac{P\{X(t + \Delta t) = y \,|\, X(t) = x\}}{\Delta t} = \lambda(x, y, t), \quad (2.2)$$

где $\lambda(x, y, t)$ или просто $\lambda(x, y)$ будем называть интенсивностью перехода системы из состояния x в состояние y.

Для такой системы справедливы следующие дифференциальные уравнения для вероятностей $p(x)$:

$$\dot{p}(x) = \sum_{y \in A} \lambda(y, x)\, p(y) - \sum_{y \in A} \lambda(x, y)\, p(x), \quad x \in A. \quad (2.3)$$

С помощью этой системы уравнений получим дифференциальное уравнение для математического ожидания произвольной функции $\varphi(X)$. Для этого умножим обе части равенства каждого уравнения системы на $\varphi(x)$ и просуммируем их по всем $x \in A$. В результате получим

$$\sum_{x \in A} \varphi(x)\, \dot{p}(x) = \sum_{x \in A} \sum_{y \in A} \varphi(x)\, \lambda(y, x)\, p(y) - \sum_{x \in A} \sum_{y \in A} \varphi(x)\, \lambda(x, y)\, p(x). \quad (2.4)$$

Стоящая в левой части этого равенства сумма представляет собой производную от математического ожидания функции $\varphi(X)$. В правой части равенства в первой двойной сумме положим $x = y'$, $y = x'$, изменим порядок суммирования и опустим штрихи у x' и y'. Это позволит, объединив суммирование по x, записать

$$\frac{d}{dt} M[\varphi(X)] = \sum_{x \in A} p(x) \sum_{y \in A} [\varphi(y) - \varphi(x)]\, \lambda(x, y). \quad (2.5)$$

Получившееся в результате в правой части равенства выражение представляет собой математическое ожидание внутренней двойной суммы. Поэтому математическое ожидание произвольной функции должно удовлетворять следующему дифференциальному уравнению

$$\frac{d}{dt} M[\varphi(X)] = M\Big[\sum_{y \in A} (\varphi(y) - \varphi(X))\lambda(X, y)\Big]. \tag{2.6}$$

Конкретизируем теперь систему. Пусть она включает l типов элементов, причем каждый из n_i $(i = 1, \ldots, l)$ элементов i-го типа может находиться в m_i состояниях. Переходы элементов из одного состояния в другое будем предполагать пуассоновскими.

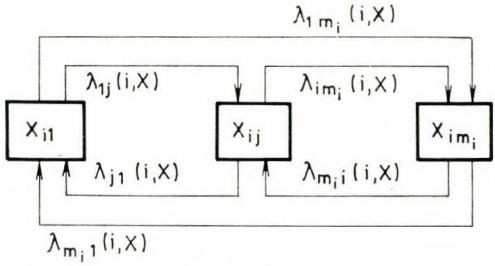


Рис. 1

Граф состояний системы представляет собой совокупность графов состояний различных типов элементов, один из которых для элементов i-го типа изображен на рис. 1. Состояния x такой системы удобно представлять в виде вектора:

$$x = (x_1, \ldots, x_i, \ldots, x_l),\ x_i = (x_{i1}, \ldots, x_{ij} \ldots, x_{i,m_i}), \tag{2.7}$$

$x_{i,j}$ — число элементов i-го типа в j-ом состоянии.

Интенсивности перехода элементов i-го типа из j-го состояния в k-е будем обозначать $\lambda_{ijk}(x)$ и считать известными функциями времени t и состояния системы x.

Ввиду ординарности пуассоновских потоков интенсивности $\lambda(x, y)$ отличны от нуля только в том случае, если все составляющие вектора y равны соответствующим составляющим вектора x, кроме каких-либо двух: $y_{ij} = x_{ij} - 1$ и $y_{ik} = x_{ik} + 1$. Следовательно, обозначая такие векторы y через x_{ijk}, можно записать

$$\lambda(x, y) = \begin{cases} \lambda_{ijk}(x) & \text{при } y = x_{ijk}, \\ 0 & \text{при } y \neq x_{ijk}. \end{cases} \tag{2.8}$$

Поэтому в выражении (2.6) суммирование по векторам $y \in A$ заменим суммированием по i, j и k, которое вынесем за знак математического ожидания. В результате получим

$$\frac{d}{dt} M[\varphi(X)] = \sum_{i=1}^{l} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} M\left[\left(\varphi(X_{ijk}) - \varphi(X)\right) \lambda_{ijk}(X)\right]. \tag{2.9}$$

В выражении (2.9) следует учитывать, что

$$\lambda_{ijk}(X) = 0 \text{ при } X_{ij} = 0,$$

т. к. в этом случае в состоянии, из которого следует поток переходов, нет ни одного элемента.

Поэтому имеет место следующее равенство

$$M[\psi(X) \cdot \lambda_{ijk}(X)] = P(X_{ij} \neq 0)\, M[\psi(X) \lambda_{ijk}(X) \mid X_{ij} \neq 0], \tag{2.10}$$

которое удобно использовать, когда ни $\lambda_{ijk}(X)$, ни $\psi(X)$ не пропорциональны составляющей X_{ij} вектора X.

Таким образом, получено дифференциальное уравнение для произвольного момента распределения вероятностей состояний системы. Подчеркнем, что при его выводе не накладывались никакие ограничения на характер изменения интенсивностей перехода как функций времени. Поэтому интенсивности перехода могут быть произвольными функциями времени. Более того, можно показать, что они могут быть линейными функциями дельта-функций, лишь бы множества моментов времени для различных интенсивностей перехода, при которых аргументы дельта-функций, входящие в них, равны нулю, не пересекались, т. е. соблюдалась ординарность совокупности всех потоков переходов.

Рассмотрим некоторые частные случаи уравнения (2.9).

3. Уравнения для математических ожиданий чисел элементов в различных состояниях и центральных моментов второго порядка

Дифференциальные уравнения для математических ожиданий чисел элементов в различных состояниях найдем, положив в (2.9) $\varphi(X) = X_{rs}$. Из определения вектора X_{ijk} следует, что в этом случае $\varphi(X_{ijk}) = X_{rs} + \delta_{ri}(\delta_{sk} - \delta_{sj})$, где $\delta_{\mu\nu}$ — индекс Кронекера. Поэтому выражение (2.9) упрощается и принимает вид:

$$\frac{d}{dt} M[X_{rs}] = \sum_{j=1}^{m_s} M[\lambda_{rjs}(X) - \lambda_{rsj}(X)], \quad r = 1, \ldots, l; s = 1, \ldots, m_r. \tag{3.1}$$

Здесь, как и в (2.9), следует учитывать возможность использования равенства (2.10).

Сравнение этих уравнений с уравнениями динамики средних показывает, что в правой части (3.1) стоят математические ожидания интенсивностей в отличие от интенсивностей как функций математических ожиданий чисел элементов в состояниях, стоящих в уравнениях динамики средних.

Поэтому, если

$$\lambda_{ijk}(X) = c_{ijk} X_{ij}, \quad i = 1, \ldots, l; j, k = 1, \ldots, m_i, \quad (3.2)$$

то уравнения динамики средних совпадают с точными уравнениями (3.1), то есть оказываются точными при любых числах элементов в системе, а не только при больших, как это обычно считается.

Теперь найдем дифференциальные уравнения для центральных моментов второго порядка. Положим в (2.9): $\varphi(X) = \dot{X}_{rs}\, \dot{X}_{\mu\nu}$, где $\dot{X}_{\alpha\beta} = X_{\alpha\beta} - M[X_{\alpha\beta}]$. В соответствии с этим $\varphi(X_{ijk}) = (\dot{X}_{rs} + \delta_{ri}(\delta_{sk} - \delta_{sj})(\dot{X}_{\mu\nu} + \delta_{\mu i}(\delta_{\nu k} - \delta_{\nu j}))$, и выражение (2.9) после упрощений принимает вид:

$$\frac{d}{dt} M[\dot{X}_{rs}\dot{X}_{\mu\nu}] = \sum_{j=1}^{m_\mu} M[\dot{X}_{rs}(\lambda_{\mu j\nu}(X) - \lambda_{\mu\nu j}(X))] +$$

$$+ \sum_{j=1}^{m_r} M[\dot{X}_{\mu\nu}(\lambda_{rjs}(X) - \lambda_{rsj}(X))] + \delta_{r\nu}\,\delta_{s\nu} \sum_{j=1}^{m_r} M[\lambda_{rjs}(x) + \lambda_{rsj}(x)] - \quad (3.3)$$

$$- \delta_{r\mu} M[\lambda_{rs\nu}(x) + \lambda_{r\nu s}(X)], \quad r, \mu = 1, \ldots, l; s = 1, \ldots, m_r; \nu = 1, \ldots, m_\mu.$$

Отсюда при $r \neq \mu$ получаем уравнения для корреляционных моментов составляющих различных типов элементов

$$\frac{d}{dt} M[\dot{X}_{rs}\dot{X}_{\mu\nu}] = \sum_{j=1}^{m_r} M[\dot{X}_{rs}(\lambda_{\mu j\nu}(X) - \lambda_{\mu\nu j}(X))] +$$

$$+ \sum_{j=1}^{m_r} M[\dot{X}_{\mu\nu}(\lambda_{rjs}(X) - \lambda_{rsj}(X))]. \quad (3.4)$$

При $r = \mu$, но $s \neq \nu$ получаются уравнения для корреляционных моментов составляющих одних и тех же типов элементов

$$\frac{d}{dt} M[\dot{X}_{rs}\dot{X}_{r\nu}] = \sum_{j=1}^{m_r} [\dot{X}_{rs}(\lambda_{rj\nu}(X) - \lambda_{r\nu j}(X)) + \dot{X}_{r\nu}(\lambda_{rjs}(X) - \lambda_{rsj}(X))] -$$

$$- M[\lambda_{r\nu s}(X) + \lambda_{rs\nu}(X)]. \quad (3.5$$

Наконец, при $r = \mu$ и $s = \nu$ получаются уравнения для дисперсий составляющих вектора X

$$\frac{d}{dt} M[(\dot{X}_{rs})^2] = \sum_{j=1}^{m} \{2M[\dot{X}_{rs}(\lambda_{rjs}(X) - \lambda_{rsj}(X))] + M[\lambda_{rjs}(X) + \lambda_{rsj}(X)]\}. \quad (3.6)$$

Во всех этих уравнениях, как и в (2.9), следует учитывать возможность использования равенства (2.10).

Аналогично могут быть получены из (3.1) дифференциальные уравнения для любых других моментов распределения вероятностей состояний системы.

Выражения (3.1), (3.5) и (3.6) позволяют найти условия, при которых $M[X_{rs}(t)] \equiv M\left[(\dot{X}_{rs}(t))^2\right]$. Эти условия определяются следующим утверждением.

Если для некоторого подмножества A'_r множества состояний A_r элементов r-го типа для любого $j \in A'_r$

$$\lambda_{rjk}(X) = c_{rjk} X_{rj}, \quad c_{rjk} \neq f(X) \; \forall k \in A_r, \qquad (3.7)$$
$$\lambda_{rkj}(X) \neq f(X) \; \forall k \notin A'_r$$

и в начальный момент времени

$$X_{rj} = 0 \; \forall j \in A'_r, \qquad (3.8)$$

то

$$\tilde{x}_{rj}(t) \equiv D_{rj}(t), \; K_{rjk}(t) \equiv 0 \;\; \forall j, k \in A'_r, \; j \neq k, \qquad (3.9)$$

где $\tilde{x}_{rj}(t) = M[X_{rj}(t)]$, $D_{rj}(t) = M\left[(\dot{X}_{rj}(t))^2\right]$, $K_{rjk}(t) = M[\dot{X}_{rj}(t)\dot{X}_{rk}(t)]$.

Доказательство. Найдем дифференциальные уравнения, которым должны удовлетворять разности $\tilde{x}_{rs} - D_{rs}$ для $s \in A'_r$, где зависимость $\tilde{x}_{rs}$ и D_{rs} от времени опустим для краткости. Вычтем (3.6) из (3.1) и подставим в правую часть получившегося равенства интенсивности перехода из (3.7)

$$\dot{\tilde{x}}_{rs} - \dot{D}_{rs} = -2 \sum_{j=1}^{m_r} M[\lambda_{rsj}(X) - \dot{X}_{rs}\lambda_{rsj}(X) + \dot{X}_{rs}\lambda_{rjs}(X)] =$$
$$= -2(\tilde{x}_{rs} - D_{rs}) \sum_{j=1}^{m_r} c_{rsj} - 2 \sum_{\substack{j \in A'_r \\ j \neq s}} c_{rjs} K_{rjs}. \qquad (3.10)$$

Теперь найдем дифференциальные уравнения, которым должны удовлетворять K_{rsv} для $s, v \in A'_r$. Подставим интенсивности перехода из (3.7) в (3.5)

$$\dot{K}_{rsv} = \sum_{\substack{j \in A'_r \\ j \neq s}} c_{rjv} K_{rsj} + \sum_{\substack{j \in A'_r \\ j \neq v}} c_{rjs} K_{rvj} - K_{rsv} \sum_{j=1}^{m_r} (c_{rvj} + c_{rsj}) -$$
$$- c_{rsv}(\tilde{x}_{rs} - D_{rs}) - c_{rvs}(\tilde{x}_{rv} - D_{rv}). \qquad (3.11)$$

Из системы уравнений (3.10—3.11) видно, что производные функций времени $\tilde{x}_{rs} - D_{rs}$ ($s \in A'_r$) и K_{rsv} ($s, v \in A'_r$, $s \neq v$) пропорциональны линейной комбинации этих же функций. Далее из (3.8) следует, что эти функции в

начальный момент времени равны нулю. Следовательно, эти функции тождественно равны нулю, что доказывает высказанное утверждение.

Отметим, что доказанное утверждение не накладывает ограничений на интенсивности как функции времени.

В работе [2] доказывается аналогичное утверждение для одной частной системы, подобной системе, рассмотренной в примере 1, с наложением дополнительных ограничений на интенсивности перехода как функции времени.

4. Аппроксимация распределения вероятностей

Выше получено дифференциальное уравнение (2.9) для произвольного момента распределения вероятностей состояний системы. С помощью этого уравнения можно записать необходимое число дифференциальных уравнений, чтобы полученная таким образом система уравнений оказалась замкнутой.

Когда распределение вероятностей состояний характеризуется небольшим числом параметров, то и число уравнений этой системы уравнений будет небольшим. В противном случае, аппроксимируя распределение вероятностей состояний каким-либо распределением, для характеристики которого требуется небольшое число параметров, можно также получить систему с небольшим числом уравнений.

Например, можно предположить, что распределение вероятностей состояний является многомерным распределением Пуассона. В этом случае для получения замкнутой системы можно ограничиться системой уравнений (3.1) для математических ожиданий, которая может рассматриваться как уточненная система уравнений динамики средних.

Допущение распределения Пуассона в качестве аппроксимирующего возможно, например, при неограниченном числе элементов в системе, когда выполняются условия (3.7—3.8) для всех состояний элементов. Такое допущение также возможно, когда числа элементов каких-либо типов ограничено, но все же достаточно велико, и множества состояний элементов каких-либо типов (тех же или других) только на одно состояние больше множества A'_r.

Уточненной системы уравнений динамики средних оказывается также достаточно, если допустить, что распределение каждого типа элементов полиномиальное, не зависящее от распределений других типов элементов.

По аналогии с уравнениями динамики средних, следуя [2], будем называть замкнутую систему уравнений для моментов распределения системой уравнений динамики моментов, а метод исследования систем элементов с ее помощью — методом динамики моментов.

Ниже на примере будет показано, что аппроксимация распределения вероятностей может позволить добиться вполне удовлетворительной точности определения необходимых моментов даже при небольших количествах элементов в системе и когда интенсивности перехода – существенно нелинейные функции вектора X.

Это отличает выражения (3.1) от уравнений динамики средних, а выражения (3.4)—(3.6), кроме того, позволяют проще получить уравнения динамики моментов для конкретных систем элементов, чем с помощью способов, предложенных в [1], [2], [3].

Пример 1. Планируется развертывание производства некоторого типа техники. Интенсивность производства единиц техники является функцией времени и не зависит от количества единиц техники, находящихся в эксплуатации. Каждая единица техники выходит из строя с интенсивностью, которая также является функцией времени и не зависит от количества единиц техники, находящихся в эксплуатации. Определить математическое ожидание и дисперсию единиц техники, находящихся в эксплуатации.

Решение. Каждая единица техники может находиться в трех состояниях: находиться в производстве, эксплуатироваться, быть неисправной. Обозначим эти состояния цифрами 1, 2 и 3 соответственно. Граф состояний единиц техники приведен на рис. 2.

Состояние системы будем характеризовать вектором $X = \{X_1, X_2, X_3\}$, где X_i — число единиц техники в i-том состоянии.

Дифференциальные уравнения для математического ожидания $\tilde{x}_2$ и дисперсии D_2 единиц техники в эксплуатации найдем с помощью систем (3.1) и (3.6)

$$\dot{\tilde{x}}_2(t) = \lambda_{12}(t) - \tilde{x}_2(t)\, c_{23}(t)\,,$$

$$\dot{D}_2(t) = -\,2\,D_2(t)\, c_{23}(t) + \lambda_{12}(t) + \tilde{x}_2(t)\, c_{23}(t)\,. \qquad (4.1)$$

Решение этой системы не вызывает трудностей. Не приводя его, заметим, что для второго состояния выполняются условия (3.7—3.8) и поэтому решения обоих уравнений совпадают, т. е. $\tilde{x}_2(t) \equiv D_2(t)$.

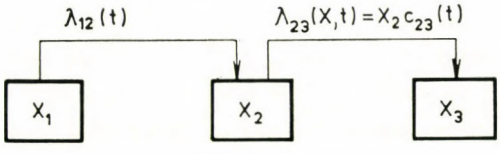


Рис. 2

Пример 2. Рассмотрим систему, выключающую два типа элементов, количества которых m и n соответственно. Элементы обоих типов могут находиться в двух состояниях. Численности первого и второго состояний элементов первого типа обозначим через X_1 и X_2, а элементов второго типа — через Y_1 и Y_2 соответственно. В начальный момент времени $X_1 = m$, $X_2 = 0$, $Y_1 = n$, $Y_2 = 0$. В дальнейшем элементы могут переходить из первого состояния во второе. Интенсивности перехода μ_{12} элементов первого типа и ν_{12} элементов второго типа из первого состояния во второе определяются следующими выражениями

$$\mu_{12} = \nu Y_1 \delta(X_1),\ \nu_{12} = \mu X_1 \delta(Y_1), \tag{4.2}$$

где μ и ν — константы.

Требуется определить математические ожидания чисел обоих типов элементов в их первых состояниях.

Применительно к биологии данная система может служить моделью борющихся популяций. Применительно к военному делу она может служить моделью боя двух группировок. Нами данная система выбрана из следующих соображений. Во-первых, интенсивности перехода элементов в ней — существенно нелинейные функции чисел элементов в состояниях. Во-вторых, еще Ланчестером [4] она исследовалась с помощью уравнений динамики средних, которые известны в литературе как уравнения Ланчестера. В-третьих, для нее автором найдены выражения для вероятностей состояний системы, что позволяет установить точность уравнений Ланчестера и метода моментов.

Уравнения динамики средних для данной системы имеет смысл записать для математических ожиданий относительных численностей элементов в первом состоянии $\xi = M[X_1]/m$, $\mu = M[Y_1]/n$ и безразмерного времени $\tau = \overline{\quad} t$, т. е. в следующем виде:

$$\dot{\xi} = -\varkappa^{-1}\eta,\ \dot{\eta} = -\varkappa\xi, \tag{4.3}$$

где $\varkappa = m\sqrt{\mu}/n\sqrt{\nu}$.

При этом необходимо учитывать ограничения $\dot{\xi} \leq 0$, $\dot{\eta} \leq 0$, $\xi \geq 0$, $\mu \geq 0$, т. к. численности элементов в первом состоянии не могут возрастать и не могут быть отрицательными.

Исследуем эту систему с помощью метода моментов. Выберем такое аппроксимирующее распределение, для характеристики которого достаточно знания математических ожиданий численностей элементов.

В этом случае будет достаточно дифференциальных уравнений для средних численностей элементов обоих типов в первом состоянии, которые в соответствии с (2.10), (3.1) и (4.2) имеют вид

$$\frac{d}{dt}M[X_1] = -M[\mu_{12}] = -\nu P(X_1 \neq 0)\,M[Y_1 \mid X_1 \neq 0] = \\ = -\nu M[Y_1] + \nu P(X_1 = 0)\,M[X_1 \mid X_1 = 0]\,, \qquad (4.4)$$

$$\frac{d}{dt}M[Y_1] = -M[\nu_{12}] = -\mu P(Y_1 \neq 0]\,M[X_1 \mid Y_1 \neq 0] = \\ = -\mu M[X_1] + \mu P(Y_1 = 0)\,M[X_1 \mid Y_1 = 0]\,.$$

Будем считать, что безусловные распределения элементов обоих типов биномиальные, т. е.

$$P(X_1 = x) = C_m^x u^x (1-u)^{m-x},\ P(Y_1 = y) = C_n^y \nu^y (1-\nu)^{n-y}, \qquad (4.5)$$

где $u(\nu)$ — вероятность перехода одного элемента первого (второго) типа из первого состояния во второе.

Кроме того, будем считать, что

$$M[X_0 \mid Y_1 = 0] = m,\ M[Y_1 \mid X_1 = 0] = n\,. \qquad (4.6)$$

Данное допущение является довольно грубым, но и в этом случае уравнения динамики моментов будут намного точнее уравнений динамики средних при малых числах элементов в системе.

Подставляя в (4.4) выражения для соответствующих вероятностей и математических ожиданий из (4.5) и (4.6) и вводя безразмерное время $\tau = \sqrt{\mu\nu}\ t$, полчим следующие дифференциальные уравнения:

$$\dot{u} = -\varkappa^{-1}[\nu - (1-u)^m],\ \dot{\nu} = -\varkappa[u - (1-\nu)^n]\,, \qquad (4.7)$$

при решении которых необходимо учитывать ограничения $u \geq 0$, $\nu \geq 0$, $\dot{u} \leq 0$, $\nu \leq 0$.

Таким образом, для данной системы имеем две системы дифференциальных уравнений: систему уравнений динамики средних (4.3) и систему уравнений (4.7) метода моментов.

На рис. 3 приведены максимальные по модулю ошибки $\delta_{\max}$ определения математических ожиданий элементов обоих типов в первых состояниях с помощью систем (4.3) и (4.7). Приведенные на рисунке ошибки соответствуют коэффициенту $\varkappa = 0{,}99$. Уравнения динамики средних (4.3), как видно из рисунка, могут иметь ошибку определения математических ожида-

ний чисел элементов в первом состоянии, достигающую 57% общего числа элементов данного типа. В то же время ошибка уравнений динамики моментов (4.7) не превышает 12%. Аналогичная картина имеет место при $\varkappa = 0{,}51$ и $\varkappa = 0{,}81$.

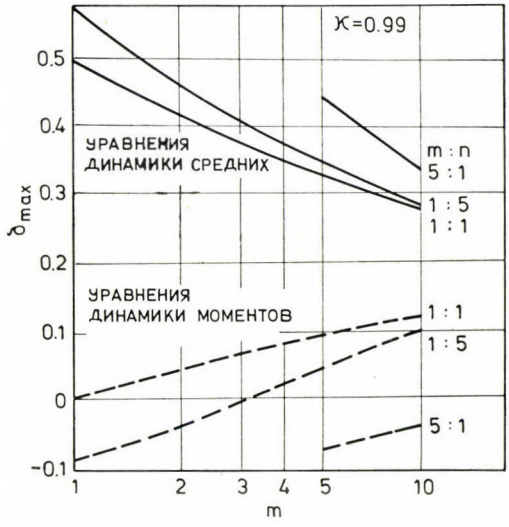


Рис. 3

Литература

1. *Вентцель Е. С.* Определение вероятностей состояний сторон в динамике боя многочисленных групп. «Морской сборник», 1962 № 10.
2. *Тараканов К. В., Овчаров Л. А., Тырышкин А. Н.* Аналитические методы исследования систем. М., «Сов. радио», 1974.
3. *Баруча-Рид А. Т.* Элементы теории марковских процессов и их приложения. М., «Наука», 1969.
4. *Lanchester, F. W.*, Aircraft in warfare. The down of Fourth Arm. London, 1916.

On method of moment dynamics

V. G. ROMANOV

(Moscow)

The systems with a countable set of states $\mathcal{A}$ which can change their state under the influence of Poisson streams are considered. The probability, $p(x)$, that the system is in the state $x \in \mathcal{A}$ at the moment t must satisfy differential equation (2.3) where $\lambda(x,y)$ is the rate of the system transition at the moment t from the state x to the state y.

It is shown that in accordance with equation (2.3) the mathematical expectation of any arbitrary function $\varphi(x)$ must satisfy differential equation (2.6).

If the system includes l types of elements and each of n_i elements of ith type may be in m_i state, then equation (2.6) takes the form of (2.9). Here $\lambda_{ijk}(x)$ is the rate

of the transition of i th type from j-th state to k-th one and x is a vector whose components are x_{il} $(i = 1, \ldots, l;\ j = 1, \ldots, m_i)$, i.e. it is a number of elements of i-th type in j-th state.

The differential equations which must be satisfied by the mathematical expectation and variances of element numbers in the different states [eqs. (3.1) and (3.6)], the correlation moments of the connection of element numbers of one type [eq. (3.5)] and different types [eq. (3.4)] are obtained by equation (2.9).

It is shown that the equations of a known method of average dynamics are in agreement with the exact ones (3.1) if $\lambda_{ijk}(x) = c_{ij\,k}x_{ij}\ \forall\ i = 1, \ldots, l;\ j = 1, \ldots, m_i$. Thereby in equations (2.9), (3.1) through (3.6) the rates of the transition may be arbitrary time functions.

It is proved that for some subset $\mathfrak{N}_r'$ of set $\mathfrak{N}_r$ of the states of elements of r-th type for which for all $j \in \mathfrak{N}_r'$ the rates of the transition satisfy conditions (3.7) and at the initial moment of time conditions (3.8) are carried out, equation (3.9) is valid.

In order to reduce the number of differential equations of the closed system of equations for the moments we suggest to approximate the probability distribution of states by some distribution which is characterized by a small number of parameters.

For instance when the element number in the system is not limited and for all states of elements, conditions (3.7 – 3.8) are performed, the probability distribution of states may be approximated by the multidimensional Poisson distribution. In some cases the distribution of each type of elements may be assumed to be polynomial without depending on the distribution of other types of elements. In both cases to get a closed system of differential equations, equation (3.1) is enough for the mathematical expectations which may be considered as a more precisely defined system of equations of average dynamics. In other cases it may be in need a greater number of differential equations including the equations for the other moments of the probability distribution of states. On the analogy of equations of average dynamics the closed system of differential equations for the moments of the probability distribution of states we, following [2], will refer to as the equations of moment dynamics.

Application of the method of moments is illustrated in two examples. In the first the development of production of some types of the technique is considered. The state graph of a such system is illustrated in Fig. 2. Differential equations which must be satisfied by the mathematical expectation $\tilde{x}_2(t)$ and variance $D_2(t)$ of the number of units of techniques being in operation are obtained.

In the second example the system with two types of elements which initial numbers are m and n is considered and each of them may be in two states. The rates of the transition of elements from the first state to the second one are essentially nonlinear functions of element numbers X_1 and Y_1 of both types in the first state. Their rates are determined by expressions (4.2). The equations of average dynamics for the given system take the form of (4.3). The equation of moment dynamics when the probability distribution of states of both types of elements is approximated by binominal distributions (4.5) connected by conditions (4.6) takes the form of (4.7).

The maximum error by modulus in defining of the mathematical expectations of elements of both types in the first states by means of systems (4.3) and (4.7) is illustrated in Fig. 3. Exact values of corresponding mathematical expectations are defined by the probability distribution of states obtained by the author. The figure shows that the method of moment dynamics permits the error in defining of mathematical expectations to be reduced up to 12% as compared to 57% that are given by the method of average dynamics without increasing of the account of differential equations.

В. Г. Романов
Научный совет по комплексной
проблеме «Кибернетика»
СССР, Москва В-333,
ул. Вавилова, д. 40

Problems of Control and Information Theory, Vol. 5 (5, 6), pp. 449—457 (1976)

AN UPPER BOUND OF ERROR PROBABILITIES FOR MULTIHYPOTHESES TESTING AND ITS APPLICATION IN ADPATIVE PATTERN RECOGNITION

L. GYÖRFI

(Budapest)

(Received September 20, 1975)

The mean square upper bound of the difference between the error probability of Bayesian decision and the decision error probability connected with some estimates of Bayesian decision function is well known for two hypotheses testing problems and the convergence of mean square minimization algorithm is also proved for weakly dependent labelled samples. This paper presents an improved upper bound given by the mean distances of Bayesian decision functions and their estimates for multihypotheses testing. A slight modification of this upper bound might actually be minimized over the space of given finite dimensional decision functions. It is an adaptive recursive algorithm, a version of stochastic approximation particularly suited to certain tasks in statistical pattern recognition and related control problems.

Introduction

In the statistical pattern recognition no really efficient algorithm is known for minimizing the misclassification probability over given finite dimensional decision functions, since the cost, viz. the misclassification probability, is usually not differentiable and has, except in some special cases, many local minima. Several authors investigated recursive algorithms which give a least mean square estimate of some Bayesian decision functions, e.g. the a posteriori probability function. For two hypotheses testing Wolverton, Wagner [9] and Csibi [5] proved that the square root of the mean square error gives bound of the difference between the error probability of Bayesian decision and the decision error probability connected with some estimates of Bayesian decision function.

The purpose of this paper is to improve this upper bound and to deal with its iterative minimization.

For multihypotheses testing we prove that the sum of mean distances gives an upper bound of error probabilities, a slight modification of which might be minimized on a linear space (e.g. under a dimensional constraint) by a version of stochastis approximation algorithm. This recursive adaptive rule is defined in a reproducing kernel Hilbert space, a specific choice of which is linear manifold defined by a finite set of linearly independent functions. This recursive finite dimensional case is usually applied in estimation and related topics in adaptive control and pattern recognition.

1. The error probabilities and the mean distances

The aim of the first part of this paper is to give some upper bounds of the error probabilities for multihypotheses testing problem when we are given only some approximations to Bayesian decision functions (e.g. a posteriori probability functions). The upper bound is the sum of the mean (L_1) distances of the Bayesian decision functions and their approximates.

Let $(\Omega, \mathcal{A}, P)$ be a probability space, $H_{ij}, H_2 \dots H_N$ a measurable partition of Ω (simple hypotheses [1]), and ξ a random variable which takes values in a measurable space $(x, \mathcal{X})$. The decision problem is to decide which of the hypotheses holds, by observing ξ.

Denote

$$D_{ij}(\xi) = P(H_i/\zeta) - P(H_j/\xi), \ (i,j = 1 \dots N) . \tag{1}$$

Lemma 1. Let $\{\tilde{A}_1, \tilde{A}_2 \dots \tilde{A}_N\}$ be an arbitrary measurable partition of X and denote by $P_e(\tilde{A}_1 \dots \tilde{A}_N)$ the error probability of the following decision: if $\xi \in A_i$ then declare H_i. Let $A_{ij} \dots A_N$ be the partition of X for Bayesian decision, viz.,

$$A_i = \{x; D_{ij}(x) \geq 0 \quad j \leq i, \quad D_{ij}(x) > 0 \quad j > i\}.$$

Then

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{B_{ij} \cup B_{ji}} |D_{ij}(\xi)| \, dP, \tag{2}$$

where

$$B_{ij} = \{\xi \in A_i \cap \tilde{A}_j\}, \qquad (i,j = 1 \dots N).$$

Proof. Since $\tilde{A}_1 \dots \tilde{A}_N$ is a partition of X

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) = \sum_{i=1}^{N} P(H_i^c, \ \xi \in \tilde{A}_i) = \sum_{i=1}^{N} \int_{\{\xi \in \tilde{A}_i\}} (1 - P(H)/\xi)) \, dP =$$

$$= 1 - \sum_{i=1}^{N} \int_{\{\xi \in \tilde{A}_i\}} P(H/\xi) \, dP, \tag{3}$$

consequently

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) = \sum_{i=1}^{N} \int_{\{\xi \in A_i\}} P(H_i/\xi) \, dP -$$

$$- \sum_{i=1}^{N} \int_{\{\xi \in \tilde{A}_i\}} P(H_i/\xi) \, dP. \tag{4}$$

Applying the facts

$$\{\xi \in A_i\} = \bigcup_{j=1}^{N} B_{ij}$$

and

$$\{\xi \in \tilde{A}_i\} = \bigcup_{j=1}^{N} B_{ji}.$$

(4) implies

$$\begin{aligned} P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) = \\ = \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{B_{ij}} P(H_i/\xi)\, dP - \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{B_{ij}} P(H_j/\xi)\, dP = \\ = \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{B_{ij}} (H_i/\xi)\, dP - \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{B_{ij}} P(H_j/\xi)\, dP = \\ = \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{B_{ij}} \big(P(H_i/\xi) - P(H_j/\xi)\big)\, dP = \sum_{i=1}^{N} \sum_{j=1}^{N} \int_{B_{ij}} D_{ij}(\xi)\, dP. \end{aligned} \tag{5}$$

Observe that $D_{ii}(\xi) \equiv 0$. $D_{ij}(\xi) = -D_{ji}(\xi)$ and $D_{ij}(\xi) \geq 0$ on B_{ij}, therefore from (5) we have

$$\begin{aligned} P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Big(\int_{B_{ij}} |D_{ij}(\xi)|\, dP + \int_{B_{ji}} |D_{ji}(\xi)|\, dP \Big) = \\ = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{B_{ij} \cup B_{ji}} |D_{ij}(\xi)|\, dP. \end{aligned} \tag{6}$$

Lemma 1 implies the following two theorems. If we give only some estimates of the Bayesian decision functions then these theorems consist of upper bounds for the error probabilities.

Theorem 1. Let $\{\tilde{D}_{ij}(x)\ i, j = 1 \dots N\}$ be a set of measurable functions on X such that $\tilde{D}_{ij}(x) = -\tilde{D}_{ji}(x)$ for each $i, j = 1 \dots N$ and put

$$\tilde{A}_i = \{x\,;\, \tilde{D}_{ij}(x) \geq 0 \quad j \leq i\,,\, \tilde{D}_{ij}(x) > 0, \quad j > i\},$$

then

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) \leq \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E\,|D_{ij}(\xi) - \tilde{D}_{ij}(\xi)|\,. \tag{7}$$

Proof. Let us use the result of Lemma 1

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{B_{ij} \cup B_{ji}} |D_{ij}(\xi)|\, dP. \tag{8}$$

$D_{ij}(\xi) \geq 0$ and $\tilde{D}_{ij}(\xi) \leq 0$ on B_{ij}; and $D_{ij}(\xi) \leq 0$ and $\tilde{D}_{ij}(\xi) \geq 0$ on B_{ji}; therefore $|D_{ij}(\xi)| \leq |D_{ij}(\xi) - D_{ij}(\xi) - \tilde{D}_{ij}(\xi)|$ on $B_{ij} \cup B_{ji}$ $(i, j = 1 \dots N)$, consequently from (8)

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) \leq \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{B_{ij} \cup B_{ji}} |D_{ij}(\xi) - \tilde{D}_{ij}(\xi)| \, dP \leq$$

$$\leq \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} E\,|D_{ij}(\xi) - \tilde{D}_{ij}(\xi)|. \tag{9}$$

Theorem 2. Let $\{\tilde{P}_i(x),\ i = 1 \dots N\}$ be a set of measurable functions on X and let

$$\tilde{A}_i = \{x\,;\, \tilde{P}_i(x) - \tilde{P}_j(x) \geq 0 \quad j \leq i, \quad \tilde{P}_i(x_j) - \tilde{P}_j(x) > 0 \quad j > i\}\,,$$

then

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) \leq \sum_{i=1}^{N} E\,|\,P(H_i/\xi) - \tilde{P}_i(\xi)\,|\,. \tag{10}$$

Proof. If $\tilde{D}_{ij}(x)$ denotes the function $\tilde{P}_i(x) - \tilde{P}_j(x)$ $(i, j = 1 \dots N)$ then we have in the same way as in the proof of Theorem 1 that

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) \leq \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{B_{ij} \cup B_{ji}} |\,D_{ij}(\xi) - \tilde{D}_{ij}(\xi)\,|\,dP =$$

$$= \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \int_{B_{ij} \cup B_{ji}} |\,P(H_i/\xi) - P(H_j/\xi) - \tilde{P}_i(\xi) + \tilde{P}_j(\xi)\,|\,dP \leq$$

$$\leq \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Big(\int_{B_{ij} \cup B_{ji}} |\,P(H_i/\xi) - \tilde{P}_i(\xi)\,|\,dP + \int_{B_{ij} \cup B_{ji}} |\,P(H_j/\xi) - \tilde{P}_j(\xi)\,|\,dP \Big) =$$

$$= \frac{1}{2} \Big[\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \int_{B_{ij} \cup B_{ji}} |P(H_i/\xi) - \tilde{P}_i(\xi)|\,dP + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \int_{B_{ij} \cup B_{ji}} |P(H_j/\zeta) - \tilde{P}_j(\xi)|\,dP \Big] =$$

$$= \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq 1}}^{N} \int_{B_{ij} \cup B_{ji}} |\,P(H_i/\xi) - \tilde{P}_i(\xi)\,|\,dP\,. \tag{11}$$

For each $i \neq k$, $i \neq l$, $l \neq k$ $B_{ik} \cup B_{ki}$ and $B_{il} \cup B_{li}$ are disjoint, therefore from (11)

$$P_e(\tilde{A}_1 \dots \tilde{A}_N) - P_e(A_1 \dots A_N) \leq \sum_{i=1}^{N} \int_{\bigcup_{\substack{j=1 \\ j \neq i}}^{N} (B_{ij} \cup B_{ji})} |\,P(H_i/\xi) - \tilde{P}_i(\xi)\,|\,dP \leq$$

$$\leq \sum_{i=1}^{N} E\,|\,P(H_i/\xi) - \tilde{P}_i(\xi)\,|\,. \tag{12}$$

If $\widetilde{P}_{ij}(x)$ is the difference $\widetilde{P}_i(x) - \widetilde{P}_j(x)$ $(i, j = 1 \ldots N)$, then, for $N = 2$, Theorem 2 follows from Theorem 1. (However, for larger N the upper bound in Theorem 2 may be much better than in Theorem 1.)

2. On minimizing the mean distance

The minimization of the classification error for the multiclass pattern classification problem is an unconvenient problem, since this cost (the misclassification probability) is usually not differentiable. However, the upper bound in Section 1 might be minimized over a linear space of functions.

Let α be an arbitrary function from the set $\{D_{ij}(x), P(H_i/x)\ \ i, j = 1 \ldots N\}$ and denote by Q the distribution of ξ, viz., by definition

$$Q(A) = P(\xi \in A), \qquad A \in \mathcal{X}. \tag{13}$$

We have to minimize the functional $\mathcal{J}(f) = E \mid f(\xi) - \mathrm{d}(\xi) \mid$ on a subspace of $L_1(X,Q)$ ($L_1(X,Q)$ is the Banach space of measurable functions $X \to R$ of finite integral with respect to Q). The corresponding subspace is usually a finite dimensional space.

For solving this problem we are given a labelled sample, a sequence of independent and identically distributed random variables $(\xi_1, \varrho_1), (\xi_2, \varrho_2) \ldots$ such that ξ_n is Q distributed, ϱ_n may take one of two or three possible values $(0, 1$ or $0, \pm 1)$ and

$$E(\varrho_n/\xi_n) = d(\xi_n), \qquad (n = 1, 2 \ldots). \tag{14}$$

The aim of this section is to propose a recursive algorithm, which minimizes the functional $\mathcal{J}^*(f) = E \mid f(\xi) - \varrho \mid$, where ϱ is a random variable such that

$$E(\varrho/\xi) = d(\xi) \tag{15}$$

and to investigate the relation of functionals $\mathcal{J}(f)$ and $\mathcal{J}^*(f)$.

Let H be an arbitrary Reproducing Kernel Hilbert Space (RKHS, [2]) of real valued measurable functions on X with the kernel $K(x,y)$. Let $\langle\ ,\ \rangle$ and $\|\ \|$ denote the inner product and the norm in H, respectively. Assume that

$$\int \sqrt{K(x, x)}\, Q(\mathrm{d}x) < +\infty \tag{16}$$

then for each $f \in H$

$$E \mid f(\xi) \mid = \int \mid f(x) \mid Q(\mathrm{d}x) = \int \mid \langle f,\ K(,x) \rangle \mid Q(\mathrm{d}x) \leq$$

$$\leq \| f \| \int \| KC\,.\,, x) \| Q(\mathrm{d}x) = \| f \| \int \sqrt{K(x, x)}\, Q(\mathrm{d}x) < +\infty \tag{17}$$

consequently H is a subset of $L_1(X, Q)$. Let $\bar{H}$ be the completion of H in $L_1(X, Q)$.

Theorem 3. Let f^* denote the element of $\bar{H}$ which minimizes the functional $J^*(f)$ on $\bar{H}$ (f^* exists and is unique). Let us assume that $K(x, y)$ is bounded Q a.s., viz.,

$$K(x, x) \leq L < +\infty \quad Q \text{ a.s.} \tag{18}$$

and suppose that $f^* \in H$. Consider the following recursive algorithm: $f_0 = 0$

$$f_{n+1} = f_n + \gamma_n \operatorname{sign}\left(\varrho_{n+1} - f(\xi_{n+1})\right) K(\,.\,, (\xi_{n+1}) \tag{19}$$

where $\gamma_n \geq 0$ is a sequence of reals satisfying the well-known condition in stochastic approximation

$$\sum_{n=0}^{\infty} \gamma_n = +\infty \qquad \sum_{n=0}^{\infty} \gamma_n^2 < +\infty . \tag{20}$$

Then

$$P\left(\lim_{n\to\infty} J^*(f) = J^*(f^*)\right) = 1 . \tag{21}$$

Proof. The proof is based on ideas due to Blum [10] and Aizerman, Braverman, Rosonoer [3]. Put $U_n = \| f^* - f_n \|^2$ then (19) implies that

$$\begin{aligned} U_{n+1} = U_n - 2\gamma_n \operatorname{sign}\left(\varrho_{n+1} - f_n(\xi_{n+1})\right) < f^* - f_n,\ K(.,\ \xi_{n+1}) > + \\ + \gamma_n^2 \| KC.,\ \zeta_{n+1}) \|^2 = U_n - 2\gamma_n \operatorname{sign}\left(\varrho_{n+1} - f_n(\xi_{n+1})\right)\left(f^*(\xi_{n+1}) - f_n(\xi_{n+1})\right) + \\ + \gamma_n^2 K(\xi_{n+1},\ \xi_{n+1}) . \end{aligned} \tag{22}$$

If $\mathcal{F}_n$ is the δ-algebra generated by (ξ_1, ϱ_1), $(\xi_2, \varrho_2) \ldots (\xi_n, \varrho_n)$ $(n = 1, 2 \ldots)$ then

$$\begin{aligned} E(U_{n+1}/\mathcal{F}_n) = U_n - 2\gamma_n E\left(\operatorname{sign}\left(\varrho_{n+1} - f_n(\xi_{n+1})\right)\left(f^*(\xi_{n+1}) - (f_n(\xi_{n+1}))/\mathcal{F}_n\right) + \\ + \gamma_n^2 \int K(x, x)\, Q(\mathrm{d}x) . \end{aligned} \tag{23}$$

It is easy to verify that

$$\begin{aligned} &\operatorname{sign}\left(\varrho_{n+1} - f_n(\xi_{n+1})\right)\left(f^*(\xi_{n+1}) - f_n(\xi_{n+1})\right) = \\ &= |\, \varrho_{n+1} - f_n(\xi_{n+1}) \,| - |\, \varrho_{n+1} - f^*(\xi_{n+1}) \,| + \\ &+ [\varrho_{n+1} - f^*(\xi_{n+1})]\left[\operatorname{sign}\left(\varrho_{n+1} - f^*(\xi_{n+1})\right) - \operatorname{sign}\left(\varrho_{n+1} - f_n(\xi_{n+1})\right)\right] \geq \\ &\geq |\, \varrho_{n+1} - f_n(\xi_{n+1}) \,| - |\, \varrho_{n+1} - f^*(\xi_{n+1}) \,| . \end{aligned} \tag{24}$$

Using (24) in (23)

$$E(U_{n+1}/F_n) \leq U_n - 2\gamma_n\left(J^*(f_n) - J^*(f^*)\right) + \gamma_n^2 \int K(x, x)\, Q(\mathrm{d}x) . \tag{25}$$

Because of (19) and (18)

$$J^*(f_{n+1}) \leq J^*(f_n) + \gamma_n L . \tag{26}$$

Thus (20), (25) and (26) imply (applying [3], Theorem II, p. 164) that

$$P\left(\lim_{n\to\infty} \mathfrak{J}^*(f_n) = \mathfrak{J}^*(f^*)\right) = 1 . \tag{27}$$

Corollary. Let $\{\varphi_1 \dots \varphi_K\}$ be a set of linearly independent bounded measurable functions on X. $K(x, y) = \sum_{i=1}^{K} \varphi_i(x)\, \varphi_i(y)$,

$$H = \left\{ \sum_{i=1}^{K} c_i \varphi_i , \quad c_i \text{ reals} \quad (i = 1 \dots k) \right\} \tag{28}$$

then $H = \overline{H}$, consequently the result of Theorem 3 might be used.

For arbitrary measurable function f we are interested in the difference of $\mathfrak{J}^*(f)$ and $\mathfrak{J}(f)$.

Lemma 2. For each $f \in L_1(X, Q)$

$$0 \leq \mathfrak{J}^*(f) - \mathfrak{J}(f) \leq E \mid \varrho - E(\varrho/\xi) \mid . \tag{29}$$

Proof. By definition

$$\mathfrak{J}^*(f) - \mathfrak{J}(f) = E \mid f(\xi) - \varrho \mid - E \mid f(\xi) - E(\varrho/\xi) \mid . \tag{30}$$

Using the properties of conditional expectation

$$\begin{aligned} \mathfrak{J}(f) = E \mid f(\xi) - E(\varrho/\xi) = E \mid E\big(f(\xi) - \varrho/\xi\big) \mid \leq \\ \leq E\big(E(\mid f(\xi) - \varrho \mid /\xi)\big) = E \mid f(\xi) - \varrho \mid = \mathfrak{J}^*(f), \end{aligned} \tag{31}$$

which gives the lower bound in (29). The upper bound in (29) follows from (30) applying the triangle inequality.

Lemma 3. If f' denotes the function which minimizes $\mathfrak{J}$ on $\overline{H}$ then

$$\mathfrak{J}(f^*) - \mathfrak{J}(f') \leq E \mid \varrho - E(\varrho/\xi) \mid . \tag{32}$$

Proof. Because of (29)

$$\mathfrak{J}(f^*) \leq \mathfrak{J}^*(f^*) . \tag{33}$$

f^* minimizes the functional $\mathfrak{J}^*$ on $\overline{H}$, consequently

$$\mathfrak{J}^*(f^*) \leq \mathfrak{J}^*(f') . \tag{34}$$

Let us use (33), (34) and (29) for $f = f'$ then

$$\mathfrak{J}(f^*) - \mathfrak{J}(f') \leq \mathfrak{J}^*(f') - \mathfrak{J}(f') \leq E \mid \varrho - E(\varrho/\xi) \mid . \tag{35}$$

Conclusion

The mean distance upper bound one may improve the mean square distance upper bound of the difference of error probabilities and still remain within the scope of simple computations. We also generalized the study to multihypotheses testing. Observe, however, in section 2 of this paper there are still some open problems, viz.,

(i) how to minimize directly the functional $\mathfrak{J}(f)$ for given labelled samples?

(ii) how to give a better upper bound of difference $\mathfrak{J}(f^*) - \mathfrak{J}(f')$ than in Lemma 3.

References

1. *Lehmann, E. L.*, Testing Statistical Hypotheses. John Wiley & Sons Inc., New York, 1959.
2. *Aronszajn, N.*, Theory of reproducing kernels, Trans. Amer. Math. Soc., **68**, 1950.
3. *Aizerman, M. A., Braverman, E. M., Rosonoer, L. I.*, The Method of Potential Function in the Theory of Machine Learning. Nauka, Moscow, 1970 (in Russian).
4. *Dvoretzky, A* , On stochastic approximation, Proc. 3rd Berkeley Symp. on Math. Statist. and Probability, Vol. **1**, 1956.
5. *Csibi, S.*, "Simple and Compound Processes in Iterative Machine Learning" CISM Summer Course, Udine, Italy, 1971.
6. *Csibi, S.*, "Stability and Complexity of Learning Algorithms" CISM Summer Course, Udine, Italy, 1972.
7. *Fritz, J., Györfi, L.*, On the minimization of classification error probability in statistical pattern recognition, Submitted to the IEEE Trans. on Information Theory, 1974.
8. *Györfi, L.*, Convergence of potential function type learning algorithms, Problems of Control and Information Theory, **1**, 1972.
9. *Wolverton, C. T., Wagner, T. J.*, Asymptotically optimal discriminant functions for pattern classification, IEEE Trans. on Information Theory, **15**, 1969.
10. *Blum, I. R.*, Multidimensional stochastic approximation methods, Ann. Math. Statistics, **25**, 1954.

Верхняя граница вероятности ошибки при исследовании с несколькими гипотезами и применение результатов для адаптивного распознавания образов

Л. ДЕРФИ

(Будапешт)

Соотношения среднеквадратичной верхней границы разности между вероятностью байесового ошибочного решения и вероятностью ошибочного решения при оценке байесовых решающих функций хорошо известны для проблемы исследования двух гипотез. Также доказана сходимость алгоритма среднеквадратичной минимизации с использованием адресованных, слабо зависимых данных проб. В настоящей работе определяется скорректированная верхняя граница, обеспечиваемая средними дистанциями байесовых решающих функций и их оценками, полученными относительно исследований проблемы с несколькими

гипотезами. Слабая модификация разработанной верхней границы минимизируется в пространстве заданной, ограниченной размерности решающих функций. Предлагаемый алгоритм представляет собой рекуррентный адаптивный алгоритм, некоторый вариант метода стохастической аппроксимации, особенно приемлимый для решения задач в области статистического распознавания образов и соответствующих проблем управления.

L. Györfi
HEI, Technical University of Budapest
H-1111 Budapest, XI. Stoczek u. 2.
Hungary

[*Problems of Control and Information Theory, Vol. (5, 6), pp. 459—471 (1976)*

COMPUTERIZED RELIABILITY STUDY OF THYRISTORIZED CONTINUOUS POWER SUPPLY SYSTEMS

A. KÁRPÁTI, E. SZENTAI, I. IPSITS, I. HERMANN

(Received October 20, 1975)

1. Introduction

The thyristorized forced-commutation current commutators are being applied ever more extensively for the continuous power supply of consumers, whose operation must be ensured even during the failure of the a. c. voltage mains. The solutions may be classified in two groups. In the first case the load is supplied in the normal state by the mains, the inverter is switched on only when it fails. In the second case the load is supplied constantly from the inverter and only the power storage battery is charged from the mains.

The thyristorized solution shows advantages and disadvantages, as compared with other emergency power supply sources. Its one disadvantage is its relatively high cost. Its advantages are that it requires no special operating personnel, its automatic starting may be solved easily, its failure probability and maintenance requirements are lower than those of other solutions.

The supply failure probability may be reduced additionally by the use of a system with reserve. In this case if one inverter fails, the second one takes over automatically the supply of the load. By reducing the period of repair of the damaged inverter the reliability of the resultant system may be increased considerably.

The redundant system may be realized in two basic forms. Figure 1 shows the functional scheme of one of the solutions, with two inverters of identical capacity serving for the supply of the load. One is supplying the load, the second one is in the idle running state of readiness. The failure of the power supply is sensed and the switch-over is effected by separate automatics (cold reserve).

In the second solution the load is supplied by both inverters connected in parallel. Figure 2 shows the layout scheme of this arrangement. The selective switch-off must be cared for by a corresponding protective arrangement (hot redundant system). The paper deals with the study of a system of this type.

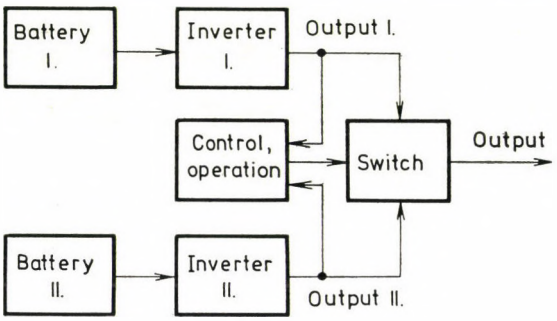


Fig. 1

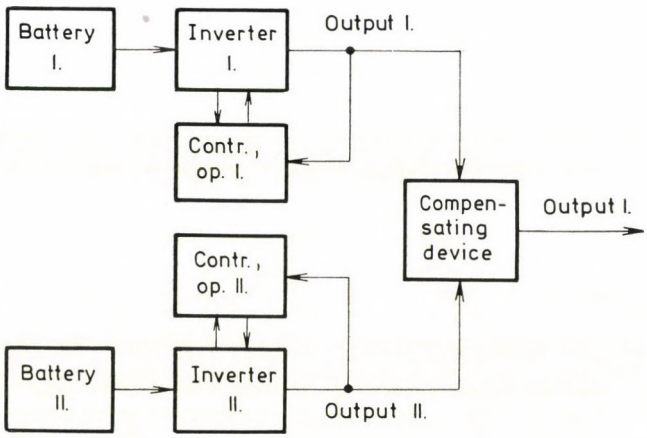


Fig. 2

2. Short summary of the basic assumptions and calculation methods used for the reliability analysis

The reliability of the studied systems is characterized by their MTBF value. The calculations are performed by the probability calculus methods in the knowledge of the failure ratios of the individual parts, assuming that the failure ratios are constant in time. The reliability analysis of the inverter systems results even with the application of simplyfying assumptions in multi-state systems, whose behaviour may be studied by the theory of the homogeneous Markov-processes of discrete values. The tasks to be solved during the reliability analysis are as follows:

a) The determination of the failure ratios of the individual elements. The failure ratios are functions of several factors. Their dependences are contained in catalogues, or may be obtained by characteristic curves found

in the literature, or by approximation formulae respectively. A problem in the analysis of inverters is the determination of the equivalent load due to the specific load of the elements.

b) Delimiting the failure classes of the system. (Elements, whose failure leads to identical effects, belong to the same failure class.)

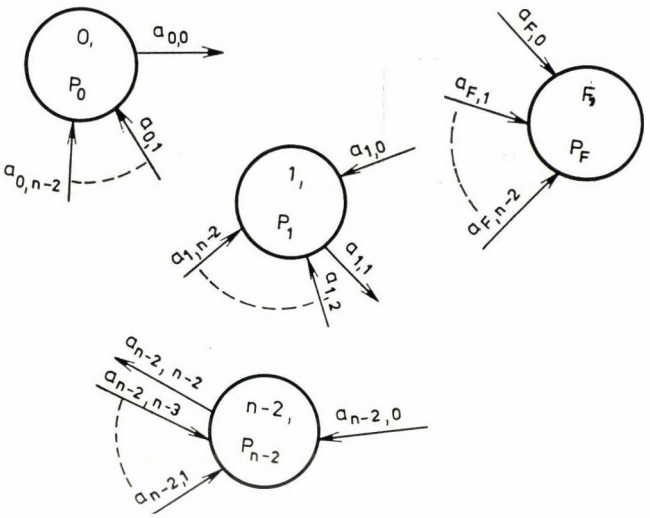


Fig. 3

c) Plotting the state graph of the system, based on the fundamental failure states corresponding to the individual failure classes and their combinations causing no complete failure.

d) Determination of the MTBF value of the system in the knowledge of the state graph. The MTBF value for the graph of arbitrary n states shown in Fig. 3 (0 = faultless system, F = failed system) is calculated by the following formula according to the literature:

$$\mathrm{MTBF} = -\frac{d}{ds}\left(s.\, P_F(s)\right)\Big|_{s=0} \tag{1}$$

where $P_F(s)$ is the Laplace transform of the probability function $P_F(t)$ of reaching the failed state. The formula may be evaluated also manually, but the determination of the function $P_F(s)$ is very labourious even for systems with 4–5 states. The variation of the formula suitable for computer programming is given in the Appendix.

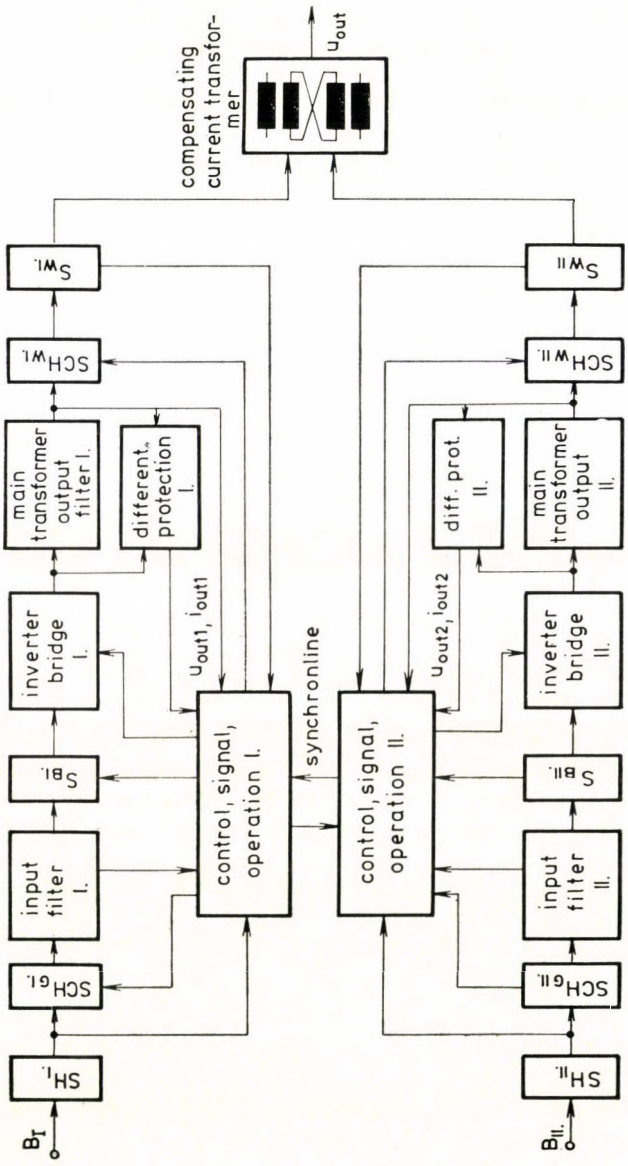

compensating current transformer
u_{out}
$S_{WI.}$
$SCH_{WI.}$
main transformer output filter I.
different. protection I.
inverter bridge I.
$S_{BI.}$
input filter I.
$SCH_{GI.}$
$SH_{I.}$
B_I
u_{out1}, i_{out1}
control, signal, operation I.
synchronline
control, signal, operation II.
u_{out2}, i_{out2}
diff. prot. II.
$S_{WII.}$
$SCH_{WII.}$
main transformer output II.
inverter bridge II.
$S_{BII.}$
input filter II.
$SCH_{GII.}$
$SH_{II.}$
$B_{II.}$

Fig. 4a

3. Reliability analysis of a hot-redundant emergency power supply system

In what follows a practically realized hot-redundant emergency power supply system is analysed with respect to its reliability. The system consists of two identical capacity inverters, each of which is capable to supply the load individually. The inverters are connected in parallel on the a.c. voltage side, as shown in the simplified functional scheme of Fig. 2. The characteristic data of one inverter are: $U_{in} = 220$ V d.c.; $U_{out} = 220$ V a.c.; $P_{out} = 6.3$ kVA; $f = 50$ c/s; $\cos \varphi = 0.9$ (capacitive) . . . 0.7 (inductive).

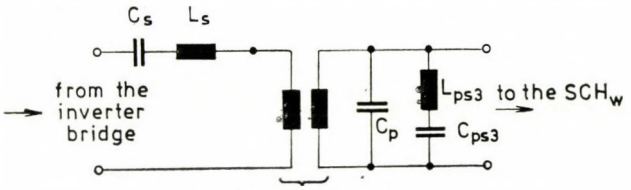
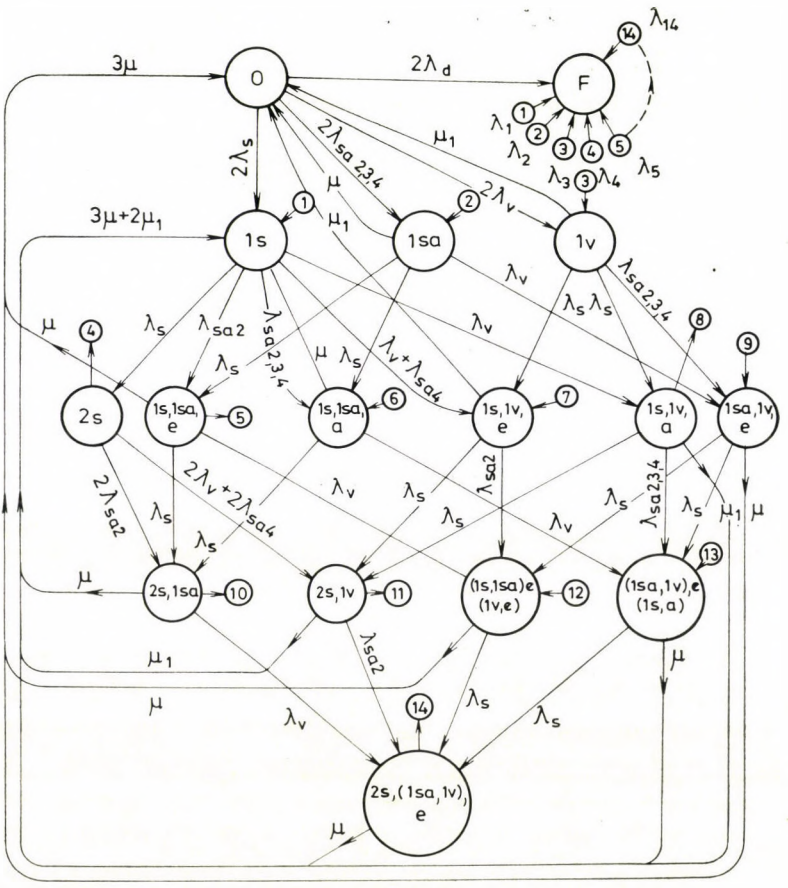


Fig. 4b

The detailed construction of the equipment is shown schematically in Fig. 4. Its operation based on this scheme may be described as follows:

The voltage of the battery is applied to the input of the bridge-connected single-phase inverter through the main switch (S_H), the d.c. side magnetic switch (SCH_G), the input filter circuit and the semiconductor fast fuse (S_B). The bridge output supplies the filter circuit transformer unit shown in Fig. 4. The filtered a.c. voltage is applied to the terminals of the compensating current transformer through the output magnetic switch (SCH_W) and the fuse (S_W).

The output voltage of the inverter is controlled by the shifted ignition of the individual bridge-halves. A signal proportional with the output voltage and the output current is fed back into the control circuit, so the individual inverters have in the current range O . . . I_N a voltage-maintaining, and over this range a current limiting characteristic curve. The main circuit of the inverter is protected by the fuses S_B and S_W with melting indication circuits connected to their terminals. If any fuse melts, the control circuits actuated by the melt-indicating circuits disconnect the failed inverter potentially from the load and the battery with the help of the magnetic switches SCH_G and SCH_W. The co-running of the rate-transmitting oscillators of both inverters is ensured by the rigid connection of their outputs. The detailed investigation of the system circuits results in the simplified graph shown in Fig. 5a, characterizing the failure relations. From among the failure

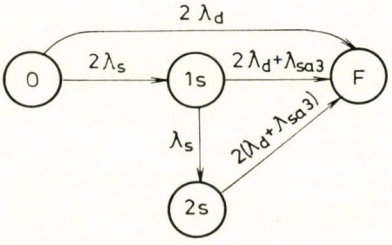


Event density to the ① ... ⑭ outputs

$\lambda_1 = 2\lambda_d + \lambda_{sa3}$; $\lambda_4 = 2(\lambda_d + \lambda_{sa3})$;

$\lambda_2 = \lambda_3 = \lambda_5 = \ldots = \lambda_{14} = 2\lambda_d + \lambda_v + \lambda_{sa\,2,3,4}$;

a) in irradiation : $\lambda_{sa\,2,3,4} = \lambda_{sa2} + \lambda_{sa3} + \lambda_{sa4}$;



b)

Fig. 5

states designated by the nodal points of the graphs the following are fundamental:

0 — faultless system,
1_s — the selectivity separating circuit of one inverter is damaged (the potential separation of both inverters is not possible),
2_s — the selectively separating circuits of both inverters are damaged,
1_v — one inverter has a failure, making it unsuitable for supplying the load, but the protection fails to sense it, so the system is not separated, there is no failure signal; (a characteristic failure of this kind is e.g. the break damage in the main circuit, or a break in the voltage feedback circuit, causing the output voltage of the faulty inverter to rise, whereby the inverter goes over into the current limiting mode of operation),
1_{sa} — failure in a group of elements, whereby the faulty unit is selectively separated, if no 1_s state has occurred before,
F — failure of the whole system.

The other states (graph points) are combinations of the states listed above, which do not lead to the failure of the complete system.

According to the state graph 2-, 3- and 4-fold variations may also occur (elements of several groups may fail) without the failure of the whole system. Here the cases must be distinguished, where various element groups fail in the same inverter, or in two of them, respectively. This distinction is made by the letters a and e beside the symbols of the states. The letter e refers to the failure of two or more element groups in the same inverter, the letter a has the opposite meaning. The new states appearing as variations of the basic states are designated by separating the participating basic states as listed by commas and designated also the occurrence of the failures in the same inverter, or in different inverters.

The system presented in the drawing is simplified for the sake of clarity. The element group causing selective failure must be subdivided namely into three more groups of elements, as if the selectively separating circuit is the one which fails then the effect of the failures leading originally to the selective drop out varies. A subgroup of the elements continues to cause selective failure, the second subgroup leads to the failure of the whole system and the effect of the third is similar to the hidden failures not observed by the protections. Here we did not introduce new states, only additional transitions to those corresponding to the main states defined above.

Failure ratios required for the interpretation of the transitions between the points of the graph:

λ_s^{μ} — resultant failure ratio of the selectively separating circuit,

λ_d — resultant failure ratio of the elements causing the failure of the whole system (the failures belonging here are in great lines those of the master oscillator and the part of the control circuit failures causing the output voltage drop by the inverter, as the output voltage of the whole system is determined always by the unit of the lower voltage),

λ_v — resultant failure ratio of the elements causing hidden failures,

λ_{sa2}, λ_{sa3}, λ_{sa4} — resultant failure of the elements causing selective failure, complete system drop out and transducing the system into the state of the hidden failures respectively.

For studying the effects of the hidden failures we introduced two types of the average correction period:

$\tau_2 = 1/\mu$ — correction period of the designated failures,

$\tau_1 = 1/\mu_1$ — correction period of the hidden failures (The hidden failures may be established by the regular control of the inverter, from the instrument readings).

As the results obtained during the calculations showed that the transitions between the individual states are — with a good approximation — independent of the load of the inverter, we have neglected the effect of the load variations on the failure of an inverter (this may be explained by the fact that for attaining a high reliability the inverter elements were greatly underloaded).

We note that in the case of *lv* type failures the compensating current transformer is unable to perform its function any more, so the distortion of the output voltage increases. In our case this distortion was still within the permitted limits, so we did not take this effect as a direct failure into account. We assumed that the protections are selective for the internal shorthings.

The maximum MTBF value attainable by the system is obtained with the assumption of zero correction time. In this case the graph describing the system is greatly simplified and assumes the form shown in Fig. 5a. After performing the calculation on this basis the maximum MTBF value may be obtained the following formula:

$$\mathrm{MTBE}_{\max} = \frac{2\lambda_d^2 + \lambda_s^2 + \lambda_{sa3}^2 + 3(\lambda_d \lambda_s + \lambda_d \lambda_{sa3} + \lambda_s \lambda_{sa3})}{2(\lambda_d + \lambda_s)(\lambda_d + \lambda_{sa3})(2\lambda_d + \lambda_s + \lambda_{sa3})}. \qquad (2)$$

The effect of the correction period was studied by a computer because of the complexity of the system. The derivation of the algorithm used for the calculation is contained in the Appendix (clause 5). The results of the calculations are summarized in Fig. 6.

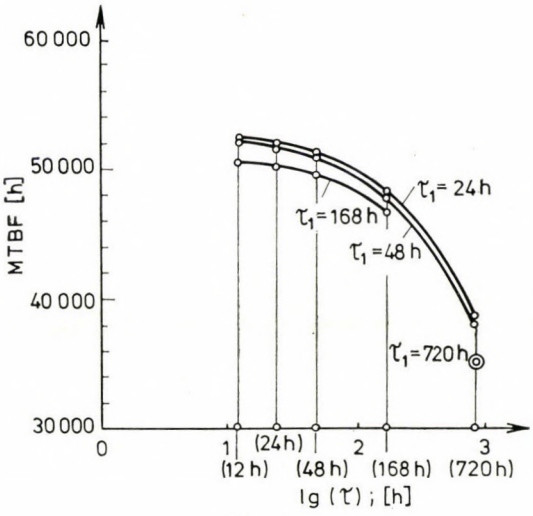


Fig. 6

4. Conclusions

On the basis of the system-technical studies and the calculations the following conclusions may be drawn:

a) The MTBF value of the system, assuming correction periods of some days, does not vary excessively, i.e. under such correction periods we are near to the attainable maximum MTBF value. Here the MTBF value is practically limited by the failure ratio λ_d causing the failure of the whole system. It is worthwhile to consider alterations in the construction in order to reduce this limitation (e.g. using one or two master oscillators respectively, to eliminate the effects causing direct failure from among those of the control circuit. In the studied case especially the latter proved to be an important factor).

b) It is advised to supply the inverter with a protection permitting to detect also the hidden failures and the disconnection of the system.

c) Our basic assumption was that on the internal failure of a shorting character of one inverter the fuses of both inverters perform selective switching off. The fulfilment of this must be very carefully investigated.

d) An interesting result, which was obtained, is that the MTBF value of the system is greatly independent of the load. The explanation of this phenomenon was already considered in the preceding clause.

5. Appendix (Description of the algorithm applicable for the evaluation also by a digital computer)

As mentioned above, the studies may be performed by the theory of the discrete homogeneous Markov-processes. Figure 3 shows the state graph of such an n-state system and the transitions between the individual states. According to the literature the behaviour of such a system may be described by the following differential equation:

$$\begin{aligned} P'_0 &= \sum_{i=0}^{F=n-1} a_{0,i} \cdot P_i \\ P'_1 &= \sum_{i=0}^{F=n-1} a_{1,i} \cdot P_i \\ &\cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \\ P'_{n-2} &= \sum_{i=0}^{F=n-1} a_{n-2,i} \cdot P_i \\ P'_F &= \sum_{i=0}^{F=n-1} a_{F,i} \cdot P_i \end{aligned} \tag{3}$$

where P_i — probability function of attaining the ith state,
a_{ij} — transition containing the effect of the jth point of the graph on the ith point of the graph (event density).

By introducing the following designations:

$$P = \begin{bmatrix} P_0 \\ P_1 \\ \cdot \\ \cdot \\ \cdot \\ P_{n-2} \\ P_F \end{bmatrix}; \mathbf{A} = \begin{bmatrix} a_{0.0} & a_{0,1} & \cdots & a_{0,n-2} & a_{0,4} \\ a_{1,0} & a_{1,1} & & a_{1,n-2} & a_{1,4} \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ a_{n-2,0} & a_{n-2,1} & \cdots & a_{n-2,n-2} & a_{n-2,4} \\ a_{F,0} & a_{F,1} & \cdots & a_{F,n-2} & a_{F,4} \end{bmatrix}$$

the system of equations (3) may be described in the following simpler form:

$$P' = \mathbf{A} \cdot P \tag{4}$$

By performing the Laplace transformation of Eq. (4) the following relationship is obtained:

$$s \cdot P(s) - e_1 = \mathbf{A} \cdot P(s) \tag{5}$$

where $e_1 = \begin{bmatrix} 1 \\ . \\ . \\ . \\ 0 \\ 0 \end{bmatrix}$ the vector of the initial values.

After arranging the equation the vector $P(s)$ is obtained in the form:

$$P(s) = \mathbf{B}^{-1} \cdot e_1 \tag{6}$$

where

$$\mathbf{B} = s, \mathbf{E} - \mathbf{A} \text{ and } \mathbf{E} \text{ is the unit matrix.}$$

For determining the MTBF value of the system by Eq. (1) the knowledge of $P_F(s)$ is necessary. This may be obtained from the vector $P(s)$ by the following multiplication:

$$P_F(s) = e_F^T \cdot P(s) = e_F^T \cdot \mathbf{B}^{-1} \cdot e_1 \tag{7}$$

where

$$e_F^T = [0,0, \ldots\ldots, 0,1].$$

Taking formula (1) into account the MTBF value of the system may be calculated in the following way:

$$\text{MTBF} = -\frac{d}{ds}(s \cdot e_F^T \cdot \mathbf{B}^{-1} \cdot e_1)\Big|_{s=0}. \tag{8}$$

After performing the matrix operations the result may be written in the following form:

$$\text{MTBF} = -\frac{d}{ds}(s \cdot \mathbf{B}_{n,1}^{-1})\Big|_{s=0} \tag{9}$$

where $\mathbf{B}_{n,1}^{-1}$ is the element $[n,1]$ of the matrix $\mathbf{B}$.

In what follows we consider the production of the inverse matrix element $\mathbf{B}_{n,1}^{-1}$.

By introducing the conception of the adjucnt matrix the inverse of matrix $\mathbf{B}$ is in the following form:

$$\mathbf{B}^{-1} = \text{adj } (\mathbf{B})/\det (\mathbf{B}) \tag{10}$$

or correspondingly as:

$$\mathbf{B}_{n,1}^{-1} = \text{adj } (\mathbf{B}_{n,1})/\det (\mathbf{B}). \tag{11}$$

Observing the definition of the adjunct matrix, the properties of matrix **A** defined by Eq. (4) and those of matrix **B** defined by Eq. (6) (for **A** it is true that $a_{0,F} = a_{1,F} \ldots = a_{F,F} = 0$) the matrix element $\mathbf{B}_{n,1}^{-1}$ may be written in the following form:

$$s \cdot \mathbf{B}_{n,1}^{-1} = (-1)^{n+1} \det\left(\mathbf{B}_{f,j}(s)\right)/\det\left(\mathbf{B}_{a,j}(s)\right) \tag{12}$$

where

$\mathbf{B}_{f,j}$ is the submatrix formed from matrix **B** by omitting its first row and its last column, respectively,

$\mathbf{B}_{a,j}$ is a subdeterminant formed from matrix **B** by omitting its last row and its last column.

By substituting Eq. (12) into Eq. (9) and performing the differentiation according to s the result is:

$$\mathrm{MTBF} = -(-1)^{(n+1)} \frac{\det'(\mathbf{B}_{f,j}(\cdot \det(\mathbf{B}_{a,j}) - \det'(\mathbf{B}_{a,j}) \cdot \det(\mathbf{B}_{f,j})}{s \to 0 \det^2(\mathbf{B}_{a,j})} \tag{13}$$

where the upper comma means the differential quotient according to s of the involved determinant.

Observing the properties of the determinants the differential quotient of the determinant may be produced in the form of a finite sum by differentiating the element of the involved matrix by rows and summing up the determinants of the partial matrices obtained by the row-by-row differentiation. This method may be easily algorithmized, if we consider the properties of the matrices to be evaluated. The limit transition $s \to 0$ may be performed simply by the substitution of $s = 0$. In the derivation care was taken to ensure that no 0/0 type result is obtained on calculating the MTBF value.

Summary

The paper deals with the reliability analysis of the thyristorized continuous power supply system; sums up in short the fundamental solution possibilities and the tasks to be solved during the reliability analysis; performs the study by utilizing the theory of the discrete-value homogeneous Markov-processes; gives the failure state graph of a hot-redundant continuous power supply system and describes the calculation results of an actual system; finally it presents an algorithm for the computer determination of the MTBF value of an arbitrary n-state system.

References

1. *Rényi, A.*, Valószínűségszámítás (Probability calculus) Tankönyvkiadó, Budapest, 1954.
2. *Szantai, E.*, Tartalékolt rendszerek megbízhatóságának számítása a felújítás figyelembevételével (Reliability calculation of redundant systems with consideration to the renewal). MTA—AKI Közlemények, 1966.

MTA Könyvtára
Példány 812 /19 77

Анализ надежности бесперебойного источника энергии на тиристорах с помощью вычислительной машины

А. КАРПАТИ, Е. СЕНТАИ, И. ИПШИЧ, И. ХЕРМАНН

(Будапешт)

Рассматриваются вопросы надежности бесперебойного источника энергии на базе тиристоров. Кратко суммируются основные возможности решений и задачи, подлежащие решению по ходу анализа надежности.

Исследования базируются на применении теории однородных Марковских процессов с дискретными значениями. Дается граф состояний отказов системы бесперебойного источника энергии с горячим резервом, а далее показаны результаты вычислений, проведенных для одной конкретной системы. В заключение показан алгоритм для определения с помощью ЭЦВМ значения СВНО системы с произвольными n состояниями.

A. Kárpáti,
E. Szentai,
I. Ipsits,
I. Hermann
Department of Automation
Technical University of Budapest
H-1111 Budapest, XI. Goldmann Gy. tér 3

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

COMMENTS ON THE THEORY OF RELAY-LINEAR CONTROL LAWS

K. A. PUPKOV and A. I. MOROZ

(USSR)

Presently, the linear theory of dynamic system stabilization is fairly well elaborated [1]. However, nonlinearities inherent to the physical systems make conclusions that the linear theory is not always valid. Saturation in the control system actuator is a typical nonlinearity of this sort. Formal approach to the synthesis of the control law requires allowance for constraints on the value of the control action, thus resulting in an essentially more complicated problem. The dynamic programming opens a general way to its solution. Yet experience gained by now shows that in this case solution of the Bellman equation is exclusively difficult. The "dimensionality curse" is still insuperable.

The present paper is devoted to the development of the theory of a special kind of relay-linear control laws (relions) which are linear within a fixed vicinity of the state space origin and relay beyond it. They are defined by piecewise-linear switching surfaces approximating switching surfaces of the time-optimal control law. Therefore, synthesis of this law may allow for constraints on the control. Introduction of the linear zone improves transients in the system owing to the elimination of self-oscillating and sliding regimes which always arise in purely systems relay in the vicinity of the origin. Under large deviations of the state from the origin, the relion manifests merits of the time-optimal relay control law by providing maximal-intensity control action transferring the system to the origin vicinity at minimal possible time, the closed system having stability region maximal in terms of constraints on the control action. In the linear zone it provides smooth damping of transient processes. Being composed of relay and linear operations only, the relion may be readily implemented by standard components.

1. Problem formulation

Consider a process described by equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = A(t)\,x + \sum_{i=1}^{m} b_i(t)\,u_i + f(t) \tag{1.1}$$

where x is an n-dimensional vector, $A(t)$ is a matrix; $b_i(t)$ is an n-dimensional vector; $f(t)$ is a known additive vector-function; and u_i is control. The region of allowable values of the control u_i is given by the condition

$$|\,u_i\,| \leq 1\,. \tag{1.2}$$

Controls $u(\tau) = \mathrm{col}\,(u_1(\tau), u_2(\tau), \ldots, u_m(\tau))$ which are measurable and satisfy the constraint

$$\operatorname*{vrai\,max}_{[t_0, t_1]} \max_{1 \leq i \leq m} \{|\,u_i(\tau)\,|\} \leq 1 \tag{1.3}$$

will be regarded as allowable. Let $\Phi(t, \tau)$ be the fundamental matrix of the system (1.1), and $B(t)$ be a matrix whose ith column is $b_i(t)$. The system (1.1) is referred to as nonsingular over interval $[t_0, t_1]$ if, for any nonzero vector l, the vector function $l^*\Phi(t_1, \tau)\, B(\tau)$ vanishes over $[t_0, t_1]$ only on the Lebesgue set of zero measure. The system (1.1) is referred to as nonsingular over $[t_0, t_1]$ with respect to control u_α if the system $\dfrac{dx}{dt} = A(t)\, x + b_\alpha(t)\, u_\alpha$ is nonsingular. The system (1.1) is referred to as t_0-differentially controllable if it is nonsingular over $[t_0, \infty)$. Such a system is apparently controllable [2] over any interval $[\tau_0, \tau_1] \subset [t_0, \infty]$.

Let $G(t_0, t_0 + \tau)$ denote a set of points $\hat{x}$ so that over the interval $[t_0, t_0 + r]$ there is an allowable control $u(t)$ for each point such that the solution $x(t)$ of (1.1) corresponding to $u = u(t)$ satisfies the boundary conditions $x(t_0) = \hat{x}$ and $x(t_0 + r) = 0$. The set $G(t_0) = \bigcup_{r=0}^{\infty} G(t_0, t_0 + \tau)$ will be referred to as the domain of t_0-controllability of the system (1.1).

The relion will be constructed as an approximation of the time-optimal control law. Therefore, consider, first the following problem: for each $t \in [a_0, a_1)$ and $x \in G(t)$ determine a vector-function $F^0(t, x)$ satisfying the constraint $|F_i^0(t, x)| \leq 1$, $i = 1, \ldots, m$, such that for any $t_0 \in [a_0, a_1]$ and $\hat{x} \in G(t_0)$ there is $T^0(\hat{x})$ and a solution $\hat{x}(t)$ of the equation

$$\frac{dx}{dt} = A(t)x + \sum_{i=1}^{m} b_i(t)\, F_i^0(t, x) + f(t) \tag{1.4}$$

satisfying boundary conditions $\hat{x}(t_0) = \hat{x}$ and $\hat{x}(t_0 + T_0(\hat{x})) = 0$, over the interval $[t_0, t_0 + T^0(\hat{x})]$, the inequality $T \geq T^0(\hat{x})$ being satisfied for any allowable control $u(t)$ transferring $\hat{x}$ to θ in the corresponding interval $[t_0, t_0 + T]$.

2. Synthesis of the time-optimal conrol law

It may be shown that if the system (1.1) is t_0-differentially controllable with respect to each component u_i of control u, the latter is defined uniquely on the basis of Pontryagin's maximum principle. Like in [3, p. 127], one can show in this case that the time-optimal programmed control $u^0(t)$ is unique. In the problem under consideration, for any pair (t_0, x^0) of initial conditions where $x^0 \in G(t_0)$, there is a time-optimal control [2]. If, in addition, values of function $f(t)$ for nearly all t belong to the interior of the set $B(t)U$, the maximum principle is a sufficient condition of time-optimality. By virtue of existence and uniqueness of optimal paths $x^0(\tau)$ ending in the origin,

$F^0(t,x)$ is determined as follows. Fix $\hat{x} \in G(t)$, and let $u^0(\tau)$ be time-optimal control such that its corresponding optimal path $\hat{x}^0(\tau)$ satisfies conditions $\hat{x}^0(t) = \hat{x}$ and $\hat{x}^0(T) = 0$ under some $T > 0$. Then give $F^0(t, \hat{x}) = u^0(t)$. According to the aforementioned observations, such a representation will be unique for each pair (t,x) if at this time t the corresponding optimal control $u^0(\tau)$ is defined uniquely. This is, in principle, a solution of the control law synthesis problem. Such a solution is, however, mathematically abstract, and it is still not clear how it can be applied to the optimal control where the value of $F^0(t,x)$ should be determined corresponding to a point x given at time t. Therefore, analyse the law $F^0(t,x)$ in more detail.

In virtue of the maximum principle, each component $u_i^0(\tau)$ of the optimal programmed control for each τ will be either $u_i = 1$ or $u_i = -1$, or control will change its sign in this point. Therefore, if at the time $\tau = t$ the optimal path goes through point x, the corresponding value of the optimal control $F^0(t,x)$ will be either equal to one of the boundary values, or control will change its sign in point x at time t while moving along the optimal path. Consequently, for fixed t, the optimal synthesizing $F^0(t, x)$ is piecewise constant with respect to x, the set $G(t)$ being broken with respect to each component u_i into two disjoint sets. In one of these sets, $G_i^+(t)$, $F_i^0(t, x) = 1$ holds at time t, and in the other, $G_i^-(t)\, t, x) = 1$ holds. Consider the set of points where $G_i^+(t)$ contacts $G_i^-(t)$.

Consider a system conjugate with (1.1):

$$\frac{\mathrm{d}\psi}{\mathrm{d}t} = -A^*(t)\,\psi \tag{2.1}$$

x will be said to belong to the switching surface $\Omega_i(t)$ of control u_i (surface corresponding to time t) if one can indicate times t_0 and t_1 such that $t_0 \leq t < t_1$, and over $[t_0, t_1]$ $\psi(\tau)$ is a non-trivial solution of Eq. (2.1) such that its corresponding (in virtue of the maximum principle) control $u^0(\tau)$ generates over $[t_0, t_1)$ the optimal path $x^0(\tau)$ of Eq. (1) satisfying $x^0(t) = x$, $x^0(t_1) = 0$, while $\psi^*(t)\, b_i(t) = 0$ is satisfied for $\tau = t$.

This definition separates in $G(t)$ those points where the ith component u_i of the control vector changes its sign while moving along the optimal path because, as it follows from the maximum principle,

$$u_i^0(\tau) = \operatorname{sign} \psi^*(\tau)\, b_i(\tau), \quad \tau \in [t_0, t_1]\,.$$

Note that assignment of the function $F_i^0(t, x)$ over the points of the set $\Omega_i(t)$ will not be unique in virtue of the rule formulated above because at time t u_i is not uniquely defined on the basis of the maximum principle. As it

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

will be demonstrated below, if a solution of the optimal closed system (1.4) is considered in A. F. Filippov's generalized sense [4], $F_i^0(t, x)$ may be assigned in points of $\Omega_i(t)$ arbitrarily. At the same time, $F_0^i(t, x)$ may be defined in $\Omega_i(t)$ so that there will be a solution of Eq. (1.4) in the classical sense. To this end, it is sufficient to give $F_i^0(t, x)$ so that the solution is continuable in the switching points along the optimal path which is obtainable on the basis of the maximum principle, i.e. to give $F_i^0(t, x) = u_i^0(t + 0)$ where $x = x^0(t; t_0, u^0(.))$ and $x^0(\tau; t_0, u^0(\,.\,))$ is the optimal path corresponding to control $u^0(\,.\,)$.

It was shown in [5] that the switching surface $\Omega_\alpha(t)$ is described for component u_α be parametric equations

$$x = -\int_t^{t+r} \Phi(t, \varphi) \sum_{i=1}^{n} b_i(\varphi) \operatorname{sign} W_\alpha^*(t) \Phi(t, \varphi)\, b_i(\varphi)\, \mathrm{d}\zeta - \int_t^{t+r} \Phi(t, \varphi)\, f(\varphi)\, \mathrm{d}\varphi \quad (2.2)$$

where vector-parameter $w_\alpha(t)$ has values so that $w_\alpha^*(t)\, b_\alpha(t) = 0$ and $||w_\alpha^*(t)|| = 1$, and parameter r varies in interval $[0, \infty)$. The surface $\Omega_\alpha(t)$ has two important properties [2, 5]. First, it is univalent in the direction of vector $b_\alpha(t)$, i.e. for any point $x \in G(t)$ there is one and only one point $w_\alpha \in \Omega_\alpha(t)$ such that decomposition $x = w_\alpha + \lambda_\alpha b_\alpha(t)$ is possible. In other words, for time t and point x, ω_α and λ_α are uniquely defined from this equation. Thus, function $\lambda_1^0(t, x)$ is defined. Second, the surface $\Omega_\alpha(t)$ is homeomorphic to the $(n-1)$-dimensional plane. It follows from these properties that, like hyperplane, it divides the domain of t-controllability into two isolated parts, and that the Lebesque measure of the set $\Omega_\alpha(t)$ is zero.

It remains to determine the sign of the law $F_\alpha^0(t, x)$ at a certain side of $\Omega_\alpha(t)$. It was shown that for an arbitrery point $x \in G(t) \setminus \Omega_\alpha(t)$, decomposition

$$x = \omega_\alpha^0(t, x) + \lambda_\alpha^0(t, x)\, b_\alpha(t) \quad (2.3)$$

may be done with $\lambda_\alpha^0(t, x) \neq 0$. As it was shown in [7], then

$$F_\alpha^0(t, x) = -\operatorname{sign} \lambda_\alpha^0(t, x)\,. \quad (2.4)$$

The resulting formula may be regarded as an abstract synthesis of the time-optimal law. Equations (2.2) and (2.3) may be used to find the value of the parameter $\lambda_\alpha^0(t, x)$ through given t and x.

3. Piecewise linear approximation of switching surfaces

Although (2.2) completely describes surfaces $\Omega_i^t(t)$, $i = 1, 2, \ldots, m$, it is not still applicable to actual control processes. Nobody succeded in obtaining analytical expression of functions $\lambda_0^\alpha(t, x)$, $\alpha = 1, 2, \ldots, m$, from (2.2) and (2.3) for more or less general systems; the only way, therefore,

is to approximate somehow solutions defined by surfaces $\Omega_\alpha(t)$. The piecewise linear surfaces $\Omega_i^A(t)$ with zero-dimensional bounds lying in $\Omega_i(t)$ may be considered as a class of approximations. To construct such an approximation, it is sufficient to compute points of the surface $\Omega_i(t)$ through (2.2). The advantage of this approximation is that the algorithm for computation of the control law in a current point x will involve only calculations of linear functions and relay operations on values assumed by these functions. To avoid unnecessary complications in approximation, assume below that $f(t) \equiv 0$. Then $\Omega_\alpha(t)$ will be symmetrical with respect to the origin.

Approximation of $\Omega_\alpha(t)$ will be done iteratively by adding some hyperplanes to the already constructed piecewise linear surface. As the first approximation $\Omega_\alpha^1(t)$, a hyperplane will be used which is stretched over $n-1$ points of the surface $\Omega_\alpha(\alpha)$ and goes through the origin. To this end, fix number $\tau_1 > 0$ and $n-1$ vectors $W_{\alpha 1}(t), \ldots, W_{\alpha n-1}(t)$, and make use of (2.2) to compute $n-1$ points $\omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)$ in $\Omega_\alpha(t)$. Note that the sense of τ_1 is the minimal time required to transfer $\omega_{\alpha i}(t)$ to 0, and for a particular problem its choice may be substantiated. When selecting $w_{\alpha i}(t)$, $i=1, \ldots, n-1$, one may come across the fact that computer calculations of various $w_{\alpha i}(t)$ through (2.2) result in the same point $\hat{\omega}_\alpha(t)$ (or in its symmetrical point $-\hat{\omega}_\alpha(t)$. This is accounted for by the following.

According to the definition, $\hat{\omega}_\alpha(t)$ is a point in the convex set boundary $G(t, t+\tau)$ where hyperplane orthogonal to vector $w_\alpha(t)$ will be supporting to $G(t, t+\tau)$ [3]. In the case of multi-dimensional systems, $G(t, t+\tau)$ is usually very elongated in one direction and flattened in others. This accounts for the failure of methods of the Neustadt–Eton-type [3] in numerical solution of the problem of time-optimal programmed control. Under such a situation, arbitrary assignment of vectors $W_\alpha(t)$ will, as a rule, evidently lead either to the same point $\hat{\omega}_\alpha(t)$, or to points close to it in some distant part of the set $G(t, t+\tau)$ (they practically coincide for numerical solution on computer). To obtain an essentially different linearly independent system $\omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)$, the following considerations may be used. Since $G(t, t+\tau)$ is a convex body if at the kth step vector $w_{\alpha k}(t)$ is selected so that its corresponding supporting hyperplane is parallel to the $(k-1)$ dimensional plane Π_{k-1} constructed over points $\omega_{\alpha 1}(t), \ldots, \omega_{\alpha k-1}(t)$ obtained during previous steps, $\omega_{\alpha k}(t)$ will be that point of surface $\Omega_\alpha(t) \cap G(t, t+\tau)$ which is at the maximal distance from Π_{k-1} in the direction of $w_{\alpha k}(t)$, i.e. points $\omega_{\alpha k}(t), \ldots, \omega_{\alpha k}(t)$ will be linearly independent. It follows from the abovesaid that $w_{\alpha k}(t)$ should be selected so that $w_{\alpha k}^*(t)\, \omega_{\alpha i}(t) = 0$, $i = 1, \ldots, k-1$, holds. The $n-1$ points $\omega_{\alpha i}(t)$ determined by such a method will define the hyperplane $x^*[\omega_{\alpha 1}(t) \times \times \omega_{\alpha 2}(t) \times \ldots \times \omega_{\alpha n-1}(t)] = 0$ where $[\omega_{\alpha 1} \times \ldots \times \omega_{\alpha n-1}(t)]$ is a vector product.

It is convenient to write the left side of this relation as $[x, \omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)]$ and to regard it as the determinant of a matrix whose columns are $x, \omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)$. Then it may be written as

$$[x, \omega_{\alpha 1}(t), \omega_{\alpha 2}(t), \ldots, \omega_{\alpha n-1}(t)] = 0 \tag{3.1}$$

Allowing for the criterion $F_{\alpha}^{k}(t, b_{\alpha}(t)) = -1$, obtain the first approximation of the law $F_{\alpha}^{0}(t,x)$:

$$u_{\alpha} = F_{\alpha}^{1}(t, x) = \operatorname{sign} P_{\alpha 1}^{*}(t)\, x \tag{3.2}$$

where

$$P_{\alpha 1}^{*}(t)\, x = -[x, \omega_{\alpha 1}(t), \omega_{\alpha 2}(t), \ldots, \omega_{\alpha n-1}(t)] \operatorname{sign} [b_{\alpha}(t), \omega_{\alpha}(t) \ldots \omega_{\alpha n-1}(t)].$$

In virtue of surface $\Omega_{\alpha}(t)$ symmetry with respect to the origin, points $-\omega_{\alpha 1}(t), \ldots, -\omega_{\alpha n-1}(t)$ will belong to $\Omega_{\alpha}(t)$ as well. Denote the resulted set of $2(n-1)$ points in $\Omega_{\alpha}(t)$ by M_{a}^{1}. Hyperplane $w^{*}\, x + d = 0$ will be said to be separating for some set M if $w^{*}z + d \geq 0$ holds for any point $z \in M$.

To construct the following approximating plane, it is necessary to fix a set of $n-1$ points $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)$ of the set M_{α}^{1} such that they do not lie in one $(n-3)$-dimensional plane (i.e. the points $\hat{\omega}_{\alpha i}(t) - \hat{\omega}_{\alpha 1}(t)$, $i = 2, \ldots, n-1$, are linearly independent) and that the hyperplane

$$\begin{gathered} [x - \hat{\omega}_{\alpha 1}(t), \hat{\omega}_{\alpha 2}(t) - \hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t) - \hat{\omega}_{\alpha 1}(t), b_{\alpha}(t)]\,. \\ \operatorname{sign} [-\hat{\omega}_{\alpha 1}(t), \hat{\omega}_{\alpha 2}(t), \ldots, \hat{\omega}_{\alpha 1-1}(t), b_{\alpha}(t)] = W_{\alpha 0}^{*}(t)\, x + A_{\alpha 0}^{0}(t) = 0 \end{gathered} \tag{3.3}$$

would be separating for M_{α}^{1}. The latter means that a part of the polyhedral surface, approximating $\Omega_{\alpha}(t)$ in vicinity of the origin and representing a convex envelope of the set of points M_{α}^{1}, lies on one side of hyperplane (3.3). This enables the construction of the next face of the approximating surface on the other side of hyperplane (3.3) by means of attaching it to a $(n-2)$-dimensional face stretched over points $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)$. To do this, it is sufficient to find through (2.2) a point $\omega_{\alpha n}(t)$ such that

$$W_{\alpha 0}^{*}(t)\, \omega_{\alpha n}(t) + d_{\alpha 0}^{0}(t) < 0 \tag{3.4}$$

and stretch a hyperplane over points $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t), \omega_{\alpha n}(t)$. To obtain a point $\omega_{\alpha n}(t)$ such that (3.4) holds, one may fix $r = r_2 \geq r_1$, take $w_{a}(t) = w_{a0}(t)$, and compute $\omega_{\alpha n}(t)$ through (2.2). In virtue of the strict convexity of $G(t, t + r)$, points $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t), \omega_{\alpha n}(t)$ will not lie in one $(n-2)$-dimensional plane, they define, therefore, the hyperplane

$$\begin{gathered} [x - \omega_{\alpha n}(t), \hat{\omega}_{\alpha 1}(t) - \omega_{\alpha n}(t), \ldots, \hat{\omega}_{n-1}(t) - \omega_{\alpha n}(t)]\,. \\ \operatorname{sign} [-\omega_{\alpha n}(t), \hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)] = P_{\alpha 2}^{*}(t)\, x + d_{\alpha 2}(t) = 0\,. \end{gathered} \tag{3.5}$$

Denote the sets $\{x : F'_{\alpha 1}(t,x) = 1\}$ and $\{x : F'_{\alpha 1}(t,x) = -1\}$ by $G^+_{\alpha 1}$ and $G^-_{\alpha 1}$, respectively. Assume, at first, that the obtained point $\omega_{\alpha n}(t)$ belongs to $G^+_{\alpha 1}$. Then a face stretched over the points $\omega_{\alpha n}(t)$, $\hat{\omega}_{\alpha 1}(t)$, . . . , $\hat{\omega}_{\alpha n-1}(t)$ will belong to $G^+_{\alpha 1}$ and the law $F^1_\alpha(t,x)$ is to be defined more correctly in $G^+_{\alpha 1}$. Note that in virtue of the fact that the points of the set M^1_α belong to the hyperplane (3.1), and the hyperplane (3.3) is separating for it, any hyperplane going through points $\hat{\omega}_{\alpha 1}(t)$, . . . , $\hat{\omega}_{\alpha n-1}(t)$ (including (3.5)) will be separating for M^1_α. Then, (3.5) divides $G^*_{\alpha 1}$ into two parts, in one of which, given by the condition $P^*_{\alpha 2}(t)x + d_{\alpha 2}(t) > 0$, the law $F^2_\alpha(t,x)$ is defined by the already constructed approximation $F^1_\alpha(t,x)$, and in another, defined by the condition $p^*_{\alpha 2}(t)x + d_{\alpha 2}(t) < 0$, it is corrected in the form of $F^2_\alpha(t,x) = 1$ in accordance to (2.4.).

Addition of the hyperplane (3.5) leads, thus, to the following law which holds in the points $x \in G^+_{\alpha 1}$:

$$F^2_\alpha(t,x) = \operatorname{sign}\left(P^*_{\alpha 2}(t)x + d_{\alpha 2}(t)\right). \tag{3.6}$$

Note that, by virtue of the symmetry of $\Omega_\alpha(t)$, the hyperplane $-p^*_{\alpha 2}(t)x + d_{\alpha 2}(t) = 0$ will be used as approximation in $G^-_{\alpha 1}$, and the law $F^1_\alpha(t,x)$ is then corrected as

$$F^2_\alpha(t,x) = -\operatorname{sign}\left(-P^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\right), \quad x \in G^-_{\alpha 1}. \tag{3.7}$$

Relations (3.6) and (3.7) may be written as a single formula

$$F^2_\alpha(t,x) = \begin{cases} \operatorname{sign}\left(P^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\right) & \text{for} \quad F^1_\alpha(t,x) = 1 \\ \operatorname{sign}\left(P^*_{\alpha 2}(t)\,x - d_{\alpha 2}(t)\right) & \text{for} \quad F^1_\alpha(t,x) = 1 \end{cases}$$

or as

$$F^2_\alpha(t,x) = \operatorname{sign}\left(P^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t)\,F^\alpha_1(t,x)\right). \tag{3.8}$$

Addition of the hyperplane (3.5) has, thus, lead to the law (3.8).

Consider now a case where $\omega_{\alpha n}(t)$ belongs to $G^-_{\alpha 1}$. In virtue of the symmetry of $\Omega_\alpha(t)$, the point $\hat{\omega}_{\alpha n} = -\omega_{\alpha n}$ will also belong to it, but $\hat{\omega}_{\alpha n} \in G^+_{\alpha 1}$ will hold. Note that (3.4) will not, however, hold for the point $\hat{\omega}_{\alpha n}(t)$, because, as it follows from $d^0_{\alpha 0}(t) > 0$, $w^*_{\alpha 0}(t)\,\hat{\omega}_{\alpha 0}(t) + d^4_{0\alpha}(t) > 0$. Then, by considering a symmetrical face stretched over points $-\hat{\omega}_{\alpha 1}(t)$, . . . , $-\hat{\omega}_{\alpha n-1}(t)$, arrive to the conditions under which the approximation (3.8) was constructed. In the given case, we have

$$P^*_{\alpha 2}(t)\,x + d_{\alpha 2}(t) = [x + \omega_{\alpha n}(t), -\hat{\omega}_{\alpha 1}(t) + \omega_{\alpha n}(t), \ldots, -\hat{\omega}_{\alpha n-1}(t) + \omega_{\alpha n}(t)]\cdot \operatorname{sign}\,[\omega_{\alpha n}(t), -\hat{\omega}_{1\alpha}(t), \ldots, -\omega_{\alpha n-1}(t)]\,. \tag{3.9}$$

Further correction of the control law with the addition of other approximating hyperplanes is done iteratively according to the procedure above. Let the law $F_\alpha^k(t, x)$ result after k steps. The set of $2(n-1)+2(k-1)$ points of the surface $\Omega_\alpha(t)$ computed for its construction will be denoted by M_α^k. To construct $F_\alpha^{k+1}(t, x)$, fix a set of $n-1$ points $\hat{\omega}_{\alpha 1}(t), \ldots, \hat{\omega}_{\alpha n-1}(t)$ belonging to M_α^k such that they do not lie in one $(n-3)$-dimensional plane and that the plane

$$[x - \hat{\omega}_{\alpha 1}(t), \hat{\omega}_{\alpha 2}(t) - \hat{\omega}_1(t), \ldots, \hat{\omega}_{\alpha n-1}(t) - \hat{\omega}_{\alpha 1}(t), b_\alpha(t)\,.$$

$$\operatorname{sign}[-\hat{\omega}_{\alpha 1}(t), \hat{\omega}_{\alpha 2}(t), \ldots, \omega_{\alpha n-1}(t), b_\alpha(t)] = W_{\alpha k-1}^*(t)x + d_{\alpha k-1}^0(t) = 0 \quad (3.10)$$

will be separating for M_α^k. This may be always done. Fix a number $r_{k+1} \geq r_k$ and compute through (2.2) a point $\omega_{\alpha n}(t)$ corresponding to this r_{k+1} and to the vector $w_{\alpha k-1}(t)$. If $\omega_{\alpha n}(t)$ belongs to $G_{\alpha k}^+$, i.e. if $F_\alpha^k(t, \omega_{\alpha n}) = 1$, compute through (3.5) the $(k+1)$-st hyperplane $p_{\alpha k+1}^*(t)\,x + d_{\alpha k+1}(t) = 0$ approximating $\Omega_\alpha(t)$. If $F_\alpha^{k+1}(t, \omega_{\alpha n}) = -1$ holds then use (3.9). If at this stage the hyperplane $p_{\alpha k+1}^*(t)x + d_{\alpha k+1}(t) = 0$ is separating for M_α^k, then by repeating the calculations above find that $F_\alpha^{k+1}(t,x)$ is determined as

$$F_\alpha^{k+1}(t,x) = \operatorname{sign}\left(P_{\alpha k+1}^*(t)\,x + d_{\alpha k+1}(t)\,F_\alpha^k(t, x)\right). \quad (3.11)$$

If for all $i = 3, \ldots, k+1$, hyperplanes $p_{\alpha i}^*(t)\,x + d_{\alpha i}(t) = 0$ are separating for sets M_α^{i-1}, one may by means of the recurrent formula (3.11) obtain $F_\alpha^{k+1}(t, x)$ in the following form

$$F_\alpha^{k+1}(t, x) = \operatorname{sign}\left(p_{\alpha k+0}^*(t)\,x + d_{\alpha k+1}(t) \operatorname{sign}\left(p_{\alpha k}^*(t)\,x + \right.\right.$$

$$+ d_{\alpha k}(t) \operatorname{sign}\left(p_{\alpha k-1}^*(t)\,x + \ldots + d_{\alpha 3}(t) \operatorname{sign}\left(p_{\alpha 2}^*(t)\,x + d_{\alpha 2}(t)\,x + \right.\right.$$

$$\left.\left.\left. + d_{\alpha 2}(t) \operatorname{sign} p_{\alpha 2}^*(t)\,x\right) \ldots\right)\right). \quad (3.12)$$

For applied problems, (3.12) is usually sufficient because approximations use small number of planes which may be selected so as to be separating. In the case of one-control systems where approximation is carried out in that domain of the state space where optimal processes are non-oscillating [6], approximating planes will be always separating.

4. Synthesis of the realy linear law

Application of relay laws often turns out to be unreasonable. Demerits of the relay law manifest themselves in vicinity of the origin, where a linear law could be substituted for it. The above method of time-optimal law approximation enables laws that are relay for large deviations from zero,

and linear for small ones. To solve this problem, one has, first, to develop a criterion for passing from one structure to another, and, second, to synthesize the linear law.

Denote the set X where control is defined by the hyperplane (14) by $G_{\alpha 0}(t)$. It follows from (3.11) that if

$$\operatorname{sign}\left(P^*_{\alpha i}(t)\,x + d_{\alpha i}(t)\right) = \operatorname{sign}\left(-\,p^*_{\alpha i}(t)\,x + d_{\alpha i}(t)\right) = 1,\ i = 2, \ldots, k+1, \quad (4.1)$$

holds in some point x, then

$$F^{k+1}_{\alpha}(t, x) = F^{k}_{\alpha}(t, x) = \ldots = F^{1}_{\alpha}(t, x) = \operatorname{sign} p^*_{\alpha 1}(t)\,x\,,$$

i.e. the law is defined only by the hyperplane approximating $\Omega_{\alpha}(t)$ in vicinity of the origin. The set of these, which satisfy (4.1), is defined by a system of inequalities

$$p^*_{\alpha i}(t)\,x + d_{\alpha i}(t) \geq 0, \quad -\,p^*_{\alpha i}(t)\,x + d_{\alpha i}(t) \geq 0, \qquad i = 2, \ldots, k+1,$$

or $-\,d_{\alpha i}(t) \leq p^*_{\alpha i}(t)\,x \leq d_{\alpha i}(t)$, $i = 2, \ldots, k+1$. Consequently, $G^{t}_{\alpha 0}$ is a convex polyhedral set. Conditions (4.1) may be used as a criterion for passing from the linear structure to the relay one, and vice versa.

Thus, the law $F^{k+1}_{\alpha}(t,x)$ is defined in $G_{\alpha 0}(t)$ by the hyperplane $p^*_{\alpha 1}(t)\,x = 0$ as separating. Note that in vicinity of the point $x = 0$ the vector of speed $A(\tau)\,x + \sum_{i=1}^{m} b_i(\tau) \operatorname{sign} p^*_{i1}(\tau)\,x$ of the path $x(\tau)$ is defined by the term $\sum_{i=1}^{m} b_i(\tau) \operatorname{sign} p^*_{i1}(\tau)\,x$ and in the points of the switching plane $p^*_{i1}(t)\,x = 0$ speed $\dfrac{dx(\tau)}{d\tau}$ will be defined by a vector which is a weighted mean of vectors $-\,b_i(\tau)$ and $b_i(\tau)$, i.e. in the points of the plane $p^*_{i1}(t)\,x = 0$ it is close to zero. Therefore, it is reasonable to use as a linear law the law $u_i = \operatorname{sat} k_i\,p^*_{i1}(t)\,x$ (where k_i is a positive constant, sat $a = 1$ for $a \geq 1$, sat $a = a$ for $|\,a\,| \leq 1$, and sat $a = -1$ for $a \leq 1$) which is intrinsically related to the initial relay law and is just an attenuation of the control action intensity in vicinity of the origin. Number k_i defines the slope of the linear characteristic and the width of the linear zone in the direction of vector $p^*_{i1}(t)$. For $k_i = \infty$, obtain the relay law sat $k_i\,p^*_{i1}(t)\,x = \operatorname{sign} p^*_{i1}(t)\,x$.

To obtain the relion's formula, introduce the following variables:

$$\xi_{\alpha i}(t, x) = \frac{1}{2}\left(1 + \operatorname{sign}\left(p^*_{\alpha i}(t)\,x + d_{\alpha i}(t)\right)\right),$$

$$\eta_{\alpha i}(t, x) = \frac{1}{2}\left(1 + \operatorname{sign}\left(-\,p^*_{\alpha i}(t)\,x + d_{\alpha i}(t)\right)\right),$$

$i = 2,3 \ldots, k+1$. Then x belongs to $G_{\alpha 0}(t)$ exactly in that case if $\xi_{\alpha i}(t,x) = \eta_{\alpha i}(t,x) = 1$ for all $i = 2, 3, \ldots, k+1$, i.e. if $\prod_{i+2}^{k+1} \xi_{\alpha i}(t,x)\, \eta_{\alpha i}(t,x) = 1$ holds, and the relion may be written as

$$Rl_{\alpha}^{k+1}(t,x) = \delta\, at \left(k_{\alpha} p_{\alpha 1}^{*}(t)\, x \prod_{i=2}^{k+1} \xi_{\alpha i}\, \eta + \left(1 - \prod_{i=2}^{k+1} \xi_{\alpha i}\, \eta_{\alpha i}\right) F_{\alpha}^{k+1}(t,x) \right). \quad (4.2)$$

5. Synthesis of the relay approximation: general case

As it was assumed in constructing the law $F_{\alpha}^{i+1}(t,x)$, the approximating hyperplane $p_{\alpha i+1}^{*}(t)x + d_{\alpha i+1}(t) = 0$ is separating for M_{α}^{i}. Consider now a case where this condition is not met.

Fix a number $r_{i+2} \geq r_i$ and $n-1$ points $\omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)$ of the set M_{α}^{i} which do not lie in the same $(n-3)$-dimensional plane and such that the hyperplane

$$[x - \omega_{\alpha 1}(t), \omega_{\alpha 2}(t) - \omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t) - \omega_{\alpha 1}(t), b_{\alpha}(t)]\,.$$
$$\text{sign}\,[-\omega_{\alpha 1}(t), \omega_{\alpha 2}(t), \ldots, \omega_{\alpha n-1}(t), b_{\alpha}(t)] = W_{\alpha i}^{*}(t)\, x + d_{\alpha i}^{0}(t) = 0 \quad (5.1)$$

will be separating for M_{α}^{i}. In virtue of the uniqueness of $\Omega_{\alpha}(t)$ in the direction of the vector $b_{\alpha}(t)$, this may be always done. Let $\omega_{\alpha n}(t) \in \Omega_{\alpha}(t)$ be determined through (2.2) under $r = r_{i+1}$ and $w_{\alpha}(t) = w_{\alpha i}(t)$. If $F_{\alpha}'(t, \omega_{\alpha n}) = 1$ holds, consider the hyperplane

$$[x - \omega_{\alpha n}(t), \omega_{\alpha 1}(t) - \omega_{\alpha n}(t), \ldots, \omega_{\alpha n-1}(t) - \omega(t)]\,. \quad (5.2)$$
$$\text{sign}\,[-\omega_{\alpha n}(t), \omega_{\alpha 1}(t), \ldots, \omega_{\alpha n-1}(t)] = p_{\alpha i+1}^{*}(t)x + d_{\alpha i}(t) = 0\,.$$

If $F_{\alpha}^{i}(t, \omega_{\alpha n}(t)) = -1$ holds, consider the hyperplane

$$[x + \omega_{\alpha n}(t), -\omega_{\alpha 1}(t) + \omega_{\alpha n}(t), \ldots, -\omega_{\alpha n-1} + \omega_{\alpha n}(t)]\,.$$
$$\text{sign}\,[\omega_{\alpha n}(t), -\omega_{\alpha 1}(t), \ldots, -\omega_{\alpha n-1}(t)] = p_{\alpha i+1}^{*}(t)\, x + d_{\alpha i+1}(t) = 0\,. \quad (5.3)$$

Let the hyperplane (24) not be separating for M_{α}^{i}. Break the space X into three sets: (1) a set for whose points the following holds

$$|\, W_{\alpha i}^{*}(t)\, x\,| \leq d_{\alpha i}^{0}(t) \quad (5.4)$$

(2) a set of those x for which the following inequality holds

$$W_{\alpha i}^{*}(t)\, x + d_{\alpha i}^{0}(t) < 0 \quad (5.5)$$

and (3) a set defined by the condition

$$W_{\alpha i}^{*}(t)\, x - d_{\alpha i}^{0}(t) > 0 \quad (5.6)$$

The hyperplane (5.3) is, by construction, separating for M_a^i, therefore all points of M_α^i belong to (5.4), i.e. the law defined by a piecewise linear switching surface stretched over the points of the set M_α^i will be used as approximation in (5.4), in other words this law is $F_\alpha^i(t, x)$. Further, the point $\omega_{\alpha n}(t)$ belongs to the set (5.5). Taking into account $F_a^{i+1}(t, b_a(t)) = -1$, we obtain that the approximation defined by the hyperplane (5.3) will, in the set (5.5), have the form of $F_a^{i+1}(t, x) = \operatorname{sign}\left(p_{\alpha i+1}^*(t)\, x + d_{\alpha i+1}(t)\right) x$ where $x = -\operatorname{sign} p_{\alpha i+1}^*(t)\, b_a(t)$. The control law $F_\alpha^{i+1}(t, x) = \operatorname{sign}\left(p_{\alpha i+1}^*(t)\, x - d_{\alpha i+1}(t)\right) x$ will be used as approximation by virtue of the symmetry in the points of the set (5.6).

Thus, in all points of the space X, the $(i + 1)$-st approximation is defined which may be written as

$$F_\alpha^{i+1}(t, x) = \begin{cases} F_\alpha^i(t, x) & \text{if } |\, W_{\alpha i}^*(t)\, x\,| \leq d_{\alpha i}^0(t) \\ \operatorname{sign}\left(p_{\alpha i+1}^*(t)\, x + d_{\alpha i+1}(t)\right) x & \text{if } W_{\alpha i}^*(t)\, x < -d_\alpha^0(t) \\ \operatorname{sign}\left(p_{\alpha i+1}^*(t)\, x - d_{\alpha i+1}(t)\right) x & \text{if } W_{\alpha i}^*(t)\, x > d_{\alpha i}^0(t) \end{cases} \quad (5.7)$$

This completes the iterative construction of piecewise linear approximations of the law $F_\alpha(t, k)$ for arbitrary systems.

6. Synthesis of the relay-linear law: General case

Let $G_{\alpha 0}(t)$ be a set in X where the law $F_a^k(t, x)$ is defined by the switching hyperplane $p_{\alpha 1}^*(t)\, x = 0$. Like above, modify in $G_{\alpha 0}(t)$ the relay law $u_a = \operatorname{sign} p_{a1}^*(t)\, x$ by $u_\alpha = \operatorname{sat} k_a\, p_{\alpha 1}^*(t)\, x$. In the general case, the fact that a currents state x belongs to the domain $G_{\alpha 0}(t)$ is established not only through the approximation hyperplanes $p_{\alpha i}^*(t)\, x + d_{\alpha i}(t) = 0$ but, in the case when they are not separating for the sets M_α^{i-1}, by hyperplanes $w_{\alpha i}^*(t)\, x + d_{\alpha i}^0(t) = 0$ as well. Let hyperplane $p_{\alpha i}^*(t)\, x + d_{\alpha i}(t) = 0$ not be separating for some i. Then functions $\xi_{\alpha i}(t,x)$ and $\eta_{\alpha i}(t, x)$ should be defined for this i as $\xi_{\alpha i}(t, x) = \frac{1}{2}\left(1 + \operatorname{sign}\left(w_{\alpha i}^*(t)\, x + d_{\alpha i}^0(t)\right)\right)$ and $\eta_{\alpha i}(t, x) = \frac{1}{2}\left(1 + \operatorname{sign}\left(-w_{\alpha i}^*(t)\, x + d_{\alpha i}^0(t)\right)\right)$. The relay-linear la wis again in the form of (4.2) where $F_\alpha^i(t, x)$ for each $i = 2,3, \ldots, k + 1$ is determined either through (3.12), or (5.7).

7. Analysis of the relay approximation

Application of an approximate control law naturally gives rise to the question of the dynamic quality of the resulting closed system. Of interest is, in particular, the problem of closeness of the optimal and suboptimal paths. In principle, this problem was solved by A. F. Filippov [4] in 1960 while working on the theory of differential equations with the discontinuous right side.

The necessity of using the notion of generalized solution of differential equation is due to the following. As it was indicated above, definition of the optimal law $F^0(t,x)$ may be completed on surfaces $\Omega_1(t), \ldots, \Omega_m(t)$ so that for each initial condition the closed system (4) will have a classical absolutely continuous solution $x(\tau)$ which is optimal and unique. However, such a completed definition is, generally speaking, unrealistic with approximate was laws $F^{k+1}(t,x)$ having piecewise linear switching surfaces, i.e. the closed system will not have a solution in the classical sense, thus leading to the necessity of extending the notion of solution to such situations. Only this extension would enable comparison of optimal and suboptimal solutions.

Let $B^n_\gamma(x, \delta)$ denote an n-dimensional (in the sense of norm γ) sphere in X with radius δ and center in the point x. The Lebesgue measure of set θ will be denoted by $\mu\theta$. Let $f(t, x)$ be a function defined and measurable over some set $[t_0, t_1] \times Q$ whose values belong to X. Then a set which is an image of the set (t, P) in X under mapping f, is understood as $f(t, P)$. The vector-function $x(t), t \in (t_0, t_1)$ is referred to as generalized solution of the equation

$$\frac{dx}{dt} = f(t, x) \tag{7.1}$$

if it is absolutely continuous and if for nearly all t's the following holds:

$$\frac{dx(t)}{dt} \in \bigcap_{\gamma<0} \bigcap_{\mu\theta=0} \overline{\operatorname{conv}} f\big(t, B^n_\gamma(x(t), \delta) \setminus \theta\big) . \tag{7.2}$$

Note that the solution in the generalized sense is independent of the values of the function $f(t, x)$ over the zero-measure set in X. Therefore, if one consideres solutions of the system (4) in this sense, they are independent of the values of the function $F(t, x)$ in the points of surfaces $\Omega_1(t), \ldots, \Omega_m(t)$; i.e. $F^0(t, x)$ may be defined over switching surfaces in an arbitrary manner. However, no a priori guarantee may be given that Filippov's solution of (4) would define optimal paths since they depend essentially on the values of the control law over the switching surfaces. As it was demonstrated by Brunovský [8], there are systems with nonoptimal generalized solutions that are not something exotic. Therefore we are going to confine the following consideration to those systems of type [1] whose generalized solutions of (1.4) coincide with the classical optimal ones when $F^0(t, x)$ is given over $\Omega_1(t), \ldots, \Omega_m(t)$ so that the latter does exist. In this case, Filippov's solution of (4) corresponding to the initial condition $(\hat{t}_0, x)$, $x \in G(\hat{t}_0)$, is unique.

According to [4], let $f(t, x)$ and $f_k(t, x)$ denote the right side of (1) for $u = F^0(t, x)$ and $u = F^k(t, x)$, respectively. Let $\varrho_c(x(\cdot))$ be a norm in space C of the continuous functions, i.e. $\varrho_c(x(\cdot)) = \max_{t_0 \le t \le t_1} \gamma(x(t))$. Differential equations

$$\frac{dx}{dt} = f(t, x) \tag{7.3}$$

and

$$\frac{dx}{dt} = f_k(t, x) \tag{7.4}$$

satisfy evidently the conditions of Theorems 3 and 11 by Filippov (here $g_k(t, x) \equiv 0$), therefore the second part of Corollary 4 of Theorem 11 may be formulated as follows:

Theorem 1. For a given point x^0 and number $\varepsilon > 0$ there is $d(x^0, \varepsilon)$ such that if

$$\Omega_\alpha(t) - \subset \bigcup_{x^0 \in \Omega_\alpha^k(t)} B_\gamma^n\big(x, d(x^0, \varepsilon)\big), \quad \alpha = 1, 2, \ldots, m, \tag{7.5}$$

holds, i.e. if $\Omega_\alpha(t)$ belongs to the d-vicinity of the surface $\Omega_\alpha^k(t)$, then $\varrho_c\big(x(\,.\,; t_0, x^0) - x^\Delta(\,.\,; t_0, x^0)\big) < \varepsilon$ holds, where $x^\Delta(t; t_0, x^0)$ is solution of Eq. (7.4), $x^\Delta(t_0; t_0, x^0) = x^0$.

This statement may be somewhat augmented. Let D be a compact set in X. For a given $\varepsilon > 0$, according to Theorem 11, for each point x^0 there is a $d(x^0, \varepsilon) > 0$ such that if $x^0 + \Delta x$ belongs to $B_\gamma^n\big(x^0, d(x^0, \varepsilon)\big)$ and (7.5) holds then $\varrho_c\big(x(\,.\,; t_0, x^0) - x^\Delta(\,.\,; t_0, x^0 + \Delta x)\big) < \frac{\varepsilon}{2}$. In virtue of the continuous dependence of the solution of (27.3) on the initial condition, there is a $\delta(x^0, \varepsilon) > 0$ such that $\varrho_c\big(x(\,.\,; t_0, x + \Delta x) - x(\,.\,; t_0, x^0)\big) < \frac{\varepsilon}{2}$ provided that $x^0 + \Delta x \in B_\gamma^n\big(x^0, \delta(x^0, \varepsilon)\big)$. If $\delta(x^0, \varepsilon) > d(x^0, \varepsilon)$, then when $x^0 + \Delta x \in B_\gamma^n\big(x^0, \delta(x^0, \varepsilon)\big)$ and the condition (7.5) is met, obtain

$$\begin{aligned} \varrho_c \quad & tx^0 + \Delta x) - x(\,.\,; t_0, x^0 + \Delta x)\big) = \varrho_c(x)\,.\,; t_0, x^0 + \Delta x) - x(\,.\,; t_0, x^0) + \\ & + x(\cdot\,; t_0, x^0) - x^\Delta(\cdot\,; t_0, x^0 + t_\Delta x) \le \varrho_c\big(x(\cdot\,; t_0, x^0 + \Delta x) - x(\cdot\,; t_0, x^0)\big) + \\ & + \varrho_c\big(x(\cdot\,; t_0, x^0) = x^0(\cdot\,; t_0, x^0 + \Delta x)\big) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\,. \end{aligned}$$

Thus, set D is covered by the family $\big\{B_\gamma^n\big(x^0, d(x^0, \varepsilon)\big) \mid \forall x^0 \in D\big\}$ so that if $\hat{x} \in B_\gamma^n\big(x^0, d(x^0, \varepsilon)\big)$, then $\varrho_c\big(x(\cdot\,; t_0, x) - \hat{x}^\Delta(\cdot\,; t_0, \hat{x})\big) < \varepsilon$. Since D is compact, from this cover one can separate a finite cover $\big\{B_\gamma^n\big(x_i^0, d(x_i^0, \varepsilon)\big) \mid i = 1, 2, \ldots, N\big\}$

$\bigcup_{x \in \Omega_\alpha^k(t)} B_\gamma^n(x, d(\varepsilon))$. Then there is $d(\varepsilon) = \min_{1 \le i \le N} \{d(x_i^0, \varepsilon)\} > 0$, and if $\Omega_\alpha(t) \subset$ then for any initial condition $x^0 \in D$, $\varrho_c\big(x(\cdot\,; t_0, x^0) - x^A(\cdot\,; t_0, x^0)\big) < \varepsilon$. This leads to the following result.

Theorem 2. For any $\varepsilon > 0$ there is a $d(\varepsilon) > 0$ such that if condition (7.5) holds for $d(x^0, \varepsilon) = d\varepsilon$, $\varrho_c\big(x(\cdot\,; t_0, x^0) - x^A(\cdot\,; t_0, x^0)\big) < \varepsilon$ for any $x^0 \in D$.

Note that the requirement that $\Omega_\alpha(t)$ belongs to the d-vicibity of $\Omega_\alpha^k(t)$ may be in this theorem substituted by the requirement that $\Omega_\alpha^k(t)$ belongs to the d-vicinity of $\Omega_\alpha(t)$. This answers, in principle, the question about the quality of the suboptimal law. In each particular case, verification of the conditions of this theorem would require a special study. At the same time, there is a special case where one can directly estimate closeness of surfaces $\Omega(t)$ and $\Omega^k(t)$. Moreover, in this case the generalized solution is optimal.

8. Stationary systems with one control

Consider a system of type (1.1) where A is a constant matrix such that all of its eigenvalues are real-valued, $f(t) \equiv 0$ and $m = 1$. In [7] for this case there was derived a law approximating $F^0(x)$ is $G(0, T)$ where T is fixed:

$$F^k(x) = \operatorname{sign}\big(\varphi_k(x) \operatorname{sign} \chi_k(x) + \operatorname{sign}\big(\varphi_{k-1}(x) \operatorname{sign} \chi_{k-1}(x) + \ldots + \\ + \ldots + \operatorname{sign}\big(\varphi_n(x) \operatorname{sign} \chi_n(x) + \operatorname{sign} \varphi_{n-1}(x)\big) \ldots\big)\big),$$

where

$$k \ge \eta,\ \varphi_j(x) = \max_{i \in N(j)} \{|\, p_{ij}^* x \,|\},\ \chi_j(x) = \min_{i \in N(j)} \{p_{ij}^* x + 1\},$$

$$N(j) = \left\{ i \;\middle|\; \binom{n-1}{j-1} + 1 \le i \le \binom{n-1}{j} \right\}, \quad j = n,\ n+1, \ldots, k;$$

$\binom{n}{j}$ is the number of combinations of j, form n;

$$p_{ij}^* x = \frac{[x, v_{q_1}, v_{q_2}, \ldots, v_{q_{n-1}}]}{\left[\sum_{\alpha=1}^{n-1} (-1)^\alpha \sum_{i=q_{\alpha-1}+1}^{q_\alpha} v_i, v_{q_1}, v_{q_2}, \ldots, v_{q_{n-1}}\right]},$$

$$v_i = g(t_i) - g(t_{i-1}),\ g(t) = \int_0^t e^{-\tau A} b\, d\tau, \quad i = 1, 2, \ldots, k;\ \text{and } t_i$$

are given numbers such that $0 = t_0 < t_1 < t_2 < \ldots < t_{k-1} < t_k = T$, parameters $q_1, q_2, \ldots, q_{n-1}$ run over all integer values, and $1 \le q_1 < q_2 < \ldots <$

$< q_{n-2} < q_{n-1} = j$ and

$$\varphi_{n-1}(x) = \frac{[x_1 v_1, v_2, \ldots, v_{n-1}]}{[v_1, v_2, v_3, \ldots, v_n]} .$$

This control law may be also represented in the form of (3.12). The piecewise linear switching surface Ω^A of this law which is a union of $(n-1)$-dimensional parallelotopes, is homeomorphic to the optimal surface Ω, and all the zero-dimensional faces of Ω^A belong to Ω.

Such a structure of Ω^A enables easy estimation of its closeness to Ω. It was shown in [7] that Ω may be represented as a union of disjoint sets, each of them corresponding to a fixed set $\{q_1, q_2, \ldots, q_{n-1}\}$ and being described parametrically as

$$\omega = \sum_{\alpha=1}^{n-2} 2(-1)^{\alpha+1} g\big(\zeta_\alpha t_{q_{\alpha-1}} + (1-\zeta_\alpha) t_{q_\alpha}\big) + \\ + (-1)^n g\big(\zeta_{n-1} t_{q_{n-1}-1} + (1-\zeta_{n-1})_{q_{n-1}}\big) \tag{8.1}$$

where parameters $\zeta_1, \ldots, \zeta_{n-1}$ vary over intervals $0 \leq \zeta_i \leq 1$. The set corresponding to $\{q_1 q_2 \ldots, q_{n-1}\}$ is approximated by a parallelotope of the following form

$$\omega^A = \sum_{\alpha=1}^{n-2} (-1)^{\alpha+1} \{2g(t_{q_\alpha}) - 2\bar{\zeta}_\alpha v_{q_\alpha}\} + (-1)^n \{g(t_{q_{n-1}}) - \bar{\zeta}_{n-1} v_{q_{n-1}}\} \tag{8.2}$$

where $0 \leq \zeta_i \leq 1$, $i = 1, 2, \ldots, n-1$. If one can demonstrate that for any point in the surface Ω a point in the surface Ω^A may be indicated such that the distance between them would be less than a prescribed d, this will mean that Ω belongs to the d-vicinity of Ω^A. However, each point of the surface Ω belongs to one of the sets (8.1). Consider, therefore, an arbitrary set from (8.1) and an arbitrary point ω in it characterized by some numbers $q_1, q_2, \ldots, q_{n-1}$ and $\zeta_1, \zeta_2, \ldots, \zeta_{n-1}$. Consider a point $\omega^A \in \Omega^A$ which is defined by the following values of parameters: $\bar{\zeta}_1 = \zeta_1, \ldots, \bar{\zeta}_{n-1} = \zeta_{n-1}$ and estimate the difference $\gamma(\omega - \omega^A)$. Obtain

$$\gamma(\omega - \omega^A) = \gamma\Big(2 \sum_{\alpha=1}^{n-2} (-1)^{\alpha-1} \{g(\zeta_\alpha t_{q_\alpha - 1} + (1-\zeta_\alpha) t_{q_\alpha}) - g(t_{q_\alpha}) + \zeta_\alpha v_{q_\alpha}\} + \\ + (-1)^n \{g(\zeta_{n-1} t_{q_{n-1}-1} + (1-\zeta_{n-1}) t_{q_{n-1}} - g(t_{q_{ne-1}}) + \zeta_{n-1} v_{q_{n-1}}\}\Big) \leq \\ \leq 2 \sum_{\alpha=1}^{n-2} \gamma \left(\int_{\zeta_\alpha t_{q_\alpha - 1} + (1-\zeta_\alpha) t_{q_\alpha}}^{t_{q_\alpha}} e^{-\tau A} b d\tau + \zeta_\alpha \int_{t_{q_\alpha - 1}}^{t_{q_\alpha}} e^{-\tau A} b d\tau \right) +$$

$$+\gamma\left(-\int_{\zeta_{n-1}t_{q_{n-1}-1}}^{t_{q_{n-1}}} e^{-\tau A}b\,d\tau+\zeta_{n-1}\int_{t_{q_{n-1}-1}}^{t_{q_\alpha-1}} e^{-\tau A}b\,d\tau\right)=$$

$$=2\sum_{\alpha=1}^{n-2}\gamma\left(\zeta_\alpha(t_{q_\alpha}-t_{q_\alpha-1})\,(e^{-\theta_{q_\alpha}A}-e^{-\hat{\theta}_{q_\alpha}A})\,b\right)+$$

$$+\gamma\left(\zeta_{n-1}(t_{q_{n-1}}-t_{q_{n-1}-1})\,(e^{-\theta_{q_{n-1}}A}-e^{-\hat{\theta}_{q_{n-1}}A})\,b\right)$$

where numbers θ_{q_α} and $\hat{\theta}_{q_\alpha}$ satisfy constraints $t_{q_\alpha-1}<\hat{\theta}_{q_{\alpha 1}}<t_{q_\alpha}$, $\zeta_\alpha t_{q_\alpha-1}+$ $+(1-\zeta_\alpha)t_{q_\alpha}<\theta_{q_\alpha}<t_{q_\alpha}$. It follows that the estimate $|\theta_{q_\alpha}-\hat{\theta}_{q_\alpha}|<t_{q_\alpha}-t_{q_\alpha-1}$ holds. Taking into account that $0\leq\theta_{q_\alpha}\leq T$ for all α, obtain inequality

$$\gamma\left((e^{-\theta_{q_\alpha}A}-e^{-\hat{\theta}_{q_\alpha}A})\,b\right)\leq\gamma\left(e^{-\theta_{q_\alpha}A}(\hat{\theta}_{q_\alpha}-\theta_{q_\alpha})\,Ab\right)+0(\theta_{q_\alpha}-\hat{\theta}_{q_\alpha})\leq$$

$$\leq\gamma(e^{-\theta_{q_\alpha}A}\,Ab)\,|\,\hat{\theta}_{q_\alpha}-\theta_{q_\alpha}\,|+0(\theta_{q_\alpha}-\hat{\theta}_{q_\alpha})\leq K_1\,|\,\hat{\theta}_{q_\alpha}-\theta_{q_\alpha}\,|+0(\theta_{q_\alpha}-\hat{\theta}_{q_\alpha})$$

where K_1 is a constant independent of q_α. Hence

$$\gamma(\omega-\omega^\Delta)\leq 2\sum_{\alpha=1}^{n-2}\zeta(t_{q_\alpha}-t_{q_\alpha-1})\,K_1\,|\,\hat{\theta}_{q_\alpha}-\theta_{q_\alpha}\,|+$$

$$+\zeta_{n-1}\,K_2(t_{q_{n-1}}-t_{q_{n-1}-1})\,|\,\theta_{q_{n-1}}-\hat{\theta}_{q_{n-1}}|+$$

$$+0\Big(\max_{1\leq\alpha\leq n-1}|\theta_{q_\alpha}-\hat{\theta}_{q_\alpha}|\max_{1\leq\alpha\leq n-1}(t_{q_\alpha}-t_{q_\alpha-1})<\hat{K}\sum_{\alpha=1}^{n-1}(t_{q_\alpha}-t_{q_\alpha-1})^2\Big)+$$

$$+0\Big(\max_{1\leq\alpha\leq n-1}(t_{q_\alpha}-t_{q_\alpha-1})^2\Big)<K\max_{1\leq\alpha\leq n-1}(t_{q_\alpha}-t_{q_\alpha-1})^2+$$

$$+0\Big(\max_{1\leq\alpha\leq n-1}t_{q_\alpha}-t_{q_\alpha-1})^2\Big).$$

Thus, in order that $\gamma(\omega-\omega^\Delta)<d$ holds for any point $\omega\in\Omega$, it is sufficient to select k and $t_1,t_2,\ldots,t_k$ so that the inequality $\max\limits_{1\leq i\leq k}(t_i-t_{i-1})^2<\dfrac{d}{K}$ is satisfied. Estimation of K may be readily obtained through the matrix A and the number T, and numbers $t_1,t_2,\ldots,t_k$ may be given on the basis of the required approximation precision d.

In the considered case Theorem 2 guarantees that for any given $\varepsilon>0$ a set of parameters $t_1,t_2,\ldots,t_k$, $k=k(\varepsilon)$, may be selected so that for any process $x^\Delta(\tau)$ under $u=F^k(x)$ beginning in $G(0,T)$ and for the optimal process $x(\tau)$, $x^\Delta(0)=x(o)$, $\max\limits_{0\leq\tau\leq T}\gamma(x(\tau)-x^\Delta(\tau))<\varepsilon$ will hold. It follows from this relation that if ε is such that $B^n_\gamma(0,\varepsilon)\subseteq G(0,T)$, then $x^\Delta(\tau)\in B^n_\gamma(G(0,T),\varepsilon)$

holds for all $\tau > 0$, i.e. the approximating law will provide for the system stability (asymptotic stability is not guaranteed here). Let $\hat{T}$ be the minimal of all T's for which $B_\gamma^n(0, \varepsilon) \subseteq G(0, T)$. Then $x^\Delta(\tau) \in B_\gamma^n(G(0, \hat{T}), \varepsilon)$ for all τ. Note, that for a given dynamic quality, i.e. for the requirement that $x_\gamma^n(\tau) \in B_\gamma^n(0, \delta)$ hold, ε may be selected so that $B_\gamma^n(G(0, \hat{T}), \varepsilon) \subseteq B_\gamma^n(0, \delta)$.

The piecewise linear method of approximation of the optimal switching surface developed in [7] enables suboptimal control laws providing the required solution precision.

Note. The results presented in this section are applicable without modifications to the case of arbitrary linear systems with one control where approximation is constructed in such $G(t_0, t_0 + T)$ that the system is non-oscillating over $[t_0, t_0 + T]$ i.e. the time-optimal control has at most $n-1$ changes of sign. In this case, like in the case of stationary systems in [7], one can show that function

$$D(t; z_1, z_2, \ldots, z_{n-1}) = [\Phi(t, z_1)\, b(z_1), \Phi(t_1 z_2)\, b(z_2), \ldots, \Phi(t, z)\, b(Z_n)]$$

will keep the same sign over the domain $t \leq Z_1 \leq Z_2 \ldots \leq Z_n \leq t + T$, whence feasibility of parallelotopic approximation of the optimal switching surface follows. Vectors V_i are defined in this case as

$$V_i(t) = \int_{t+t_{i-1}}^{t+t_i} \Phi(t, \varphi)\, b(y)\, d\varphi, \quad i = 1, 2, \ldots, k, \text{ where } 0 = t_0 < t_1 < t_2 < \ldots < t \leq T.$$

9. Relion Analysis

The Filippov's theory enables determination of the quality of the time-optimal law approximations not only in the relay, but in the relion form as well. Corollary 2 of Theorem 11 [4] establishes that if $f^\Delta(t, x)$ is the mean value (with any non-negative weight) of function $f(t,x)$ over the domain $B_\gamma^n(x, \Delta)$, solutions of the equation $\dfrac{dx}{dt} = f^\Delta(t, x)$ converge uniformly under $\Delta \to 0$ to the corresponding solutions of the equations $\dfrac{dx}{dt} = f(t, x)$. The linear zone always contains both the point x where $F_i^0(t, x) = 1$, and the one where $F_i^0(t, x) = -1$. For relion approximation, the linear law with saturation is in each point x an average, with non-negative weight, of these boundary values. On the basis of the abovesaid, one can therefore, take a sufficiently small linear zone and provide the required closeness of paths to the time-optimal one. Like in the case of linear approximations it may be shown that by an appropriate choice of Δ a linear zone may be constructed such that relion's linear part will provide for the system stability.

References

1. *N. N. Krasovskii*, Motion Control Theory, Nauka, Moscow, 1968 (in Russian).
2. *A. I. Moroz*, Report on research project "Development of a logical-dynamical theory of cybernetic systems", Moscow Inst. of Electronic Equipment, State Registration No. P—003554, inv. No. 6309154, 1974 (in Russian).
3. *V. G. Boltyanskii*, Mathematical Methods for Optimal Control, Nauka, Moscow, 1969 (in Russian).
4. *A. F. Filippov*, Differential Equations with Discontinuous Right Side, Matematicheskii Sbornik, v. 51, No. 1, 1960 (in Russian).
5. *A. I. Moroz*, Time-Optimal Control Synthesis Problem, Automation and Remote Control, No. 1, 1970.
6. *E. L. Tonkov*, Nonoscillation and Number of Switchings in Linear Nonstationary Speed-Optimal System, Differentisial'nye Uravneniya, No. 12, 1973 (in Russian).
7. *A. I. Moroz*, Elements of the Optimal System Theory, MIEM Publ., Moscow, 1974 (in Russian).
8. *P. Brunosky*, The Closed-Loop Time-Optimal Control. I. Optimality. "SIAM J. Control", 12, No. 4, 1974.

SITUATION RECOGNITION IN THE PURSUIT PROBLEM

D. A. POSPELOV, N. N. CHEREZOVA, D. I. SHAPIRO

(Moscow, USSR)

The paper is devoted to the problem of pursuit represented as a sequence of situations. A situation recognition technique is proposed based on a special set of logic functions and decision tables.

The paper proposes a recognition algorithm for situations arising in pursuit of two objects by a single object. Objects are moving over a plane, have speed, orientation, and, in addition, the pursuer has some spare resource. The algorithm results in decisions about attainability/unattainability of each of the pursued objects under a given situation.

For the sake of simplicity, refer to pursuer as "hound", and to pursued objects as "hares"; although, as it should be clear to the reader, any other objects with the above parameters may be used.

1. Problem formulation

The pursuit problem is formulated as follows: a hound and two hares are moving over a plane. The hound should catch one or both hares, if possible. At each fixed time, the hound should make decision depending on the current situation. The following decisions are possible: both hares are unattainable; the first hare is attainable (the second one being unattainable) and vice versa; both hares are attainable. If both are attainable, the hound should decide which hare to pursuit.

At each decision time the following situation is analysed:

$$S^h(\varrho_e^h, \vartheta_e^h, \nu_e^h, m_e^h, \vartheta_{31}^h, \nu_{31}^h, m_{31}^h, \vartheta_{32}^h, \nu_{32}^h, m_{32}^h) \tag{1.1}$$

where h is the number of a situation (or of the time of decision making);

S^h is situation with number h;

ϱ_e^h is the hound's resource at time h (it decreases/increases in direct proportion to time);

ϑ_e^h is the hound's speed at time h;

ν_e^h is the hound's direction at time h;

m_e^h is the hound's location at time h which is described by two coordinates on the plane (x_e^h, y_e^h);

ϑ_{31}^h is the speed of Hare 1 at time h;

ϑ_{31} is the direction of Hare 1 at time h;
m_{31} is the location of Hare 1 at time h which is described by two coordinates on the plane (x^h_{31}, y^h_{31});
ϑ^h_{31} is the speed of Hare 2 at time h;
v^h_{31} is the direction of Hare 2 at time h;
m^h_{31} is the location of Hare 2 at time h which is described by two coordinates on the plane (x^h_{32}, y^h_{32}).

For situations represented by a set of characteristic parameters (1)' procedure of decision making is as follows:

Particular situation S is broken into two situations S^{h1}_3, and S^{h2} containing sets of parameters describing the hound's position with respect to each hare:

$$\begin{aligned} &S^{h1}(\varrho^h_e, \vartheta^h_e, v^h_e, m^h_e, \vartheta^h_{31}, v^h_{31}, m^h_{31}) \\ &S^{h2}(\varrho^h_e, \vartheta^h_e, v^h_e, m^h_e, \vartheta^h_{32}, v^h_{32}, m^h_{32}). \end{aligned} \tag{1.2}$$

Such a partitioning is feasible since the hound estimates the position of each hare taking into account its own possibilities and, consequently, makes decision upon each hare successively. The hares are assumed to move independently of each other.

The problem of 1st/2nd hare attainability is solved in a regular manner by means of the characteristic parameters of situations.

2. Method of regular problem solution

In particular, a geometrical method may be suggested. Let a situation S^h be fixed at some time. Connect hound's location (x^h_e, y^h_e) with those of the 1st (x^h_{31}, y^h_{31}) and 2nd (x^h_{32}, y^h_{32}) hares by intercepts AB and AM, respectively. Draw lines $CD \perp AB$ and $KL \perp AM$. Line CD divides the plane where the hound and Hare 1 move, into two half-planes A_1 and B_1. Line KL divides the plane of the hound and the second hare into half-planes A_2 and B_2. In accordance with this, a set of possible decisions about attainability/unattainability of the 1st (2nd) hare is partitioned with due regard to the parameters of particular situations S^{h_1} and S^{h_2}. If direction of the 1st (2nd) hare lies in the plane $A_1(A_2)$ and if $\vartheta_{31} \geq \vartheta_e(\vartheta_{32} \geq \vartheta_e)$, the hare is unattainable. If $\vartheta_e > \vartheta_{31}$ $(\vartheta_e > \vartheta_{32})$, the first condition being valid, it is necessary to compute the point R of possible meeting of the hound and the hare, i.e. to estimate the distance AR (direction of the hound). The hare will be attainable if the hound has enough resource to cover the distance AR (Hound's resource decreases/increase in direct proportion to time). The hare is attainable

if $\varrho_e^h - \Delta\varrho_e t_e^h \geq 0$, where $\Delta\varrho$ is resource decrement for a time unit, and t_e is the time of hound's run. If direction of the first/second hare lies in the half-plane B_1/B_2, the first (second) hare is attainable provided that $\vartheta_e \geq \vartheta_{31}$ ($\vartheta_e > \vartheta_{32}$) and the hound has resource to cover the shortest distance AS to the point of its possible meeting with the hare. Moreover, the 1st/2nd hare may be attainable if $\vartheta_e \leq \vartheta_{31}(\vartheta_e \leq \vartheta_{32})$. It is important, in this case, for the hound to cover the distance AS at time equal (shorter) to the time of hare's movement and to have enough resource.

If a situation S^{h_1}/S^{h_2} has decision "hare attainable", its parameters are tabulated in a "hare attainable" table which is a matrix with rows representing sets of characteristic parameters of each situation and seven columns representing:

1 — hound's resources,
2 — hound's speeds,
3 — hound's directions,
4 — hound's locations,
5 — hare's speeds,
6 — hare's directions, and
7 — hare's locations.

In a similar manner a "hare unattainable" table is formed.

3. Analysis and application of the "Hare Attainable" and "Hare Unattainable" tables

Generation of such tables is education of computer, parameters of the tables being used for recognition and analysis of particular situations, and making decisions for each of them. The tables are analysed and used in the following manner. Let the "hare unattainable" and "hare attainable" tables have r and k rows-situations, respectively, i.e. let the tables be $(r \times 7)$ and $(k \times 7)$ matrices. Call the tables "unattainability matrix" (A_{r_7}) and "attainability matrix" (A_{k_7}). With the aim of using it as educational material in the process of recognizing a particular situation, the unattainability matrix A_{r_7} is transformed as follows: rows-situations in the matrix are put in the decreasing order of resource

$$\varrho_e^1 \leq \varrho_e^2 \leq \ldots \leq \varrho_e^r \tag{3.1}$$

where 1, 2, . . ., r are numbers of rows-situations in the unattainability matrix. The attainability matrix A_{k_7} is transformed as follows: its rows-situations are placed in the increasing order of the first characteristic parameter (hound's

resource), i.e.

$$\varrho_e^1 \geq \varrho_e^2 \geq, \ldots, \varrho_e^k \tag{3.2}$$

where $1, 2, \ldots, k$ are numbers of situations in the attainability matrix.

Let during the pursuit situations S^h be fixed that is to be analysed and about which a decision is to be made. Processing of a new particular situation S^h proceeds from analysis of its first characteristic parameter (hound's resource) and looking for the location of this situation in the unattainability matrix:

$$\varrho_e^{l-1} \leq \varrho_e^h \leq \varrho_e^l \tag{3.3}$$

where $1 \leq l-1$, $l \leq r$ are numbers of the unattainability matrix rows. Thus, a truncated matrix A_{r_7} is separated having rows with numbers $l, l+1, \ldots, r$. The rows-situations of the truncated matrix contain those parameters of the hound's resource that are greater than the resource in the particular situation S^h. Thus, all other parameters in the rows-situations being regarded as fixed, decision about hare's unattainability which is true for all situations of the truncated matrix, will be true for the particular situations S^h. Moreover, parameters in the final row-situation are interrelated: each resource has appropriate hound's speed, hare's speed, etc. This interrelation is manifested by the fact that all the parameters are in one situation, and it is attributed to the intelligent behaviour of the hound and the hare.

Thus, if the hound's resource under a particular situation S^h is some generalizations of the resources in the rows-situations of the truncated matrix (in the sense that a conclusion about hare's unattainability which is true for the rows-situations of the truncated matrix, is also true for the situation S^h provided that only resource changes), it is necessary to make some generalizations of the hound's and hare's speeds, and other parameters of the truncated matrix in order to compare the generalized parameters with those a particular situation and to see as to what degree they are similar. Similarity of situations depends on that of conclusions which may be drawn upon hare's attainability or unattainability. Parameters of the truncated matrix are analysed and generalized in the following manner:

$$\underline{\vartheta}_e^l = \min(\vartheta_e^l, \ldots, \vartheta_e^r); \tag{3.4}$$

$$\bar{\vartheta}_e^l = \max(\vartheta_e^l, \ldots, \vartheta_e^r); \tag{3.5}$$

$$\underline{\vartheta}_3 = \min(\vartheta_3^l, \ldots, \vartheta_3^r); \tag{3.6}$$

$$\bar{\vartheta}_3 = \max(\vartheta_3^l, \ldots, \vartheta_3^r); \tag{3.7}$$

$$\underline{f}^l = \min(f^l, \ldots, f^r); \tag{3.8}$$

$$\bar{f}^l = \max(f^l, \ldots, f^r),$$

where

$$f^l = \sqrt{(x_e^l - x_3^l)^2 + (y_e^l - y_3^l)^2} \tag{3.9}$$
$$v^l = \{v_3, \ldots, v_3^r\}.$$

By means of successive computation of the logic functions $M_1, M_2, \ldots, M_8$ described below, similarity is verified of a particular situation and generalizations based upon the truncated matrix. The attainability matrix A_{k_7} is used in a similar fashion. Its rows-situations are put in the increasing order the first characteristic parameter, the hound's resource as indicated in page 5.

Under a particular situation S^h with

$$\varrho_e^{d-1} \geq \varrho_e^h \geq \varrho_e^d \tag{3.10}$$

where $1 \leq d - 1$, $d \leq k$ are numbers of the attainability matrix rows, a truncated matrix $A_{k_7}^1$ is separated having rows with numbers $d, d+1, \ldots, k$, and parameters of the matrix are generalized row-wise:

$$\underline{\vartheta}_e^d = \min(\vartheta_e^d, \ldots, \vartheta_e^k) \tag{3.11}$$
$$\bar{\vartheta}_e^d = \max(\vartheta_e^d, \ldots, \vartheta_e^k); \tag{3.12}$$
$$\underline{\vartheta}_3^d = \min(\vartheta_3^d, \ldots, \vartheta_3^k); \tag{3.13}$$
$$\bar{\vartheta}_3^d = \max(\vartheta_3^d, \ldots, \vartheta_3^k); \tag{3.14}$$
$$\underline{f}^d = \min(f^d, \ldots, f^k); \tag{3.15}$$
$$\bar{f}^d = \max(f^d, \ldots, f^k);$$

where

$$f^d = \sqrt{(x_e^d - x_3^d)^2 + (y_e - y_3)^2},$$
$$v_3^d = \{v_3^d, \ldots, v_3^k\}, \tag{3.16}$$

This enables generation of some generalized situations which are compared with the particular one. Similarity between the generalized and particular situations is established through logic functions $F_1, \ldots, F_8$ described below.

4. Description of the situation recognition process

Let a situation S^h arise. It is broken into two situations S^{h_1} and S^{h_2} as indicated in page 2. Each situation S^{h_1} and S^{h_2} expresses hound's attitude toward the first and the second hares, respectively. At first, unattainability of the first hare is considered. Situation S^{h_1} is analysed by means of the "hare unattainable" table, an appropriate cross-section of the table is found, and parameters of the situations involved into the truncated matrix are generalized by formulae (3.4)–(3.9). Now, parameters of the particular situation S^{h_1} are compared with those resulting from the generalization. To this end, values

of some logic functions $(M_1, \ldots, M_8)$ are computed successively, functions being generated through the parameters of the particular situation and generalizations making use of some logic operators [1]. If at least one of the logic functions is 1, the first hare is regarded as unattainable, and one proceeds to analysing the situation S^{h_2} through the "hare unattainable" table. If all logic functions are equal, attainability of the first hare under the situation S^{h_1} is analysed through the "hare attainable" table.

By means of some logic operators [1], functions $F_1, \ldots, F_8$ are generated on the basis of the situation S^{h_1} parameters and generalizations of the truncated "hare attainable" matrix. If at least one of the logic functions $F_1, \ldots, F_2$ is 1, the first hare is regarded as attainable, and one proceeds to analysing the situation S^{h_2} by means of the "hare unattainable" table.

The situation S^{h_2} is analysed by means of the "hare unattainable" table like the situation S^{h_1}. The same logic functions $M_1, \ldots, M_8$ are used to make decision about hare's unattainability. If this analysis does not give some positive results (i.e. no one function is 1), the situation S^{h_2} is analysed by means of the "hare attainable" table like situation S^{h_1}. Decision about attainability of the second hare is made by means of functions $F_1, \ldots, F_8$. It may happen that at least one of the components of the particular situation $S^h(S^{h_1}$ or $S^{h_2})$ is not recognizable through the "hare attainable" and "hare unattainable" tables. This happens when comparison of the parameters of a particular situation S^{h_1} and generalizations of the truncated "hare unattainable" matrix by means of logic functions $M_1, \ldots, M_8$ does not give positive results, i.e. when all successively computed logic functions $M_1, \ldots, M_8$ are 0. Assume that subsequent attainability checking of S^{h_1} does not lead to any positive result, i.e. that all logic functions $F_1, \ldots, F_8$ are 0. Therefore, decision about S^{h_1} cannot be made, and, on the whole, there is nothing to say about the situation S^h.

In this case S^h is estimated by a regular method resulting in a decision about attainability or unattainability of both hares.

Decision in the form of situation parameters is entered into appropriate "hare attainable" or "hare unattainable" tables, thus updating them.

5. Situation recognition algorithm

The following logic operators are introduced for analysis of particular situations by means of the unattainability (A_{r_τ}) and attainability (A_{k_τ}) matrices:

$$\alpha_1(Y) = \begin{cases} 1 & \text{if} \quad Y \leq k \\ 0 & \text{if} \quad Y > k, \; k \text{ is given} \end{cases} \tag{5.1}$$

$$\alpha_2(Y) = \begin{cases} 1 & \text{if} \quad Y \subset \Omega \\ 0 & \text{if} \quad Y \not\subset \Omega,\ \Omega \text{ is given} \end{cases} \tag{5.2}$$

$$\alpha_3(Y) = \begin{cases} 1 & \text{is} \quad P \leq Y \leq k \\ 0 & \text{if} \quad Y > k \vee Y < p;\ k,\ P \text{ are given} \end{cases} \tag{5.3}$$

$$\alpha_4(Y) = \begin{cases} 1 & \text{if} \quad Y \geq Z \\ 0 & \text{if} \quad Y < Z,\ Z \text{ is given}. \end{cases} \tag{5.4}$$

Consider functions $M_1, \ldots, M_8$ used for comparison of the particular situation parameters with, say, the generalized parameters of a truncated "hare unattainable" matrix.

$$\begin{aligned} M_1[\alpha_1(\varrho_e^{h_1} \leq \underline{\varrho}_e^l) \wedge \alpha_1(\vartheta_e^{h_1} \subseteq \vartheta^l) \wedge \\ \alpha_2(v_1^{h_1} \subset v_3^l) \wedge \alpha_4(\vartheta_{3_1}^{h_1} \geq \vartheta_3^l) \wedge \alpha_4(f^{h_1} \geq f_1^l)] \end{aligned} \tag{5.5}$$

If $M_1 = 1$, hare 1 is unattainable because hound's resource is small, the speed is low, and the hare has too great speed and is far.

In this case, analyse S^{h_2} of the second hare by means of the unattainability table. If $M_1 = 0$, compute the following function

$$\begin{aligned} M_2[\alpha_1(\varrho_e^{h_1} \leq \underline{\varrho}_e^l) \wedge \alpha_3(\vartheta_e^l \leq \vartheta_e^{h_1} \leq \vartheta_e^l) \wedge \alpha_2(\vartheta_{3_1}^{k_1} \subset v^l) \wedge \\ \alpha_4(\vartheta_{3_1}^{h_1} \geq \bar{\vartheta}^l) \wedge \alpha_4(f^{h_1} \geq \bar{f}^l)] \end{aligned} \tag{5.6}$$

If $M_2 = 1$, hare 1 is unattainable because the hound has small resource, and the hare is far and has great speed. Now pass to analysis of the second hare's S^{h_2} by means of the unattainability table. If $M_2 = 0$, for the first hare compute the following function:

$$\begin{aligned} M_3[\alpha_1(\varrho_e^{h_1} \leq \underline{\varrho}_e^l) \wedge \alpha_1(\vartheta_e^{h_1} \leq \underline{\vartheta}_e^l) \wedge \\ \alpha_2(v_1^{h_1} \subset v_3) \wedge \alpha_3(\underline{\vartheta}_3^l \leq \vartheta_{:}^{h_1} \leq \vartheta_3^l) \wedge \alpha_4(f^{h_1} \geq \bar{f}^l)]. \end{aligned} \tag{5.7}$$

If $M_3 = 1$, the first hare is unattainable because the hound has low speed, and the hare is far away.

Pass to analysis of the second hare's S^{h_2} by means of the unattainability table. If $M_3 = 0$, compute the following function for the second hare:

$$\begin{aligned} M_4[\alpha_1(\varrho_e^{h_1} \leq \underline{\varrho}_e^l) \wedge \alpha_1(\vartheta_e^{h_1} \leq \underline{\vartheta}_e^l) \wedge \alpha_2(v_{3_1}^{h_1} \subset v_3) \wedge \\ \alpha_4(\vartheta_3^{h_1} \geq \bar{\vartheta}_3) \wedge \alpha_3(\underline{f}^l \leq f^{h_1} \leq \bar{f}^l)]. \end{aligned} \tag{5.8}$$

If $M_4 = 1$, the first hare is unattainable. The hound has small resource and low speed, and the hare's speed is great. By menas of the unattainability

table analyse the second hare's S^h. If $M_4 = 0$, compute M_5 for the first hare:

$$M_5[\alpha_1(\varrho_e^{h_1} \leq \varrho_e^l) \wedge \alpha_1(\vartheta_e^{h_1} \leq \vartheta_e^l) \wedge \alpha_2(v_3^{h_1} \subset v_3) \wedge \alpha_3(\vartheta_3^l \leq \vartheta_3^{h_1} \leq \bar{\vartheta}_3^l) \wedge \alpha_3(\underline{f}^l \leq f^{h_1} \leq \bar{f}^l)]. \tag{5.9}$$

If $M_5 = 1$, the first hare is unattainable because hound's resource is small and speed is low. Now, analyse the second hare's S^{h_2} by means of the unattainability table. If $M_5 = 0$, compute M_6 for the first hare:

$$M_6[\alpha_1(\varrho_e^{h_1} \leq \varrho_e^l) \wedge \alpha_3(\delta_e^l \leq \vartheta_e^{h_1} \leq \bar{\vartheta}_e^l) \wedge \alpha_2(v_3^{h_1} \subset v_3^l) \wedge \alpha_4(\vartheta_3^{h_1} \geq \bar{\vartheta}_3^l) \wedge \alpha_3(\underline{f}^l \leq f^{h_1} \leq \bar{f}^l)]. \tag{5.10}$$

If $M_6 = 1$, the first hare is unattainable because hound's resource is small, and hare's speed is very high. Further, analyse by means of the unattainability table the second hare's S^{h_2}.

If $M_6 = 0$, compute M_7 for the first hare:

$$M_7[\alpha_1(\varrho_e^{h_1} \leq \varrho_e^l) \wedge \alpha_3(\underline{\vartheta}_e^l \leq \vartheta_e^{h_1} \leq \bar{\vartheta}_c^l) \wedge \alpha_2(v_3^{h_1} \subset v_3^l) \wedge \alpha_3(\underline{\vartheta}_3^l \leq \vartheta_3^{h_1} \leq \bar{\vartheta}_3^l) \wedge \alpha_4(f^{h_1} \geq \bar{f}^l)]. \tag{5.11}$$

If $M_7 = 1$, the first hare is unattainable because hound's resource is small, and the hare is far.

Now, pass to the analysis of the second hare's substituations S^{h_2} through the unattainability table. If $M_8 = 0$, compute for the first hare:

$$M_8[\alpha_1(\varrho_e^{h_1} \leq \varrho_e^l) \wedge \alpha_3(\underline{\vartheta}_e^l \leq \bar{\vartheta}_e^l) \wedge \alpha_2(v_3^{h_1} \subset v_3^l) \wedge \alpha_3(\underline{\vartheta}_3^l \leq \vartheta_3^{h_1} \leq \bar{\vartheta}_3^l) \wedge \alpha_3(f^l \leq f^{h_1} \leq \bar{f}^l)]. \tag{5.12}$$

If $M_8 = 1$, the first hare is unattainable because hound's resource is small. Pass to analysis of subsituation S^{h_2} of the second hare by means of the unattainability table. If $M_8 = 0$, situation describing the hound's position with respect to the first hare is regarded as unrecognizable through the unattainability table.

It may happen that the first hare is attainable. Situation is analysed through an appropriate truncated matrix of the attainability table by means of the following functions:

$$F_1[\alpha_4(\varrho_e^h \geq \varrho_e^d) \wedge \alpha_4(\vartheta_e^{h_1} \geq \bar{\vartheta}_e^l) \wedge \alpha_2(v_3^{h_1} \subset v_3^l) \wedge \alpha_1(\vartheta_3^{h_1} \leq \vartheta_3^d) \wedge \alpha_1(f^{h_1} \leq \underline{f}^d)]. \tag{5.13}$$

If $F_1 = 1$, the first hare is attainable because the hound has enough resource, very high speed, and the hare is slow and near.

Now analyse second hare's S^{h_2} through the unattainability table. If $F_1 = 0$, analyse S^{h_1} through the function:

$$F_2[\alpha_4(\varrho_e^{h_1} \geq \varrho_e^d) \wedge \alpha_3(\underline{\vartheta}_e^d \leq \vartheta_e^{h_2} \leq \bar{\vartheta}_e^d) \wedge \alpha_2(\nu_3^{h_1} \subset \nu_3^d) \wedge \alpha_1(\vartheta_3^{h_1} \leq \bar{\vartheta}_3^d) \wedge \alpha_1(f^{h_1} \leq \underline{f}^d)]. \tag{5.14}$$

If $F_2 = 1$, the first hare is attainable because the hound's resource is sufficient, and the hare's speed is low and the hare is close. Analyse second hare's S^{h_2} through the unattainability table.

If $F_2 = 0$, analyse S^{h_1} by means of function F_3:

$$F_3[\alpha_4(\varrho_e^{h_1} \geq \varrho_e^d) \wedge \alpha_4(\vartheta_e^{h_1} \geq \bar{\vartheta}_e^d) \wedge \alpha_2(\nu_3^{h_1} \subset \nu_3^d) \wedge \alpha_3(\underline{\vartheta}_3^d \leq \vartheta_3^{h_2} \leq \bar{\vartheta}_3^d) \wedge \alpha_1(f^{h_1} \leq \underline{f}^d)]. \tag{5.15}$$

If $F_3 = 1$, the first hare is attainable because the bound's resource is sufficient and speed is very great, and the hare is near. Now we pass to the analysis of the second hare through the the unattainability table. If $F_3 = 0$, compute for the first hare:

$$F_4[\alpha_4(\varrho_e^{h_1} \geq \varrho_e^d) \wedge \alpha_4(\vartheta_e^{h_1} \geq \bar{\vartheta}_c^d) \wedge \alpha_2(\nu_3^{h_1} \subset \nu_3^d) \wedge \alpha_1(\nu_3^{h_1} \leq \vartheta_3^d) \wedge \alpha_3(f^d \leq f^{h_1} \leq f^d)]. \tag{5.16}$$

If $F_4 = 1$, the first hare is attainable because the hound's resource is sufficient and speed is very high, and the hare's speed is low. Now analyse the second hare by means of the unattainability table.

If $F_4 = 0$, compute F_5 for the first hare:

$$F_5[\alpha_4(\varrho_e^{h_1} \geq \varrho_e^d) \wedge \alpha_3(\underline{\vartheta}_e^d \leq \vartheta_e^{h_1} \leq \bar{\vartheta}_e^d) \wedge \alpha_2(\nu_3^{h_1} \subset \nu_3^d) \wedge \alpha_3(\underline{\vartheta}_3^d \leq \vartheta_3^{h_1} \leq \bar{\vartheta}_3^d) \wedge \alpha_1(f^{h_1} \leq \underline{f}^d)]. \tag{5.17}$$

If $F_5 = 1$, the first hare is attainable because the hound's resource is sufficient and the hare is near.

Pass to analysis of the second hare through the unattainability table.

If $F_5 = 0$, compute F_6 for the first hare:

$$F_6[\alpha_4(\varrho_e^{h_1} > \varrho_e^d) \wedge \alpha_3(\underline{\vartheta}_e^d \leq \vartheta_e^{h_1} \leq \bar{\vartheta}_e^d) \wedge \alpha_2(\nu_3^{h_1} \subset \nu_3^d) \wedge \alpha_1(\vartheta_3^{h_1} \leq \vartheta_3^d) \wedge \alpha_3(\underline{f}^d \leq f^{h_1} \leq \bar{f}^d)]. \tag{5.18}$$

If $F_6 = 1$, the first hare is attainable because the hound's resource is sufficient, and the hare's speed is low.

Analyse now the second hare through the unattainability table.

If $F_6 = 0$, compute for the first hare the following logic function:

$$F_7[\alpha_4(\varrho_e^{h_1} \geq \varrho_e^d) \wedge \alpha_4(\vartheta_e^{h_1} \geq \bar{\vartheta}_e^d) \wedge \alpha_2(\nu_3^{h_1} \subset \nu_3^d) \wedge \alpha_3(\underline{\vartheta}_3^d \leq \vartheta_3^{h_1} \leq \bar{\vartheta}_3^d) \wedge \alpha_3(\underline{f}^d \leq f^{h_1} \leq \bar{f}^d)] \tag{5.19}$$

If $F_7 = 1$, the first hare is attainable because the hound has sufficient resource and very great speed.

Analyse the second hare by the unattainability table.

If $F_7 = 0$, compute for the first hare the following logic function:

$$F_8[\alpha_4(\varrho_e^{h_1} \geq \varrho_e^d) \wedge \alpha_3(\delta_e \leq \vartheta_e^{h_1} \leq \bar{\vartheta}_e^d) \wedge \alpha_2(\nu_3^{h_1} \subset \nu^d) \wedge \alpha_3(\underline{\vartheta}_3^d \leq \vartheta_3^{h_1} \leq \bar{\vartheta}_3^d) \wedge \alpha_4(\vartheta_e^{h_1} > \vartheta_3^{h_1}) \wedge \alpha_3(\underline{f}^d \leq f^{h_1} \leq \bar{f}^d)] \tag{5.20}$$

If $F_8 = 1$, the first hare is attainable because the hound has sufficient resource. Now analyse second hare's position by the unattainability tabel.

If $F_8 = 0$, situation S^{h_1} is regarded as unrecognizable by the attainability table.

Thus, situation S^h is deemed unrecognizable if at least one of its components S^{h_1} or S^{h_2} is not recognizable through the attainability and unattainability tables by means of the above logic functions.

In this case, the question of hare's attainability /unattainability is solved by a regular method. Results of the solution are entered in an appropriate table. Thus, this is the problem of table updating.

Moreover, when passing from one situation to another in the simulated process and recognizing them by means of the unattainability M_i $(i = 1, \ldots, 8)$ and attainability $F_j (j = 1, \ldots, 8)$ functions, one can estimate the pursuit process qualitatively in the sense of its improvement or deterioration.*

"Worse-better" notions are formed on the basis of semantic and qualitative analysis of the parameters of a particular situation which is to be recognized by the above logic functions.

If $M_k \rightarrow M_e$ and $l > k$, recognition of situations by the unattainability functions (M_l) reveals deterioration of the pursuit in terms of attainability. If $F_m \rightarrow F_n$ and $m > n$, recognition of situations by the attainability functions reveals improvement of the pursuit in terms of attainability. Transfer from any function M to any function F means that the pursuit process has improved.

* Estimation of qualitative parameters "worse-better" may be done by means of Zadeh's fuzzy sets [3] as shown, for example, in [4].

6. Example

The above pursuit situation recognition algorithm was written as a program in ALGOL—60 for the BESM–6 computer.

Situations-input successively in the computer were recognized by the program using the above set of logic functions and decision tables kept in the computer memory. A test problem is shown below requiring recognition of 14 situations by means of decision tables involving 6 situations.

Hare Unattainable:

ϱ	ϑ_e	v_e	x_e	y_e	ϑ_3	v_3	x_3	y_3
(100;	11;	0.2000;	5;	7;	14;	0.5236;	10;	10;)
(140;	15;	0.4000;	3;	3;	20;	1.0647;	13;	6;)
(170;	5;	0.3665;	13;	12;	20;	1.0472;	21;	14;)

Hare Attainable:

(180;	20;	0.5236;	3;	3;	10;	1.0472;	11;	5;)
(130;	80;	0.5236;	3;	3;	10;	2.6180;	8;	9;)
(100;	20;	0.4014;	4;	9;	5;	2.0944;	10;	10;)

The pursuit is described by a series of the following situations:

1. (135; 10; 0.4000; 2; 1; 24; 1.0472; 15; 2; 28; 1.0647; 25.4); Hare 1 unattainable, Hare 2 attainable, recognition by M_2

2. (130; 30; 0.4014; 2; 1; 6; 2.0944; 6; 6; 7; 2.0944; 7; 6;); Hare 1 attainable, Hare 2 attainable, recognition by F_8;

3. (180; 60; 0.5236; 2; 1; 9; 1.0472; 6; 7; 9; 1.0472; 5; 8); Both hares are attainable, recognition by F_8;

4. (170; 60; 0.5236; 3; 2; 9; 2.6180; 4; 6; 9; 2.6180; 5; 8); Both are attainable, recognition by F_5;

5. (160; 90; 0.5236; 3; 4; 4; 2.6180; 6; 10; 9; 2.6180; 8; 9); Hare 1 attainable, recognition by F_4. Hare 2 attainable, recognition by F_5;

6. (150; 90; 0.4014; 4; 9; 6; 2.6180; 2; 4; 9; 2.6180; 9; 4); Hare 1 attainable, recognition by F_3. Hare 2 attainable, recognition by F_4;

7. (140; 70; 0.5236; 5; 1; 4; 2.0944; 7; 6; 4; 2.0944; 2; 7); Hare 1 attainable, recognition by F_2. Hare 2 attainable, recognition by F_3;

8. (130; 30; 0.4014; 2; 1; 6; 2.0944; 6; 6; 7; 2.0944; 7; 6); Both attainable, recognition by F_8;

9. (110; 30; 0.4014; 4; 9; 3; 2.0944; 8; 10; 5; 2.0944; 9; 10); Both attainable, recognition by F_1;

10. (105; 30; 0.4014; 4; 9; 3; 2.0944; 4; 9; 5; 2.0944; 9; 10); Hare 1 is caught. Hare 2 attainable, recognition by F_1;

11. (108; 25; 0.4014; 5; 8; 0; 0; 8; 5; 2.0944; 8; 6); Hare 1 caught. Hare 2 attainable, recognition by F_1;

12. (105; 25; 0.4014; 3; 8; 0; 0; 3; 8; 5; 2.0944; 5; 4); Hare 1 caught. Hare 2 attainable, recognition by F_1;

13. (102; 23; 0.4014; 2; 2; 0; 0; 2; 3; 5; 2.0944; 1; 8); Hare 1 caught. Hare 2 attainable, recognition by F_1;

14. (100; 23; 0.4014; 2; 1; 0; 0; 2; 1; 5; 2.0944; 2; 1); Both are caught.

If hare's direction and speed parameters are equal to zero and the hound' and hare's coordinates coincide, the hare caught is.

The presented graphs were drawn by plotter, they illustrate the process if pursuit consisting of 14 situations.

Presently, works are being carried out on compilation of (i) operating programs generating decision tables on the basis of one of the regular methods, and (ii) also of table updating programs that enable interfacing the recognition program with the decision table generating one.

It may happen at situation recognition and table updating that the majority of situations would be recognized through the tables without further updating. Determination of the minimal size of such tables is an important result in solution of this problem.

References

1. *M. M. Bongard*, The Problem of Recognition, Moscow, Nauka, 1967 (in Russian).
2. *D. A. Pospelov*, *V. N. Pushkin*, Thinking and Automata, Sovetskoe Radio, Moscow, 1972 (in Russian).
3. *Z. Zadeh*, Fuzzy sets. Inf. and Control, No. 8, 1965, pp. 338—352.
4. *D. I. Shapiro*, Mathematical Methods in Decision Making (preprint). Council for Cybernetics, Academy of Sciences USSR, Moscow, 1974 (in Russian).

COMMENTS ON THE METHOD OF THE MOMENT DYNAMICS

V. G. ROMANOV

(Moscow)

Improvements are introduced into the moment dynamics method which is intended for the study of systems involving elements of various types whose states belong to discrete ests and change under the action of Poisson flows.

1. Introduction

The study of large systems and the development of automatic control systems involve simulation of their internal processes. Simulation of data aquisition, accumulation and handling, operation of separate subsystems and the system as a whole under various controls may be guoted as as examples.

The scope of this paper is systems consisting of various elements, each of them assuming a state from a discrete state set and altering it under the effect of Poisson flows.

For small numbers of elements and states, such systems have been studied by means of equations of system state probabilities. For greater number of system elements, the method of the means dynamics finds extensive application. Moreover, the method of the moment dynamics [1, 2] is also applicable here, and moment dynamics equations may be obtained through generating functions [3] if element transition intensities are simple (mostly, linear) functions of the number of elements having different states.

However, (i) each of the above methods is applicable to a restricted subsystem, and (ii) in the majority of cases, to generate moment dynamics equations, the problem is to be solved a new.

Differential moment equations derived in the present paper for an arbitrary enough system are adjustable to a particular system under study in a rather simple manner.

2. Differential moment equations

Consider a system which at each time $t \geq 0$ is in one of states forming a countable set A. The probability that at time $t \geq 0$ the system is in a state $x \in A$ will be denoted by $p(x)$.

With the course of time the system may pass from one state to another. This process will be described by a stochastic process $\{X(t), t \geq 0\}$, assuming that $X(t) = x$ if the system is in state x at time t.

The process in the system is assumed to be Markovian, i.e. for any $x, y(\sigma) \in A$ under $0 \leq \sigma \leq \tau$ and $\tau < t$ the following equality is assumed to hold:

$$\mathcal{P}\{X(t) = x \mid X(\sigma) = y(\sigma), 0 \leq \tau \leq \tau\} = \mathcal{P}(X(t) = x \mid X(\tau) = y(\tau)\}. \quad (2.1)$$

Moreover, it will be assumed below that for all $t \geq 0$ and $x, y \in A$ there is a limit

$$\lim_{\Delta t \to 0} \frac{\mathcal{P}\{X(t + \Delta t) = y \mid x(t) = x\}}{\Delta t} = \lambda(x, y, t) \quad (2.2)$$

where $\lambda(x, y, t)$ or, simpler, $\lambda(x, y)$ will be referred to as the intensity of system's transition from state x to y.

In such a system, the following differential equations for probabilities $p(x)$ hold:

$$\dot{p}(x) = \sum_{y \in A} \lambda(y, x)\, p(y) - \sum_{y \in A} \lambda(x, y)\, p(x), \; x \in A. \quad (2.3)$$

Through this set of equations obtain a differential equation of mathematical expectation of an arbitrary function $\varphi(x)$. To this end, multiply both sides of each equation of the system by $\varphi(x)$ and sum them over all $x \in A$. Obtain

$$\sum_{x \in A} \varphi(x)\, \dot{p}(x) = \sum_{x \in A} \sum_{y \in A} \varphi(x)\, \lambda(y, x)\, p(y) - \sum_{x \in A} \sum_{y \in A} \varphi(x)\, \lambda(x, y)\, p(x). \quad (2.4)$$

The sum in the left side of the above equation is the derivative of the mathematical expectation of the function $\varphi(x)$. By assuming in the right side of the equation in the first double sum that $x = y'$ and $y = x'$, by changing the order of summation, and by omitting the strokes of x' and y', one can combine summation with respect to x and to obtain the following:

$$\frac{d}{dt} M[\varphi(x)] = \sum_{x \in A} p(x) \sum_{y \in A} [\varphi(y) - \varphi(x)]\, \lambda(x, y). \quad (2.5)$$

The expression resulting in the right side of the equation is the mathematical expectation of the internal double sum. Mathematical expectation of an arbitrary function should, therefore, satisfy the following differential equation

$$\frac{d}{dt} M[\varphi(x)] = M\Big[\sum_{y \in A} \big(\varphi(y) - \varphi(x)\big)\, \lambda(x, y)\Big]. \quad (2.6)$$

Now, particularize, the system. Let it involve l types of elements, each of $n_i (i = 1,2, \ldots, l)$ elements of ith type being allowed to be in m_i states. State transitions will be assumed to the Poisson.

The state graph of the system is a totality of state graphs of elements of various types. Figure 1 examplifies such a graph for the elements of ith type. It is convenient to represent the states x of such a system as vectors

$$x = (x_1, \ldots, x_i, \ldots, x_l), x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{i,m_i}), \tag{2.7}$$

where x_{ij} is the number of elements of ith type in jth state.

Denote intensities of transition of ith type elements from state "j" to state "k" by $\lambda_{ijk}(x)$ and consider them as known functions of time t and system state x.

Since the Poisson flows are ordinary, intensities $\lambda(x, y)$ are non-zero only if all the components of the vector y are equal to the respective components of the vector x, apart from any two ones: $y_{ij} = x_{ij} - 1$ and $y_{ik} = x_{ik} + 1$. Denoting such vectors y as x_{ijk}, obtain the following writting

$$\lambda(x, y) = \begin{cases} \lambda_{ijk}(x) & \text{at } y = x_{ijk} \\ 0 & \text{at } y \neq x_{ijk}. \end{cases} \tag{2.8}$$

Therefore, summation over vectors $y \in A$ will be substituted in (2.6) by summation over i,j, and k which will be taken out of the sign of mathematical expectation, thus resulting in

$$\frac{d}{dt} M[\varphi(X)] = \sum_{i=1}^{l} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} M[\big(\varphi(x_{ijk}) - \varphi(X)\big) \lambda_{ijk}(X)]. \tag{2.9}$$

It should be taken into account in (2.9) that

$$\lambda_{ijk}(X) = 0 \quad \text{at} \quad X_{ij} = 0$$

because in this case there is not a single element in that state from which the transition flow proceeds.

Therefore, the following equality holds

$$M[\Psi(X) \cdot \lambda_{ijk}(X)] = \mathcal{P}(X_{ij} \neq 0)\, M[\psi(x)\, \lambda_{ijk}(x) \mid X_{ij} \neq 0] \tag{2.10}$$

which may be conveniently applied where neither $\lambda_{ijk}(X)$, nor $\psi(X)$ are proportional to the component X_{ij} of vector X.

Thus, differential equation has been obtained for an arbitrary moment of system state probability distribution. It should be emphasized that its derivation did not involve any constraints on the transition intensities as time finctions. The transition intensities, therefore, mav be arbitrary time

functions. Moreover, they even may be linear functions of delta-functions provided that sets of instants are disjoint for various transition intensities where arguments of delta-functions involved are zero, i.e. that the ordinariness of the totality of all flows is observed.

Consider some special cases of Eq. (2.9).

3. Equations for mathematical exceptions of the number of elements in different states and for central second-order moments

Differential equations for mathematical expectations of the number of elements in different states may be found by assuming in (2.9) that $\varphi(X) = x_{r\varrho}$. As it follows from the definition of vector X_{ijk}, in this case

$$\varphi(X_{ijk}) = X_{r\varrho} + \delta_{ri}(\delta_{\varrho k} - \delta_{\varrho j})$$

where $\delta_{\mu\nu}$ is Kronecker's index. Therefore, (2.10) becomes simpler:

$$\frac{d}{dt} M[X_{rs}] = \sum_{j=1}^{m_r} M[\lambda_{rjs}(x) - \lambda_{rsj}(x), \; r = 1, \ldots, s = 1, \ldots, m_r. \qquad (3.1)$$

Like in (2.9), the possibility of using here (2.10) should be taken into account.

Comparison of these equations with those for the means dynamics reveals that in the right side of (3.1) there are mathematical expectations of intensities as opposed to the intensities as functions of mathematical expectations of the number of elements in states involved into the equations for the means dynamics.

Therefore, if

$$\lambda_{ijk}(x) = c_{ijk} X_{ij}, \quad i = 1, \ldots, l; \; j, \; k = 1, \ldots, m_i, \qquad (3.2)$$

equations for the means dynamics coincide with precise equations (3.1), i.e. they turn out to be precise for any number of elements in the system, not only for great numbers as it was commonly believed.

Find now differential equations for central second-order moments. Assume in (2.9) that $\varphi(X) = \mathring{X}_{rs}\mathring{X}_{\mu\nu}$ where $\mathring{X}_{\alpha\beta} = X_{\alpha\beta} - M[X_{\alpha\varrho}]$. Then $\varphi(X_{ijk}) = \left(\mathring{X}_{rs} + \delta_{ri}\,(\delta_{sk} - \delta_{\varrho j})\right)\left(\mathring{X}_{\mu\nu} + \delta_{\mu\nu}(\delta_{\nu k} - \delta_{\nu k})\right)$ and, after simplifications, (2.9) takes the following form

$$\frac{d}{dt} M[\mathring{X}_{rs}\,\mathring{X}_{\mu\nu}] = \sum_{j=1}^{m_\nu} M[\mathring{X}_{rs}\left(\lambda_{\mu j\nu}(x) - \lambda_{\mu\nu j}(x)\right)] +$$

$$\sum_{j=1}^{m_r} M[\mathring{X}_{\mu\nu}\left(\lambda_{rjs}(X) - \lambda_{rsj}(x)\right)] + \delta_{r\mu}\delta_{s\nu} \sum_{j=1}^{m_r} M[\lambda_{rjs}(x) + \lambda_{rsj}(x)] - \qquad (3.3)$$

$$\delta_{r\mu} M[\lambda_{rs\nu}(X) - \lambda_{r\nu s}(X)], \; r, \mu = 1, \ldots, l; \; \varrho = 1, \ldots, m_r; \; \nu = 1, \ldots, m_\mu.$$

Hence, for $r \neq \mu$ obtain equations for correlation moments of components of elements of various types

$$\frac{d}{dt} M[X_{rs} X_{\mu\nu}] = \sum_{j=1}^{m_\mu} M[\dot{X}_{rs}(\lambda_{\mu j\nu}(x) - \lambda_{\mu\nu j}(x))] + \sum_{j=1}^{m_r} M[\dot{X}_{\mu\nu}(\lambda_{rjs}(x) - \lambda_{rsj}(x))]. \tag{3.4}$$

For $r = \mu$ and $\varrho \neq \nu$, equations result for correlation moments of components of the same types of elements

$$\frac{d}{dt} M[\dot{X}_{rs} \dot{X}_{r\nu}] = \sum_{j=1}^{m_r} \dot{X}_{rs}(\lambda_{rj\nu}(x) - \lambda_{r\nu j}(x)) + \dot{X}_{r\nu}(\lambda_{rjs}(x) - \lambda_{rsj}(x))] - \\ - M[\lambda_{r\nu s}(x) + \lambda_{rs\nu}(x)]. \tag{3.5}$$

Finally, for $r = \mu$ and $s = \nu$ equations result for variances of vector X components

$$\frac{\mathrm{d}}{\mathrm{d}t} M[(\dot{X}_{rs})^2] = \sum_{j=1}^{m_r} \{2M[\dot{X}_{rs}(\lambda_{rjs}(x) - \lambda_{rsj}(x))] + M[\lambda_{rjs}(x) + \lambda_{rsj}(x)]\}. \tag{3.6}$$

Like in (2.9), in all the equations above the possibility of using (2.10) should be taken into account.

Similarly, differential equations for any other system state probability distribution moments may be obtained from (3.1).

Expressions (3.1), (3.5) and (3.6) enable establishment of conditions under which $M[X_{rs}(t)] \equiv M[(\dot{X}_{rs}(t))^2]$ and which are defined by the following statement.

If for some subset A'_r of state set A_r of elements of rth type the following holds for any $j \in A'_r$

$$\lambda_{rjk}(x) = c_{rjk} X_{rj}, \; c_{rjk} \neq f(x) \; \forall \; k \in A_r, \quad \lambda_{rkj}(x) \neq f(x) \; \forall \; k \notin A'_r, \tag{3.7}$$

and if at the initial time

$$X_{rj} = 0 \; \forall \; j \in A'_r \tag{3.8}$$

then

$$\tilde{x}_{rj}(t) \equiv D_{rj}(t), \; K_{rjk}(t) = 0 \; \forall \; j, \; k \in A'_r, \quad j \neq k \tag{3.9}$$

where

$$\tilde{x}_{rj}(t) = M[X_{rj}(t)], \; D_{rj}(t) = M[(\dot{X}_{rj}(t))^2], \; k_{rjk}(t) = M[\dot{X}_{rj}(t) \dot{X}_{rk}(t)]$$

Proof. Find differential equations which should be satisfied by differences $\tilde{x}_{rs} - D_{rs}$ for $\varrho \in A'_r$ where dependence of $\tilde{x}_{rs}$ and D_{rs} on time is omitted for

the sake of briefness. Subtract (3.6) from (3.1) and substitute the transition intensities from (3.7) and (3.8) into the right side of the resulting equation

$$\dot{\tilde{x}}_{rs} - \dot{D}_{rs} = -2 \sum_{j=1}^{m_r} M[\lambda_{rsj}(X) - \mathring{X}_{rs}\lambda_{rsj}(X) + \mathring{X}_{rs}\lambda_{rjs}(X)] =$$

$$- (\tilde{x}_{rs} - D_{rs}) \sum_{j=1}^{m_r} c_{rsj} - \sum_{j \in A_2'} c_{rjs} K_{rjs}. \tag{3.10}$$

Further, find differential equations which should be satisfied by $K_{rs\nu}$ for $\varrho, \nu \in A_r'$. Substitute transition intensities from (3.7) and (3.8) into (3.5)

$$\dot{K}_{rs\nu} = \sum_{j \in A_2'} c_{rjk} K_{rsj} + \sum_{j \in A'} c_{rjs} K_{r\nu j} - K_{rs\nu} \sum_{j=1}^{m_r} (c_{r\nu j} - c_{r\varrho j}) -$$

$$C_{rs\nu}(\tilde{x}_{rs} - D_{rs}) - C_{r\nu s}(\tilde{x}_{r\nu} - D_{r\nu}). \tag{3.11}$$

It may be seen from the set of equations (3.10) and (3.11) that the derivatives of the time functions $\tilde{x}_{rs} - D_{r\varrho}(\varrho \in A_r')$ and $k_{r\varrho\nu}(\varrho, \nu \in A_r', \ \varrho \neq \nu)$ are proportional to linear combinations of the same functions. Further, it follows from (3.8) that at the initial time these functions are zero, Consequently, these functions are identical to zero, and the statement above is proved.

Mention, that this statement does not constraint intensities as functions of time.

A similar statement is proved in [2] for a particular system similar to that of Example 1, additional constraints being imposed on transition intensities as time functions.

4. Approximation of probability distribution

We have derived above Eq. (2.9) for an arbitrary moment of system state probability distribution. This equation enables writing required number of differential equations and making the resulting set of equations closed.

When the state probability distribution is characterized by few parameters, the number of equations involved into the system will be small. Otherwise, by approximating the state probability distribution by a distribution characterized by a small number of parameters, obtain as well a set with few equations.

One may assume, for example, that the state probability distribution is the multidimensional Poisson distribution. In this case, to obtain a closed system one can confine oneself to the set of equations (3.1) for mathematical expectations which may be regarded as a refined set of equations of the means dynamics.

The Poisson distribution is allowable as approximating, for instance, for unlimited number of elements in the system when conditions (3.7), 3(.8) are met for all the states of elements. Such an assumption is also allowable if number of elements of some type are limited but large enough, and state sets of elements'. of some types (the same or other) have only one state more than the set A_2'.

The refined set of equations of the means dynamics is sufficient also under the assumption that distribution of each type of elements if polynomial, independent of distributions of other types of elements.

By analogy with the equations of the means dynamics and according to [2], a closed set of equations for distribution moments will be referred to as moment dynamics equation set, and a method for investigation of element systems based on it will be referred to as the moment dynamics method.

As it will be demonstrated below, approximation of probability distribution may give fairly satisfactory accuracy in determining the required moments even for small numbers of elements in a system and for transition intensities being essentially nonlinear functions of the vector X.

This determines the difference between (3.1) and the equations of the means dynamics; moreover, (3.4)–(3.5) enable simpler generation of moment dynamics equations than methods suggested in [1, 2, 3].

Example 1. Production of some item is contemplated. Intensity of production of units of this item is a time function independent of the number of units in use. Failure intensity of each unit is also time dependent and independent of the number of units in use. It is required to determine the mathematical expectation and variance of the units in use.

Solution. Each unit may be in three states, be in production, be in use, or be failed. Denote these states as 1,2,3, respectively. A graph of the states of the units is shown in Fig. 2.

The state of the system will be characterized by a vector $X = \{X_1, X_2, X_3\}$ where X_i is the number of units in ith state.

Differential equations for the mathematical expectation $\tilde{x}_2$ and variance D_2 of units in use may be found through (3.1) and (3.6)

$$\dot{\tilde{x}}_2(t) = \lambda_{12}(t) - \tilde{x}_2(t)\, C_{23}(t),$$
$$\dot{D}_2(t) = -2D_2(t)\, C_{23}(t) + \lambda_{12}(t) + \tilde{x}_2(t)\, C_{23}(t) \qquad (4.1)$$

Solutions of this system do not present any difficulties, and will not be written here. Just mention that, for the second state, conditions (3.7),

(3.8) are met and, therefore, solutions of both equations coincide, i.e. $\tilde{x}_2(t) = D_2(t)$.

Example 2. Consider a system involving m and n pieces of elements of two types, respectively. Elements of both types may be in two states. Numbers of elements of the first/second type in both states will be denoted, respectively, X_1/y_1 and x_2/y_2. At the initial moment, $X_1 = m$, $x_2 = 0$, $y_2 = 0$. Further, the elements may pass from the first state to the second one. Intensities of transition of the elements of the first (μ_{12}) and the second (ν_{12}) types from the first state to the second one are as follows:

$$\mu_{12} = \nu Y_i \delta(x_1), \quad \nu_{12} = \mu X_i \delta(Y_1) \tag{4.2}$$

where μ and ν are constants.

It is required to determine mathematical exceptations of the numbers of elements of both types in their initial states.

As applied to the biology, the system serves as a model of fighting populations. As applied to the military science, it may be used as a model of fight of two groups. We have selected this system on the basis of the following reasons. First, element transition intensities here are essentially nonlinear functions of the number of elements in the states. Second, it was previously considered by Lanchester [4] by means of the equations of the means dynamics which are known in the literature as the equations of Lanchester. Third, the author has determined system state probabilities for this system thus enabling accuracy evaluation of the Lanchester equations and the method of moments.

There is reason to write for this system the equations of the means dynamics for mathematical expectations of relative numbers of elements in the first state $\xi = M[X_1]/m$, $\eta = M[Y_1]/n$ and for dimensioneless time $\tau = \sqrt{\mu\nu}\, t$, i.e. in the following form:

$$\dot{\xi} = -\varkappa^{-1}\eta, \quad \dot{\eta} = -\varkappa\xi \tag{4.3}$$

where $\varkappa = m\sqrt{\mu}/n\sqrt{\nu}$. One must allow here for the constraints $\dot{\xi} \leq 0$, $\dot{\eta} \leq 0$, $\xi \geq 0$, $\eta \geq 0$, i.e. the number of elements in the first state cannot grow and be negative.

Study the system by means of the moment method. Take such an approximating distribution for whose characterization mathematical expectations of the number of elements suffice.

In this case, differential equations for mean numbers of elements of both types in the first state will be sufficient. According to (2.10), (3.1) and

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

(1 2), they are as follows:

$$\begin{aligned}\frac{d}{dt}M[X_1] &= -M[\mu_{12}] = -\nu\mathcal{P}(x_1 \neq 0)\,M[Y_1 \mid X_1 \neq 0] = \\ &= \nu M[Y_1] + \nu\mathcal{P}(X_1 = 0)\,M[Y_1 \mid X_1 = 0], \\ \frac{d}{dt}M[Y_1] &= -M[\nu_{12}] = \mu\mathcal{P}(Y_1 \neq 0)\,M[X_1 \mid Y_1 \neq 0] = \\ &= -\mu M(X_1] + \mu\mathcal{P}(Y_1 = 0)M([X_1 \mid Y_1 = 0].\end{aligned} \tag{4.4}$$

The absolute distributions of elements of both types will be regarded as binomial, i.e.

$$\mathcal{P}(X_1 = x) = C_m^x U^x(1-u)^{m-x};\ \mathcal{P}(Y_1 = y)\,C_n^y\vartheta^y(1-\vartheta)^{n-y} \tag{4.5}$$

where $u(\vartheta)$ is the probability of transition of an element of the first (second) type from the first state into the second one.

Moreover, assume that

$$M[X_1 \mid Y_1 = 0] = m,\ M[y, \mid X_1 = 0] = n \tag{4.6}$$

This assumption is fairly rough, but even so the moment dynamics equations will be, as it will be seen below, much more accurate than those of the means dynamics for small numbers of elements in the system.

By substituting in (4.4) expressions for corresponding probabilities and expectations from (4.5) and (4.6) and by introducing simensionless time $\tau = \sqrt{\mu\nu}\,t$, obtain the following differential equations:

$$\dot{u} = -\varkappa^{-1}[v - (1-u)^m],\ \dot{v} = -\varkappa[u - (1-v)^n] \tag{4.7}$$

whose solution should allow for the constraints $u \geq 0$, $\dot{v} \geq 0$, $\dot{u} \leq 0$, $\dot{v} \leq 0$.

Thus, for this system there are two sets of differential equations, the means dynamics (4.3) and the moment method (4.7).

Depicted in Fig. 3 are the maximal-by-modulo errors $\delta_{\max}$ in determination of the mathematical expectations of elements of both types in the first states by means of (4.3) and (4.7). Errors shown in the figure correspond to coefficient $\varkappa = 0.99$. As it may be seen from the figure, the error of the equations of the means dynamics in determining mathematical expectations of the numbers of elements in the first state may run into 57 per cent of the total number of the elements of the given type. At the same time, the error of the moment dynamics method (4.7) is at most 12 per cent. A similar patterns are observed for $\varkappa = 0.51$ and $\varkappa = 0.81$.

References

1. *Ventsel'*, *E. S.*, Determination of party state probabilities in the dynamics of fight of numerous groups, "Morskoi Sbornik", 1962, *10* (in (Russian).
2. *Tarakanov, K. V., Ovcharov, L. A., Tyryshkin, A. N.*, Analytical methods for system study, Moscow, "Sovetskoe Radio", 1974 (in Russian).
3. *Bharucha-Reid, A. T.*, Elements of the theory of Markov Processes and their Applications. Mc. Graw-Hill, 1960.
4. *Lanchester, F. W.*, Aircraft in warfare. The down of Fourth Arm. London, 1916.

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

H—1111 Budapest XI., Goldmann György tér 3, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4—5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary — possibly in Russian if the paper is in English and *vice versa* — should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10—15 per cent of the paper.

The authors will be sent sheet-proofs which they are to eturn by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которых должен содержать рисунки, таблицы и литературу) направляются

Е. Д. Теряеву — Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский проспект, 14, Москва В-71, СССР

или

Я. Кочишу — Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета
H—1111 Будапешт XI, площадь Голдманн Д., 3,
Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре.

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4—5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру .

Статье должна предшествовать аннотация объемом 50—100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме — на русском языке, если статья написана на английском, и наоборот — должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем резюме не должен превышать 10—15% объема статьи.

Авторам высылается корректура статьи, которую они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаютс авторам.

CONTENTS • СОДЕРЖАНИЕ

Emelyanov, S. V., Larichev. O. I.: A multi-criteria approach to applied *R* and *D* planning: the case of qualitative criteria (*Емельянов С. В., Ларичев О. И.* Многокритериальный подход к проблеме планирования прикладных научных исследований и разработок при качественных критериях) 385

Пупков К. А., Мороз А. И. К теории релейно-линейных законов управления (*Pupkov, K. A., Moroz, A. I.:* Contribution to the theory of relay-linear control laws) 401

Поспелов Д. А., Черезова Н. Н., Шапиро Д. И. Распознавание ситуаций в задаче о преследовании (*Pospelov, D. A., Cherezova, N. N., Shapiro, D. I.:* Situation recognition in the pursuit problem) 421

Романов В. Г. К вопросу о методе динамики моментов (*Romanov, V. G.:* On method of moment dynamics) 437

Györfi, L.: An upper bound of error probabilities for multihypotheses testing and its application in adaptive pattern recognition (*Дерфи Л.* Верхняя граница вероятности ошибки при исследовании с несколькими гипотезами и применение результатов для адаптивного распознавания образов) 449

Kárpáti, A., Szentai, E., Ipsits, I., Hermann, I.: Computerized reliability study of thyristorized continuous power supply systems (*Карпати А., Сентаи Е., Ипшич И., Херманн И.* Анализ надежности бесперебойного источника энергии на тиристорах с помощью вычислительной машины) 459