J316.920
J4
1875

VOL. 4 * NUMBER 1
ТОМ 4 * НОМЕР 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

1975

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 four issues published per year make up a volume of some 320 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. Box 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

Президиумами Академии Наук СССР и Академии Наук ВНР было принято решение о совместном издании журнала

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал создан для обеспечения быстрейшей публикации материалов о новейших результатах научных исследований, разработок и кратких сообщений 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 издается четыре раза в год общим объёмом около 320 печатных страниц.

Журнал в первую очередь будет публиковать научные достижения 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

Volume 4, 1975

PAPERS

Almásy, G. A., Sztanó, T.: Checking and correlation of measurement on the basis of linear system model. **4,** 1, pp. 57—69 57
Bányász, Cs., Keviczky, L.: Some experiences with the PKF discrete-time identification method. **4,** 3, pp. 257—274 274
Bársony, A.: A linear learning algorithm for the purpose of identification. **4,** 1, pp. 23—33 23
Belavkin, V. P.: Optimal observation of boson signals in quantum Gaussian channels. **4,** 3, pp. 241—257 241
Bliznakov, J.: The efficiency of complex hierarchical multilevel systems with special regard to reliability. **4,** 4, pp. 327—337 327
Breuer, P., Vajta, M., Jr.: Structural character recognition by forming projections. **4,** 4, pp. 339—352 339
Bukhalev, V. A.: The analysis of the accuracy of dynamic systems changing their structure in the random time moment. **4,** 2, pp. 139—146 139
Csáki, F., El-Said Hamada: Method for design of linear optimal systems in frequency domain. **4,** 2, pp. 103—120 103
Csibi, S.: Learning under computational constraints from weakly dependent samples. **4,** 1, pp. 3—21 3
Dashevsky, M. L.: A semiinvariant method of closing the equations for moments in statistical analysis of nonlinear systems. **4,** 4, pp. 317—326 317
Gerencsér, L.: The extension of the extrapolation technique in the SUMT method. **4,** 4, pp. 311—316 311
Goppa, V. D.: Nonprobabilistic mutual information without memory. **4,** 2, pp. 97—102 97
Hetthéssy, J., Haber, R. Keviczky, L.: A separating transformation to the minimum variance control. **4,** 2, pp. 147—154 147
Jávor, A.: An approach to the modelling of uncertainties in the simulation of quasideterministic discrete event systems. **4,** 3, pp. 219—229 219
Kalyaev, A. V.: Programming structure automata. **4,** 1, pp. 35—56 35
Khodak, G. L.: On the cost sequence transmission. **4,** 4, pp. 299—310 299
Kolmanovsky, V. B.: Some problems of time optimal control for stochastic systems. **4,** 4, pp. 353—367 353
Kurzhanskii, A. B.: On minimum control and estimation strategies under incomplete information. **4,** 3, pp. 205—218 205
Marton, K.: On the rate distortion function of stationary sources. **4,** 4, pp. 289—296 289
Mikhailov, F. A., Tykhevich, O. F., Khadzhinov, M. K.: Application of structural transforms to the analysis of time-varying linear systems. **4,** 1, pp. 71—94 71
Muszély, Gy.: The reduction of the dimension of input pattern recognition vectors by searching the optimal parameter set for linear regression. **4,** 2, pp. 165—181 165
Prékopa, A.: Optimal control of a storage level using stochastic programming. **4,** 3, pp. 193—204 193
Söderström, T,: Ergodicity results for sample covariances. **4,** 2, pp. 131—138 131
Tarnay, K.: Remarks on the analysis of a stabilized analog-digital converter. **4,** 3, pp. 231—239 231

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Tsonkov, S. M.: Estimating the dynamics of modified systems with variable structure. **4**, 2, pp. 155—163 155
Voronov, A. A., Petrushinin, E. P.: Control optimization of multiple operation resources. **4**, 4, pp. 275—287 275
Zakharov, V. N., Pospelov, D. A.: An approach to estimation problems. **4**, 2, pp. 121—130 121

*

JUBILEE NUMBER

Bognár, G.: The Hungarian-Soviet scientific and technical cooperation, pp. 3—11 3
Csáki, F.: The development of the work connections between the Soviet and the Hungarian research schools in the field of control and information theory, pp. 33—49 33
Csibi, S.: Results of co-work in information theory and recognition problems, advances in the development of computer assistance to heart and vascular system examinations, pp. 65—76 65
Emelyanov, S. V., Ozernoi, V. M.: Decision-making in multi-objective problems: A survey. pp. 51—64 51
Gavrilov, M. A.: Development of the finite automata theory in the discrete circuit design. pp. 77—96 77
Pugachev, V. S.: A survey of statistical control theory. pp. 13—32 13

AUTHORS

Almásy, G. A. **4**, 1, pp. 57—69
Bányász, Cs. **4**, 3, pp. 259—274
Bársony, A. **4**,, 1, pp. 23—33
Belavkin, V. P. **4**, 3, pp. 241—257
Bliznakov, J. **4**, 4, pp. 327—337
Breuer, P. **4**, 4, pp. 339—352
Bukhalev, V. A. **4**, 2, pp. 139—146
Csáki, F. **4**, 2, pp. 103—120
Csibi, S. **4**, 1, pp. 3—21 ;
Dashevsky, M. L. **4**, 4, pp. 317—326
Gerencsér, L. **4**, 4, pp. 311—316
Goppa, V. D. **4**, 2, pp. 97—102
Haber, R. **4**, 2, pp. 147—154
Hamada, El-Said **4**, 2, pp. 103—120
Hetthéssy, J. **4**, 2, pp. 147—154
Jávor, A. **4**, 3, pp. 219—229
Kalyaev, A. V. **4**, 1, pp. 35—56
Keviczky, L. **4**, 2, pp. 147—154; **4**, 3, pp. 259—274
Khadzhinov, M. K. **4**, 1, pp. 71—96
Khodak, G. L. **4**, 4, pp. 299—310
Kolmanovsky, V. B. **4**, 4, pp. 353—367
Kurzhanskii, A. B. **4**, 3, pp. 205—218
Marton, K. **4**, 4, pp. 289—296
Mikhailov, F. A. **4**, 1, pp. 71—94
Muszély, Gy. **4**, 2, pp. 165—181
Petrushinin, E. P. **4**, 4, pp. 275—287
Pospelov, D. A. **4**, 2, pp. 121—130
Prékopa, A. **4**, 3, pp. 193—204
Söderström, T. **4**, 2, pp. 131—138
Sztanó, T. **4**, 1, pp. 57—69
Tarnay, K. **4**, 3, pp. 231—239
Tsonkov, S. M. **4**, 2, pp. 155—163
Tykhevich, O. F. **4**, 1, pp. 71—94
Vajta, M. Jr. **4**, 4, pp. 339—352
Voronov, A. A. **4**, 4, pp. 275—287
Zakharov, V. N. **4**, 2, pp. 121—130

*

JUBILEE NUMBER

Bognár, G. pp. 3—11
Csáki, F. pp. 33—49
Csibi, S. pp. 65—76
Emelyanov, S. V. pp. 51—64
Cavrilov, M. A. pp. 77—96
Ozernoi, V. M. pp. 51—64
Pugachev, V. S. pp. 13—32

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Том 4. 1975

СТАТЬИ

Алмаши Г. А., Стано Т.: Проверка и поправка результатов измерений на основе линейной модели системы. **4,** 1, стр. 57—69 57
Баняс Ч., Кевицки Л.: Применение метода дискретной идентификации, использующего дискретную информацию о структуре объекта: **4,** 3, стр. 259—274 259
Баршонь А.: Применение линейного алгоритма обучения для идентификации динамических характеристик. **4,** 1, стр. 23—33 23
Белавкин В. П.: Оптимальное наблюдение бозонных сигналов в квантовых гауссовских каналах. **4,** 3, стр. 241—259 241
Близнаков Й.: Эффективность сложных иерархических многоуровневых систем с учетом надежности. **4,** 4, стр. 327—337 327
Брейер П., Вайта М.: Структурное распознавание букв путем проектирования. **4,** 4, стр. 339—352 339
Бухалев В. А.: Анализ точности динамических систем, изменяющих структуру в случайный момент времени. **4,** 2, стр. 139—146 139
Воронов А. А., Петрушинин Е. П.: Оптимизация управления ресурсами многократного действия. **4,** 4, стр. 275—287 275
Геренчер Л.: Развитие экстраполяционных алгоритмов в методе штрафных функций. **4,** 4, стр. 311—316 311
Гоппа В. Д.: Невероятностная взаимная информация без памяти. **4,** 2, стр. 97—102 ... 97
Дашевский М. Л.: Семиинвариантный метод замыкан ияуравнений для моментов в задачах статистического анализа нелинейных систем. **4,** 4, стр. 317—326 317
Захаров В. Н., Поспелов Д. А.: Об одном подходе к решению оценочных задач. **4,** 2, стр. 121—130 121
Каляев А. В.: Автоматы с программируемой структурой. **4,** 1, стр. 35—56 35
Колмановский В. Б.: О некоторых вопросах оптимального быстродействия в стохастических системах. **4,** 4, стр. 353—367 353
Куржанский А. Б.: О минимаксных стратегиях управления — оценивания при неполной информации. **4,** 3, стр. 205—218 205
Мартон К.: Об ε-энтропии стационарных источников. **4,** 4, стр. 289—296 289
Михайлов Ф. А., Тыхевич О. Ф., Хаджинов М. К.: Применение структурных преобразований к анализу нестационарных линейных систем. **4,** 4, стр. 71—94 71
Мусели Д.: Сокращение исходного описания в задаче распознавания образов с помощью метода поиска оптимального подмножества параметров линейной регрессии. **4,** 2, стр. 165—181 165
Прекопа А.: Применение стохастического программирования для оптимального управления уровнем запаса. **4,** 3, стр. 193—204 193
Содерстром Т.: Эргодические результаты для ковариации, полученные на основе реализаций. **4,** 2, стр. 131—138 131
Тарнаи К.: Замечания к анализу стабилизированных аналого-цифровых преобразователей. **4,** 3, стр. 231—239 231
Хетеши Й., Хабер Р., Кевицки Л.: О сепарабельном преобразовании в задачах управления с минимальной среднеквадратической ошибкой. **4,** 2, стр. 147—154 ... 147

Ходак Г. Л.: О стоимости передачи последовательностей. **4**, 4, стр. 299—310 299
Цонков С. М.: Ощенка динамикки модифищированных систем с переменной структурой. **4**, 2, стр. 155—163 155
Чаки Ф., Эл-Саид Хамада: Метод конструкции линейных оптимальных систем в области частот. **4**, 2, стр. 103—120 103
Чиби Ш.: Обучение при вычислительных ограничениях на основе статистически зависимых данных. **4**, 1, стр. 3—21 3
Явор А.: О приближенном моделировании квазидетерминированных дискретных систем. **4**, 3, стр. 219—229 219

ЮБИЛЕЙНЫЙ НОМЕР

Богнар Г.: Венгеро-советское научное и техническое сотрудничество стр. 3—11 3
Гаврилов М. А.: Развитие теории конечных автоматов в области проектирования дискретных цепей. стр. 77—96 77
Емельянов, С. В., Озерной В. М.: Принятие решений в области многокритериальных проблем. Обзор. стр. 51—64 51
Пугачев В. С.: Обзор теории стохастического управления. стр. 13—32 13
Чаки Ф.: Развитие рабочих контактов между советскими и венгерскими школами научных исследований в области теории управления и информации стр. 33—49 33
Чиби Ш.: Результаты сотрудничества в области теории информации и проблем распознавания, развитие применения ЭВЦМ для исследования сердечной и кровососудной системы. стр. 65—76 65

АВТОРЫ

Алмаши Г. А. **4**, 1, стр. 57—69
Баняс Ч. **4**, 3, стр. 259—274
Баршонь А. **4**, 1, стр. 23—33
Белавкин В. П. **4**, 3, стр. 241—257
Близнаков Й. **4**, 4, стр. 327—337
Брейер П. **4**, 4, стр. 339—352
Бухалев В. А. **4**, 2, стр. 139—146
Вайта М. **4**, 4, стр. 339—352
Воронов А. А., **4**, 4, стр. 275—287
Геренчер Л. **4**, 4, стр. 311—316
Гоппа В. Д. **4**, 2, стр. 97—102
Дашевский М. Л. **4**, 4, стр. 317—326
Захаров В. Н. **4**, 2, стр. 121—130
Каляев А. В. **4**, 1, стр. 35—56
Кевицки Л. **4**, 2, стр. 147—154; **4**, 3, стр. 259—274
Колмановский В. Б. **4**, 4, стр. 353—367
Куржанский А. Б. **4**, 3, стр. 205—218
Мартон Ф. **4**, 4, стр. 289—296
Михайлов Ф. А. **4**, 1, стр. 71—94
Мусели Д. **4**, 2, стр 165—181
Петрушинин Е. П. **4**, 4, стр. 275—287
Поспелов Д. А. **4**, 2, стр. 121—130
Прекопа А. **4**, 3, стр. 193—204
Содерстром Т. **4**, 2, стр. 131—138
Стано Т. **4**, 1, стр. 57—69
Тарнаи К. **4**, 3, стр. 231—239
Тыхевич О. Ф. **4**, 1, стр. 71—94
Хабер Р. **4**, 2, стр. 147—154
Хаджинов М. Ф. **4**, 1, стр. 71—96
Хетеши Й. **4**, 2, стр. 147—154
Ходак Г. Л. **4**, 4, стр. 299—310
Цонков С. М. **4**, 2. стр. 155—163
Чаки Ф. **4**, 2, стр. 103—120
Чиби Ш. **4**, 1, стр. 3—21
Эл-Саид Хамада **4**, 2, стр. 103—120
Явор А. **4**, 3, стр. 219—229

*

ЮБИЛЕЙНЫЙ НОМЕР

Богнар Г. стр. 3—11
Гаврилов М. А. стр. 77—96
Емелянов С. В. стр. 51—64
Озерной В. М. стр. 51—64
Пугачев В. С. стр. 13—32
Чаки Ф. стр. 33—49
Чиби Ш. стр. 65—76

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

S. V. EMELIANOV (Moscow)
L. KALMÁR (Szeged)

M. A. GAVRILOV (Moscow)
I. CSISZÁR (Budapest)

A. M. LETOV (Moscow)
A. PRÉKOPA (Budapest)

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

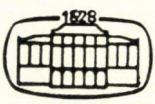
С. В. ЕМЕЛЬЯНОВ (Москва)
Л. КАЛМАР (Сегед)

М. А. ГАВРИЛОВ (Москва)
И. ЧИСАР (Будапешт)

А. М. ЛЕТОВ (Москва)
А. ПРЕКОПА (Будапешт)

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. 4 (1), pp. 3—21 (1975)

LEARNING UNDER COMPUTATIONAL CONSTRAINTS FROM WEAKLY DEPENDENT SAMPLES*

S. CSIBI

(Budapest)

(Received June 15, 1974)

Assume a classification rule has to be learned from samples for which the correct hypotheses are known, however the sample sequence is weakly dependent. Admit, for the sake of simplicity, just m_0-dependence. Suppose the relation between hypotheses and samples to be memoryless. (Both notions are specified more precisely in the paper.) We show, by extending well-known stability studies due to Braverman and Rozonoer to perturbations, including also real valued but centered components, how such sort of learning problems may be reduced to the disciplines of learning from independent observations. We show, as an illustration, how may be finite dimensional projections of Bayesian discrimination functions, defined with respect to first-order probability distributions, estimated, under such conditions, by means of the very same algorithm, independently of how the subsequent samples precisely depend on each other.

1. Introduction

Let our problem be to learn a classification rule from labelled samples $Z_n = (X_n, Y_n)$, $n = 1, 2, \ldots$ Assume $Y_n = 1$ if, for X_n, some hypothesis H_1, and $Y_n = 0$ if its complement H_0 holds true.

Suppose $X = \{X_n;\ n = 1, 2, \ldots\}$ to be a random process defined in some probability space $(\Omega, \mathfrak{F}_\Omega, \mathbf{P})$, taking values in $(\mathfrak{X}, \mathfrak{B}_\mathfrak{X})$. Assume X to be strict sense stationary and m_0-dependent.

m_0-dependence is used in the usual sense. I.e., we assume there exists some positive integer m_0 for which

$$\mathbf{P}(A'A'') = \mathbf{P}(A')\,\mathbf{P}(A'')$$

for any $A' \in \mathfrak{F}(X^n)$ and $A'' \in \mathfrak{F}({}_{n+m_0}X)$ and $n = 1, 2, \ldots$ ($X^n = \{X_i,\ i \leq n\}$ and ${}_{n-m_0}X = \{X_i,\ i > n + m_0\}$. $\mathfrak{F}(X^n)$ stands for the σ-algebra generated by X^n, and $\mathfrak{F}({}_{n+m_0}X)$ is defined in a similar way.)

We confine ourselves to m_0-dependence just for the sake of simplicity. For the probability measure of X only the afore-mentioned loose properties are assumed to be a priori known.

* Partly presented at the 3rd International Symposium on Information Theory, URSI—USSR Ac. Sc. (Tallinn, June 15—19, 1973).

Suppose X and $Y = \{Y_n;\ n = 1, 2, \ldots\}$ to be related in a memoryless manner, in the following sense

$$\mathbf{P}(Y_n = j \mid X_n, Z_{n-1}) = \mathbf{P}(Y_n = j \mid X_n) = \Pi_j(X_n)$$

for $j = 0, 1$. ($Z^n = \{Z_i, i \leq n\}$. We call, according to the usual terminology, $\Pi_j = \{\Pi_j(x), x \in \mathfrak{X}\}$ a posteriori probability function with respect to H_j.

In this paper we show how this problem, concerning weakly dependent observations, may be reduced to disciplines of learning from totally independent sample sequences. In so doing we first extend two well-known theorems, due to Braverman and Rozonoer [5], in a way appropriate for our further purposes. Next we illustrate how to use these results for learning from weakly dependent samples under a set of usual computational constraints.

2. Extending stability theorems

We adopt, concerning a pair of sequences $\widetilde{S} = \{\widetilde{S}_n;\ n = 1, 2, \ldots\}$ and Z of random variables, the following well-known [8] notions:

Definition A [8]. We say $\widetilde{S}$ to be adopted to Z if, for any n, $\widetilde{S}_n$ is measurable with respect to $\mathfrak{F}(Z^n)$. (Here and in what follows we refer by $A, B, \ldots$ and $1, 2, \ldots$ to cited and new notions, results, etc. respectively.)

Definition B [8]. A sequence $\widetilde{S}$ of random variables, adapted to Z, is centered (with respect to Z) if $\mathbf{E}(\widetilde{S}_{n+1} \mid Z^n) = 0$ for any n. (Here and in what follows $\mathbf{E}$ stands for the expectation and $\mathbf{E}(.\mid.)$, specifically, for the conditional expectation.)

Let us now turn, concerning Braverman's and Rozonoer's stability studies [5], to a fundamental result which underlies a number of theorems on learning either functions or classification rules from labelled samples.

Theorem A [5]. Given a sequence $G = \{G_n,\ n = 1, 2, \ldots\}$ of nonnegative valued random variables adapted to Z. Let (i) $\mathbf{E}G_1 < \infty$ and (ii)

$$\mathbf{E}(G_{n+1} \mid Z^n) < G_n + S_n^{(0)}. \tag{1}$$

($S^{(0)} = \{S_n;\ n = 1, 2, \ldots\}$ stands for a sequence of nonnegative valued random variables adapted to Z, such that $\sum_{n=1}^{\infty} \mathbf{E}S_n^{(0)} < \infty$.) *Then* there exists a real valued random variable ϱ for which $\mathbf{P}(\lim_{n \to \infty} G_n = \varrho) = 1$.

Let us extend Theorem A according to our further purposes.

Theorem 1. For the sequence G there still exists some ϱ for which $\mathbf{P}(\lim_{n\to\infty} G_n = \varrho) = 1$ if we replace (1) by

$$\mathbf{E}(G_{n+1} \mid Z^n) \leq G_n + S_n \tag{1'}$$

provided all other conditions in Theorem A hold. ($S_n = S_n^{(0)} + S_n^{(1)}$. $S^{(1)}$ denotes a real valued centered sequence of random variables w.r.t. Z, for which $\sum_{n=1}^{\infty} \mathbf{E}(S_n^{(1)})^2 < \infty$.)

Proof. Let

$$\tilde{G}_n = G_n - \sum_{i=1}^{n-1} S_i \tag{2}$$

for any $n > 1$, and $\tilde{G}_1 = G_1$ as in the proof of Theorem A [5]. One may readily show, exactly as in the proof of Theorem A, that $\tilde{G}$ is a supermartingale.

However, some additional considerations are needed to show that, for the supermartingale $\tilde{G}$, convergence conditions hold. From (1'), (2) and $S_n^{(0)} \geq 0$ follows:

$$\mathbf{E}\,|\tilde{G}_n| \leq \mathbf{E}G_1 \tag{3}$$

for any $n > 1$.

From (3) and $\tilde{G}_1 = G_1$ follows

$$\mathbf{E}\mid \tilde{G}_n \mid < C_0 \tag{4}$$

for any $n \geq 1$. (Here and in what follows C_i and n_i denote, for $i = 0, 1, \ldots$ a priori fixed positive, reals and integers, respectively.)

From (4) follows that for the supermartingale $\tilde{G}$ the well-known convergence theorem [8] holds. I.e., there exists some real valued random variable $\tilde{\varrho}$ for which

$$\mathbf{P}\,(\lim_{n\to\infty} \tilde{G}_n = \tilde{\varrho}\,) = 1\,. \tag{5}$$

Observe that $\mathbf{E}G_1 < \infty$ and $\sum_{n=1}^{\infty} \mathbf{E}S_n^{(0)} < \infty$. Obviously, for any real valued random variable $\tilde{S}$

$$|\,\tilde{S}\,| \leq 1 + \tilde{S}^2. \tag{6}$$

Notice that a centered sequence is also orthogonal [8]. Thus:

$$\mathbf{E}\left(\sum_{i=1}^{n} S_i^{(1)}\right)^2 = \sum_{i=1}^{n} \mathbf{E}(S_i^{(1)})^2 \tag{7}$$

for any $n > 1$.

From the monotone convergence theorem and the convergence theorem of centered processes [8], follows the existence of random variables $\varrho^{(0)}$ and $\varrho^{(1)}$, to which $\sum_{i=1}^{n-1} S_i^{(0)}$, resp. $\sum_{i=1}^{n-1} S_i^{(1)}$ tend, as $n \to \infty$, almost surely. From this, (5) and (2) follows Theorem 1.

Concerning the scope of the admitted perturbations S, one may go even slightly further. Viz. one may prove, in an almost similar way

Theorem 1′. Theorem 1 still holds if, for some positive integer $N < \infty$

$$S_n = S_n^{(0)} + \sum_{j=1}^{N} S_{n-j}^{(1,j)} \tag{8}$$

where $S^{(1,j)}$ $(j = \overline{1, N})$ stand for centered sequences of real valued random variables, adapted to Z, such that $\sum_{i=1}^{\infty} \mathbf{E}(S_i^{(1,j)})^2 < \infty$, for $j = \overline{1, N}$. ($S_i^{(1,j)} = 0$ if $i \leq 0$.)

By Theorem 1 a number of results within Braverman's and Rozonoer's stability theory [5] may be extended also to perturbations of the form (8). Here we confine ourselves just to the extention of one of these results. Viz., to

Theorem 2. Let (i) $U = \{U_n, n \geq 1\}$ and $V = \{V_n, n \geq 1\}$ be two sequences of nonnegative valued random variables, both adapted to Z, and (ii) $\mathbf{E}U_1 < \infty$. Let U and V be related by (iii) and (iv), as follows: (iii) Assume that

$$\mathbf{E}(U_{n+1} \mid Z_n) < (1 + \gamma_n) U_n - \beta_n V_n + S_n. \tag{9}$$

Here $\gamma = \{\gamma_n, n \geq 1\}$ and $\beta = \{\beta_n, n \geq 1\}$ stand for a priori given sequences of positive reals, such that

$$\sum_{n=1}^{\infty} \beta_n = \infty\,, \quad \sum_{n=1}^{\infty} \gamma_n < \infty\,. \tag{10}$$

$S = \{S_n, n \geq 1\}$ stands for a sequence of real valued random variables adapted to Z.

$$S_n = S_n^{(0)} + \sum_{j=1}^{N} S_{n-j}^{(1,j)}\,.$$

$S^{(0)}$ is nonnegative valued, and $\sum_{n=1}^{\infty} \mathbf{E}S_n^{(0)} < \infty$, $S^{(1,j)}$ $(j = 1, N)$ are centered sequences of real valued random variables, such that $\sum_{n=1}^{\infty} \mathbf{E}(S_{n-j}^{(1,j)})^2 < \infty$, for $j = \overline{1, N}$.

(iv) Suppose that, for any subsequence $\{n_i;\ i = 1, 2, \ldots\}$ of subscripts for which $\mathbf{P}(\lim_{i\to\infty} V_{n_i} = 0) = 1$ we have $\lim_{i\to\infty} \mathrm{p}\ U_{n_i} = 0$. ($\lim p$ stands for convergence in probability.)

Then

$$\mathbf{P}\,(\lim_{n\to\infty} U_n = 0) = 1\,.$$

Remark [5], [11]. It is well known that (iv) is met if

$$\inf_{\varepsilon<n<\varepsilon^{-1}} V_n > 0 \tag{11}$$

for any $\varepsilon > 0$ and n.

(11) is met within a broad class of regression problems with unimodal risk functions [5], [11].

Theorem 1 is a typical case of that sort of results in which (i) a pair U and V of nonnegative valued sequences of random variables, adapted to Z, are considered, and (ii) the joint properties of U and V guarantee that either U or V, or both tend to zero, almost surely. We may call, because of obvious reasons, such sort of results Lyapunov type theorems, according also to a remark by Braverman and Rozonoer [5].

Sketch of the proof. One may show, in a way almost similar to Braverman's and Rozonoer's proof [5] concerning specifically $S = S^{(0)}$, that for

$$G_n = U_n \prod_{i=n}^{\infty} (1 + \gamma_i) \tag{12}$$

Theorem 1′ holds. Thus there exists some real valued random variable ϱ for which $\mathbf{E}(\lim_{n\to\infty} G_n = \varrho) = 1$ almost surely. From this, (10) and (12) follows, in the well known way [5], [11]

$$\mathbf{P}\,(\lim_{n\to\infty} U_n = \varrho) = 1\,. \tag{13}$$

However, some additional considerations are still needed for proving that $\varrho = 0$, almost surely. From (12), (9) and (10)

$$\mathbf{E}G_{n+1} = \mathbf{E}[\mathbf{E}(G_{n+1} \mid Z^n)] < \mathbf{E}G_n - \beta_n \mathbf{E}V_n + C_1 \mathbf{E}S_n\,. \tag{14}$$

Let us use (14) for overbounding $\mathbf{E}G_{n-1}, \ldots, \mathbf{E}G_1$ after another. Then

$$\mathbf{E}G_{n+1} \leq \mathbf{E}G_1 - \sum_{i=1}^{n} \beta_i\, \mathbf{E}V_i + C_1 \sum_{i=1}^{n} \mathbf{E}S_i\,. \tag{15}$$

Observe that $S_n = S_n^{(0)} + \sum_{j=1}^{N} S_{n-j}^{(1,j)}$ where $S_n^{(0)} \geq 0$. Thus

$$\sum_{i=1}^{n} \mathbf{E}S_i \leq \sum_{i=1}^{\infty} \mathbf{E}S_1^{(0)} + \mathbf{E} \sum_{i=1}^{n} \sum_{j=1}^{N} S_{i-j}^{(i,j)} \leq \sum_{i=1}^{\infty} \mathbf{E}S^{(0)} + \sum_{j=1}^{n} \mathbf{E} \left| \sum_{i=1}^{N} S_{i-j}^{(1,j)} \right| . \quad (16)$$

We recall $\sum_{n=1}^{\infty} \mathbf{E}S_n^{(0)} < \infty$ and $\sum_{n=1}^{\infty} \mathbf{E}(S_n^{(1,j)})^2 < \infty$.

From these, (16) and (6)

$$\sum_{n=1}^{\infty} \mathbf{E}S_i \leq \sum_{i=1}^{\infty} \mathbf{E}S_i^{(0)} + N + \sum_{j=1}^{N} \sum_{i=1}^{\infty} \mathbf{E}(S_{i-j}^{(1,j)})^2 < C^2 < \infty . \quad (17)$$

Observe that $\mathbf{E}U_1 < \infty$ and $\sum_{n=1}^{\infty} \gamma_n < \infty$. Then from (12) we have $\mathbf{E}G_1 < \infty$. From this, (17) and (15) follows

$$\sum_{n=1}^{\infty} \beta_n \mathbf{E}V_n < \infty .$$

From this point on the proof exactly follow well-known ideas [3], [5]. Viz., by observing $\sum_{n=1}^{\infty} \beta_n = \infty$ and $V_n \geq 0$ $(n = 1, 2, \ldots)$ follows the existence of some subsequence $\{n_k, k \geq 1\}$ for which $\lim_{k \to \infty} \mathbf{E}V_{n_k} = 0$. By Chebyshev's theorem we obtain $\lim_{k \to \infty} \mathrm{p}\ V_{n_k} = 0$. From this, however, follows that some subsequence $\{n_{k_i}, i \geq 1\}$ exists for which $\mathbf{P}(\lim_{i \to \infty} V_{n_{k_i}} = 0) = 1$.

By this and Condition (iv) (Theorem 2) follows $\lim_{k \to \infty} \mathrm{p}\ U_{n_{k_i}} = 0$. I.e., there exists some $\{n_{k_i}(l), l \geq 1\}$ for which $\mathbf{P}(\lim_{l \to \infty} U_{n_{k_i}(l)} = 0) = 1$. This fact and (13) imply $\varrho = 0$, almost surely. This completes the proof of Theorem 2.

3. Learning projections from weakly dependent samples

Let $\mathbf{P}_{X_1}$ be the probability distribution of, say, X_1. Assume that $\mathbf{P}_{X_1}$ is absolutely continuous with respect to some a priori given finite measure μ, defined on $(\mathfrak{X}, \mathfrak{B}_{\mathfrak{X}})$. I.e., $p = \mathbf{P}_{X_1}(\mathrm{d}x)/\mu(\mathrm{d}x) < C_2 < \infty$.

Assume that a finite set $\Phi = \{\varphi_j, j = 1, \overline{M}\}$ of linearly independent functions is given on $\mathfrak{X}$, such that $\varphi_j \in \mathfrak{L}_2(\mu)$ and $|\varphi_j(x)| < C_3 < \infty$ for all j and x. Denote by $\mathfrak{M}(\Phi)$ the linear manifold spanned by Φ.

Recall that g^* is the projection of any function $g = \{g(x), x \in \mathfrak{X}\} \in \mathfrak{L}_2(\mathbf{P}_{X_1})$ onto $\mathfrak{M}(\Phi)$ iff

$$\int_{\mathfrak{X}} [\tilde{g}(x) - g(x)]^2 \mathbf{P}_{X_1}(dx) = \min \tag{19}$$

for $\tilde{g} = g^*$ (for any $\tilde{g} \in \mathfrak{M}(\Phi) < \mathfrak{L}_2(\mathbf{P}_{X_1})$). Observe that, because of $\mathbf{P}_{X_1} \ll \mu$ and $p(x) < C_2(x \in \mathfrak{X})$ we have, for any $g \in \mathfrak{L}_2(\mu)$, $g \in \mathfrak{L}_2(\mathbf{P}_{X_1})$.

Let us associate with $\mathfrak{M}(\Phi)$ also another inner product, in addition to that of $\mathfrak{L}_2(\mu)$. Let us define this, as usually [6], [1], by $\langle g, \tilde{g} \rangle = \sum_{j=1}^{M} C_j \tilde{C}_j$, for any $g = \sum_{j=1}^{M} C_j \varphi_j$ and $\tilde{g} = \sum_{j=1}^{\mu} C_j \tilde{C}_j$ in $\mathfrak{M}(\Phi)$. Define $\mathfrak{H}$ by $\mathfrak{M}(\Phi)$ and $\langle .,. \rangle$ (Let $\| g \|^2 = \langle g, g \rangle$.)

Let our goal be to estimate, at any $n \geq 1$, the projection Π^* of $\Pi = \Pi_1 - \Pi_0$ consistently, in the sense of $\mathfrak{H}$. Viz., let $\mathbf{E}(\lim_{n \to \infty} \| F_n - \Pi^* \| = 0) = 1$. ($F_n$ stands for the estimate we propose at n.)

Let $B(g) = \{x : g(x) \geq 0\}$. Consider the following classification rule: Accept H_1 if $x \in B(g)$ and H_0 otherwise. Denote by $\mathbf{P}_e(g)$ the probability of misclassification, meant specifically with respect to the first order probability distribution $\mathbf{P}_{X_1}$. Viz.,

$$\mathbf{P}_e(g) = \int_{B(g)} \Pi_0(x)\, p(x)\, \mu(dx) + \int_{B^c(g)} \Pi_1(x)\, p(x)\, \mu(dx)\, M =$$

$$= P_1 - \int_{\mathfrak{X}} \chi_{B(g)}\, \Pi(x)\, p(x)\, \mu(dx)\,.$$

($P_1 = \mathbf{P}(\omega : Y_0 = 1)$. $\chi_{B(g)}$ stands for the indicator of $B(g)$, and superscript c denotes the complement.))

Assessing the classification rule, specifically in terms of this very probability of misclassification, appears frequently as meaningful if the rule, learned from samples up to n, has to be used only far behind n.

We consider, specifically, Π just for illustration. One could equally well take any other real valued Bayesian discrimination function, e.g. Πp, by which a decision rule (i.e. a subdivision of $\mathfrak{X}$) with a minimum misclassification probability may be defined, in a similar way, by level crossing.

Consistency of $F = \{F_n,\ n \geq 1\}$ with respect to $\mathfrak{H}$ is not meaningless in testing H_1 against H_0. Viz., we have

Lemma A [12]. *If* $\mathbf{P}(\lim_{n \to \infty} \| F_n - \Pi^* \| = 0) = 1$, then

$$\mathbf{P}\big(\lim_{n \to \infty} [P_e(F_n) - P_e(\Pi)] \leq C_3 \| \Pi - \Pi^* \|\big) = 1\,.$$

Remark. Lemma A follows from the following two facts, summarized as Lemmas B and C.

Lemma B [15], [12]

$$P_e(g) - P_e(\Pi) \leq \sqrt{\int_{\mathcal{X}} (g(x) - \Pi(x))^2 \mathbf{P}_{X_1}(dx)}$$

for any $g \in \mathfrak{L}_2(\mu)$.

Lemma C [7], [10]. If $| \varphi_j(x) | < C_3$ for all $x \in \mathcal{X}$ and $j = \overline{1, M}$ *then*

$$g(x) - \Pi^*(x) < C_3 \, || \, g - \Pi^* \, ||. \tag{20}$$

Remark. Lemma B follows from the definition of $P_e(g)$, the property that $[\chi_{B(g)}(x) - \chi_{B(\Pi)}(x)]\, g(x) \geq 0$ and Schwartz's inequality. Lemma C readily follows from $| \varphi_j(x) | < C_2 (x \in \mathcal{X})$, the definition of $\mathfrak{M}(\Phi)$ and $\langle .,. \rangle$. (By the way, it represents a well-known property of Reproducing Kernel Hilbert Spaces [2], [7] in a very simple context. Viz., observe that $\mathcal{H}$ is the Reproducing Kernel Hilbert Space generated by $K_0(x, \tilde{x}) = \sum_{j=1}^{M} \varphi_j(x)\, \varphi_j(\tilde{x})$ and that $||K_0|| = K_0(x, x) < C_4$. (20) represents a typical property of such spaces.) Lemma A immediately follows from Lemmas B and C, taking $g = F_n$ and observing the orthogonality of $\Pi - \Pi^*$ to $g - \Pi^*$ with respect to $\mathbf{P}_{X_1}$.

Assume, some subset $\mathcal{A}$ in $\mathcal{H}$ is also a priori given for which $\Pi^* \in \mathcal{A}$. (It is essential that $\mathcal{A}$ need not be a subspace of $\mathcal{H}$. E.g., $|| \mathcal{A} || < \infty$ is also admitted, which choice obviously offers the possibility of considering also range limitations appearing in actual computations.)

Consider estimates $F = \{F_n, n \geq 0\}$ generated recursively, taking some initial estimate $F_0 \in \mathfrak{M}(\Phi)$ $(\mathbf{E} \, || \, F_0 \, ||^2 < \infty)$ and proceeding as

$$F_{n+1} = \psi(F_n + \alpha_n \, W_{n+1}) \tag{21}$$

for any $n \geq 0$.

By $W = \{W_n, \; n \geq 1\}$, which we confine to be adapted to Z, the estimate may be updated, at each n, according to recent experience. $\alpha = \{\alpha_n, \; n \geq 0\}$ stands for an a priori given sequence of positive integers, by the choice of which the corrections may, at each step, be weighted. By the mapping ψ: $\mathcal{H} \to \mathcal{A} \subset \mathcal{H}$ the estimates $F_n (n \geq 1)$ may be kept within $\mathcal{A}$. Let us interpret F_0 as the first term of the sequence Z, as one arrives, in this way, at an F which is adapted to Z.

Condition A [11], [15]. We consider, as previously, such an α for which

$$\sum_{n=0}^{\infty} \alpha_n = \infty, \quad \sum_{n=0}^{\infty} \alpha_n^2 < \infty$$

and

$$\sum_{n=0}^{\infty} \alpha_n n^{-\varepsilon} < \infty, \quad \sum_{n=0}^{\infty} n^{2C_1} \alpha_n \left(\max_{n-[n^{C_1}] \leq i \leq n} \alpha_i \right) < \infty$$

for any $\varepsilon > 0$ and some $0 < C_s < 1$. ($\lceil \tilde{x} \rceil = 1 + \text{ent } \tilde{x}$, for any real x.)

Remark. Condition A is obviously met for, say, $\alpha_n = C_6/(C_7 + n)$.

Definition C [11], [15]. We say that ψ is a non-expansive truncation with respect to $\mathcal{A}$ iff

$$|| \psi(\xi) - \eta || < || \xi - \eta || \text{ and } \psi(\eta) = \eta$$

for any $\xi \in \mathcal{H}$ and $\eta \in \mathcal{A}$.

Example [11], [15]. Let $\mathcal{A}$ be a hypercube, defined by

$$\mathcal{A} = \{\eta : a_j \leqq \langle \eta, \varphi_j \rangle \leqq b_j; j = \overline{1, M}\}$$

for a priori given reals a_j, b_j ($j = \overline{1, M}$). A non-expensive truncation ψ is defined with respect to $\mathcal{A}$ if we take, as $\psi(\xi)$, the nearest neighbour of ξ on the boundary surface of $\mathcal{A}$, provided $\xi \notin \mathcal{A}$.

It is well known how to estimate the projection Π^* of Π consistently, provided X is totally independent. Viz., a Robbins—Monro type process with a correction term

$$W_{n+1} = W_{n+1}^{(0)} = [2Y_{n+1} - 1 - F_n(X_{n+1})]\, K_0(\cdot, X_{n+1}) \tag{22}$$

turns out to be appropriate for such purpose [6]. $K_0(x, y) = \sum_{j=1}^{M} \varphi_j(x)\, \varphi_j(y)$. Recall that one may arrive at $W^{(0)}$ by (i) taking in (19) $g = \Pi$, and (ii) observing, in such equalities as (70), that Y_{n+1} is a conditionally unbiased estimate of $\Pi(X_{n+1})$ in the sense:

$$\mathbf{E}(Y_{n+1} \mid X_{n+1}, Z^n) = \mathbf{E}(Y_{n+1} \mid X_{n+1}) = \Pi(X_{n+1}).$$

In this paper we prove that the same learning goal, viz. Π^*, is actually learned, by $W^{(0)}$, for any X which is m_0-dependent (including of course, also total independence).

Theorem 3. Let (i) $\Pi^* \in \mathcal{A} \subset \mathfrak{M}(\Phi) \subset \mathfrak{L}_2(\mu)$ and $|| \mathcal{A} || < \infty$. (ii) Generate estimates, taking at $F_0 \in \mathfrak{M}(\Phi)$ ($\mathbf{E} || F_0 ||^2 < \infty$) and adopting (21), $W = W_{(0)}$ and an α according to Condition A. (iii) Let ψ be a non-expansive truncation,

with respect to $\mathcal{A}$. *Then*

$$\mathbf{P}(\lim_{n\to\infty} \| F_n - \Pi^* \| = 0) = 1 \tag{23a}$$

and

$$\mathbf{P}\Big(\overline{\lim_{n\to\infty}} [P_e(F_n) - P_e(\Pi)] \leq \sqrt{\int_{\mathfrak{X}} [\Pi(x) - \Pi^*(x)]^2 \mathbf{P}_{X_1}(dx)}\Big) = 1 . \tag{23b}$$

Remark. For the proof see Section 6. We are concerned in the present paper only with m_0-dependent observations just for the sake of a simple illustration. One may show, almost in the same way, that the very same learning goal, viz. Π^* , may also be estimated in the sense of (23a) and (23b) by $W^{(0)}$ even if the sample sequence is uniformly strongly mixing (with some given decay bound [15]).

4. On consistent rounded estimates

As all studies have been confined until now only to $F_n \in \mathfrak{M}(\Phi)$ and $\| \mathcal{A} \| < \infty$, one may obviously be content with conditions in Theorem 3, also from a computational point of view, as far as analog computations are considered.

However, one may also adopt some of our previous studies also for deriving consistent estimates $F = \{F_n, n \geq 0\}$ of Π^* which take, at each step, values in some finite set $\tilde{\mathcal{A}}_n \in \mathcal{A} \in \mathfrak{M}(\Phi)$ ($|\tilde{\mathcal{A}}_n| < \infty$), and make in this way the process, at each step, also digitally computable [13].

We recall some notions of [13]. Let $\{\tilde{\mathcal{A}}_n;\ n = 0, 1, \ldots\}$ be a sequence of finite sets, all within $\mathcal{A}$. Assume that (i) $\tilde{\mathcal{A}} \supset \tilde{\mathcal{A}}_{n+1} \supset \mathcal{A}_n$ and (ii) ψ is such that $\psi(\eta) \in \mathcal{A}_n$ if $\eta \in \mathcal{A}_n$ $(n = 0, 1, \ldots)$. (Constraint (ii) may obviously be met by an appropriate specialization of the pair $(\mathcal{A}, \psi)$, devised according to the Example.)

Let $\tilde{W}^{(0)} = \{\tilde{W}_n^{(0)},\ n \geq 0\}$ be such that $\alpha_n \tilde{W}_{n+1}^{(0)} \in \tilde{\mathcal{A}}_n$ and

$$\alpha_n \tilde{W}_{n+1}^{(0)} = \alpha_n W_{n+1}^{(0)} + \tilde{T}_{n+1} . \tag{24}$$

Let the sequence $\tilde{T}$ of round-off errors be such that

$$\sum_{n=0}^{\infty} \mathbf{E} \| \mathbf{E}(\tilde{T}_{n+1} | W_{n+1}^{(0)}) \| < \infty \tag{25}$$

$$\sum_{n=0}^{\infty} \mathbf{E} \| \tilde{T}_{n+1} \|^2 < \infty . \tag{26}$$

Theorem 4. Let $F_0 \in \widetilde{\mathcal{R}}_0$. We still have $\mathbf{P}(\lim_{n\to\infty} ||F_n - \Pi^*|| = 0) = 1$ if $W_n^{(0)}$, in Theorem 3, is replaced, at each n, by $\widetilde{W}_n^{(0)}$.

Remark. For the proof see Section 6. We have already pointed out in [13] how conditions on the rounded correction term $\widetilde{W}^{(0)}$ may be met by well-known simple rules for the round-off. (Here we just refer to [13]).

5. Admitting more general estimation rules

Next we admit, as a correction term, not just $W^{(0)}$ but also more general sort of recursive estimates of the form (21). Viz., we admit within (21) any correction term of the following form:

$$W_{n+1} = w_{n+1}(Z_{n+1}, F^n)\,. \tag{27}$$

Here $w = (\{w_{n+1}(z, f^n), z \in \mathfrak{Z} = (\mathfrak{X} \times \mathfrak{Y}, f^n \in \mathbf{X}_0^n \mathcal{H}\}, n \geq 0)$ ($\mathfrak{Y} = \{0, 1\}$ denotes some a priori fixed sequence of functions, taking values in $\mathcal{H}$. More specifically, we admit any w according to Conditions 1–4. (One may, of course, have also specifically $w_n = w_0$, for all $n \geq 1$. See, e.g., $W = W^{(0)}$.)

Condition 1. We confine ourselves to functions $w_n (n \geq 1)$ which are uniformly bounded in $\mathcal{H}$. Viz.,

$$||w_n(z, f^n)|| < C_8 \tag{28}$$

(for all $z \in \mathfrak{Z}$, $f^n \in \mathbf{X}_0^n \mathcal{H}$ and $n \geq 1$).

Condition 2. $w_n(z, f^n)$ $(n \geq 1)$ are assumed to be of decaying memory, with respect to f^n, in the following sense:

$$||w_n(z, f^n) - w_n(z, f_{n-\nu}^n)|| \leq C_9\, \nu^{-C_{10}} \tag{29}$$

for some C_9 and C_{10} and any $z \in \mathfrak{Z}$, $f^n \in \mathbf{X}_0^n \mathcal{H}$ and $\nu \geq n_1$. (Here $f_{n-\nu}^n = \{\tilde{f}_i, i \leq n\}$ assuming $\tilde{f}_i = f_i$ for $n - \nu < i \leq n$, and $\tilde{f}_i = 0$ for $i \leq n - \nu$.)

Definition 1. Let, for any $z \in \mathfrak{Z}$, $f^n \in \mathbf{X}_0^n \mathcal{H}$, $g \in \mathcal{H}$ and $n \geq 1$:

$$v_{n+1}(z, f^n) = \langle f_n - \Pi^*, w_{n+1}(z, f^n) \rangle\,. \tag{30}$$

Condition 3. $v_n(z, f^n)$ $(n \geq 1)$ are assumed to meet, with respect to f^n, the following Lipschitz type condition:

$$|v_{n+1}(z, f^n) - v_{n+1}(z, \tilde{f}^n)| \leq C_{11}\, \delta_n \tag{31}$$

(for any $z \in \mathfrak{Z}$, a pair f^n, $\tilde{f}^n \in \mathbf{X}_n^0 \mathcal{H}$ and some C_{11}). Here

$$\delta_n = \max_{0 \leq i \leq n} ||f_i - \tilde{f}_i||\,. \tag{32}$$

Definition 2. Let

$$U_n = ||F_n - \Pi^*||^2 \tag{33}$$

and

$$V_n^{(0)} = -\int_{\mathfrak{Z}} v_{n+1}(z, f^n[F_{\tilde{n}}])\, \mathbf{Q}(dz)\,. \tag{34}$$

Here $\mathbf{Q}$ denotes the probability distribution of, say, Z_1 and

$$f_n[F_{\tilde{n}}] = \{f_i = F_{\tilde{n},i} \leq n,\ \tilde{n} \leq n\}\,.$$

Condition 4. Assume that one may associate with $V^{(0)}$ a pair of real valued sequences, viz. $V^{(1)} = \{V_n^{(1)}, n \geq 1\}$ and $V = \{V_n,\ n \geq 1\}$, of real valued random variables such that

$$V_n^{(0)} = V_n^{(1)} + \tilde{G}_n$$

$$\sum_{n=0}^{\infty} \alpha_n \mathbf{E}\,|\,\tilde{G}_n\,| < \infty \tag{35}$$

and, for some C_{100},

$$0 < V_n C_{100} \leq V_n^{(1)}$$

and V is according to (11).

Theorem 5. Theorem 3 still holds if $W^{(0)}$ is replaced by any W, according to Conditions 1—4, however the conditions imposed in Theorem 3 are left otherwise unchanged.

Remark. For the proof see Section 6. We confine ourselves to $\mathcal{H} = [\mathfrak{M}(\Phi), \langle .\,, .\rangle]$, because of the particular computational significance of this very choice. However a more general study of Theorem 5, concerning any separable Hilbert space, could also be possible, along the lines followed in two of our previous studies [11], [15].

6. Proving Theorems 3—5

We start at proving Theorem 5. Next to this we will prove that Theorems 3 and 4 are just specific cases. The proof of Theorem 5 relies mainly upon Lemma 1.

Lemma 1. If (i) F is generated, specifically, by some rule of the form (21), (ii) α is according to Condition A, (iii) $||\mathcal{A}|| < \infty$ and ψ is non-expansive, with respect to $\mathcal{A}$, and (iv) W is according to Conditions 1—3, *then*

$$\mathbf{E}(U_{n+1}\,|\,Z^n) \leq U_n - 2\alpha_n V_n^{(0)} + S_n\,. \tag{36}$$

Here $S = S^{(0)} + S^{(1)}$. $S^{(0)}$ stands for a sequence of nonnegative valued random variables adapted to Z. $S_n^{(1)} = \Sigma_{j=1}^{m_0} S_{n-j}^{(1,j)}$ denotes, for any $j = \overline{1, m_0}$, a real valued centered sequence of random variables adapted to Z. $S^{(0)}$ and $S^{(1)}$ are such that

$$\sum_{n=0}^{\infty} \mathbf{E} S_n^{(0)} < \infty \tag{37}$$

and

$$\sum_{n=0}^{\infty} \mathbf{E}(S_{n-j}^{(1,j)})^2 < \infty \,. \tag{38}$$

($j = \overline{1, m_0}$.)

Proof. Let $\tilde{F}_n = F_n - \Pi^*$ and $\hat{F}_{n+1} = F_n + \alpha_n W_{n+1}$. As (i) ψ is nonexpansive with respect to $\mathcal{A}$. (ii) $F_{n+1} = \psi(\hat{F}_{n+1})$ and (iii) $\Pi^* \in \mathcal{A}$, we have:

$$\|\tilde{F}_{n+1}\|^2 \leq \|\hat{F}_{n+1} - \Pi^*\| \,.$$

Thus

$$\|\tilde{F}_{n+1}\|^2 \leq \|\tilde{F}_n + \alpha_n W_{n+1}\|^2 \,. \tag{39}$$

As F_n is $\mathfrak{F}(Z^n)$ measurable (under the convention $Z_0 = F_0$).

$$\mathbf{E}(U_{n+1} \,|\, Z^n) \leq U_n + 2\alpha_n \mathbf{E}(\langle F_n, W_{n+1}\rangle \,|\, Z^n) + \alpha_n^2 \mathbf{E}(\|W_{n+1}\|^2 \,|\, Z^n) \tag{40}$$

and, because of the same property

$$\mathbf{E}(\langle \tilde{F}_n, W_{n+1}\rangle \,|\, Z^n) = \langle \tilde{F}_n, \mathbf{E}(W_{n+1} \,|\, Z^n) \,. \tag{41}$$

Let

$$H_n = \mathbf{E}\big([w_{n+1}(Z_{n+1}, F^n) - w_{n+1}(Z_{n+1}, F_{n-\nu}^n)] \,|\, Z^n\big) \tag{42}$$

$$G_n = \mathbf{E}\big(\big[v_{n+1}(Z_{n+1}, F_{n-\nu}^n) - v_{n+1}(Z_{n+1}, f_{n-\nu}^n [F_{n-m_0}])\big] \,|\, Z^n\big) \tag{43}$$

$$\tilde{H}_n = \mathbf{E}\big(\big[w_{n+1}(Z_{n+1}, f_{n-\nu}^n [F_{n-m_0}]) - w_{n+1}(Z_{n+1}, f^n [F_{n-m_0}])\big] \,|\, Z^n\big) \,. \tag{44}$$

Here $f_{n-\nu}^n[F_{\tilde{n}}] = \{f_i,\ i \leq n\}$ assuming $f_i = F_{\tilde{n}}$ for $n - \nu < i \leq n$, and $f_i = 0$ for $i \leq n - \nu$. ($\tilde{n} = n - m_0$, $\nu \geq m_0$.)

Accordingly, one may introduce, concerning (40), the following approximation:

$$\mathbf{E}(U_{n+1} \,|\, Z^n) \leq U_n + 2\alpha_n \langle \tilde{F}_n, \mathbf{E}\{w_{n+1}(Z_{n+1}, f^n [F_{n-m_0}]) \,|\, Z^n\}\rangle +$$

$$+ 2\alpha_n [\langle \tilde{F}_n, H_n + \tilde{H}_n\rangle + \hat{G}_n] + \alpha_n^2 \mathbf{E}(\|W_{n+1}\|^2 \,|\, Z^n) \,. \tag{45}$$

Let us estimate the third term (viz., the approximation error) within the bound in (45). Consider first $\langle \tilde{F}_n, H_n + \tilde{H}_n\rangle$. Let $\nu = \lceil n^{C_5} \rceil$. ($0 < C_5 < 1/2$).

By this and Condition 2:

$$\left.\begin{matrix}\|H_n\| \\ \|\tilde{H}_n\|\end{matrix}\right\} \leq C_9\, n^{-C_{11}} \tag{46}$$

(for all $n \geq n_2 > \max\,(n_1^{1/C_5},\ m_0^{1/C_5})$ and $C_{11} = C_5\, C_{10}$).

By Schwartz's inequality and (46):

$$|\langle \tilde{F}_n, H_n + \tilde{H}_n \rangle| \leq C_9\, \|\tilde{F}_n\|\, n^{-C_{11}} \tag{47}$$

(for all $n \geq n_2$). As $F_n \in \mathfrak{H}$, for $n \geq 1$, and $\|\mathfrak{H}\| < \infty$, we have:

$$|\langle \tilde{F}_n, H_n + \tilde{H}_n \rangle| \leq C_{12}\, n^{-C_{11}} \tag{48}$$

(for $n \geq \max(n_2, 1)$).

Consider the other error term G_n. By the Triangle Inequality and because ψ is non-expansive with respect to $\mathfrak{H}$:

$$\|F_n - F_i\| \leq \sum_{k=i}^{n-1} \|F_{k+1} - F_k\| \leq \sum_{k=i}^{n-1} \alpha_k\, \|W_{n+1}\|\,. \tag{49}$$

As all members of F_{n-v}^n and $f_{n-v}^n[F_{n-m_0}]$ equal, for $i \leq n - v$, zero:

$$\delta_n = \max_{n-v<i\leq n} \|F_{n-m_0} - F_i\| \leq \sum_{i=n-v+1}^{n-1} \|F_{n-m_0} - F_i\|\,. \tag{50}$$

By (50), (51), Condition 1 and $v = \lceil n^{C_5} \rceil$:

$$\delta_n \leq C_8\, n^{2C_6} \max_{n-\lceil c_5 \rceil \leq i \leq n} \alpha_i\,. \tag{51}$$

Considering (43), (51) and Condition 3:

$$|\hat{G}_n| \leq C_{13}\, n^{2C_6} \max_{n-\lceil n^{C_5} \rceil \leq i \leq n} \alpha_i \quad (C_{13} = C_8 \cdot C_{12})\,. \tag{52}$$

By (48) and (52)

$$\alpha_n(|\langle F_n, H_n + \tilde{H}_n \rangle| + |\hat{G}_n|) \leq S_n^{(0,0)}. \tag{53}$$

Here

$$S_n^{(0,0)} = C_{12}\, \alpha_n\, n^{-C_{11}} + C_{13}\, \alpha_n\, n^{2C_5} \max_{n-\lceil n^{C_5} \rceil < i \leq n} \alpha_i\,. \tag{54}$$

$S^{(0,0)}$ is a nonnegative valued sequence of a priori fixed terms, and is, therefore, trivially adapted to Z. By (54) and Condition A: $\sum_{n=0}^{\infty} \mathbf{E}\, S_n^{(0,0)} < \infty$.

Let us now proceed by approximating, in (45), $\mathbf{E}\{w_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]|Z^n\}$ by conditional expectations of of the very same random variable, replacing the condition Z^n by $Z^{n-1}, \ldots, Z^{n-m_0}$ after another.

By the notion of the Conditional Expectation we have, for any integrable random variable $\tilde{S}$:

$$\mathbf{E}[\mathbf{E}(\tilde{S} \mid Z^n) \mid Z^{n-1}] = \mathbf{E}(\tilde{S} \mid Z^{n-1}) \tag{55}$$

almost surely.

Adopt (55) to the following random variable:

$$\tilde{S}_n = 2\alpha_n \langle \tilde{F}_n, w_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]) \rangle \tag{56}$$

(for any n).

Let

$$\begin{aligned} S_n^{(1.1)} &= \mathbf{E}(\tilde{S}_n \mid Z^n) - \mathbf{E}(\tilde{S}_n \mid Z^{n-1}) \\ &\vdots \\ S_{n-i+1}^{(1,i)} &= \mathbf{E}(\tilde{S}_n \mid Z^{u-i+1}) - \mathbf{E}(\tilde{S}_n \mid Z^{n-i}) \\ &\vdots \\ S_{n-m_0+1}^{(1.\,m_0)} &= \mathbf{E}(\tilde{S}_n \mid Z^{n-m_0+1}) - \mathbf{E}(\tilde{S}_n \mid Z^{n-m_0}) \end{aligned} \tag{57}$$

for $n > m_0$.

By (55) follows that the sequence $S^{(1,i)} = \{S_n^{(1,i)}, n \geq i-1\}$ (i) is adapted to Z and (ii) it is also centered with respect to Z (for any $1 \leq i \leq m_0$).

By Schwartz's inequality:

$$|\langle \tilde{F}_n, w_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]) \rangle| \leq \|\tilde{F}_n\| \, \|w_n(Z_{n+1}, f^n[F_{n-m_0}])\| \,. \tag{58}$$

As $\|\mathfrak{R}\| < \infty$ we have $\|\tilde{F}_n\| < C_{14} < \infty$. By Condition 1:

$$\|w_n(Z_{n+1}, f^n[F_{n-m_0}]\| \leq C_8 \,. \tag{59}$$

Thus

$$(S_{n-i+1}^{(1,i)})^2 \leq \alpha_n^2 C_{15} \,. \tag{60}$$

By this and Condition A:

$$\sum_{n=m+1}^{\infty} \mathbf{E}(S_{n-i+1}^{(1,i)})^2 < \infty. \tag{61}$$

Thus

$$\begin{aligned} &2\alpha_n \langle \tilde{F}_n, \mathbf{E}\{w_{n+1}(Z_{n+1}, f^n[F_{n-m_0}] \mid Z^n)\} \rangle = \\ &\quad = 2\alpha_n \langle \tilde{F}_n, \mathbf{E}\{w_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]) \mid Z^{n-m_0}\} \rangle + S_n^{(1)} \,. \end{aligned} \tag{62}$$

(Here $S^{(1)}$ is adapted to Z, $S_n^{(1)} = \sum_{i=1}^{m_0} S_{n-i+1}^{(1,i)}$. $S^{(1,i)}$ is, for any $i = \overline{1, m_0}$, $n > m_0$, centered with respect to Z, and is according to (61).)

By (49), Condition 1 and Definition 1

$$2\alpha_n \langle F_n, \mathbf{E}\{w_{n+1}(Z_{n+1}, f^u[F_{n-m_0}]) \,|\, Z^{n-m_0}\} \rangle \leq$$
$$\leq 2\alpha_n \mathbf{E}\{v_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]) \,|\, Z^{n-m_0}\} + S_n^{(0,1)}.$$

Here

$$0 \leq S_n^{(0.1)} \leq C_8 \alpha_n \max_{n-[n^{C_5}] \leq i \leq n} \alpha_i .$$

As X is m_0-dependent, we have, from Lemma 2:

$$\mathbf{E}[v_{n+1}(Z_{n+1}, f_u[F_{n-m_0}] \,|\, Z^{n-m_0}] = \int_{\mathcal{Z}} v_{n+1}(z, f^n[F_{n-m_0}])\, \mathbf{Q}(dz) = -V_n^{(0)}. \quad (63)$$

Consider the following upper estimation:

$$2\alpha_n \int_{\mathcal{Z}} v_{n+1}(z, f^u[F_{n-m_0}])\, \mathbf{Q}(dz) \leq$$
$$\leq 2\alpha_n \int_{\mathcal{Z}} v_{n+1}(z, f^n[F_n])\, \mathbf{Q}(dz) + S_n^{(0,2)} \quad (S_n^{(0,2)} \geq 0). \quad (64)$$

By Condition 2, in a similar way as for (52):

$$S_n^{(0.1)} \leq C_{13} n^{2C_5} \max_{n-[n^{C_5}] \leq i \leq n} \alpha_i . \quad (65)$$

In addition observe that $\alpha_n^2 \mathbf{E}(\|W_{n+1}\|^2 \,|\, Z^n) \leq C_{17} \alpha_n^2 = S_n^{(0,3)}$. Obviously $S_n^{(0.j)} \geq 0$, and $\sum_{n=0}^{\infty} \mathbf{E} S_n^{(0,j)} < \infty$ (for $j = \overline{0,3}$).

Consider these facts, (53) and (62) through (65). By these and (45)

$$\mathbf{E}(U_{n+1} \,|\, Z^n) \leq U_n - 2\alpha_n V_n^{(0)} + S_n \quad (66)$$

with $S_n = S_n^{(0)} + S_n^{(1)}$, $S_n^{(0)} = \sum_{j=0}^{3} S_n^{(0,j)}$, $S_n^{(1)} = \sum_{i=1}^{m_0} S_{n-i+1}^{(i,j)}$. Obviously S meets all the conditions imposed in Lemma 1. This completes the proof.

Lemma 2

$$\mathbf{E}\{v_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]) \,|\, Z^{n-m_0}\} = \int_{\mathcal{Z}} v_{n+1}(z, f^n[F_{n-m_0}])\, \mathbf{Q}(dz) \quad (67)$$

almost surely. ($\mathbf{Q}$ stands for the probability distribution of any Z_n, $n \geq 1$, say Z_1.)

Proof. Consider the notion of the Conditional Expectation. Observe that X is m_0-dependent, and (X, Y) are related in a memoryless manner. Accord-

ingly, for any $A^{n-m_0} \in \mathfrak{F}(Z^{n-m_0})$ one has:

$$\int_{A^{n-m_0}} \mathbf{E}\{v_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]) \mid Z^{n-m_0}\}\, \mathbf{P}(d\omega) =$$

$$= \int_{A^{n-m_0}} v_{n+1}\big(Z_{n+1}[\omega], f^n[F_{n-m_0}(\omega)]\big)\, \mathbf{P}(d\omega) = \tag{68}$$

$$= \int_{A^{n-m_0}} \Big[\int_{\Omega} v_{n+1}\big(Z_{n+1}[\omega], f^n[F_{n-m_0}(\lambda)]\big)\, \mathbf{P}_{n+1}(d\omega)\Big]\, \mathbf{P}^{n-m_0}(d\lambda)\,.$$

(Here $\mathbf{P}_{n+1}$ and $\mathbf{P}^{n-m_0}$ are the restrictions of $\mathbf{P}$ to $\mathfrak{F}(Z_{n+1})$ and $\mathfrak{F}(Z^{n-m_0})$ respectively.)

As (i) the integrands in the first and last integrals in (68) are both $\mathfrak{F}(Z^{n-m_0})$ measurable, and $\{Z_n,\ n \geq 1\}$ is stationary in the strict sense:

$$\mathbf{E}\,\{v_{n+1}(Z_{n+1}, f^n[F_{n-m_0}]) \mid Z^{n-m_0}\} =$$

$$= \int_{\Omega} v_{n+1}(Z_{n+1}[\omega], f^x[F_{n-m_0}])\, \mathbf{P}_{n+1}(d\omega) = \tag{69}$$

$$= \int_{\mathfrak{Z}} v_{n+1}(z, f^n[F_{n-m_0}])\, \mathbf{Q}(dz).$$

This completes the proof of Lemma 2.

Proof of Theorem 5. Observe that all conditions imposed in Lemma 1 hold. Let us have a V according to Condition 4. Obviously inequality (36) as well as the the assumed feature of S is retained if $V_n^{(0)}$ is replaced by $V_n^{(1)} + \widetilde{G}_n$, and $V_n^{(1)}$ is overbounded by $-V_n C_{100}$. From which follows Condition (iii) in Theorem 2 (and, specifically, $\gamma_n = 0$ and $2\alpha_n = \beta_n$).

According to Condition 4, for the sequence V, (11) holds, which implies, in the well-known way [5], [11], Condition (iv). From $\mathbf{E} \mid\mid F_0 \mid\mid^2 < \infty$, (21), the non-expensiveness of ψ, with respect to $\mathfrak{A}$, (27) and Condition 1 follows Condition (ii) in Theorem 2. Thus all conditions imposed in Theorem 2 hold. From this, and the definition of U, follows (23a). (23b) is obtained from (23a), Lemma B and the very property that Π^* denotes the projection of Π, with respect to the first order probability distribution $\mathbf{P}_{X_1}$. This completes the proof of Theorem 5.

Proof of Theorem 3. As $\mid \varphi_j(x) \mid < C_3$, $\mid K(x, y) \mid < C_3$, for all $x,\ y \in \mathfrak{X}$. From this and $\mid\mid \mathfrak{A} \mid\mid < \infty$ follows, for $W = W^{(0)}$, Condition 1. For $W^{(0)}$ Conditions 2 and 3 also readily follow.

Observe that, for $W = W^{(0)}$,

$$V_n^{(0)} = -\int_{\mathfrak{Z}} [2y - F_n(x)]\, [F_n(x) - \Pi^*(x)]\, \mathbf{Q}(dz) =$$

$$= -\int_{\mathfrak{Z}} [\Pi(x) - F_n(x)]\, [F_n(x) - \Pi^*(x)]\, p(x)\, dx. \tag{70}$$

(Here $(x, y) = z$.) Observe that Π^* is a projection with respect to $\mathbf{P}_{X_1}$. Accordingly

$$V_n^{(0)} = \int_{\mathfrak{X}} [F_n(x) - \Pi^*(x)]^2 \, p(x) \, dx \,. \tag{71}$$

It may be readily seen that, for $V^{(0)}$, (11) and, therefore, also Condition (iv) in Theorem 2 holds. Thus all conditions imposed in Theorem 5 follow. This completes the proof.

Proof of Theorem 4. Approximate, at n, the actually used correction term $\widetilde{W}_{n+1}^{(0)}$ by $W_{n+1}^{(0)}$ according to (24). One arrives, in this way in (36), at an error term $S_n + T_{n+1}$ instead of S_n (as would be the case if $W = W^{(0)}$). However, by this the feature of S remains still according to Lemma 1. Theorem 5, therefore, holds, from which Theorem 4 obviously follows.

7. Concluding remarks

In this paper we have shown, specifically for one of the Bayesian discrimination functions, viz. for Π, that if one constrains learning to finite dimensional estimates, e.g. adopts $W = W^{(0)}$, still the finite dimensional projection Π^* of Π is consistently estimated, even if the sample sequence X is m_0-dependent.

We did so simply to demonstrate conditions under which stability properties of learning processes prove to be stable (i.e. the processes correctly defined) with respect to statistical dependence.

We proved the consistency of the sequence of estimates by an appropriate extention of Braverman's and Rozonoer's stability studies to real valued but centered perturbations, and also illustrated by this an interesting further use of centered sequence properties.

While our studies have mainly been devoted to learning under dimensional constraints, we also pointed out under what conditions may one devise consistent estimates under range limitation (viz., $|| \mathfrak{R} || < \infty$) and round-off.

We have confined ourselves just to an appropriate sort of consistency of the considered estimates. Obviously the principles presented in this paper offer also simple means to extend the investigation to the finite step behaviour of learning processes [4,] [14]. We do not enter, however, within the present context into this further topic.

References

1. *Aizerman, M. A., Braverman, E. M., Rozonoer, L. I.*, "Method potentsial'nykh funktsii i teorii obucheniya mashin", Nauka, M., 1970.
2. *Aronszajn, N.*, "Theory of reproducing kernels", Trans. Amer. Math. Soc. **63** (1950) 337—404.

3. *Blum, J. R.*, "Multidimensional stochastic approximation", Amer. Math. Stat. **25** 4 (1954) 737—744.
4. *Braverman, E. M., Pyatnitski, E. C.*, "Otsenka skorosti skhodimosti algoritmov osnovanykh na metode potentsialnykh funktsii." Avtomatika i Telemekhanika (1966) 95—112.
5. *Braverman, E. M., Rozonoer, L.*, "Skhodimost sluchoinykh protsessov v teori obucheniya mashin I." Avtomatika i Telemekhanika (1969) 57—77.
6. *Tsypkin, Ya. Z.*, "Osnovy teorii obuchayushchikhsya sistem". Nauka, M., 1970.
7. *Meschkowski, H.*, "Hilbertsche Räume mit Kernfunktion". Springer, 1962.
8. *Neveu, J.*, "Bases mathematiques du calcul des probabilities". Masson et Cie, 1964.
9. *Nevel'son, M. B., Khasminskii, R. Z.*, "Stchasticheskaya approksimatsiya i rekurrentnoe otsenivaniye". Nauka, M., 1972.
10. *Wolverton, Ch. T., Wagner, T. J.*, "Asymptotically optimal discriminant functions for pattern recognition", IEEE Trans. Inf. Theory **15** 2 (1969) 258—265.
11. *Csibi, S.*, "On iteration rules with memory in machine learning", Problems of Control and Information Theory, **1** 1 (1972) 37—50.
12. *Csibi, S.*, "Simple and Compound Processes in Machine Learning", Lecture Notes, International Center for Mechanical Sciences (to appear) (Preprint, ibid, 1971).
13. *Csibi, S.*, "Approximations in learning decision functions recursively", Trans. 6th Prague Conf. on Inf. Theory, Statistical Decision Functions, Random Processes, Czech. Acad. Sci., Prague, 1973 pp. 149—158.
14. *Csibi, S.*, "Stability and complexity of learning processes", Lecture Notes, International Center for Mechanical Sciences (to appear) (Preprint, ibid, 1972).
15. *Csibi, S.*, "Statistical learning processes", Preprint, Research Institute, for Telecommunication, 1973 (in Hungarian).

Обучение при вычислительных ограничениях на основе статилтически зависимых данных

Ш. ЧИБИ

(Будапешт)

Путем обобщения работ Бравермана и Розоноэра по устойчивости алгоритмов обучения в условиях помех показано, что большинство практически интересных задач обучения при зависимых данных могут быть сведены к задаче обучения с независимыми данными. В качестве иллюстрации рассмотрена задача реализации Бейесовых дискриминантных функций. Приводятся соответствующие примеры.

S. Csibi
Technical University of Budapest
H—1111 Budapest, XI., Stoczek u. 2.

and

Research Institute for Telecommunication
H—1026 Budapest, II., Gábor Áron u. 65.
Hungary

Problems of Control and Information Theory, Vol. 4 (1), pp. 23—33 (1975)

A LINEAR LEARNING ALGORITHM FOR THE PURPOSE OF IDENTIFICATION

A. BÁRSONY

(Budapest)

(Received May 1, 1974)

The direct aim of most of the identification procedures is to determine the parameters (coefficients) in the relationships between the input and output signals of a given system. If the data regarding both the input and output signals have measurement noises — which always occurs in the practice — then the relationships which describe the behaviour of the system are of approximate validity — besides the simplifications made necessarily for the construction of a mathematical model of the system. In these cases it is reasonable to supplement the parameter estimation procedures with the "prefiltering" — smoothing — of the data required to the estimation. After smoothing we can also get continuous functions regarding both the input and output signals which may be characterized by parameter vectors obtained in the smoothing procedure. Then the parameter estimation procedures may be formulated by assuming that a certain approximate continuous function of the input and output signals exists. In the paper an estimation procedure using polynomial approximate functions and a learning algorithm — based on this procedure — are considered for systems linear in parameters. The application of the learning algorithms — besides the prospective improvement of the statistical properties for on-line real-time identification is treated, too.

The learning algorithm

The procedure is shown for single input-single output case. The method can be generalized for a multivariable case without difficulties.

Let $L(\mathbf{a}, \mathbf{y})$ and $R(\mathbf{b}, \mathbf{u})$ be linear functions of vectors $\mathbf{a}$ and $\mathbf{b}$, respectively, furthermore let

$$\mathbf{y}^T = [y, y', \ldots, y^{(i)}, \ldots, y^{(n)}]$$

and (1)

$$\mathbf{u}^T = [u, u', \ldots, u^{(j)}, \ldots, u^{(m)}]$$

be vectors of derivatives of output and input signals, and $n \geq m$.

Let the differential equation of the system be

$$L(\mathbf{a}, \mathbf{y}) - R(\mathbf{b}, \mathbf{u}) = 0. \tag{2}$$

Let a polynomial approximation of the time function of input and output signals be known

$$\hat{y}_0(t) = \hat{y}_0 = \alpha_0^T t_N$$

and (3)

$$\hat{u}_0(t) = u_0 = \beta_1^T t_M$$

where

$$\alpha_0^T = [\alpha_{00}, \alpha_{10}, \ldots, \alpha_{N0}]$$

is the parameter vector;

$t_N^T = [1, t, t^2, \ldots, t^N]$ is $(N \times 1)$ vector of powers of the time-coordinate;

$\beta_0^T = [\beta_{00}, \beta_{10}, \ldots, \beta_{MO}]$ is the parameter vector;

$\mathbf{t}_M^T = [1, t, t^2, \ldots, t^M]$ is $(M \times 1)$ vector of powers of time-coordinate.

Thus, according to the above, the output signal is approximated by a polynomial of Nth degree, on the other hand the approximation of the input signal, is achieved by a polynomial of Mth degree.

Now, let us apply a certain procedure for estimation of parameter vectors **a** and **b**. Let the result of the estimation be $\mathbf{a}_0$, $\mathbf{b}_0$. Since the parameter vectors are estimated from the approximate values of the input and output signals, furthermore certain faults are likely to be found with the parameter estimation procedure itself, it is generally true that

$$L(\mathbf{a}_0, \mathbf{y}_0) - R(\mathbf{b}_0, \hat{\mathbf{u}}_0) \neq 0 \tag{4}$$

therefore, after substituting the estimated parameter values into the differential equation of the system, we cannot say anything about the goodness of the parameter estimation.

If the estimation of the input and output is valid in the time interval $t_{20} - t_{10}$, then — considering this interval — several measures may be formulated for the characterization of the goodness of the estimation.

Such possibilities are, e.g.:

$$h = N_0 \left| \int_{t_{10}}^{t_{20}} \left[L(\mathbf{a}_0, \hat{\mathbf{y}}_0) - R(\mathbf{b}_0, \hat{\mathbf{u}}_0) \right) dt \right| \tag{5a}$$

$$h_{abs} = N_0 \int_{t_{10}}^{t_{20}} \left| L(\mathbf{a}_0, \hat{\mathbf{y}}) - R(\mathbf{b}_0 \hat{\mathbf{u}}_0) \right| dt \tag{5b}$$

$$h_n = N_0 \int_{t_{10}}^{t_{20}} \left(L(\mathbf{a}_0, \hat{\mathbf{y}}_0) - R(\mathbf{b}_0, \hat{\mathbf{u}}_0) \right)^2 dt \tag{5c}$$

where

$$N_0 = \frac{1}{(t_{20} - t_{10}) \left| (\mathbf{a}_0^T, \mathbf{b}_0^T) \right|}$$

is the normalizing factor.

From the above criteria (5a) and (5b) can be evaluated very simply. Our method, however, may be theoretically applied for all the three criteria.

Since — mostly in the case of linear systems — the evaluation of criterion (5a) is the simplest, we intend to deal hereinafter only with this.

Using (5a) the line of figures may be separated into two parts:

$$\begin{aligned} &\text{1st class: } h \leq h_0, \\ &\text{2nd class: } h > h_0, \end{aligned} \tag{6a}$$

where h_Φ is a chosen limit.

If value h, computed from the estimated $\mathbf{a}_0$, $\mathbf{b}_0$, satisifes (6a) then the estimation is considered good, i.e. vectors $\mathbf{a}_0$, $\mathbf{b}_0$ are put to the 1st class, otherwise to the 2nd class.

Let us consider the sequence of the output and input signals, $\hat{y}_1(t)$, $\hat{u}_1(t)$, $\hat{y}_2(t)$, $\hat{u}_2(t)$ . . ., $\hat{y}_p(t)$, $\hat{u}_p(t)$ belonging to the time intervals t_{11}—t_{21}, t_{12}—t_{22}, . . ., . . ., t_{1p}—t_{2p}, and the estimates of the parameter vectors $\mathbf{a}_1$, $\mathbf{b}_1$, $\mathbf{a}_2$, $\mathbf{b}_2$, . . ., $\mathbf{a}_p$, $\mathbf{b}_p$ which are obtained by polynomials estimating the output and input signals of the same subscript.

Let the accepted value of $\mathbf{a}$ and $\mathbf{b}$ be $\hat{\mathbf{a}}_{p-1}$ and $\hat{\mathbf{b}}_{p-1}$ in the $(p-1)$-th step. Using the pth estimated values of the output and input signals, let us compute the following measure:

$$h_p = N_p \left| \int_{t_{1p}}^{t_{2p}} \left(L(\hat{\mathbf{a}}_{p-1}, \mathbf{y}_p) - R(\hat{\mathbf{b}}_{p-1}, \mathbf{u}_p) \right) dt \right| \tag{7}$$

where

$$N_p = \frac{1}{(t_{2p} - t_{1p}) \left| (\hat{\mathbf{a}}_{p-1}^T, \hat{\mathbf{b}}_{p-1}^T) \right|} .$$

According to criterion (6a) the following decision is made:

$$\hat{\mathbf{a}}_p = \begin{cases} \hat{\mathbf{a}}_{p-1} & , \text{ if } h_p \leq h_0 \\ n_p(\hat{\mathbf{a}}_{p-1} + \lambda_p \mathbf{a}_p), & \text{if } h_p > h_0 \end{cases} \tag{8a}$$

and

$$\hat{\mathbf{b}}_p = \begin{cases} \hat{\mathbf{b}}_{p-1} & , \text{ if } h_p \leq h_0 \\ n_p(\hat{\mathbf{b}}_{p-1} + \lambda_p \mathbf{b}_p), & \text{if } h_p > h_0 \end{cases} \tag{8b}$$

where

$$n_p = \frac{\left| (\hat{\mathbf{a}}_{p-1}^T, \hat{\mathbf{b}}_{p-1}) \right|}{\left| \left((\hat{\mathbf{a}}_{p-1} + \lambda_p \mathbf{a}_p)^T, (\hat{\mathbf{b}}_{p-1} + \lambda_p \mathbf{b}_p)^T \right) \right|} .$$

The value of λ_p in decisions (8a), (8b) may be chosen in different ways:

(i) Computing

$$I_p = \int_{t_{1p}}^{t_{2p}} \left(L(\mathbf{a}_p, \mathbf{y}_p) - R(\mathbf{b}_p, \mathbf{u}_p)\right) dt \tag{9}$$

and denoting the absolute value of the integral in Eq. (7) with I_{p-1},

$$\lambda_p = -\operatorname{sign}(I_p) \operatorname{sign}(I_{p-1})\, \lambda \tag{10}$$

where λ is a suitably chosen constant.

(ii) The coefficient is chosen in such a way that condition

$$\int_{t_{1p}}^{t_{2p}} \left(L(\hat{\mathbf{a}}_{p-1} + \lambda_p \mathbf{a}_p, \mathbf{y}_p) - R(\hat{\mathbf{b}}_{p-1} + \lambda_p \mathbf{b}_p, \mathbf{u}_p)\right) dt = 0$$

should be observed. Hence — by notations in (7) and (9)

$$\lambda_p = \frac{I_{p-1}}{I_p}. \tag{11}$$

(iii) If during $p-k$ steps $h_r > h$ $(r = p-k, p-k+1, \ldots, p)$, then let

$$\lambda_p = \frac{\lambda_p^0}{k+1} \tag{12}$$

where λ_p^0 is the correction factor, for example, by (ii).

For the off-line application of the learning algorithm the procedure may be finished if the decisions (8a) and (8b) give $\lambda_p = 0$ in K cases in succession (K can be given in advance), i.e. K times over condition $h_p \leq h_0$ is valid.

Linear differential equation

In case of a linear differential equation

$$L = \mathbf{y}^T \mathbf{a} \quad \text{and} \quad R = \mathbf{u}^T \mathbf{b} \tag{13}$$

Applying the learning algorithm, in the pth step

$$\hat{y}_p = \boldsymbol{\alpha}_p^T \mathbf{t}_N \quad \text{and} \quad \hat{u} = \boldsymbol{\beta}_p^T \mathbf{t}_M \tag{14}$$

according to (3).

From this

$$\hat{y}_p^{(i)} = \sum_{k=i}^{N} \alpha_{p,k} \frac{k!}{(k-1)!} t^{k-i} \tag{15}$$

$$\hat{u}_p^{(j)} = \sum_{k=j}^{M} \beta_{p,k} \frac{k!}{(k-j)!} t^{k-j}. \tag{16}$$

By substitutions $k - i = r$, or $k - j = r$, let us define matrices α_p, β_p in the following way:

Let the elements of the ith row (jth row) and the rth column of these matrices be:

$$\alpha_{p,ir} = \alpha_{p,i+r} \frac{(i+r)!}{r!}; \quad \alpha_{p,i+r} = 0, \text{ if } i + r > N; \tag{17}$$

$$\beta_{p,jr} = \beta_{p,j+r} \frac{(j+r)!}{r!}; \quad \beta_{p,j+r} = 0, \text{ if } j + r > N.$$

Using the above matrices we can write

$$\hat{\mathbf{y}}_p = \alpha_p \mathbf{t}_N \quad \text{and} \quad \hat{\mathbf{u}}_p = \beta_p \mathbf{t}_M. \tag{18}$$

Taking Eqs. (9) and (13) into consideration:

$$I_p = \int_{t_{1p}}^{t_{2p}} (\mathbf{t}_N^T \alpha_p^T \mathbf{a}_p - \mathbf{t}_M^T \beta_p^T \mathbf{b}_p)\, dt \tag{19a}$$

or on the basis of (7) and (13):

$$I_{p-1} = \int_{t_{1p}}^{t_{2p}} (\mathbf{t}_N^T \alpha_p^T \hat{\mathbf{a}}_{p-1} - \mathbf{t}_M^T \beta_p^T \mathbf{b}_{p-1})\, dt. \tag{19b}$$

Let us introduce the following vectors to the computation of (19a), (19b)

$$\begin{aligned} \mathbf{I}_{Np}^T &= \left[t_{2p} - t_{1p}, \ldots, \frac{t_{2p}^{r+1} - t_{1p}^{r+1}}{r+1}, \ldots, \frac{t_{2p}^{N+1} - t_{1p}^{N+1}}{N+1} \right], \\ \mathbf{I}_{Mp}^T &= \left[t_{2p} - t_{1p}, \ldots, \frac{t_{2p}^{r+1} - t_{1p}^{r+1}}{r+1}, \ldots, \frac{t_{2p}^{M+1} - t_{1p}^{M+1}}{M+1} \right]. \end{aligned} \tag{20}$$

By these vectors

$$I_p = \mathbf{I}_{Np}^T \alpha_p^T \mathbf{a}_p - \mathbf{I}_{Mp}^T \beta_p^T \mathbf{b}_p \tag{21a}$$

and

$$I_{p-1} = \mathbf{I}_{Np}^T \alpha_p^T \mathbf{a}_{p-1} - \mathbf{I}_{Mp}^T \beta_p^T \mathbf{b}_{p-1}. \tag{21b}$$

Using (21b) the decisions (8a) or (8b) can be made; in case of $\lambda_p \neq 0$, if λ_p is chosen according to (ii) or (iii) then (21a) is also applied.

Determination of $\mathbf{a}_p$ and $\mathbf{b}_p$ in case of a linear differential equation

The estimated values of the parameter vectors may be determined, for example, by minimization of functional

$$\Phi = \int_{t_{1p}}^{t_{2p}} \left(L(\mathbf{a}_p, \hat{\mathbf{y}}_p) - R(\mathbf{b}_p, \hat{\mathbf{u}})\right)^3 dt, \tag{22}$$

The parameter vectors which ensure the minimum — as it is known — may be determined by solving the system of equations

$$\begin{aligned} \frac{\partial \Phi}{\partial a_{p,i}} &= 0; \qquad i = 0, 1, \ldots, n \\ \frac{\partial \Phi}{\partial b_{p,j}} &= 0; \qquad j = 0, 1, \ldots, m. \end{aligned} \tag{23}$$

Since $n + m + 2$ equations are available for determination of $n + m + 2$ parameters which have zeros on their right sides, in most cases — under linear differential equations in every case — one of the vector components may be admitted arbitrarily, or the system equation $L - R = 0$ may be divided by a component. In the latter case, assuming a linear differential equation we can divide by

$a_{p,0}$ if the plant has no integrator;

$b_{p,0}$ if the plant has an integrator, since then $a_{p,0} = 0$,

and it should be noted that in case of a plant, having a differentiating term, $b_{p,0} = 0$.

Later on these obligations are not investigated at the formulation of equations for the computation of parameter vectors, since these special cases mean only the striking out of the corresponding elements of matrices and vectors. We intend to deal with only the first case (dividing by $a_{p,0}$). According to this the first equation of (23) is omitted ($i = 0$).

After simplification the system (23) is in detail:

$$\begin{aligned} &\int_{t_{1p}}^{t_{2p}} \frac{\partial L(\mathbf{a}_p, \mathbf{y}_p)}{\partial a_{p,i}} \left(L(\mathbf{a}_p, \hat{\mathbf{y}}) - R(\mathbf{b}_p, \hat{\mathbf{u}}_p)\right) dt = 0 \qquad i = 1, 2, \ldots, n; \\ &\int_{t_{1p}}^{t_{2p}} \frac{\partial R(\mathbf{b}_p, \hat{\mathbf{u}}_p)}{\partial b_{p,j}} \left(L(\mathbf{a}_p, \hat{\mathbf{y}}) - R(\mathbf{b}_p, \hat{\mathbf{u}}_p)\right) dt = 0 \qquad j = 0, 1, \ldots, m. \end{aligned} \tag{24}$$

In case of linear differential equations these partial derivations may be computed very easily:

$$\frac{\partial L(\mathbf{a}_p, \hat{\mathbf{y}}_p)}{\partial a_{p,i}} = \frac{\partial \hat{\mathbf{y}}_p^T \mathbf{a}_p}{\partial a_{p,i}} = \hat{y}_{p,i} \tag{25a}$$

or

$$\frac{\partial R(\mathbf{b}_p, \hat{\mathbf{u}}_p)}{\partial b_{p,j}} = \frac{\partial \hat{\mathbf{u}}_p^T \mathbf{b}_p}{\partial b_{p,j}} = \hat{u}_{p,j} \tag{25b}$$

where

$\hat{y}_{p,i}$ is the ith component of vector $\hat{\mathbf{y}}_p$,
$u_{p,j}$ however, is the jth element of vector $\mathbf{u}_p$.

Taking Eq. (18) into account, the right sides of (25a) and (25b) will be:

$$\hat{y}_{p,i} = \boldsymbol{\alpha}_{p,i}^T \mathbf{t}_N \quad \text{and} \quad \hat{u}_{p,j} = \boldsymbol{\beta}_{p,j}^T \mathbf{t}_M \tag{26}$$

where

$\boldsymbol{\alpha}_{p,i}^T$ is the ith row vector of matrix α_p,
$\boldsymbol{\beta}_{p,j}^T$ is the jth row vector of matrix β_p.

After substitutions the system (24) will have the following form:

$$\int_{t_{1p}}^{t_{2p}} \boldsymbol{\alpha}_{p.i}^T t_N (\mathbf{t}_N^T \boldsymbol{\alpha}_p^T \mathbf{a}_p - \mathbf{t}_M^T \boldsymbol{\beta}_p \mathbf{b}_p)\, dt = 0 \qquad i = 1, 2, \ldots, n$$

$$\int_{t_{1p}}^{t_{2p}} \boldsymbol{\beta}_{p.j}^T \mathbf{t}_M^T (\mathbf{t}_N^T \boldsymbol{\alpha}_p^T \mathbf{a}_p - \mathbf{t}_M^T \boldsymbol{\beta}_p^T \mathbf{b}_p)\, dt = 0 \qquad j = 0, 1, \ldots, m. \tag{27}$$

Let us introduce the following notations:

$$\mathbf{I}_{NN} = \int_{t_{1p}}^{t_{2p}} \mathbf{t}_N \mathbf{t}_N^T\, dt$$

$$\mathbf{I}_{NM} = \int_{t_{1p}}^{t_{2p}} \mathbf{t}_N \mathbf{t}_N^T\, dt \tag{28}$$

$$\mathbf{I}_{MM} = \int_{t_{1p}}^{t_{2p}} \mathbf{t}_M^T \mathbf{t}_M^T\, dt\,,$$

and let us partition the matrix $\boldsymbol{\alpha}_p^T$ and vector $\mathbf{a}_p$ in the following way:

$$\boldsymbol{\alpha}_p^T = [\boldsymbol{\alpha}_{p,0}, \boldsymbol{\alpha}_p'^T]\,; \quad \mathbf{a}_p = \begin{bmatrix} 1 \\ \mathbf{a}_p' \end{bmatrix}. \tag{29}$$

By the substitution of Eqs. (28) and (29) into (27) we get the following hyper-matrix equation:

$$\begin{bmatrix} \boldsymbol{\alpha}_p' \mathbf{I}_{NN}, \boldsymbol{\alpha}_p'^T & \boldsymbol{\alpha}_p' \mathbf{I}_{NM} \boldsymbol{\beta}_p^T \\ \boldsymbol{\beta}_p \mathbf{I}_{NM}^T \boldsymbol{\alpha}_p'^T & \boldsymbol{\beta}_p \mathbf{I}_{MM} \boldsymbol{\beta}_p^T \end{bmatrix} \begin{bmatrix} \mathbf{a}_p' \\ -\mathbf{b}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_p' \mathbf{I}_{NN} \\ \boldsymbol{\beta}_p \mathbf{I}_{NM} \end{bmatrix} \boldsymbol{\alpha}_{p,0}\,. \tag{30}$$

This is a first order linear equation system with respect to the parameter vector $\mathbf{P}^T = [\mathbf{a}_p'^T, -\mathbf{b}_p^T]$. Since the number of rows of matrix $\boldsymbol{\alpha}_p'$ is n, the

number of its columns is $N + 1$, the number of rows of matrix β_p is $m + 1$, the number of its columns is $M + 1$, the inverse of the hypermatrix on the left side of the equation system exists only if $N \geq n$, $M \geq m$. Results of the first attempts have shown that the order of the approximate polynomials should be 2—3 times higher than the order of the left or right side of the differential equation.

Some simplifying conditions

When applying the above generally formulated algorithm it is reasonable to introduce some simplifying conditions.

(i) It is reasonable to choose the validity region of the estimation of output and input signals in such a way that the starting time point should be zero for all estimations. This means the linear transformation

$$t' = t - t_{1p}, \tag{31}$$

according to which

$$t'_{1p} = 0, \quad \text{or} \quad t'_{2p} = t_{2p} - t_{2p}. \tag{32}$$

(ii) It is also reasonable to introduce the linear transformation of the time coordinate:

$$t' = d_p + c_p \tau. \tag{33}$$

The constants of the transformation c_p and d_p have to be chosen in such a way that the region $(0, t'_{2p})$ should be transformed to the region $(-1, 1)$. This condition may be fulfilled by

$$c_p = d_p = \frac{t'_{2p}}{2},$$

therefore the transformation (33) may be written as:

$$t' = \frac{t'_{pp}}{2}(1 + \tau). \tag{35}$$

From the previous simplifying conditions (i) is obvious, because the values of the elements of the parameter vector **P** are not influenced by the time (in case of time-invariant system). Therefore this type of transformation may be performed and is justified at all events.

The transformation (ii) simplifies or standardizes the forms of vectors $\mathbf{I}_N$ an $\mathbf{I}_M$, or matrices $\mathbf{I}_{NN}$, $\mathbf{I}_{NM}$ and $\mathbf{I}_{MM}$. It can easily be seen that the vectors

$\mathbf{t}_N$ and $\mathbf{t}_M$ in (3) may be rewritten as (taking the transformation (31) into account):

$$\mathbf{t}'_N = \mathbf{C}_{Np}\,\mathbf{T}_{NN}\,\tau_N$$
$$\mathbf{t}'_M = \mathbf{C}_{Mp}\,\mathbf{T}_{MM}\,\tau_M \tag{36}$$

where

$$\mathbf{C}_{Np} = \left\langle 1, \frac{t'_{2p}}{2}, \ldots, \left(\frac{t'_{2p}}{2}\right)^N \right\rangle \text{ is a diagonal matrix;}$$

$\mathbf{T}_{NN}$ and $\mathbf{T}_{MM}$ are the quadratic matrices of the transformation whose elements in their kth row ($k = 0, 1, \ldots,$) and rth column ($r = 0, 1, \ldots,$) are:

$$T_{k,r} = \binom{k}{r}, \quad \text{if} \quad r \leq k \quad \text{and}$$
$$T_{k,r} = 0 \quad , \quad \text{if} \quad r > k;$$
$$\boldsymbol{\tau}_N^T = (1, \tau, \tau^2, \ldots, \tau^N),$$

$\boldsymbol{\tau}_M^T = (1, \tau, \tau^2, \ldots, \tau^M)$ are $(N+1)$ or $(M+1)$ vectors. Using relationships in (36):

$$\hat{y}_p(t) = \boldsymbol{\alpha}_p^T\,\mathbf{t}_N = \boldsymbol{\alpha}_p^T\,\mathbf{C}_{Np}\,\mathbf{T}_{NN}\,\boldsymbol{\tau}_N \quad = \boldsymbol{\gamma}_p^T\,\boldsymbol{\tau}_N = \hat{y}_p(\tau);$$
$$\hat{u}_p(t) = \boldsymbol{\beta}_p^T\,\mathbf{t}_M = \boldsymbol{\beta}_p^T\,\mathbf{C}_{Mp}\,\mathbf{T}_{MM}\,\boldsymbol{\tau}_M = \boldsymbol{\delta}_p^T\,\boldsymbol{\tau}_M = \hat{u}(\tau), \tag{37}$$

where

$$\boldsymbol{\gamma}_p^T = \boldsymbol{\alpha}_p^T\,\mathbf{C}_{Np}\,\mathbf{T}_{NN},$$

$\boldsymbol{\delta}_p^T = \boldsymbol{\beta}_p^T\,\mathbf{C}_{Mp}\,\mathbf{T}_{MM}$ are the transformed coefficient vectors.

Using $\boldsymbol{\gamma}_p$ and $\boldsymbol{\delta}_p$ the matrices $\boldsymbol{\gamma}_p$ and $\boldsymbol{\delta}_p$ may be defined, whose elements in the ith (or jth) row and rth column are:

$$\gamma_{p,ir} = \gamma_{p,i+r}\,\frac{(i+r)!}{r!}; \quad \gamma_{p,i+r} = 0, \quad \text{if} \quad i + r > N; \tag{38}$$

$$\delta_{p,jr} = \delta_{p,j+r}\,\frac{(j+r)!}{r!}; \quad \delta_{p,j+r} = 0, \quad \text{if} \quad j + r > M.$$

Taking into account that $dt' = \frac{t'_{2p}}{2}\,d\tau$, because of (35), the criterion h_p may be written as:

$$h_p = N_p \left| \int_{-1}^{1} (\boldsymbol{\tau}_N^T\,\boldsymbol{\gamma}_p^T\,\mathbf{a}_p^0 - \boldsymbol{\tau}_p^T\,\mathbf{b}_0^0)\,d\tau \right| \tag{39}$$

where

$$N_p = \frac{1}{2\,|(\mathbf{a}_p^0, \mathbf{b}_p^0|}.$$

Because of the special choise of the integrating limits

$$\mathbf{I}_N^T = \int_{-1}^{1} \boldsymbol{\tau}_M^T \, d\tau = \left(2, 0, \frac{2}{3}, 0, \ldots\right) \tag{40}$$

$$\mathbf{I}_M^T = \int_{-1}^{1} \boldsymbol{\tau}_M^T \, d\tau = \left(2, 0, \frac{2}{3}, 0, \ldots\right)$$

are $(N + 1)$ or $(M + 1)$ vectors, respectively, whose last element is zero or $\frac{2}{N+1}$ $\left(\text{or } \frac{2}{M+1}\right)$ depending on whether $N(M)$ is an odd or even member.

Considering definitions (28) the matrices $\mathbf{I}_{NN}$, $\mathbf{I}_{NM}$ and $\mathbf{I}_{MM}$ may be constructed in a similar way.

In the expression of h_p in (39) $\mathbf{a}_p^0$ and $\mathbf{b}_p^0$ had to be written instead of the original $\mathbf{a}_p$ and $\mathbf{b}_p$ because of the transformation (35).

The following relationship exists between the original and modified parameter vector components:

$$\mathbf{a}_p^0 = \mathbf{C}_{np}^{-1} \mathbf{a}_p \quad \text{and} \quad \mathbf{b}_p^0 = \mathbf{C}_{mp}^{-1} \mathbf{b}_p \tag{41}$$

where

$$\mathbf{C}_{np} = \left\langle 1, \frac{t'_{2p}}{2}, \ldots, \left(\frac{t'_{2p}}{2}\right)^n \right\rangle;$$

$$\mathbf{C}_{mp} = \left\langle 1, \frac{t'_{2p}}{2}, \ldots, \left(\frac{t'_{2p}}{2}\right)^m \right\rangle$$

are diagonal matrices.

It should be noted that the learning algorithm does not need $\mathbf{a}_p^0$ and $\mathbf{b}_p^0$ but the original parameter vector. Therefore at the transformation of the time domain the parameter vector compounds obtained by solving the equation system (30) have to be recalculated because the elements of matrices $\mathbf{C}_{np}$ and $\mathbf{C}_{mp}$ also depend on p — theoretically. The recalculation — as it is seen from (41) — may be performed easily:

$$\mathbf{a}_p = \mathbf{C}_{np} \mathbf{a}_p^0 \quad \text{and} \quad \mathbf{b}_p = \mathbf{C}_{mp} \mathbf{b}_p^0 . \tag{42}$$

At the calculation of the correcting factor λ_p — if alternatives (ii) or (iii) are chosen — the operations in (42) may be performed after correction, too.

Применение линейного алгоритма обучения для идентификации динамических характеристик

А. ВАРШОНЬ

(Будапешт)

Рассматривается применение линейного алгоритма обучения для идентификации параметров линейной динамической системы. Алгоритм основан на использовании полиномиального приближения входных и выходных сигналов. В этом случае открываются возможности применения однолинейной вычислительной процедуры, реализуемой в реальном масштабе времени.

A. Bársony
Department of Process Control
Technical University of Budapest
H—1111 Budapest XI., Műegyetem rkprt. 5—7.
Hungary

Problems of Control and Information Theory, Vol. 4 (1), pp. 35—56 (1975)

АВТОМАТЫ С ПРОГРАММИРУЕМОЙ СТРУКТУРОЙ

А. В. КАЛЯЕВ

(Таганрог)

(Поступила в редакцию 1 октября 1973 г.)

В статье рассматриваются особенности автоматов с программируемой структурой, их свойства и отличия от автоматов с программируемым процессом. Доказываются теоремы о поведении автоматов с различными графами и о поведении автоматов с программируемой структурой и программируемым процессом.

Рассматривается важный класс автоматов с программируемой структурой — автоматы с программируемой коммутацией. Вводится понятие функции коммутации. Анализируется система зависимостей, описывающих автомат с программируемой коммутацией. Показывается зависимость функций переходов и выходов, а также множества состояний и выходного алфавита автомата с программируемой коммутацией от функции коммутации.

Проводится классификация автоматов по принципу программирования процесса, структуры и коммутации.

Введение

Однородные вычислительные среды и структуры, интенсивно разрабатываемые в последнее время и успешно реализуемые на основе достижений микроэлектроники, представляют собой новый класс автоматов, графы которых заранее не заданы, а задаются и изменяются под воздействием управляющих сигналов.

Поведение и свойства подобных автоматов существенно отличаются от поведения и свойств обычных автоматов с жесткими неизменными графами.

В статье ставится задача проанализировать особенности автоматов, структура и графы которых могут изменяться программным способом,рассмотреть их свойства и отличия от автоматов с неизменными графами и жесткой структурой, наметить принципы управления структурой и графами подобных автоматов.

1. Программирование процесса, структуры и коммутации автоматов

Большинство автоматов, с которыми приходится встречаться практически, имеют жесткую неизменную структуру. Программировать поведение такого автомата можно только путем задания процесса его переходов из одного состояния в другое за счет выбора начального состояния и конкрет-

ного входного слова. Подобные автоматы условно отнесем к классу автоматов с программируемым процессом.

Автомат с программируемым процессом (рис. 1) имеет всегда жесткий неизменный граф g, а его поведение определяется программой работы $\pi\{a(0)\}$, задающей начальную вершину графа $a(0)$, и текущей входной информацией $y(t)$, $t = 1, 2, \ldots$ Здесь и в дальнейшем под графом автомата понимается граф, образованный множеством состояний и функциями переходов и вы-

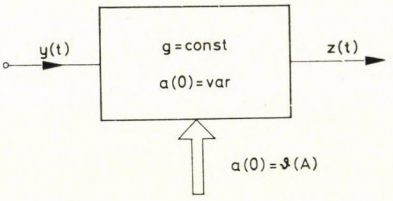


Рис. 1

ходов автомата. Функции переходов и выходов, входной и выходной алфавиты, множество состояний, а также структура такого автомата, задаются при его синтезе и не изменяются в течение всей его жизни.

Выбор начального состояния автомата $a(0) \in A$ из множества состояний $A = \{a_1, a_2, \ldots a_\gamma\}$ осуществляется в автомате с программируемым процессом с помощью оператора управления, который соответствует программе $\pi\{a(0)\} = \{\vartheta(A)\}$, причем $\vartheta(A) \in A$. В простейшем варианте оператор управления может быть задан в виде множества сигналов или команд управления $\theta = \{\vartheta_1, \vartheta_2, \ldots, \vartheta_\xi\}$, $\vartheta \in \theta$, каждая из которых определяет выбор конкретной начальной вершины графа автомата.

Наряду с автоматами с программируемым процессом будем различать автоматы с программируемой структурой. Под автоматом с программируемой структурой будем понимать такой автомат, поведение которого определяется заданием его графа g. Известно, что каждому конкретному графу g_i соответствует некоторое подмножество различных структур автомата S_i. Множеству графов $G = \{g_1, g_2, \ldots g_\gamma\}$ будет соответствовать множество подмножеств структур $S = \{S_1, S_2, \ldots, S_\gamma\}$. Для того, чтобы задать граф автомата $g_i \in G$, необходимо и достаточно выбрать структуру автомата S_{ij} таким образом, чтобы $S_{ij} \in S_i$, где $S_i \in S$.

Следовательно, автомат с программируемой структурой это такой автомат, в котором возможно изменение графа его переходов и выходов $g_i \in G$ за счет выбора из множества структур S соответствующего подмножества S_i и последующего произвольного выбора в этом подмножестве конкретной структуры S_{ij}.

В простейшем случае каждому графу $g_i \in G$ можно заранее сопоставить одну конкретную структуру $S_i \in S_i$, исключив из рассмотрения все остальные структуры, входящие в подмножество S_i. Тогда элементам множества графов $G = \{g_1, g_2, \ldots, g_\gamma\}$ будут взаимно однозначно соответствовать элементы множества структур $S^* = \{S_1, S_2, \ldots, S_\gamma\}$. В этом случае задача программирования структуры автомата с заданным графом $g_i \in G$ сводится к выбору соответствующей графу g_i структуры $S_i \in S^*$.

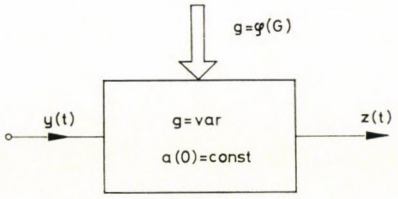


Рис. 2

В автомате с программируемой структурой (рис. 2) внешняя программа $\pi\{g\}$ задает множество состояний, функции переходов и выходов, а также входной и выходной алфавиты автомата.

Выбор графа $g \in G$ из множества графов $G = \{g_1, g_2, \ldots g_\gamma\}$ осуществляется в автомате с программируемой структурой с помощью оператора управления φ, который определяется программой $\pi\{g\} = \{\varphi(G)\}$, где $\varphi(G) \in G$. Оператор управления, предназначенный для выбора графа, может задаваться множеством сигналов или команд управления $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_\tau\}$, $\varphi \in \Phi$.

По существу при структурном программировании автомат в исходном состоянии остается как бы неполностью синтезированным и синтезируется окончательно лишь под воздействием управляющих входных сигналов в зависимости от решаемой задачи и условий работы.

Определенной разновидностью структурного программирования является коммутационное программирование, заключающееся в коммутации достаточно крупных функциональных блоков, графы каждого из которых заранее заданы. Система автоматов с программируемой коммутацией (рис. 3) это такая система, поведение которой задается программой коммутации $\pi\{k\}$, где k — функция коммутации, и входной текущей информацией, причем граф каждого подавтомата, входящего в систему, остается неизменным при любой коммутации.

Функция коммутации $y(t) = k[z(t-1), x(t)]$ определяет зависимость между выходными сигналами коммутатора $y(t)$, внешними входными сигналами коммутатора $x(t)$ и вектором выходных сигналов подавтоматов $z(t-1)$, $t = 1, 2, 3, \ldots$, подаваемых на вход коммутатора с задержкой на такт (рис. 3).

Выбор функции коммутации k из множества функций коммутации $K = \{k_1, k_2, \ldots, k_\varrho\}$ осуществляется с помощью оператора управления ψ, соответствующего программе $\pi\{k\} = \{\psi(K)\}$, где $\psi(K) \in K$. Оператор управления, предназначенный для выбора функции коммутации, может быть представлен в виде множества сигналов или команд управления $\Psi = \{\psi_1, \psi_2, \ldots, \psi_\zeta\}$, $\psi \in \Psi$.

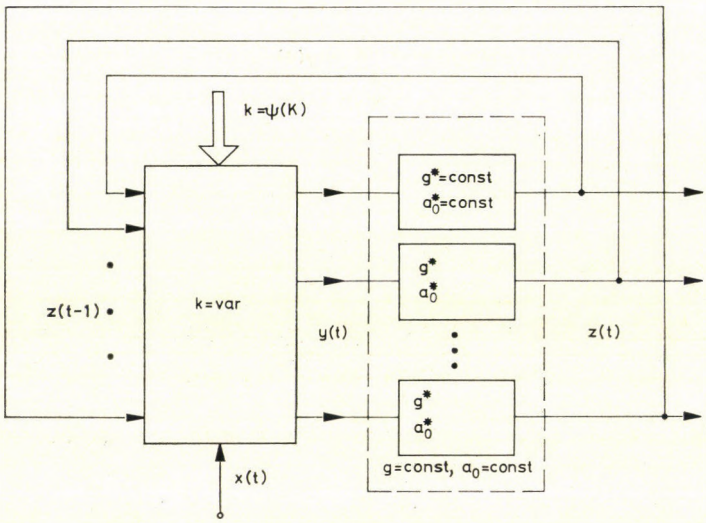


Рис. 3

В автоматах возможно, а в ряде случаев необходимо сочетание процессорного, структурного и коммутационного принципов программирования, которое понимается здесь в том смысле, что процессорное и структурное программирование осуществляется в отдельности для каждого достаточно крупного функционального блока (подавтомата), входящего в автомат, а коммутационное программирование определяет соединение указанных подавтоматов между собой в соответствии с решаемой задачей.

Таким образом, в самом общем случае программа поведения автомата может задаваться одновременно начальным состоянием, графом и функцией коммутации $\pi\{a(0), g, k\}$ (рис. 4). Возможны и другие сочетания принципов программирования: $\pi\{a(0), g\}$, $\pi\{a(0), k\}$, $\pi\{g, k\}$ (табл. 1).

Наибольшей гибкостью и наибольшими возможностями обладает автомат, в котором сочетаются одновременно все три принципа программирования, то есть автомат с программируемыми структурой, коммутацией и процессом (рис. 4). Такой автомат может быть построен из однотипных стандартных подавтоматов, структура каждого из которых может дискретным

способом перестраиваться на несколько графов. Каждый стандартный подавтомат соединяется с одним или несколькими подавтоматами системы с помощью коммутирующей структуры, причем соединение определяется функцией коммутации.

Программирование подобного автомата заключается в задании графа подавтоматов $g = g\,[A, Y, Z, \delta(a, y), \lambda(a, y)]$, который выбирается на множест-

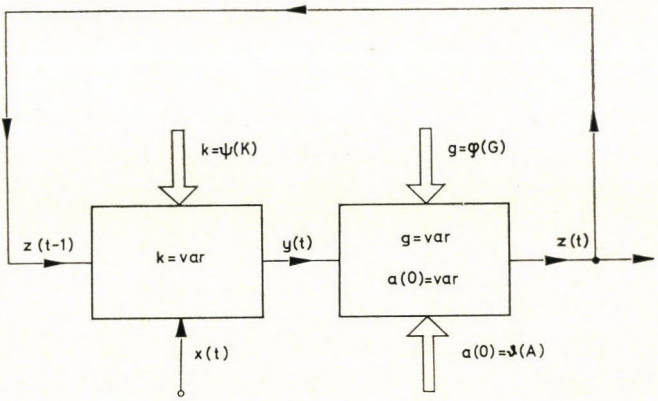


Рис. 4

ва графов G оператором управления φ, функции коммутации $y = k(z, x)$, которая определяется на множество функций коммутации K оператором управления ψ, и начального состояния $a(0)$, которое выбирается из множества состояний автомата A с помощью оператора управления ϑ. Таким образом, поведение рассматриваемого автомата определяется программой $\pi\{a(0), g, k\} = \{\vartheta(A), \varphi(G), \psi(K)\}$, где $\vartheta(A) \in A$, $\varphi(G) \in G$, $\psi(K) \in K$.

Отсюда следует, что для полного задания автомата с программируемыми процессом, структурой и коммутацией необходимо и достаточно задать совокупность следующих данных:

множество состояний	$A = \{a_1, a_2, \ldots, a_k\}$,
входной алфавит	$X = \{x_1, x_2, \ldots, x_n\}$,
промежуточный алфавит	$Y = \{y_1, y_2, \ldots, y_m\}$,
выходной алфавит	$Z = \{z_1, z_2, \ldots, z_l\}$,
множество функций переходов	$\Delta = \{\delta_1, \delta_2, \ldots, \delta_\mu\}$,
множество функций выходов	$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_\nu\}$,
множество функций коммутации	$K = \{k_1, k_2, \ldots, k_\varrho\}$,
операторы управления структурой	$\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_\tau\}$,

операторы управления коммутацией $\Psi = \{\psi_1, \psi_2, \ldots, \psi_\xi\}$,

операторы управления процессом $\Theta = \{\vartheta_1, \vartheta_2, \ldots, \vartheta_\xi\}$,

программу поведения автомата

$$\pi\{a_i(0),\ g_i,\ k_i\} = \{\vartheta_i(A),\ \varphi_i(G),\ \psi_i(K)\},$$

где

$$\vartheta_i \in \theta,\ \varphi_i \in \Phi,\ \psi_i \in \Psi,$$

$$\vartheta_i(A) \in A,\ \varphi_i(G) \in G,\ \psi_i(K) \in K.$$

Иными словами, для определения автомата с программируемыми процессом, структурой и коммутацией следует задать множество графов $G = \{g_1, g_2, \ldots, g_j\}$, множество функций коммутации K, входной алфавит X, множества операторов управления Φ, Ψ и Θ и программу $\pi\{a(0), g, k\}$.

Конкретная программа

$$\pi\{a_i(0),\ g_i,\ k_i\} = \{\vartheta_i(A),\ \varphi_i(G),\ \psi_i(K)\}$$

определяет на множестве графов конкретный граф

$$g_i = g_i[A_i,\ Y_i,\ Z_i,\ \delta_i(a, y),\ \lambda_i(a, y)],\ g_i \in G,$$

для которого

$$A_i\ A,\ Y_i\ Y,\ Z_i\ Z,\ \delta_i \in \Delta,\ \lambda_i \in \Lambda,$$

а на множестве функций коммутации K — конкретную функцию коммутации $y = k_i(z, x)$, $k_i \in K$, и задает начальное состояние $a_i(0) \in A_i$. В результате полностью задается конкретный автомат, который в процессе функционирования при поступлении входных слов вырабатывает соответствующие им выходные слова.

2. Различие в поведении автоматов с программируемым процессом и программируемой структурой

Реакцию автомата на любое входное слово $p = y(1)\ y(2) \ldots y(t)$ при заданном начальном состоянии $a(0)$, представляющую собой последовательность пар состояний и выходных букв

$$r = \{a(1),\ z(1)\}\ \{a(2),\ z(2)\} \ldots \{a(t),\ z(t)\},$$

$$\{a(t),\ z(t)\} \in A \times Z,\ r \in R,$$

соответствующую входным буквам $y(t)$ ($t = 1, 2, 3, \ldots$), назовем поведением автомата. Здесь $R = \{r_1, r_2, r_3 \ldots\}$ — множество реакций.

В общем случае поведение автомата однозначно зависит от программы и от входного слова

$$r_i = r\{\pi[a(0), g, k\}, p_i],$$

т. е. в конечном счете от графа, коммутации, начального состояния и входного слова.

Рассмотрим условия, при которых поведение двух автоматов может быть одинаковым и условия, при которых их поведение будет различным. С этой целью введем вспомогательные понятия совпадающих и различимых графов и подграфов.

Любой граф g состоит из множества полных дуг $d = \{a_н, y, a_к, z\}$, каждая из которых характеризуется начальным состоянием $a_п$, входной буквой y, конечным состоянием $a_к = \delta(a_п, y)$ и выходной буквой $z = \lambda(a_н, y)$, где δ и λ, соответственно — функции переходов и выходов. Множество всех полных дуг $D = \{d_1, d_2, \ldots, d_\zeta\}$ полностью определяет граф g.

Пусть имеется граф g_i, который характеризуется множеством полных дуг $D_i = \{d_{i1}, d_{i2}, \ldots, d_{i\zeta}\}$ и граф g_j, характеризующийся множеством полных дуг $D_j = \{d_{j1}, d_{j2}, \ldots, d_{j\xi}\}$. Два графа g_i и g_i являются полностью совпадающими, если для любой дуги каждого из этих графов: $d_i \in D_i$, $d_i = \{a_{iн}, y_i, a_{ik}, z_i\}$ или $d_j \in D_j$, $d_j = \{a_{jн}, y_j, a_{jк}, z_j\}$, можно подобрать одну и только одну дугу другого графа такую, что для каждой полученной пары дуг (d_i, d_j) выполняются условия

$$\left.\begin{aligned} a_{iн} &= a_{jн}, \\ y_i &= y_j, \\ a_{ik} &= a_{jк}, \\ z_i &= z_j. \end{aligned}\right\} \qquad (1)$$

К полностью совпадающим относятся равные графы $g_i = g_j$ и изоморфные графы.

Если условия (1) выполняются для каждой пары дуг (d_i, d_j) лишь частично, в частности, если из четырех равенств (1) имеют место лишь три, два или одно равенство, то соответствующие графы g_i и g_j будем называть частично совпадающими.

В том случае, когда для любой пары дуг (d_i, d_j) ни одно из равенств (1) не выполняется, графы g_i и g_j являются полностью несовпадающими.

Несовпадающие и частично совпадающие графы будем называть различимыми графами. В общем случае любая пара графов g_i и g_j распадается на полностью совпадающие подграфы $g_{iп} = g_{jп} = g_п = g_i \cup g_j$, для которых условия (1) выполняются, и на различимые подграфы $g_{ip} = g_i \setminus g_j$ и $g_{jp} =$

$= g_j \setminus g_i$, для которых условия (1) частично или полностью не выполняются.

Сформулируем теперь условия совпадения поведения автоматов. Пусть даны два автомата, характеризующиеся графами g_i и g_j. Для того, чтобы при одинаковых входных словах $p_i = p_j$ поведение r_i автомата g_j совпадало с поведением r_j автомата g_i, необходимо и достаточно, чтобы на каждом шаге их работы совпадали пары состояний и выходных букв автоматов: $\{a_i(t), z_i(t)\} = \{a_j(t), z_j(t)\}$. Отсюда следует, что для совпадения поведения автоматов g_i и g_j необходимо и достаточно, чтобы на каждом шаге работы автоматов при совпадающих входных буквах $y_i(t) = y_j(t)$, $t = 1, 2, 3, \ldots$, выполнялись условия

$$a_{i_н}(t) = a_{j_н}(t),$$
$$a_{ik}(t) = a_{jk}(t),$$
$$z_i(t) = z_j(t).$$

Изложенное позволяет сформулировать теорему о поведении двух различных автоматов.

Теорема 1. О поведении автоматов с различными графами.

Поведение двух полностью определенных автоматов с различными графами g_i и g_j совпадает при одинаковых входных словах $p_i = p_j$ и равных начальных состояниях $a_i(0) = a_j(0)$ лишь в области полностью совпадающих подграфов $g_{i_п} = g_{j_п} = g_п = g_i \cap g_j$ и не совпадает в области различимых подграфов.

Доказательство. Рассмотрим совпадающие подграфы $g_{i_п} = g_{j_п}$, которым соответствуют совпадающие подмножества полных дуг $D_{i_п} = D_{j_п} = D_i \cap D_j$. Пусть заданы начальные состояния $a_i(0) = a_j(0)$ и входные буквы $y_j(1) = y_j(1)$, принадлежащие подграфам $g_{i_п}$ и $g_{j_п}$. В этом случае всегда найдутся совпадающие дуги $d_i(1) = \{a_i(0), y_i(1), a_i(1), z_i(1)\}$, $d_i(1) \in D_{i_п}$ и $d_j(1) = \{a_j(0), y_j(1), a_j(1), z_j(1)\}$, $d_j(1) \in D_{j_п}$, для которых выполняются условия

$$y_i(1) = y_j(1),$$
$$a_i(0) = a_j(0),$$
$$a_i(1) = a_j(1),$$
$$z_i(1) = z_j(1).$$

Аналогично для второго шага работы автомата при входных буквах $y_i(2) = y_j(2)$, принадлежащих подграфам $g_{i_п}$ и $g_{j_п}$, и при исходных состояниях

$a_i(1) = a_j(1)$ обязательно найдется пара совпадающих дуг $d_i(2) = \{a_i(1), y_i(2), a_i(2), z_i(2)\}$, $d_i(2) \in D_{i\text{п}}$ и $d_j(2) = \{a_j(1), y_j(2), a_j(2), z_j(2)\}$ $d_j(2) \in D_{j\text{п}}$, для которых

$$\begin{aligned} y_i(2) &= y_j(2), \\ a_i(1) &= a_j(1), \\ a_i(2) &= a_j(2), \\ z_i(2) &= z_j(2). \end{aligned}$$

Продолжая подобные рассуждения для следующих шагов работы автоматов g_i и g_j, приходим к выводу, что в области совпадающих подграфов $g_{i\text{п}}$ и $g_{j\text{п}}$ для любых допустимых совпадающих слов $p_i = y_i(1), y_i(2), \ldots y_i(t)$ и $p_j = y_j(1)\, y_j(2) \ldots y_j(t)$, $p_i = p_j$, для которых $y_i(t) = y_j(t)$, $t = 1, 2, 3, \ldots$, всегда будут выполняться условия совпадения поведения

$$\begin{aligned} a_i(t-1) &= a_j(t-1), \\ a_i(t) &= a_j(t), \\ z_i(t) &= z_j(t), \quad t = 1, 2, 3, \ldots \end{aligned}$$

Отсюда следует, что в области совпадающих подграфов поведение автоматов при одинаковых входных словах и одинаковых начальных состояниях совпадает.

Рассмотрим, далее различимые подграфы g_{ip} и g_{jp}, которым соответствуют подмножества различимых полных дуг D_{ip} и D_{jp}. Подмножество различимых полных дуг D_{ip} содержит дуги, каждая из которых либо полностью не совпадает, либо совпадает только частично с любой дугой множества полных дуг D_j графа g_j. То же можно сказать и о подмножестве D_{jp}.

Разобьем все дуги входящие в подмножество D_{ip} на два подмножества. К одному из них D^*_{ip} отнесем все дуги, которые не совпадают с дугами D_j по букве входного алфавита y, а по другим параметрам либо совпадают, либо не совпадают. Вследствие несовпадения входных букв для подмножества D^*_{ip} нельзя образовать входные слова, совпадающие с входными словами в области D_j. Поэтому в области D_j не может возникнуть поведение автомата g_i, совпадающее при одинаковых входных словах с поведением автомата g_j.

Ко второму подмножеству $D^{**}_{ip} = D_{ip}\ D^*_{ip}$ отнесем все дуги подмножества D_{ip}, которые совпадают по букве входного алфавита с одной или несколькими дугами D_j, но не совпадают по крайней мере по одному из других параметров. В области D^{**}_{ip} могут быть образованы входные слова p_i автомата g_i, которые совпадают с входными словами p_j автомата g_j. Однако для любой дуги $d^{**}_i \in D^{**}_{ip}$, $d^{**}_i = \{a_{i\text{н}}, y_i, a_{i\text{к}}, z_i\}$ любая дуга $d^{**}_j \in D_j\, d_j = \{a_{j\text{н}}, y_j, a_{j\text{к}}, z_j\}$,

которая совпадает с дугой d_i^{**} по входной букве $y_i = y_j$, будет отличаться от нее хотя бы по одному другому параметру. Это означает, что условия, необходимые и достаточные для совпадения поведения двух автоматов, в области D_{ip}^{**} никогда не выполняются.

Таким образом, в области $D_{ip}^{*} \cup D_{ip}^{**} = D_{ip}$ не может возникнуть при одинаковых входных словах $p_i = p_j$ одинаковое поведение автоматов g_i и g_j. Отсюда следует, что в области различимых подграфов g_{ip} и g_{jp} поведение автоматов g_i и g_j при одинаковых входных словах никогда не совпадает. Теорема доказана.

Обратимся к анализу поведения автоматов с программируемым процессом и программируемой структурой. Автоматы с программируемым процессом имеют принципиальное отличие от автоматов с программируемой структурой, которое определяется двумя теоремами о поведении указанных классов автоматов.

Теорема 2. О поведении автомата с программируемым процессом.

Если задан автомат с программируемым процессом, имеющий жесткий граф $g = g[A, Y, Z, \delta(a, y), \lambda(a, y)] = \text{const}$, для которого при любом начальном состоянии $a(0) \in A$ достижимы все состояния $a \in A$, и если заданы различные программы поведения автомата $\pi_i\{a(0)\} = \{\vartheta_i(A)\}$ и $\pi_j\{a(0)\} = \{\vartheta_j(A)\}$, то для программы $\pi_i\{a(0)\}$ существует входное слово p_j, а для программы $\pi_j\{a(0)\}$ входное слово p_j, такие, что после их окончания обе программы при любых последующих одинаковых словах p^* будут генерировать идентичное поведение автомата r^*:

$$r_i r^* = r[\{\vartheta_i(A)\}, p_i p^*],$$
$$r_j r^* = r[\{\vartheta_j(A)\}, p_j p^*],$$

где

$$r_i = r[\{\vartheta_i(A)\}, p_i],$$
$$r_j = r[\{\vartheta_j(A)\}, p_j].$$

Доказательство. Рассмотрим некоторое состояние автомата $a^* \in A$. Так как все состояния автомата достижимы, то достижимым является и состояние a^* при любых $a(0) \in A$. Поэтому для программы $\pi_i = \{\vartheta_i(A)\}$ всегда найдется некоторое конечное входное слово p_i, которое переведет автомат из состояния $a_i(0) = \vartheta_i(A)$ в состояние $a^* = a[\{\vartheta_i(A)\}, p_i]$. Точно также для программы $\pi_j = \{\vartheta_j(A)\}$ всегда найдется такое конечное входное слово p_j, которое переведет автомат из начального состояния $a_j(0) = \vartheta_j(A)$ в то же самое состояние $a^* = a\{\vartheta_i(A)\}, p_j]$. Соответствующие поведения автомата будут при этом

определяться соотношениями

$$r_i = r[\{\vartheta_j(A)\},\ p_i],$$
$$r_j = r[\{\vartheta_i(A)\},\ p_j], \qquad r_i \neq r_j.$$

Достигнутое состояние a^* можно рассматривать как новое начальное состояние $a(0) = a^*$, соответствующее некоторой программе $\pi^* = \{\vartheta^*(A)\}$, для которой при произвольном допустимом слове p^* поведение автомата будет определяться выражением $r^* = r[\{\vartheta^*(A)\},\ p^*]$.

Образуем из последовательных поведений r_i и r^* поведение $r_i\, r^*$. Так как программа $\{\vartheta^*(A)\}$ является продолжением программы $\{\vartheta_i(A)\}$ после окончания слова p_i, то в случае объединенного поведения $r_i\, r^*$ программа $\{\vartheta_i(A)\}$ поглощает программу $\{\vartheta^*(A)\}$: $r_i r^* = r[\{\vartheta_i(A)\}, p_i]\ r[\{\vartheta^*(A)\},\ p^*] = r[\{\vartheta_i(A),\ p_i\, p^*]$.

Аналогично образуем из последовательности поведений r_j и r^* поведение $r_j r^*$:

$$r_j\, r^* = r[\{\vartheta_j(A)\},\ p_j]\ r[\{\vartheta^*(A)\},\ p^*] = r[\{\vartheta_j(A),\ p_j\, p^*].$$

Отсюда следует, что программа $\pi_i = \{\vartheta_i(A)\}$ после окончания слова p_i и программа $\pi_j = \{\vartheta_j(A)\}$ после окончания слова p_j будут в дальнейшем при любых одинаковых входных словах p^* генерировать идентичное поведение автомата r^*. Теорема доказана.

Теорема 3. О поведении автомата с программируемой структурой.

Если задан автомат с программируемой структурой, имеющий перестраиваемый связный граф $g_0 = g_0\,[A, Y, Z, \vartheta_0(a, y), \lambda_0(a, y)] = \mathrm{var}$, где $g_0 = g_1 \cup g_2 \cup g_3 \ldots \cup g_j$, $G = \{g_1, g_2, \ldots, g_j\}$, для которого грады $g_i \in G$ и $g_j \in G$ $(i, j = 1, 2, \ldots, \gamma)$, $i \neq j$ попарно не совпадают $g_i \setminus g_j = \varnothing$, $g_j \setminus g_i \neq \varnothing$, различимые подграфы $g_{jp} = (g_i \setminus g_j) \subset g_i$ и $g_{jp} = (g_j \quad g_i) \subset g_j$ не изоморфны как между собой, так и по отношению к совпадающим подграфам $g_{i_\Pi} = g_{j_\Pi} = g_i \cap g_j$, а любое состояние $a \in A_i \subset A$ достижимо при любом начальном состоянии $a(0) \in A_i$, $(i = 1, 2, \ldots, \gamma)$; и если заданы две различимые программы поведения автомата $\pi_i\{g\} = \{\varphi_i(G) \in G\}$ и $\pi_j\{g\} = \{\varphi_j(G) \in G,\ i \neq j,$ $(i, j = 1, 2, \ldots, \gamma)$, то для каждой программы существует по крайней мере одно входное слово p^* такое, что для двух программ π_i и π_j нельзя подобрать пары слов p_i и p_j, после окончания которых разные программы π_i и π_j будут генерировать при одинаковых входных словах p^* одинаковое поведение автомата r^*:

$$r_i\, r^* = r[\{\varphi_i(G)\},\ a_i(0),\ p_i\, p^*],$$
$$r_j\, r^* \neq r[\{\varphi_j(G)\},\ a_j(0),\ p_j\, p^*\},$$

где $r_i = r[\{\varphi_i(G)\},\ a_i(0),\ p_i],$

$r_j = r[\{\varphi_j(G)\},\ a_j(0),\ p_j].$

Доказательство. Пусть заданы некоторые программы $\pi_i = \{\varphi_j(G)\}$ и $\pi_j = \{\varphi_j(G)\}$, которые настраивают автомат с программируемой структурой соответственно на графы g_i и g_j. Каждый из графов распадается на различимый и полностью совпадающий подграфы $g_i = g_{ip} \cup g_{in}$, $g_j = g_{jp} \cup g_{jn}$ где $g_{in} = g_{jn} = g_i \cup g_j$, $g_{ip} = g_i \setminus g_j$, $g_{jp} = g_j \setminus g_i$.

Рассмотрим поведение автомата в случае программы $\pi_i = \{\varphi_i(G)\}$, $\varphi_i(G) = g_i \in G$. В каком бы из подграфов g_{ip} или $g_{\text{п}}$ ни находилось начальное состояние $a_i(0) \in A_i$, вследствие достижимости любого состояния $a \in A$ всегда существует такое слово p_i, которое при программе π_i переведет автомат в состояние a_{pi}, содержащееся в различимом подграфе g_{ip}. По тем же соображениям существует слово p^*, которое будет после окончания слова p_i удерживать автомат в течение конечного интервала дискретного времени в области различимого подграфа g_{ip}. При этом будет генерироваться поведение автомата $r_i\, r^* = r[\{\varphi_i(G)\},\ a_i(0),\ p_i\, p^*]$.

На основании теоремы 1 поведение r^*, воспроизводимое в области различного подграфа g_{ip}, ни при каких условиях не может совпасть при том же входном слове p^* с поведением в области графа g_{jp}. Поэтому независимо от того, какое слово p_j предшествовало в области графа g_j слову p^* всегда будем иметь $r_j^* \neq r^*$, где $r_i\, r_j^* = r[\{\varphi_j(G), a_j(0),\ p_j\, p^*]$. Значит поведение r^* не может быть воспроизведено в случае программы $\pi_j = \{\varphi_j(G)\}$ ни при одном из входных слов $p_j \in P_j$. Теорема доказана.

Из указанных теорем вытекают весьма важные следствия.

Следствие 1. Для связного автомата с программируемым процессом существуют такие конечные входные слова, которые приводят после их окончания к неразличимости двух различных программ, то есть к совпадению поведения автомата при любых одинаковых последующих входных словах. Это означает, что если связный автомат с программируемым процессом обучить выполнению некоторых функций, в частности, функций распознавания или функций принятия решений, то навыки, полученные таким автоматом в процессе обучения, могут быть разрушены в процессе его дальнейшей работы при определенных типах входной информации, поступающей на информационный вход автомата.

Следствие 2. Для связного автомата с программируемой структурой не существуют такие конечные входные слова, которые приводили бы к неразличимости двух различных его программ, то есть к совпадению поведения автомата при любых одинаковых последующих входных словах. Это озна-

чает, что если связный автомат с программируемой структурой обучить выполнению некоторых функций, например функций распознавания или функций принятия решений, то навыки, полученные таким автоматом в процессе обучения, не могут быть разрушены в процессе последующей работы никакой информацией, поступающей на информационный вход автомата.

3. Автоматы с программируемой коммутацией

Автомат с программируемой коммутацией (рис. 5) состоит в общем случае из набора отдельных автоматов, входы и выходы которых могут соединяться в соответствии с заданной программой $\pi\{k\} = \{\psi(K)\}$, где $k = \psi(K)$ —

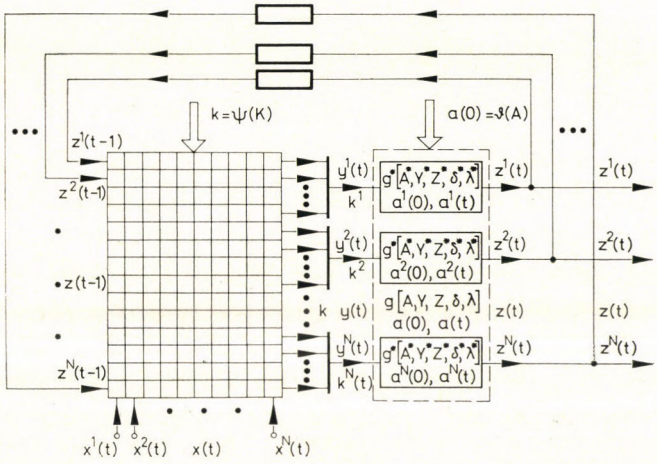


Рис. 5

функция коммутации, $K = \{k_0, k_1, k_2, \ldots, k_e\}$ — множество функций коммутации, а $\psi \in \Psi$ — оператор управления коммутацией, принадлежащий множеству операторов $\Psi = \{\psi_1, \psi_2, \ldots \psi_e\}$. Функция коммутации реализуется коммутатором (рис. 5).

Основной интерес представляет однородная система стандартных автоматов с программируемой коммутацией. Поэтому в дальнейшем будем считать, что все отдельные стандартные автоматы, входящие в систему (рис. 5), однотипны и имеют одинаковые графы, то есть одинаковые множества состояний, входные и выходные алфавиты, а также одинаковые функции переходов и выходов:

$$g^* = g^*[A^*, Y^*, Z^*, \delta^*(a^p, y^p), \lambda^*(a^p, y^p)],$$
$$A^* = \{a_1^*, a_2^*, \ldots, a_b^*\}, \; a^p(t) \in A^*,$$

$$Y^* = \{e, y_1^*, y_2^*, \ldots, y_d^*\},\ y^p(t) \in Y^*,$$
$$Z^* = \{e, z_1^*, z_2^*, \ldots, z_l^*\},\ z^p(t) \in Z,$$
$$a^p(t) = \delta^*[a^p(t-1),\ y^p(t)],$$
$$z^p(t) = \lambda^*[a^p(t-1),\ y^p(t)].$$

Здесь p — номер стандартного отдельного автомата, входящего в систему автоматов с программируемой коммутацией, e — пустая буква, означающая отсутствие сигнала.

Рассмотрим совокупность параллельно работающих стандартных автоматов (рис. 5). Эту совокупность можно интерпретировать как некоторый единый объединенный автомат (рис. 6), представляющий собой декартово произведение всех стандартных автоматов. Для объединенного автомата имеем следующие зависимости

$$g_0 = g_0[A,\ Y,\ Z,\ \delta_0(a,\ y),\ \lambda_0(a,\ y)],$$
$$A = \{a_1,\ a_2,\ \ldots,\ a_k\}, \qquad a(t) \in A,$$
$$Y = \{e, y_1, y_2, \ldots, y_m\}, \qquad y(t) \in Y,$$
$$Z = \{e, z_1, z_2, \ldots, z_l\}, \qquad z(t) \in Z,$$
$$a(t) = \delta_0[a(t-1), y(t)],$$
$$z(t) = \lambda_0[a(t-1), y(t)],$$

где граф, множество состояний и алфавиты объединенного автомата образуются как соответствующие декартовы произведения графов, множеств состояний и алфавитов стандартных автоматов:

$$g_0 = g^* \times g^* \times \ldots g^* = (g^*)^N,$$
$$A = A^* \times A^* \times \ldots A^* = (A^*)^N,$$
$$Y = Y^* \times Y^* \times \ldots Y^* = (Y^*)^N,$$
$$Z = Z^* \times Z^* \times \ldots Z^* = (Z^*)^N.$$

Рассмотрим далее коммутатор системы автоматов с программируемой коммутацией (рис. 5). На входы коммутатора поступают с задержкой на один такт выходные буквы $z(t-1)$, вырабатываемые на выходе объединенного автомата, и внешние сигналы $x(t) \in X$. Внешний алфавит $X = \{e, x_1, x_2, \ldots, x_n\}$, очевидно, образуется как декартово произведение $X = (X^*)^M$ частичных внешних входных алфавитов коммутатора $X^* = \{e, x_1^*, x_2^*, \ldots, x_g^*\}$. На каждом p-том выходе коммутатора, соединенном с входом p-го стандартного автомата, образуется в соответствии с частичной функцией коммутации k^p буква входного алфавита стандартного автомата $y^p(t) \in Y^*$, зависящая от входных букв коммутатора $y^p(t) = k^p[z(t-1), x(t)]$.

Для любого p-го выхода коммутатора, то есть для любого стандартного автомата существует некоторое конечное множество частичных функций коммутации $K^* = \{k_0^*, k_1^*, k_2^*, \ldots k_\nu^*\}$, $k^p \in K^*$. Совокупность частичных функций коммутации $k^p (p = 1, 2, \ldots N)$ для всех стандартных автоматов образует вектор общей функции коммутации $k = \{k^1, k^2, \ldots k^N\}$, $k \in K$. Множество общих или полных функций коммутации $K = \{k_0, k_1, k_2, \ldots k_\varrho\}$ представляет собой декартово произведение $K = (K^*)^N$ множеств частичных функций коммутации. Полные функции коммутации определяют связь между входными и выходными сигналами коммутатора $y(t) = k[z(t-1), x(t)]$.

Резюмируя сказанное, для коммутатора (рис. 7), входящего в систему автоматов с программируемой структурой (рис. 5), можно составить следующие соотношения

$$y^p(t) = k^p\{z(t-1), x(t)\},$$
$$K^* = \{k_0^*, k_1^*, k_2^*, \ldots k_\nu^*\},\ k^p \in K^*,$$
$$y(t) = kz(t-1),\ x(t)],$$
$$K = \{k_0, k_1, k_2, \ldots k_\varrho\},\ k \in K,$$
$$y^p(t) \in Y^*,\ y(t) \in Y,\ z(t) \in Z,\ x(t) \in X,$$
$$X^* = \{e, x_1^*, x_2^*, \ldots x_g^*\},$$
$$X = \{e, x_1, x_2, \ldots x_n\},$$
$$X = X^* \times X^* \times \ldots X^* = (X^*)^M.$$

Объединенный автомат (рис. 6) и коммутатор (рис. 7) образуют в целом некоторый более сложный автомат (рис. 8), характеристики и свойства которого зависят не только от характеристик стандартных автоматов, входящих в

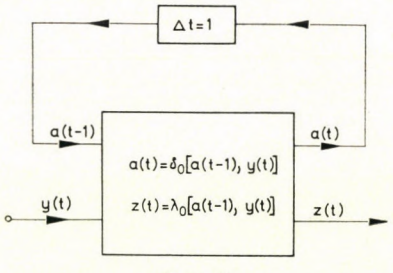


Рис. 6

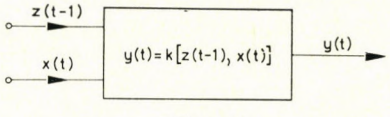


Рис. 7

систему, но в значительной степени и от конкретных функций коммутации, на которые настраивается коммутатор. Функция коммутации определяет в конечном счете функции переходов и выходов автомата с программируемой коммутацией в целом, множество его состояний, а также входной и выходной алфавиты.

Определение алфавитов, множеств состояний и функций переходов и выходов автомата с программируемой коммутацией в зависимости от функции

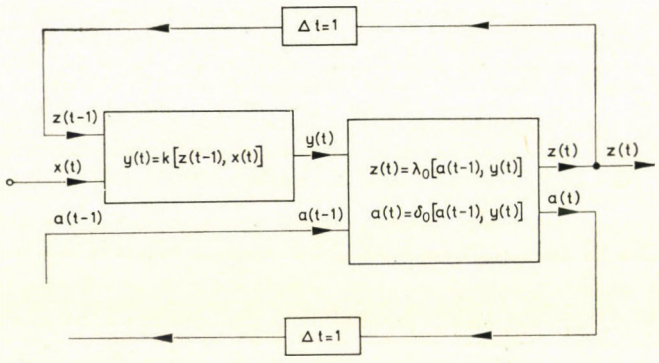


Рис. 8

коммутации составляет одну из основных задач исследования автоматов подобного типа. Рассмотрим эту проблему подробнее. Поведение системы автоматов в целом (рис. 8) полностью описывается функциями переходов и выходов объединенного автомата и функцией коммутации коммутатора

$$a(t) = \delta_0[a(t-1), y(t)],$$
$$z(t) = \lambda_0[a(t-1), y(t)],$$
$$y(t) = k[z(t-1), x(t)].$$

Пользуясь этими выражениями, можно определить функции переходов и выходов системы автоматов в целом, если рассматривать автомат с программируемой коммутацией (рис. 8), состоящий из объединенного автомата и коммутатора, как единый автомат, на вход которого поступают входные буквы $x(t) \in X$, а на выходе вырабатываются выходные буквы $z(t) \in Z$, причем автомат последовательно проходит состояния $a(t) \in A$. Подставляя, с учетом сказанного, в функции переходов δ_0 (a, y) и выходов $\lambda_0(c, y)$ объединенного автомата значения букв промежуточного алфавита $y(t) = k[z(t-1), x(t)]$, получим

$$a(t) = \delta_0\{a(t-1),\ k[z(t-1),\ x(t)]\},$$
$$z(t) = \lambda_0\{a(t-1),\ k[z(t-1),\ x(t)]\}.$$

Соответствующая схема, реализующая данные выражения, изображена на рис. 9.

Как следует из полученных выражений и рис. 9, буквы $z(t)$ поступают с выхода автомата на его вход с задержкой на один такт и с этой точки зрения не отличаются в функциональном отношении от состояний $a(t)$. Это дает нам право объединить состояния $a(t)$ и буквы $z(t)$ в единые пары и описывать состояния автомата в целом вектором $v(t) = \{a(t),\ z(t)\} = a(t) \cup z(t)$, $v(t) \in V = A \times Z$.

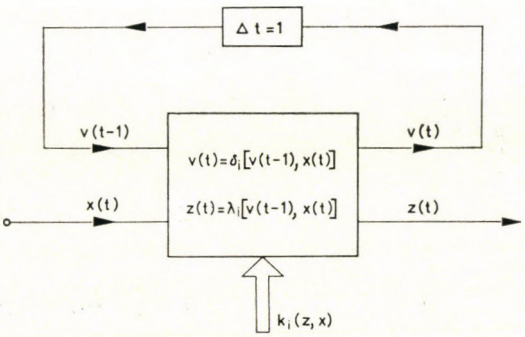


Рис. 9

В результате функции переходов и выходов для автомата с программируемой коммутацией приобретают вид

$$v(t) = \delta_i[v(t-1),\ x(t)],$$
$$z(t) = \lambda_i[v(t-1),\ x(t)],$$

где

$$\delta_i[v(t-1),\ x(t)] = \delta_0\{a(t-1),\ k_i[z(t-1)\ x(t)]\} \cup \lambda_0\ \{a(t-1),\ k_i[z(t-1),\ x(t)]\},$$
$$\lambda_i[v(t-1),\ x(t)] = \lambda_0\{a(t-1),\ k_i[z(t-1),\ x(t)]\}.$$

Как видно, функция переходов $\delta_i(v,\ x)$ и функция выходов $\lambda_i(v,\ x)$ существенно зависит от функции коммутации $k_i(z,\ x)$. Поэтому, изменяя функцию коммутации, можно перестраивать автомат с программируемой коммутацией на различные функции переходов и выходов. От функции коммутации зависят не только функции переходов и выходов, но в общем случае также множество состояний V_i и выходной алфавит Z_i автомата в целом. Остановимся на определении указанной зависимости.

Любая конкретная коммутация $k_i(z, x) = \{k_i^1, k_i^2, \ldots k_i^N\}$, где $k_i \in K$, $k_i^p \in K^*$, $p = 1, 2, \ldots, N$, приводит к тому, что в каждом конкретном стандартном автомате используется лишь часть входного алфавита $Y_i^p \subset Y^*$,

часть выходного алфавита $Z_i^p \subset Z^*$ и часть множества состояний $A_i^p \subset A^*$ (рис. 10). В соответствии с этим лишь частично используются множество состояний и алфавиты объединенного автомата (рис. 6):

$$A_i = \coprod_{p=1}^{N} A_i^p, \qquad A_i \subset A,$$

$$Y_i = \coprod_{p=1}^{N} Y_i^p, \qquad Y_i \subset Y,$$

$$Z_i = \coprod_{p=1}^{N} Z_i^p, \qquad Z_i \subset Z.$$

Покажем, что это действительно имеет место и для определения частичных множеств A_i^p, Y_i^p, Z_i^p, а также множеств A_i, Y_i, Z_i проведем следующие рассуждения. Частичная функция коммутации $y^p(t) = k_i^p[z(t-1), x(t)]$ задана на множестве $(Z_i \times X)$ и однозначно определяет множество Y_i^p. Символически это обстоятельство можно записать в следующей форме $Y_i^p = k_i^p[Z_i \times X]$, или, учитывая, что $Z_i = \coprod_{p=1}^{N} Z_i^p$ в виде $Y_i^p = k_i^p[(\coprod_{j=1}^{N} Z_i^j) \times X]$.

С другой стороны, функции переходов $a^p(t) = \delta^*\ [a^p(t-1),\ y^p(t)]$ и выходов $z^p(t) = \lambda^*[a^p(t-1), y^p(t)]$ стандартных автоматов заданы на множестве $(A_i^p \times Y_i^p)$ и однозначно определяют соответствующие множества A_i^p и Z_i^p, что можно символически записать в форме

$$A_i^p = \delta^*[A_i^p \times Y_i^p],$$
$$Z_i^p = \lambda^*[A_i^p \times Y_i^p].$$

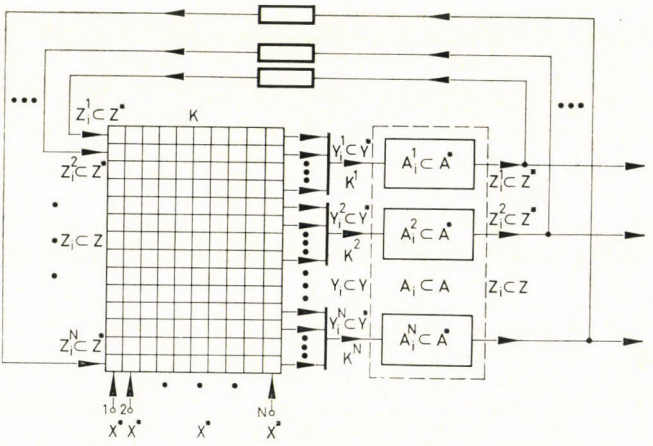


Рис. 10

В результате получаем систему соотношений, которые определяют зависимость частичных множеств A_i^p, Y_i^p и Z_i^p от частичных функций коммутации

$$\left.\begin{aligned} Y_i^p &= k_i^p\left[\left(\prod_{j=1}^{N} Z_i^j\right)\times X\right], \\ A_i^p &= \delta^*[A_i^p \times Y_i^p], \\ Z_i^p &= \lambda^*[A_i^p \times Y_i^p], \\ p &= 1, 2, \ldots, N. \end{aligned}\right\}$$

Проводя аналогичные рассуждения для объединенного автомата (рис. 11) найдем следующие связи между частичными множествами A_i, Y_i и Z_i и полной функцией коммутации $k_i(z, x)$:

$$\left.\begin{aligned} Y_i &= k_i[Z_i \times X], \\ A_i &= \delta_0[A_i \times Y_i], \\ Z_i &= \lambda_0[A_i \times Y_i]. \end{aligned}\right\}$$

Остается получить соответствующие зависимости для системы автоматов с программируемой коммутацией в целом (рис. 12). Множество состояний системы автоматов в целом образуется как декартово произведение множества состояний и входного алфавита объединенного автомата $V_i = A_i \times Z_i$. Кроме этого, функции переходов $v(t) = \delta_i[v(t-1), x(t)]$ и выходов $z(t) = \lambda_i[v(t-1), x(t)]$ автомата в целом заданы на множестве $(V_i \times X)$ и однозначно определяют множества V_i и Z_i. В результате получаем следующие связи, определяющие множество состояний и выходной алфавит системы автоматов и программируемой коммутацией

$$\begin{aligned} V_i &= A_i \times Z_i, \\ Y_i &= \delta_i[V_i \times X], \\ Z_i &= \lambda_i[V_i \times X], \end{aligned}$$

причем $A_i \subset A$, $V_i \subset V$, $Z_i \subset Z$. Так как функции $\delta_i(v, x)$ и $\lambda_i(v, x)$ существенно зависят от функции коммутации $k_i(z, x)$, то множество состояний V_i и выходной алфавит Z_i также зависят от указанной функции.

Таким образом, как следует из полученных соотношений, в автомате с программируемой коммутацией множество состояний, выходной алфавит, а также функции переходов и выходов зависят, с одной стороны, от множеств состояний, алфавитов и функций переходов и выходов стандартных автоматов, а с другой стороны, в значительной степени от функции коммутации. В результате, управляя функцией коммутации, мы получаем мощное средство для управления структурой автомата.

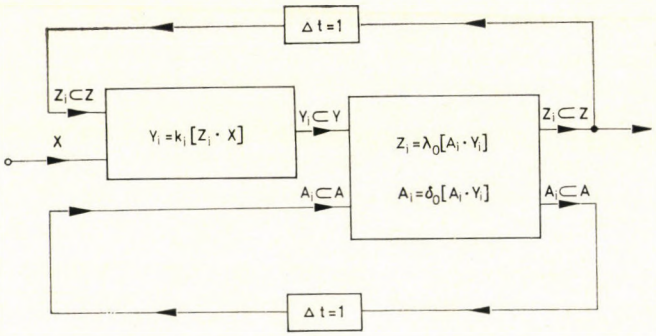


Рис. 11

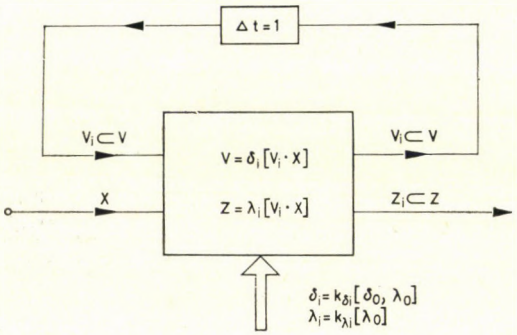


Рис. 12

В заключение приведем все зависимости, характеризующие поведение автомата с программируемой коммутацией:

$$g_i = g_i[V_i, X, Z_i, \delta_i(v, x), \lambda_i(v, x)],$$

$$\delta_i(v, x) = \delta_0\{a(t-1), k_i[z(t-1), x(t)]\} \cup \lambda_0\{a(t-1), k_i[z(t-1), x(t)]\},$$

$$\lambda_i = \lambda_0\{a(t-1), k_i[z(t-1), x(t)]\},$$

$$Y_i = k_i[Z_i \times X_i],$$

$$A_i = \delta_0[A_i \times Y_i],$$

$$Z_i = \lambda_0[A_i \times Y_i],$$

$$V_i = A_i \times Z_i.$$

Основная характерная особенность автомата с программируемой коммутацией заключается в том, что программа функционирования такого автомата $\pi\{k\} = \{\psi(K)\}$ задается наиболее простым способом, но позволяет при этом в широких пределах перестраивать граф автомата, изменять основные его параметры и, как следствие, существенно изменять свойства и возможности автомата в целом.

Заключение

Автоматы с программируемой структурой, в частности автоматы с программируемой коммутацией, графы которых могут перестраиваться под воздействием операторов управления, имеют значительно более широкие функциональные возможности, чем обычные автоматы с неизменными графами.

Возможность перестройки графа переходов придает автомату особые свойства и особую гибкость в приспособлении к внешней среде, позволяет ему наиболее эффективно реагировать на внешние воздействия.

Автомат с перестраиваемым графом имеет потенциальную возможность самонастраиваться, имеет способность приспосабливать свое поведение к внешней обстановке, способность организовывать свою структуру под влиянием обучения. В таких автоматах возникает возможность случайного поиска и отбора наиболее жизнеспособных, наиболее приспособленных к обстановке структур и графов.

При этом принципы и методы программирования и перестройки графов и структур подобных автоматов, как показано в настоящей работе, оказываются достаточно простыми и могут быть легко технически реализованы.

Таблица 1

Классификация автоматов по принципам программирования

	Тип программирования	$\pi = \{g, k, a(0)\}$	
		var	const
Отдельный автомат	Автомат с программируемым процессом	$a(0)$	g, k
	Автомат с программируемой структурой	g	$a(0)$, k
	Автомат с программируемыми структурой и процессом	$a(0)$, g	k
Система автоматов	Система автоматов с программируемой коммутацией	k	$a(0)$, g
	Система автоматов с программируемыми коммутацией и процессом	$a(0)$, k	g
	Система автоматов с программируемыми коммутацией и структурой	g, k	$a(0)$
	Система автоматов с программируемыми коммутацией, структурой и процессом	$a(0)$, g, k	—

Programmable structure automata

A. V. KALYAYEV

(Taganrog)

In the paper the features of the programmable structure automata, their properties and the distinctions from those with programmable processes are considered. The theorems of the behaviour of different graphs automata and the behaviour of the programmable structure and programmable process automata are proved.

An important class of programmable structure automata, i.e. programmable commutation automata is given as well. The notion of commutation function is introduced. The dependenc esystem describing programmable commutation automata is analysed. The dependence of junction and output functions, a set of states and the output alphabet of the automata with the programmable commutation upon the commutation function are considered.

Automata classification according to the principle of programmable process, structure and commutation is given.

А. В. Каляев
Таганрогский радиотехнический институт
СССР, Таганрог,
ул. Чехова, д. 22

Problems of Control and Information Theory, Vol. 4 (1), pp. 57—69 (1975)

CHECKING AND CORRECTION OF MEASUREMENTS ON THE BASIS OF LINEAR SYSTEM MODEL

G. A. ALMÁSY, T. SZTANÓ

(Budapest)

(Received June 5, 1974)

Introduction

Reliability of data is of great importance if they are used for process control or identification. Data burdened with rough measurement errors result in quite false control actions or in false (inadequate) models respectively, if they are not detected.

While data acquisition and process control is a human activity, the measurement data are checked upon by the operator on the basis of his practice and experience. When, however, this is performed by a computer, this checking is missing, so that some programmed checking algorithms are necessary. Such algorithms make use of some information about the process substituting the human experience.

Two further problems arise in connection with data acquisition. First of them is the problem of maximum likelihood state estimation. The second, if the checking algorithm proposes rejecting the measurement, then there is need of an algorithm promoting to encircle the defected measurement point.

Minimum previous knowledge about the system is needed by the usual checking, testing whether each measurement is between certain lower and upper bounds. This way of checking is quick and simple and this fact is a very important advantage from the point of view of on-line computer control but it is far not equivalent with the subjective critics on the basis of experience. It has namely the drawback that a rather wide range has to be specified as acceptable for each variable in order to avoid false error messages when the value of the measured value actually changes, due to the varying conditions of the process. Therefore, the permissible range must include the entire range of realistic values of the variable in question. Evidently, this way of checking is capable to indicate rough errors only, like short circuit or rupture in the wiring, but errors, influencing the measurement to a lesser degree cannot be indicated thereby. For more rigorous checking other, more sophisticated methods are needed, where the feasibility limits change with the change in the process, so that the tolerance limits can be specified actually according to the permissible measurement error. A heuristic algorithm, considering some quali-

tative information about the process was published by Hartman and his co-workers [1].

The least squares estimation of composed measurements is well known long ago (e.g. [2]). The application of this idea to component balance equations in chemical systems was first proposed by Swenker [3]. This topic was developed by Hoffman [4] and Václavek [5]. The authors and their co-workers outlined an iterative solution of second order balance equations [6], applied the algorithm to linear dynamic systems [7] and recently they extended the deterministic methods of balance error smoothing to stochastic systems [8].

All these algorithms yield the least squares or more generally the Markov estimation of the system variables. It is easy to see that with the supposition of zero mean Gaussian distribution of errors this corresponds to the maximum likelihood estimation as well.

The practice, however, shows that in several cases these least squares or Markov estimators yield unacceptable results, like negative concentrations or other senseless states from the point of view of the technology. This problem is known as that of extreme errors and some attemps have been published to solve it [9, 10], but all these are without any theoretical basis, i.e. regardless of the distribution of extreme errors.

Outline

Linearity of the system and the knowledge of the mathematical model are assumed for measurement error analysis.

A quantity, expressing the distance of measured point from the linear subspace of feasible states, is defined. With certain assumptions, regarding the measurement error distribution, complex measurements can be checked upon reliability thereby.

If a complex measurement is not rejected by the above test, then it is regarded to be regular and maximum likelihood estimation, fulfilling the process model, is possible. In the other case another algorithm is recommended on another supposition about the error distribution to facilitate the search of defected instruments.

Finally the same problems are treated for the case when instead of a priori knowledge of the error variance matrix the empirical variance matrix is available only.

Definitions and fundamental concepts

1. Let be an elementary event characterized by an n-component vector $\mathbf{x}$, composed of measurable system variables

$$\mathbf{x} \in \mathfrak{X} \equiv E^n$$

where $\mathfrak{X}$ denotes the (Euclidean) space of states. The indices of elements of $\mathbf{x}$ are called measurement points.

2. Let the subset of feasible (physically realizable) system variables be denoted by $\breve{\mathfrak{X}}$. Thus for a realization, characterized by $\breve{\mathbf{x}}$

$$\breve{\mathbf{x}} \in \breve{\mathfrak{X}}$$

and

$$\breve{\mathfrak{X}} \subset \mathfrak{X}$$

is valid.

In our treatment we restrict to observable discrete linear systems. Thus the set of feasible states is characterized by a set of linear equations, the so-called system equations, so that

$$\breve{\mathfrak{X}} = \{\mathbf{x} : \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}\}$$

where $\mathbf{A}$ is a $p \times n$ coefficient matrix ($p < n$; $\mathrm{r}(\mathbf{A}) = p$) and $\mathbf{b}$ is a p-component coefficient vector.

It is known that if the elements of $\mathbf{x}$ are not restricted to belong to simultaneous states of the system, then this set of equations is equivalent to the discrete input-output model of linear systems.

3. Measurements of $\mathbf{x}$ are denoted by $\tilde{\mathbf{x}}$, each of its elements corresponding to that of $\mathbf{x}$. Since measured values are burdened by random errors

$$\tilde{\mathbf{x}} \notin \breve{\mathfrak{X}}$$

except on a zero measure subset of $\{\tilde{\mathbf{x}}\}$.

4. An estimation of $\breve{\mathbf{x}}$ is denoted by $\hat{\mathbf{x}}$. Details are given later.

5. The difference of any (measured or estimated) system variable and its true value is called error, denoted by $\mathbf{d}$:

$$\mathbf{d} = \mathbf{x} - \breve{\mathbf{x}},$$

especially

$$\tilde{\mathbf{d}} = \tilde{\mathbf{x}} - \breve{\mathbf{x}} \tag{1}$$

and

$$\hat{\mathbf{d}} = \hat{\mathbf{x}} - \mathbf{x}, \tag{2}$$

respectively.

6. The difference of any state $\mathbf{x}$ and its measured value is called correction. Thus the estimated correction is defined as

$$\breve{\mathbf{c}} = \hat{\mathbf{x}} - \tilde{\mathbf{x}}. \tag{3}$$

7. The p dimensional vector $\mathbf{f}$ $(\check{\mathbf{f}}, \tilde{\mathbf{f}}, \check{\mathbf{f}})$, defined as

$$\mathbf{Ax} - \mathbf{b} = \mathbf{f}$$

with any $\mathbf{x}$ $(\check{\mathbf{x}}, \tilde{\mathbf{x}}, \hat{\mathbf{x}})$ is called model defect. Obviously

$$\check{\mathbf{f}} = \mathbf{A}\check{\mathbf{x}} - \mathbf{b} = \mathbf{0}. \tag{4}$$

Measurement error distribution

Two kinds of measurement errors viz. regular and extreme errors are assumed.

Regular errors are supposed to be of n-variate Gaussian distribution with zero expectation and nonsingular $\mathbf{V}_{\tilde{d}}$ dispersion matrix

$$\tilde{\mathbf{d}} \sim N_n(\mathbf{0}, \mathbf{V}_{\tilde{d}}). \tag{5}$$

Further it is supposed that errors are uncorrelated with the true values of the process variables, i.e.:

$$C(\tilde{\mathbf{d}}, \check{\mathbf{x}}) = \mathbf{0}. \tag{6}$$

The assumptions about the extreme errors are given later.

Error analysis

Model defect

According to the definition of $\mathbf{f}$, the apparent model defect is

$$\tilde{\mathbf{f}} = \mathbf{A}\tilde{\mathbf{x}} - \mathbf{b}. \tag{7}$$

Now, making use of Eqs. (1) and (4), we have

$$\tilde{\mathbf{f}} = \mathbf{A}\tilde{\mathbf{d}} \tag{8}$$

so that, from the Gaussian distribution of $\tilde{\mathbf{d}}$ according to (5), the distribution of $\tilde{\mathbf{f}}$ is

$$\mathbf{f} \sim N_p(\mathbf{0}, \mathbf{V}_{\tilde{f}}) \tag{9}$$

where

$$\mathbf{V}_{\tilde{f}} = \mathbf{A}\mathbf{V}_{\tilde{d}}\mathbf{A}^T \tag{10}$$

is the dispersion matrix of error defects. It is easy to see that

$$\mathrm{r}(\mathbf{V}_{\tilde{f}}) = p\,.$$

It is important from the point of view of practical application that $\tilde{\mathbf{f}}$ can be calculated for each measurement from $\tilde{\mathbf{x}}$, representing a set of linear constraints on $\tilde{\mathbf{d}}$ according to Eq. (8), independently of the system variables $\check{\mathbf{x}}$. In particular, concerning a measurement, indicated by k

$$\tilde{\mathbf{f}}_k = \mathbf{A}\tilde{\mathbf{x}}_k - \mathbf{b} \tag{11}$$

and

$$\tilde{\mathbf{f}}_k = \mathbf{A}\tilde{\mathbf{d}}_k \tag{12}$$

are valid.

Feasibility checking with known dispersion of errors

Let

$$t^2 = \tilde{\mathbf{f}}^T \mathbf{V}_{\tilde{f}}^{-1} \tilde{\mathbf{f}}. \tag{13}$$

Then

$$t^2 \sim \chi^2(p) \tag{14}$$

i.e. t^2 has χ^2 distribution on p degrees of freedom. Thus, at a specified significance level α

$$\mathcal{P}(t^2 \geq \chi^2_{1-\alpha}(p)) = \alpha. \tag{15}$$

This means that the probability of the event that a particular t_k^2 exceeds the critical χ^2 value belonging to $1 - \alpha$ level and p degrees of freedom is α. If we choose α sufficiently small, then the probability of exceeding the critical χ^2 value under normal conditions is so small that we may well infere that at least one of our suppositions (either the supposition of zero mean error or its matrix of variance or the Gaussian character of its distribution) does not hold.

Note that t^2 has a geometric meaning. It is the square of the distance of $\tilde{\mathbf{x}}$ from the linear subset of feasible system variables, in the space normed by the square root of $\mathbf{V}_{\tilde{d}}$. Thus, the use of t^2 as the measure of the measurement error is quite reasonable.

Error smoothing with known dispersion of errors

If the feasibility checking does not reject a measurement $\tilde{\mathbf{x}}_k$ then there is a possibility to estimate the maximum likelihood state $\hat{\mathbf{x}}_k$, satisfying the system equations.

Supposing the Gaussian distribution of measurement error the density function of $\tilde{\mathbf{x}}$ with fixed true value of $\check{\mathbf{x}}$ is

$$\varphi(\tilde{\mathbf{x}} \,|\, \check{\mathbf{x}}_k) = (2\pi)^{-\frac{n}{2}} \cdot \det(\mathbf{V}_{\tilde{d}})^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}} - \mathbf{x}_k)^{\cdot} \mathbf{V}_{\tilde{d}}^{-1}(\tilde{\mathbf{x}} - \check{\mathbf{x}}_k)\right). \tag{16}$$

In this sense the maximum likelihood estimation of $\check{\mathbf{x}}_k$ denoted by $\hat{\mathbf{x}}_k$ is the vector $\mathbf{x}$ which maximizes the likelihood function

$$\varphi(\tilde{\mathbf{x}}_k \mid \mathbf{x}) = (2\pi)^{-\frac{n}{2}} \cdot \det(\mathbf{V}_{\tilde{d}})^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2}(\tilde{\mathbf{x}}_k - \mathbf{x})^T \mathbf{V}_{\tilde{d}}^{-1}(\tilde{\mathbf{x}}_k - \mathbf{x})\right) \tag{17}$$

under the constrain

$$\mathbf{f} = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}. \tag{18}$$

It is easy to see that this is equivalent with minimizing the quadratic form

$$\psi_k = \mathbf{c}_k^T \mathbf{V}_{\tilde{d}}^{-1} \mathbf{c}_k \tag{19}$$

due $\mathbf{c}_k$, under the constrain

$$\mathbf{A}\mathbf{c}_k = -\tilde{\mathbf{f}}_k. \tag{20}$$

The conditional extremum point is

$$\hat{\mathbf{c}}_k = -\mathbf{V}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{V}_{\tilde{d}} \mathbf{A}^T)^{-1} \tilde{\mathbf{f}}_k. \tag{21}$$

Hence the maximum likelihood estimator of the state vector $\hat{\mathbf{x}}_k$ is

$$\| \hat{\mathbf{x}}_k = (\mathbf{I} - \mathbf{V}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{V}_{\tilde{d}} \mathbf{A}^T)^{-1} \mathbf{A}) \tilde{\mathbf{x}}_k + \mathbf{V}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{V}_{\tilde{d}} \mathbf{A}^T)^{-1} \mathbf{b} \| . \tag{22}$$

Note that the maximum likelihood estimation is equivalent with a constrained Markov estimation in this case. It follows from Eq. (20) that

$$\hat{\mathbf{c}} \sim N_n(\mathbf{0}, \mathbf{V}_{\hat{c}}) \tag{23}$$

with

$$\mathbf{V}_{\hat{c}} = \mathbf{V}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{V}_{\tilde{d}} \mathbf{A}^T)^{-1} \mathbf{A}\mathbf{V}_{\tilde{d}}. \tag{24}$$

As a consequence of the definitions of $\hat{\mathbf{d}}$, $\tilde{\mathbf{d}}$ and $\hat{\mathbf{c}}$ we obtain that

$$\hat{\mathbf{d}} = \tilde{\mathbf{d}} + \hat{\mathbf{c}}. \tag{25}$$

Thus, from Eq. (21)

$$\hat{\mathbf{d}}_k = (\mathbf{I} - \mathbf{V}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{V}_{\tilde{d}} \mathbf{A}^T)^{-1} \mathbf{A}) \tilde{\mathbf{d}}_k \tag{26}$$

and hence

$$\hat{\mathbf{d}} \sim N_n(\mathbf{0}, \mathbf{V}_{\hat{d}}) \tag{27}$$

with

$$\mathbf{V}_{\hat{d}} = \mathbf{V}_{\tilde{d}} - \mathbf{V}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{V}_{\tilde{d}} \mathbf{A}^T)^{-1} \mathbf{A}\mathbf{V}_{\tilde{d}}. \tag{28}$$

The zero expectation of $\hat{\mathbf{d}}$ means that $\hat{\mathbf{x}}$ is an unbiased estimator of $\check{\mathbf{x}}$.

Search for extreme error

If a measurement is found to be unacceptable by feasibility checking, then we need some algorithm helping the search of the source of extreme error.

Further on we analyse the case when all of the suppositions on error distribution are kept, except the zero expectation of errors. The later is supposed to be parallel with one of the coordinate axes, i.e. non-zero expectation is supposed for a single measurement point. Let be the rth the only measurement point for which

$$\mathcal{E}(\tilde{d}^r) \neq 0$$

and

$$\mathcal{E}(\tilde{d}^j) = 0 \quad \text{if} \quad j \neq r, \tag{29}$$

where $\tilde{d}^r$ and $\tilde{d}^j$ denote the rth and jth element of vector $\tilde{\mathbf{d}}$, resp. This assumption approximates the common case when one of the measuring instruments goes wrong after a faultless period of operation. The evident question is: which of the instruments is defective.

Let ${}^j\mathbf{a}$ be the jth column of $\mathbf{A}$ and

$$\mathbf{D} = \operatorname{diag}\left({}^j\mathbf{a}^T(\mathbf{A}\mathbf{V}_{\tilde{d}}\,\mathbf{A}^T)^{-1}\,{}^j\mathbf{a}\right)^{-\frac{1}{2}} \tag{30}$$

and further

$$\mathbf{g} = \mathbf{D}\mathbf{A}^T(\mathbf{A}\mathbf{V}_{\tilde{d}}\,\mathbf{A}^T)^{-1}\tilde{\mathbf{f}}. \tag{31}$$

As it is proved in Appendix 1, vector $\mathbf{g}$ has the following property:

If supposition (29) is valid, then

$$\operatorname{abs}\left(\mathcal{E}(g^r)\right) \geq \operatorname{abs}\left(\mathcal{E}(g^j)\right) \tag{32}$$

where equality

$$\operatorname{abs}\left(\mathcal{E}(g^r)\right) = \operatorname{abs}\left(\mathcal{E}(g^{r'})\right), \quad r \neq r'$$

is valid if ${}^r\mathbf{a} = h \cdot {}^{r'}\mathbf{a}$ and there is no possibility of distinction between error sources in r and r'.

It follows from the above facts that, supposing that only a single measurement point has non-zero expectation of error, one of the measurement points with maximum absolute value of $\mathcal{E}(g^j)$ is defect.

Our practical aim is to use the above algorithm to error detecting, using averages or even single measurement data instead of expectations. At the moment there is no exactly proved theorem concerning the confidence for these cases. Computer simulations, however, showed that the use of Eq. (32) for error source detection with single $\tilde{\mathbf{f}}_k$ values instead of expectations is sufficiently effective if the ratio of the extreme error and the dispersion of regular error is not too small. Some results of these are given in Appendix 2 for two particular systems. Such computer simulation investigations may be done concerning every actual case, in order to determine the confidence of the error detecting algorithm as a function of the magnitude of extreme error.

Feasibility checking when the dispersion of errors is unknown

There is no direct information about the error dispersion from measurements but model defects. It is not possible to estimate $\mathbf{V}_{\tilde{d}}$ therefrom without additional information. Estimation of $\mathbf{V}_{\tilde{d}}$ is not dealt with in this paper but an algorithm is presented here for feasibility checking that is on the basis of empirical dispersion matrix of $\tilde{\mathbf{f}}$ model defects.

Let be denoted the empirical dispersion matrix by $\tilde{\mathbf{f}}$, determined from a set of $\tilde{\mathbf{f}}_k$'s ($k \in \mathcal{K}$) by $\widetilde{\mathbf{V}}_{\tilde{f},\mathcal{K}}$ and the empirical expectation by $\widetilde{\mathbf{M}}_{\tilde{f},\mathcal{K}}$. $\widetilde{\mathbf{V}}_{\tilde{f},\mathcal{K}}$ has a p-variate central Wishart distribution:

$$\mathbf{V}_{\tilde{f},\mathcal{K}} \sim \mathbf{W}_r(s, \mathbf{V}_{\tilde{f}}) \tag{33}$$

where s is the number of elements of $\mathcal{K}$. On the other hand the empirical expectation of $\tilde{\mathbf{f}}$, determined on the basis of $\mathcal{K}$ set of measurements, has a central Gaussian distribution with $\frac{1}{\sqrt{s}} \cdot \mathbf{V}_{\tilde{f}}$ matrix of dispersion:

$$\widetilde{\mathbf{M}}_{\tilde{f},\mathcal{K}} \sim N_p\left(\mathbf{0}, \frac{1}{\sqrt{s}} \mathbf{V}_{\tilde{f}}\right). \tag{34}$$

As it is known (e.g. [11]) $\widetilde{\mathbf{V}}_{\tilde{f},\mathcal{K}}$ and $\widetilde{\mathbf{M}}_{\tilde{f},\mathcal{K}}$ are independently distributed since $\tilde{\mathbf{f}}$ is of Gaussian distribution, and hence the scalar test variable

$$t^2_{\mathcal{K},\mathcal{K}} = \overset{3}{s^2} \widetilde{\mathbf{M}}^T_{\tilde{f},\mathcal{K}} \cdot \widetilde{\mathbf{V}}^{\mathcal{K}}_{\tilde{f},\mathcal{K}} \cdot \widetilde{\mathbf{M}}_{\tilde{f},\mathcal{K}} \tag{35}$$

is that of Hotelling's generalized T^2 distribution. This latter can easily be transformed into the well-known and tabelled Fisher's F distribution since

$$\frac{s-p+1}{p \cdot s} \cdot T^2 \sim F(p \cdot s).$$

Thus

$$F_{\mathcal{K},\mathcal{K}} = \frac{s-p+1}{p} \sqrt{s} \cdot \widetilde{\mathbf{M}}^T_{\tilde{f},\mathcal{K}} \cdot \widetilde{\mathbf{V}}^{-1}_{\tilde{f},\mathcal{K}} \cdot \widetilde{\mathbf{M}}_{\tilde{f},\mathcal{K}} \sim F(p, s). \tag{36}$$

Another test variable can be formed on the basis of the empirical dispersion matrix on $\mathcal{K}$ set of measurements and any single measurement $\tilde{\mathbf{f}}_l$ where $l \notin \mathcal{K}$ and $\tilde{\mathbf{f}}_l$ is independent of any $\tilde{\mathbf{f}}_k$, $k \in \mathcal{K}$.

Utilizing the above ideas on $\mathbf{f}_l$ instead of $\widetilde{\mathbf{M}}_{\tilde{f},\mathcal{K}}$ the scalar test variable

$$t^2_{l,\mathcal{K}} = s \cdot \tilde{\mathbf{f}}^T_l \cdot \widetilde{\mathbf{V}}^{-1}_{\tilde{f},\mathcal{K}} \cdot \tilde{\mathbf{f}}_l \tag{37}$$

is defined. $t^2_{l,\mathfrak{K}}$ is of Hotelling's generalized T^2 distribution and may be transformed into the F distributed variable

$$F_{l,\mathfrak{K}} = \frac{s-p+1}{p} \tilde{\mathbf{f}}_T^l \cdot \tilde{\mathbf{V}}^{-1}_{\tilde{f},\mathfrak{K}} \cdot \tilde{\mathbf{f}}_l \sim F(p,s) \,. \tag{38}$$

Both $F_{\mathfrak{K},\mathfrak{K}}$ or $F_{l,\mathfrak{K}}$ are suitable to feasibility checking, comparing them with critical values corresponding to certain significance level.

The effect of errors in the dispersion matrix on error smoothing

Practically we have no exact dispersion matrix of errors but an approximation thereof.

Let $\mathbf{U}_{\tilde{d}}$ denote the approximating matrix of dispersion and the estimations, obtained using $\mathbf{U}_{\tilde{d}}$ instead of $\mathbf{V}_{\tilde{d}}$ by $\hat{\mathbf{x}}_k$ and $\hat{\mathbf{d}}_k$.

Thus the estimator for $\check{\mathbf{x}}$ is

$$\hat{\mathbf{x}}_k = (\mathbf{I} - \mathbf{U}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{U}_{\tilde{d}} \mathbf{A}^T)^{-1} \mathbf{A}) \tilde{\mathbf{x}}_k + \mathbf{U}_{\tilde{d}} \mathbf{A}^T (\mathbf{A}\mathbf{U}_{\tilde{d}} \mathbf{A}^T)^{-1} \tag{39}$$

and the error of estimation is

$$\hat{\mathbf{d}}_k = (\mathbf{I} - \mathbf{U}_d \mathbf{A}^T (\mathbf{A}\mathbf{U}_{\tilde{d}} \mathbf{A}^T)^{-1} \mathbf{A}) \tilde{\mathbf{d}}_k \,. \tag{40}$$

It is easy to see that $\hat{\mathbf{x}}_k$ is unbiased with any nonsingular $\mathbf{U}_{\tilde{d}}$:

$$\mathscr{E}(\hat{\mathbf{x}}) = \check{\mathbf{x}} \,. \tag{41}$$

As it is proved in Appendix 3 the dispersion matrix of the errors of estimation is

$$\mathbf{V}_{\hat{d}} = \mathbf{V}_{\hat{d}} + \mathrm{o}(\mathbf{U}_{\tilde{d}} - \mathbf{V}_{\tilde{d}}) \tag{42}$$

where $\mathrm{o}(\mathbf{U}_{\tilde{d}} - \mathbf{V}_{\tilde{d}})$ denotes a matrix polynomial having no lower than second order terms in $\mathbf{U}_{\tilde{d}} - \mathbf{V}_{\tilde{d}}$.

Eq. (42) means that the error of estimation is not sensitive on small errors in the estimation of $\mathbf{V}_{\tilde{d}}$.

Conclusion

All formulas for feasibility checking and error smoothing contain simple matrix operations. The operations with measurement independent quantities can be performed in off-line computations so that the feasiblity checking needs the computation of a quadratic form and the error smoothing needs that of a matrix-vector product only. Thus the algorithms proposed can be applied to on-line measurement analysis advantageously.

Notation

$\mathbf{a}$	a column of matrix $\mathbf{A}$
$\mathbf{A}$	coefficient matrix of the system equation
$\mathbf{b}$	constant vector of the system equation
$\mathbf{c}$	vector of corrections of system variables
$\mathbf{d}$	vector of errors of system variables
$\det(\cdot)$	determinant of matrix $\cdot$
$\mathbf{D}$	matrix defined by Eq. (30)
$\exp(\cdot)$	exponential function of $\cdot$
$\mathcal{E}(\cdot)$	expectation of random variable $\cdot$
$\mathbf{f}$	vector of model defects
$\mathrm{F}(\cdot,\cdot)$	symbol of Fisher's distribution on $\cdot$ and $\cdot$ degrees of freedom
$\mathbf{g}$	vector defined by Eq. (31)
h	scalar constant
$\mathbf{i}$	unit vector
$\mathbf{I}$	unit matrix
j	measurement point index
k	measurement index
$\mathcal{K}$	a set of measurements
l	measurement index
n	number of system variables
$N_v(\cdot,\cdot)$	symbol of v-variate Gaussian distribution with $\cdot$ vector of expectation and $\cdot$ matrix of dispersion
$\mathbf{M}_{\cdot,\cdot}$	expectation of $\cdot$, on the basis of $\cdot$
p	number of system equations
$\mathcal{P}(\cdot)$	probability of event $\cdot$
q	scalar coefficient
r	measurement point index
$\mathrm{r}(\cdot)$	rank of matrix $\cdot$
s	number of measurements in $\mathcal{K}$
t	test variable
$\mathbf{U}_{\cdot}$	approximating dispersion matrix of random vector variable $\cdot$
$\mathbf{V}_{\cdot}$	dispersion matrix of random vector variable $\cdot$
$\mathrm{W}_v(\cdot,\cdot)$	symbol of v-variate central Wishart distribution on $\cdot$ degrees of freedom with $\cdot$ parameter matrix
$\mathbf{x}$	vector of system variables
$\mathcal{X}$	space of system variables
α	significance level
$\varphi(\cdot)$	density function of random variable $\cdot$
$\chi^2(\cdot)$	symbol of central chi square distribution on $\cdot$ degrees of freedom

Appendix 1

Let the rth measurement point be that with extreme error, i.e. let the distribution of errors be

$$\tilde{\mathbf{d}} \sim N_n(q \cdot {}^r\mathbf{i},\ \mathbf{V}_{\tilde{d}})$$

where ${}^r\mathbf{i}$ is the rth unit vector.

Thus, the observed model defect $\tilde{\mathbf{f}}$ has the distribution

$$\tilde{\mathbf{f}} \sim N_p(q \cdot {}^r\mathbf{a},\ \mathbf{V}_{\tilde{f}})$$

where ${}^r\mathbf{a} = \mathbf{A} \cdot {}^r\mathbf{i}$ denotes the rth column of matrix $\mathbf{A}$.

The vector $\mathbf{g}$, constructed according to Eq. (31) is of Gaussian distribution as well, since its jth component is

$$g^j = \frac{{}^j\mathbf{a}^T \cdot \mathbf{V}_{\tilde{f}}^{-1}}{({}^j\mathbf{a}^T \mathbf{V}_{\tilde{f}} \cdot {}^j\mathbf{a})^{1/2}} \cdot \tilde{\mathbf{f}}.$$

The expectation of g^j is

$$\mathcal{E}(g^j) = q \frac{{}^j\mathbf{a}^T \cdot \mathbf{V}_{\tilde{f}}^{-1} \cdot {}^r\mathbf{a}}{({}^j\mathbf{a}^T \cdot \mathbf{V}_{\tilde{f}}^{-1} \cdot {}^j\mathbf{a})^{1/2}},$$

wherefrom, utilizing Schwartz's inequality

$$|\mathcal{E}(g^j)| \leq q \cdot ({}^r\mathbf{a}^T \cdot \mathbf{V}_{\tilde{f}}^{-1} \cdot {}^r\mathbf{a})^{1/2}$$

is obtained. The equality is valid, iff ${}^j\mathbf{a} = c \cdot {}^r\mathbf{a}$ with scalar c. This means that $|\mathcal{E}(g^j)|$ takes his maximum value if $j = r$. By the way, it is easy to see that every g^j is of unit dispersion, independently of r and factor q.

Appendix 2

The efficiency of the error detecting algorithm according to the relations (29)—(32) but with $\mathbf{g}_k$ values instead of $\mathcal{E}(\mathbf{g})$ is illustrated by computer simulation. Two systems were studied, both with various dispersion of regular errors. (For the sake of simplicity, errors were generated independently, i.e. $\mathbf{V}_{\tilde{d}}$ was taken as diagonal matrix.) Extreme error was generated alternately at all measurement points with an absolute value of 2, 4, 8, 16 and 32fold of the correspondent dispersion of regular error. Table 1 shows the relative number of hits, for several dispersion of regular and several values of extreme errors.

The table is indicating the obvious dependence of the effectivity on the magnitude of extreme error. For the two cases in question we can conclude that extreme errors of double magnitude cannot be detected by this way, but those of 8fold magnitude are detected fairly well.

Table 1

The relative number of hits in error detection

System 1

dispersion of measurement errors	relative magnitude of extreme error: 2	4	8	16	32
1 1 1 1 1 1 1 1 1 1	.24	.58	.98	1.00	1.00
10 1 1 1 1 1 1 1 1 1	.22	.55	.91	.99	1.00
63 1 1 1 1 1 1 1 1 1	.23	.57	.98	1.00	1.00

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 0 & 0 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

System 2

dispersion of measurement errors				relative magnitude of extreme error:		2	4	8	16	32
1	1	1	1	1	1	.43	.77	1.00	1.00	1.00
1	1	1	1	1	0.01	.35	.73	.85	.92	.95
1	1	1	1	1	100	.38	.60	.72	.85	.90
1	2	4	8	16	32	.32	.58	.85	.98	1.00

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix}; \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Appendix 3

We can suppose, without loss of generality that $\mathbf{V}_{\tilde{d}} = \mathbf{I}$. Let be further denoted $\mathbf{E} = \mathbf{U}_{\tilde{d}} - \mathbf{V}_{\tilde{d}}$ where $\mathbf{E}$ is the error in $\mathbf{U}_{\tilde{d}}$. $\mathbf{E}$ is supposed to be symmetric and $||\,\mathbf{E}\,||$ to be small. Thus from Eq. (40)

$$\hat{\mathbf{d}}_k = \left(\mathbf{I} - (\mathbf{I} + \mathbf{E})\,\mathbf{A}^T(\mathbf{A}(\mathbf{I} + \mathbf{E})\,\mathbf{A}^T)^{-1}\mathbf{A}\right)\tilde{\mathbf{d}}_k.$$

Utilizing the known formula for the inverse of a sum we obtain

$$\mathbf{A}^T\left(\mathbf{A}(\mathbf{I} + \mathbf{E})\,\mathbf{A}^T\right)^{-1}\mathbf{A} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\,\mathbf{A}\mathbf{E}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\,\mathbf{A} + \\ + \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\,\mathbf{A}\mathbf{E}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\,\mathbf{A}\mathbf{E}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\,\mathbf{A} - \ldots$$

With the notation

$$\mathbf{P} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}$$

it is equivalent to

$$\mathbf{A}^T\left(\mathbf{A}(\mathbf{I} + \mathbf{E})\,\mathbf{A}^T\right)^{-1}\mathbf{A} = \mathbf{P} - \mathbf{P}\mathbf{E}\mathbf{P} + \mathbf{P}\mathbf{E}\mathbf{P}\mathbf{E}\mathbf{P} - \ldots$$

It is easy to see that $\mathbf{P}^T = \mathbf{P}$ and $\mathbf{P}^2 = \mathbf{P}$. Dropping the terms of third and higher order in $\mathbf{E}$ we obtain by an elementary way that

$$\mathbf{V}_{\hat{d}} = \mathcal{E}(\hat{\tilde{\mathbf{d}}}\,\hat{\tilde{\mathbf{d}}}^T) = \left(\mathbf{I} - (\mathbf{I} + \mathbf{E})\,\mathbf{A}^T(\mathbf{A}(\mathbf{I} + \mathbf{E})\,\mathbf{A}^T)^{-1}\mathbf{A}\right) \cdot \mathcal{E}(\tilde{\mathbf{d}}\,\tilde{\mathbf{d}}^T) \cdot \\ \cdot \left(\mathbf{I} - (\mathbf{I} + \mathbf{E})\mathbf{A}^T(\mathbf{A}(\mathbf{I} + \mathbf{E})\mathbf{A}^T)^{-1}\mathbf{A}\right)^T \approx \mathbf{I} - \mathbf{P} + (\mathbf{I} - \mathbf{P})(\mathbf{E}\mathbf{P}\mathbf{E})(\mathbf{I} - \mathbf{P}) = \\ = \mathbf{V}_{\hat{d}} + \mathbf{V}_{\hat{d}}\,\mathbf{E}(\mathbf{I} - \mathbf{V}_{\hat{d}})\,\mathbf{E}\mathbf{V}_{\hat{d}}.$$

References

1. *Hartman, K.*, Analyse und Steuerung von Prozessen der Stoffwirtschaft. Akademie Verlag, Berlin 1971, pp. 144—149.
2. *Grosmann, W.*, Grundzüge der Ausgleichungsrechnung, Springer-Verlag, Berlin (2nd ed.) 1961, p. 322.
3. *Swenker, A. G.*, Acta IMEKO, Budapest, 1964. Vol. I. Proceedings of the 3rd International Measurement Conference.
4. *Hoffmann, R., Müller, R.*, Messen—Steuern—Regeln **9** (1966) 233.
5. *Václavek, V., Kubicek, M., Hlavácek, V., Marek, M.*, Coll. Czech. Chem. Commun., **33**, 3653 (1968).

6. *Almásy, G., Sztanó, T., Csillag, P., Veress, G., Holderith, J.*, 3rd Congress CHISA, Marianské Lázne, 1969.
7. *Gertler, J., Almásy, G.*, Automatica **9** (1973) 79.
8. *Sztanó, T., Almásy, G.*, IFAC Symp. on Stochastic Control, Budapest, 1974.
9. *Ripps, D. L.*, Chem. Eng. Progr. Symp. Ser. No. 55. **61** (1965) 8—13.
10. *Václavek, V.*, 4th Congress CHISA, Prague, 1972. K 5. 4.
11. *Rao, C. R.*, Linear Statistical Inference and its Applications. J. Wiley and Sons (2nd ed.), 1973. p. 524, p. 537.

Проверка и поправка результатов измерений на основе линейной модели системы

Г. А. Алмаши, Т. Стано

(Будапешт)

Рассматривается линейная система с известной математической моделью. Для оценки ошибок результатов измерений вводится специальная величина, характеризующая расстояние между результатом измерения параметра и его эталонным значением.

Для суждения о приемлемости сигнала ошибки системы и использовании коррекции применяется принцип максимального правдоподобия.

G. A. Almásy and T. Sztanó
Computer and Automation Institute
Hungarian Academy of Sciences
H-1502 Budapest, XI., Kende u. 13—17
Hungary

Problems of Control and Information Theory, Vol. 4 (1), pp. 71—94 (1975)

ПРИМЕНЕНИЕ СТРУКТУРНЫХ ПРЕОБРАЗОВАНИЙ К АНАЛИЗУ НЕСТАЦИОНАРНЫХ ЛИНЕЙНЫХ СИСТЕМ

Ф. А. МИХАЙЛОВ, О. Ф. ТЫХЕВИЧ, М. К. ХАДЖИНОВ

(Москва)

(Поступила в редакцию 2 июля 1973 г.)

Рассматривается задача вычисления передаточных функций различных соединений нестационарных линейных систем при известных передаточных функциях систем, входящих в соединение.

Используя левое преобразование Лапласа, получены точные формулы передаточных функций для параллельного, последовательного и циклического соединений. Отмечается, что формулы для последовательного и циклического соединений сложны для практического применения, и предлагается приближенный способ вычисления их передаточных функций. В этом способе передаточные функции системы и её звеньев аппроксимируются дробно-рациональными функциями комплексной переменной s с коэффициентами, зависящими от времени. Получены формулы, связывающие коэффициенты передаточной функции соединения с коэффициентами передаточных функций звеньев.

Приводится пример, иллюстрирующий применение полученных результатов.

Введение

При анализе сложных систем автоматического управления значительное удобство представляют структурные схемы, характеризующие систему как соединение более простых систем (звеньев). При исследовании стационарных линейных систем в качестве характеристик звеньев обычно используются передаточные функции, легко получаемые по дифференциальным уравнениям звеньев. Для нестационарных линейных систем в качестве характеристик звеньев также могут использоваться передаточные функции. Однако вопрос осложняется тем, что не все звенья имеют передаточную функцию [1], и что для нестационарных систем нет простой связи между дифференциальными уравнениями и передаточными функциями. Как следствие этого, в известных в настоящее время методах структурного анализа линейных нестационарных систем звенья характеризуются дифференциальными уравнениями [2, 3] или импульсными переходными функциями [4]. Вместе с тем представляет несомненный практический интерес метод анализа структурных схем линейных нестационарных систем, основанный на использовании передаточных функций так же, как и в случае стационарных систем.

В данной статье рассматривается задача вычисления передаточной функции системы по передаточным функциям её звеньев. При этом предполагается, что система и рассматриваемый интервал времени таковы, что передаточные функции звеньев и системы в целом существуют.

1. Структурные преобразования и формулы передаточных функций соединений

Всякая сложная нестационарная линейная система может быть представлена в виде соединения более простых систем, или звеньев. В общем случае систему можно представить в виде комбинации трех видов соединений: параллельного, последовательного и «циклического» (рис. 1). Найдём формулы для передаточных функций этих соединений.

Случай параллельного соединения двух звеньев показан на рис. 1а. На основании принципа суперпозиции передаточная функция этого соединения находится по формуле

$$W_n(s, t) = W_1(s, t) + W_2(s, t). \tag{1.1}$$

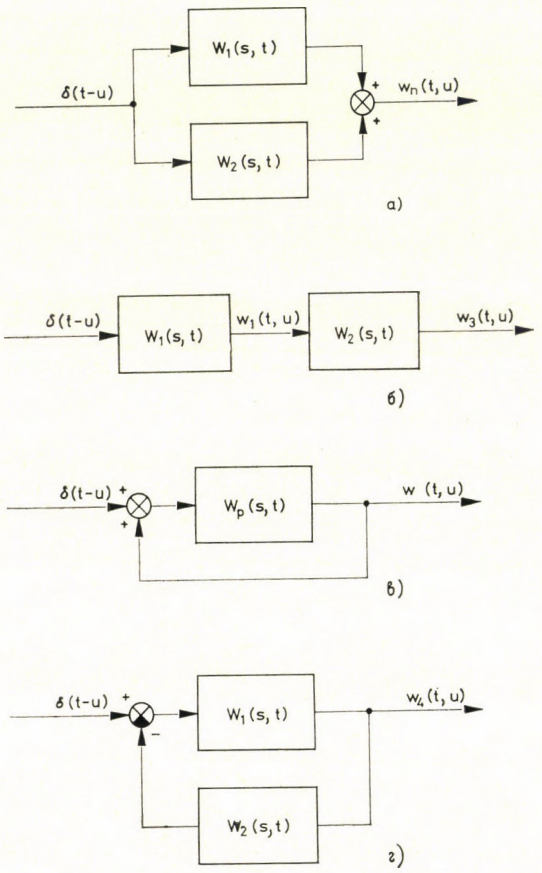


Рис. 1. Виды элементарных соединений

Случай последовательного соединения двух звеньев показан на рис. 1б. Для вычисления передаточной функции соединения $W_3(s, t)$ воспользуемся формулой для импульсной переходной функции последовательного соединения [4, стр. 1589, 10]

$$w_3(t, u) = \int_u^t w_2(t, \eta) \cdot w_1(\eta, u) \cdot d\eta. \tag{1.2}$$

Заменив переменную интегрирования η на $\xi = t - \eta$, получим

$$w_3(t, u) = \int_0^{t-u} w_2(t, t - \xi) \cdot w_1 (t - \xi, u) \cdot d\xi. \tag{1.3}$$

Подставим в (1.3) вместо функции $w_2(t, t- \xi)$ её выражение через передаточную функцию второго звена

$$w_2(t, t-\xi) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma, t) \cdot \exp(\sigma\xi) \cdot d\sigma, \tag{1.4}$$

где c — положительное число, выбранное с таким расчетом, чтобы интеграл абсолютно сходился при всех $Re\sigma \geq c$. В результате найдем

$$w_3(t, u) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma, t) \cdot \int_0^{t-u} w_1 (t - \xi, u) \cdot \exp(\sigma\xi) \cdot d\xi \cdot d\sigma. \tag{1.5}$$

Произведём замену u на $t-\tau$ и применим к обеим частям равенства (1.5) преобразование Лапласа по аргументу τ. После преобразования получим

$$W_3(s, t) = \frac{1}{2\pi i} \int_0^{\infty} \exp(-s\tau) \int_{c-i\infty}^{c+i\infty} W_2(\sigma, t) \int_0^{\tau} w_1 (t - \xi, t - \tau) \cdot$$

$$\cdot \exp(\sigma\xi)\, d\xi\, d\sigma\, d\tau\,. \tag{1.6}$$

Учитывая, что $w_1(t-\xi, t-\tau) = 0$ при $\xi > \tau$ верхний предел интегрирования по ξ заменим на ∞. Предположим, что все интегралы в (1.6) сходятся равномерно. В силу этого предположения ([11], стр. 62) можно изменить порядок интегрирования

$$W_3(s, t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma, t) \int_0^{\infty} \exp(\sigma\xi) \int_0^{\infty} w_1(t - \xi, t - \tau) \cdot$$

$$\cdot \exp(-s\tau)\, d\tau\, d\xi\, d\sigma. \tag{1.7}$$

Представим внутренний интеграл в виде

$$\int_0^\infty w_1[t-\xi, (t-\xi)-(\tau-\xi)] \cdot \exp(-s\tau)\, d\tau,$$

и ограничиваясь случаем $\xi \leq \tau$ (т. к. в области $\xi > \tau$ подынтегральное выражение равно нулю), по первой теореме смещения ([5], стр. 39) приведем его к виду

$$\exp(-s\xi) \cdot W_1(s,\ t-\xi), \tag{1.8}$$

где $W_1(s, t-\xi)$ — передаточная функция первого звена с комплексным аргументом s и вещественным $t-\xi$.

Теперь интеграл по ξ в формуле (1.7) можно рассматривать как левое изображение по Лапласу ([1], стр. 493) от функции $W_1(s, t)$, рассматриваемой как функция от t_1 со смещенным на комплексное число σ комплексным аргументом

$$\int_0^\infty W_1(s, t-\xi) \cdot \exp[-\xi(s-\sigma)]\, d\xi = \Lambda_{t/s-\sigma}[W_1(s, t)]. \tag{1.9}$$

Здесь нижние индексы при знаке левого преобразования Лапласа указывают соответственно вещественный и комплексный аргументы изображения.

Перепишем формулу (1.7) с учетом (1.8) и (1.9)

$$W_3(s, t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma, t) \cdot \Lambda_{t/s-\sigma}[W_1(s, t)]\, d\sigma. \tag{1.10}$$

Применяя формулу обратного левого преобразования Лапласа (см. приложение) и рассматривая формально подынтегральную функцию как левое изображение с комплексным аргументом σ некоторой функции от t с параметром s, можем записать

$$W_3(s, t) = \Lambda^{-1}_{t/\sigma}\{W_2(\sigma, t) \cdot \Lambda_{t/s-\sigma}[W_1(s, t)]\}. \tag{1.11}$$

Таким образом, передаточная функция последовательного соединения $W_3(s, t)$ выражается в явном виде через передаточные функции звеньев $W_1(s, t)$ и $W_2(s, t)$. Заметим, что эта зависимость является очень сложной.

Формула (1.11), очевидно, должна быть справедлива и для случая стационарной системы с передаточными функциями звеньев $W_1(s)$ и $W_2(s)$. Действительно,

$$\Lambda_{t/s-\sigma}[W_1(s)] = \frac{W_1(s)}{(s-\sigma)}, \tag{1.12}$$

и тогда

$$W_3(s) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} W_2(\sigma)\cdot\frac{W_1(s)}{s-\sigma}\,d\sigma = \frac{W_1(s)}{2\pi i}\int_{c+i\infty}^{c-i\infty}\frac{W_2(\sigma)}{\sigma-s}\,d\sigma\,. \quad (1.13)$$

Это равенство определено для $Re\sigma \geq c$. Но в этом случае на основании интегральной формулы Коши интеграл (1.13) равен $2\pi i W_2(s)$. Отсюда

$$W_3(s) = W_1(s)\cdot W_2(s). \quad (1.14)$$

«Случай «циклического» соединения показан на рис. 1в. Для нахождения передаточной функции такого соединения воспользуемся результатом, полученным для последовательного соединения. Импульсную переходную функцию $w_{ц}(t, u)$ можно определить соотношением ([4], стр. 1589)

$$w_{ц}(t, u) = w_p(t, u) + \int_u^t w_p(t, \eta)\cdot w_{ц}(\eta, u)\,d\eta. \quad (1.15)$$

По аналогии с (1.11) сразу запишем

$$W_{ц}(s, t) = W_p(s, t) + \Lambda^{-1}_{t/\sigma}\{W_p(\sigma, t)\cdot\Lambda_{t/s-\sigma}[W_{ц}(s, t)]\}. \quad (1.16)$$

Уравнение (1.16) является сложным интегральным уравнением, в котором $W_{ц}(s, t)$ — неизвестная функция. Получить из него в явном виде передаточную функцию соединения $W_{ц}(s, t)$ не удается. Так же, как и (1.11), формула (1.16) для случая стационарной системы превращается в известное соотношение

$$W_{ц}(s) = W_p(s) + W_p(s)\cdot W_{ц}(s), \quad (1.17)$$

которое разрешимо относительно функции $W_{ц}(s)$.

Таким образом, для всех трех видов соединений мы располагаем формулами, выражающими передаточную функцию соединения через передаточные функции звеньев либо явно — (1.1) и (1.11), либо неявно — в виде интегрального уравнения (1.16). Однако формулы (1.11) и (1.16) неудобны для практического применения. Поэтому для последовательного и циклического соединений возникает задача о нахождении приближенных, но удобных в применении формул.

2. Степенное разложение импульсной переходной функции

Рассмотрим разложение импульсной переходной функции в степенной ряд по аргументу $\tau = t - u$

$$w(t, t-\tau) = \sum_{k=0}^{\infty} w_k(t)\cdot\frac{\tau^k}{k!}, \quad (2.1)$$

предположив, что разложение сходится в некоторой области $\tau \leq Q$.

В [1, стр. 245] показано, что в случае системы, описываемой дифференциальным уравнением

$$\frac{d^n x}{dt^n} + b_1(t)\frac{d^{n-1}x}{dt^{n-1}} + \ldots + b_n(t)\,x = a_0(t)\frac{d^{n-1}y}{dt^{n-1}} + \ldots + a_{n-1}(t)\,y \qquad (2.2)$$

с коэффициентами $a_{i-1}(t)$ и $b_i(t)$, $i = 1, \ldots, n$, являющимися аналитическими функциями, коэффициенты $w_k(\tau)$ могут быть выражены через эти функции.

Импульсная переходная функция $w(t, u)$ может быть найдена ([2], стр. 114) по формуле

$$w(t, u) = \sum_{i=0}^{n-1} (-1)^{n-i-1} \cdot \frac{d^{n-i-1}}{du^{n-i-1}} [a_i(u) \cdot g(t, u)], \qquad (2.3)$$

в правой части которой фигурирует функция $g(t, u)$ и коэффициенты правой части уравнения (2.2). В силу того, что коэффициенты уравнения (2.2) —аналитические функции, можно разложить $a_i(u)$ в сходящиеся ряды Тейлора в некоторой окрестности точки $u = t$, а именно

$$a_i(u) = \sum_{j=0}^{\infty} a_i^{(j)}(t) \cdot \frac{(t-u)^j}{j!} (-1)^j, \qquad (2.4)$$

где $a_i^{(j)}(t)$ — производная порядка j по t от функции $a_i(t)$. Здесь и всюду в дальнейшем под производной нулевого порядка $a_i^{(0)}(t)$ от функции $a_i(t)$ подразумевается сама функция $a_i(t)$.

Функция $g(t,u)$ — импульсная переходная функция интегрирующей части системы (т. е. системы, описываемой уравнением (2.2) с правой частью, замененной на $y(t)$). Для нее справедливо разложение [1, стр. 244]

$$g(t, t-\tau) = \sum_{l=n=1}^{\infty} g_{n,l}(t) \cdot \frac{\tau^l}{l!}, \qquad (2.5)$$

коэффициенты которого могут быть выражены через $b_1(t), \ldots, b_n(t)$. Эти зависимости для $n \leq 4$, $l \leq n+4$ приведены в таблице 1.

Подставив (2.3) и (2.4) в (2.5) с учетом того, что произведение двух сходящихся степенных рядов есть тоже сходящийся ряд ([9], стр. 51), получим

$$w(t, u) = \sum_{i=0}^{n-1} \sum_{l=n-1}^{\infty} \sum_{j=0}^{\infty} (-1)^{n-1-i+j} \cdot a_i^{(j)}(t) \cdot g_{n,l}(t) \cdot \frac{d^{n-1-i}}{dt^{n-1-i}} \left[\frac{(t-u)^{l+j}}{l!\,j!}\right].$$

Таблица 1

n	$g_{n,e}$
1	$g_{1,0} = 1$ $g_{1,1} = -b_1$ $g_{1,2} = b_1^2 + \dot{b}_1$ $g_{1,3} = -b_1^3 - 3b_1^2\dot{b}_1 - \ddot{b}_1$ $g_{1,4} = b_1^4 + 6b_1^2\dot{b}_1 + 4b_1\ddot{b}_1 + 3\dot{b}_1^2 + \dddot{b}_1$ $g_{1,5} = -b_1^5 - 10b_1^3\dot{b}_1 - 10b_1^2\ddot{b}_1 - 15b_1\dot{b}_1^2 - 5b_1\dddot{b}_1 - 10\dot{b}_1\ddot{b}_1$
2	$g_{2,1} = 1$ $g_{2,2} = -b_1$ $g_{2,3} = b_1^2 + 2\dot{b}_1 - b_2$ $g_{2,4} = -b_1^3 + 2b_1b_2 - 5b_1\dot{b}_1 + 2\dot{b}_2 - 3\ddot{b}_1$ $g_{2,5} = b_1^4 - 3b_1^2b_2 + b_2^2 + 9b_1^2\dot{b}_1 - 5b_1\dot{b}_2 + 9b_1\ddot{b}_1 + 8\dot{b}_1^2 - 6\dot{b}_1b_2 + 4\dddot{b}_1 - 3\ddot{b}_2$ $g_{2,6} = -8b_1^5 + 4b_1^3b_2 - 3b_1b_2^2 - 14b_1^3\dot{b}_1 + 9b_1^2\dot{b}_2 - 19b_1^2\ddot{b}_1 - 15b_1\dot{b}_1^2 +$ $+ 21b_1\dot{b}_1b_2 - 14b_1\dddot{b}_1 + 9b_1\ddot{b}_2 - 6b_2\dot{b}_2 + 3\dot{b}_1\dot{b}_2 + 13\ddot{b}_1b_2 + 4\dddot{b}_2 - 35\dot{b}_1\ddot{b}_1 - 4\overset{(VI)}{b}_1$
3	$g_{3,2} = 1 \qquad g_{3,3} = -b_1$ $g_{3,4} = b_1^2 - b_2 + 3\dot{b}_1$ $g_{3,5} = -b_1^3 + 2b_1b_2 - b_3 - 7b_1\dot{b}_1 + 3\dot{b}_2 - 6\ddot{b}_1$ $g_{3,6} = b_1^4 - 3b_1^2b_2 + 2b_1b_3 + b_2^2 + 12b_1^2\dot{b}_1 - 7b_1\dot{b}_2 + 16b_1\ddot{b}_1 + 15\dot{b}_1^2 - 8\dot{b}_1b_2 + 10\dddot{b}_1 -$ $- 6\ddot{b}_2 + 3\dot{b}_3$ $g_{3,7} = -b_1^5 + 4b_1^3b_2 - 3b_1^2b_3 - 3b_1b_2^2 + 2b_2b_3 - 18b_1^3\dot{b}_1 + 12b_1^2\dot{b}_2 - 31b_1^2\ddot{b}_1 - 57b_1\dot{b}_1^2 +$ $+ 27b_1\dot{b}_1b_2 - 30b_1\dddot{b}_1 + 16b_1\ddot{b}_2 - 7b_1\dot{b}_1 - 8b_2\dot{b}_2 - 9\dot{b}_2b_3 + 33\dot{b}_1\dot{b}_2 + 21\ddot{b}_1b_2 -$ $- 6\ddot{b}_3 + 10\dddot{b}_2 - 81\dot{b}_1\ddot{b}_1 - 14\overset{(IV)}{b}_1$
4	$g_{4,3} = 1 \qquad g_{4,4} = -b_1$ $g_{4,5} = b_1^2 - b_2 + 4\dot{b}_1$ $g_{4,6} = -b_1^3 + 2b_1b_2 - b_3 - 9b_1\dot{b}_1 + 4\dot{b}_2 - 10\ddot{b}_1$ $g_{4,7} = b_1^4 - 3b_1^2b_2 + 2b_1b_3 + b_2^2 - b_4 + 15b_1^2\dot{b}_1 - 9b_1\dot{b}_2 + 25b_1\ddot{b}_1 + 24\dot{b}_1^2 - 10\dot{b}_2\dot{b}_1 +$ $+ 20\dddot{b}_1 - 10\ddot{b}_2 + 4\dot{b}_3$ $g_{4,8} = -b_1^5 + 4b_1^3b_2 - 3b_1^2b_3 - 3b_1b_2^2 + 2b_1b_4 + 2b_2b_3 - 22b_1^3\dot{b}_1 + 15b_1^2\dot{b}_2 - 87b_1\dot{b}_1^2 +$ $+ 33b_1b_2\dot{b}_1 - 55b_1\dddot{b}_1 + 25b_1\ddot{b}_2 - 9b_1\dot{b}_3 - 10b_2\dot{b}_2 - 11\dot{b}_1\dot{b}_3 + 4\dot{b}_4 - 46b_1^3\ddot{b}_1 +$ $+ 52\dot{b}_1\dot{b}_2 + 31b_2\ddot{b}_1 - 10\ddot{b}_3 + 20\dddot{b}_2 - 154\dot{b}_1\ddot{b}_1 - 34\overset{(IV)}{b}_1$

Полагая, что $t-u=\tau$ и приравнивая коэффициенты при одинаковых степенях τ, получим

$$w_k(t)=\sum_{i=0}^{n-1}\sum_{l=n-1}^{n-1+k-i}(-1)^{n-1+k-i-l}\cdot\binom{n-1+k-i}{l}\cdot a_i^{(n-1+k-i-l)}(t)\cdot g_{n,l}(t), \quad (2.6)$$

где $\binom{n-1+k-i}{l}$ — биномиальные коэффициенты.

3. Степенное разложение передаточной функции

Рассматривая в правой части (2.1) τ как независимую переменную, t — как параметр, и преобразуя почленно ряд по Лапласу, получим формальное разложение в ряд для передаточной функции

$$W(s,t)\sim\sum_{k=0}^{\infty}\frac{w_k(t)}{s^{k+1}}. \quad (3.1)$$

Как показано в [5, стр. 188], ряд (3.1) не всегда может оказаться сходящимся, но всегда представляет собой асимптотическое разложение функции $W(s, t)$ при $s\to\infty$. Обратно, по разложению (3.1) легко восстановить ряд (2.1), т. е. значения $w(t, t-\tau)$ при малых τ.

4. Аппроксимация передаточной функции дробно-рациональной функцией

В общем случае для нестационарных систем найти передаточную функцию в аналитическом виде не удается.

Для ее аппроксимации удобнее всего дробно-рациональная функция вида

$$\tilde{W}(s,t)=\frac{c_0(t)\cdot s^{q-1}+\ldots+c_{q-2}(t)\cdot s+c_{q-1}(t)}{s^q+d_1(t)\cdot s^{q-1}+\ldots+d_q(t)}, \quad (4.1)$$

где $c_i(t)$, $i=0,1,\ldots,q-1$; $d_i(t)$, $i=1,2,\ldots,q$ — вещественные непрерывные функции времени.

Передаточная функция $W(s, t)$ вида (4.1) однозначная и аналитическая для всех достаточно больших $|s|$, поэтому она всегда допускает в этой окрестности разложение в ряд Лорана

$$\tilde{W}(s,t)=\sum_{k=0}^{\infty}\frac{\tilde{w}_k(t)}{s^{k+1}}.$$

Коэффициенты $\tilde{w}_k(t)$ этого ряда определяются через коэффициенты передаточной функции (4.1) по рекуррентной формуле

$$\tilde{w}_k(t) = c_k(t) - \sum_{i=1}^{q} d_i(t) \cdot \tilde{w}_{k-i}(t), \qquad (4.2)$$

$$k = 0, 1, \ldots, \infty,$$

причем $c_k = 0$ при $k > q - 1$; $\tilde{w}_k(t) = 0$ при $i > k$.

Согласно формуле (4.2) все коэффициенты передаточной функции (4.1) могут быть определены, если известны $2q$ первых коэффициентов $\tilde{w}_k(t)$. Поэтому в основу аппроксимации положим требование

$$\tilde{w}_k(t) = w_k(t), \qquad k = 0, 1, \ldots, 2q-1, \qquad (4.3)$$

где $w_k(t)$ — коэффициенты разложения (3.1).

Решение задачи аппроксимации передаточной функции таким способом неоднозначно, так как зависит от выбора q — степени полинома знаменателя аппроксимирующей дробно-рациональной функции. К вопросу выбора q обратимся позже, а пока примем, что q уже выбрано. Тогда коэффициенты передаточной функции (4.1) в дробно-рациональном виде связаны с первыми $2q$ коэффициентами аппроксимирующего ряда системой линейных алгебраических уравнений, получаемых из формулы (4.2) ($k = 0, 1, \ldots, 2q-1$).

Путь решения системы следующий. Выделив систему уравнений относительно коэффициентов $d_i(t)$ ($k = q, q + 1, \ldots, 2q-1$), находим из неё все $d_i(t)$. Из оставшихся уравнений ($k = 0, 1, \ldots, q-1$), подставив в них $d_i(t)$, получим формулы для всех коэффициентов $c_k(t)$.

Возможен случай, когда система алгебраических уравнений относительно коэффициентов $d_i(t)$ не имеет решения, так как определитель системы тождественно равен нулю. В таком случае следует составить новую систему уравнений, увеличив q.

5. Приближенные передаточные функции соединений

Представим импульсные переходные функции звеньев и их соединений в виде ряда (2.1). Воспользовавшись этой формулой, можно получить формулы для функции $W_3(s, t)$, аппроксимирующей передаточную функцию последовательного соединения, коэффициенты которой выражаются через коэффициенты передаточных функций звеньев. С этой целью можно использовать полученные в [6, стр. 119] соотношения, связывающие импульсные переходные функции соединений с импульсными переходными функциями звеньев для значений аргумента τ из общей области сходимости для соединения и звеньев.

Коэффициенты $w_{3,r}(t)$, $r = 0, 1, \ldots$ (первый индекс — индекс звена или соединения) разложения импульсной переходной функции $w_3(t, u)$ последовательного соединения двух звеньев (рис. 1б) находятся по формуле ([6], стр. 119)

$$w_{3,r}(t) = \sum_{l=0}^{r-1} \sum_{k=0}^{r-1-l} (-1)^{r-1-l-k} \cdot \binom{r-1-l}{k} \cdot w_{2,k}(t) \cdot w_{1,l}^{(r-1-l-k)}(t). \quad (5.1)$$

Эта формула позволяет синтезировать приближенную передаточную функцию последовательного соединения по заданным приближенным передаточным функциям звеньев. При этом учитываются равенства (4.3).

Последовательность операций при выводе формул для коэффициентов передаточной функции соединения следующая.

А) Находим выражения коэффициентов $w_{i,j}(t)$ степенного разложения заданной приближенной передаточной функции каждого звена через коэффициенты $c_i(t)$ и $d_i(t)$.

Б) Задаемся степенью q знаменателя приближенной передаточной функции соединения и по формуле (5.1) находим первые $2q$ коэффициентов её степенного разложения.

В) Полученные коэффициенты степенного разложения передаточной функции соединения подставляем в систему (4.2) уравнений относительно коэффициентов $c_i(t)$ и $d_i(t)$ искомой передаточной функции.

Решив эту систему уравнений, получим формулы, выражающие коэффициенты передаточной функции соединения через передаточные функции звеньев, входящих в соединение.

Для циклического соединения (рис. 1в) коэффициенты $w_{u,r}(t)$ разложения импульсной переходной функции соединения выражаются в виде рекуррентной формулы через коэффициенты разложения импульсной переходной функции разомкнутой системы

$$w_{u,r}(t) = w_{p,r}(t) + \sum_{l=0}^{r-1} \sum_{k=0}^{r-1-l} (-1)^{r-l-1-k} \cdot \binom{r-1-l}{k} w_{p,k}(t) \cdot w_{u,l}^{(r-1-l-k)}(t), \quad (5.2)$$

$$w_{u,0}(t) = w_{p,0}(t), \qquad r = 1, 2, \ldots$$

Последовательность операций нахождения коэффициентов передаточной функции циклического соединения такая же, как и для последовательного соединения.

В случае, когда приближенные передаточные функции звеньев, входящих в соединение, неизвестны, и звенья характеризуются дифференциальными уравнениями, заданными на интервале $(-\infty, T]$ с коэффициентами, являющимися аналитическими функциями, вышеизложенная методика также применима. Только в этом случае коэффициенты степенного разложе-

ния для звеньев определяются на основании формулы (2.6) через коэффициенты их уравнений. Все последующие операции не отличаются от описанных выше.

По описанной выше методике были выведены формулы для коэффициентов передаточной функции последовательного и циклического соединения звеньев для случая, когда показатель степени полинома знаменателя приближенной передаточной функции соединения берется равным сумме показателей степеней знаменателей передаточных функций звеньев. Формулы сведены в таблицы.

В таблице 2 через $W_1(s, t)$ и $W_2(s, t)$ обозначены передаточные функции звеньев, через $W_3(s, t)$ — передаточная функция их последовательного соединения (рис. 1б). В таблице 3 передаточная функция разомкнутой цепи обозначена $W_p(s, t)$, а циклического соединения $-W_ц(s, t)$ (рис. 1в). В таблице 4 приведены формулы для других видов соединений с обратной связью. Здесь $W_1(s, t)$ — передаточная функция прямой цепи, $W_2(s, t)$ — цепи обратной связи, $W_4(s, t)$ — всего соединения (рис. 1г).

6. Условия применения метода и пример

Предложенный метод основан на разложении в степенной ряд импульсной переходной функции $w(t, u)$ в некоторой окрестности точки $t = u$ по степеням $\tau = t-u$. Быстрота сходимости ряда внутри круга сходимости понижается с удалением от точки разложения, поэтому точность формул (5.1) и (5.2) уменьшается с увеличением $\tau = t-u$. В то же время аппроксимация передаточной функции соединения производится из условия совпадения некоторого числа первых членов этого разложения. Поэтому метод дает хорошие результаты для систем, сигнал на выходе которых определяется, главным образом, значениями входного сигнала в достаточно близкие к моменту наблюдения реакций моменты времени, т. е. для систем с быстрозатухающей по аргументу τ импульсной переходной функцией. Расчет по полученным формулам в случае стационарных систем дает точные результаты. Поэтому предложенный метод структурных преобразований также эффективен для систем с медленно меняющимися параметрами.

Пример. Воспользуемся результатами данной работы для получения передаточной функции и дисперсии выходного сигнала нестационарной системы, представляющей собой простейшую математическую модель одной задачи управления движением.

Рассмотрим задачу сближения двух тел в плоскости, из которых одно, неуправляемое, движется равномерно и прямолинейно, а второе имеет постоянную скорость движения и управляется с целью сближения с первым

Таблица 2

$\frac{N}{n/n}$	Последовательное соединение	
1	$\widetilde{W}_1 = \frac{c}{s+d}$ $\widetilde{W}_2 = \frac{c_0}{s+d_1}$ $\widetilde{W}_3 = \frac{c_0^* s + c_1^*}{s^2 + d_1^* s + d_2^*}$	$d_1^* = d + d_1 + \frac{\dot{c}}{c}$; $d_2^* = dd_1 + d\frac{\dot{c}}{c} - \dot{d} - \frac{\ddot{c}}{c} + \left(\frac{\dot{c}}{c}\right)^2$; $c_0^* = 0; \quad c_1^* = cc_0$;
2	$\widetilde{W}_1 = \frac{c}{s+d}$ $\widetilde{W}_2 = \frac{c_1}{s^2 + d_1 s + d_2}$ $\widetilde{W}_3 = \frac{c_0^* s^2 + c_1^* s + c_2^*}{s^3 + d_1^* s^2 + d_2^* s + d_3^*}$	$d_1^* = d + d_1 + 2\frac{\dot{c}}{c}$; $d_2^* = d_2 + dd_1 + \frac{\dot{c}}{c}(2d + d_1) + 4\left(\frac{\dot{c}}{c}\right)^2 - 3\frac{\ddot{c}}{c} - 2\dot{d}$; $d_3^* = dd_2 + 2dd_1\frac{\dot{c}}{c} - d_1\dot{d} - 3d\frac{\ddot{c}}{c} + 3\ddot{d} - 4\frac{\dddot{c}}{c} - 4d\left(\frac{\dot{c}}{c}\right)^2 - 12\frac{\dot{c}\ddot{c}}{c^2} - 2\frac{\dot{c}}{c}\dot{d}$; $c_0^* = c_1^* = 0; \quad c_2^* = cc_1$;
3	$\widetilde{W}_1 = \frac{c_1}{s^2 + d_1 s + d_2}$ $\widetilde{W}_2 = \frac{c}{s+d}$ $\widetilde{W}_3 = \frac{c_0^* s^2 + c_1^* s + c_2^*}{s^3 + d_1^* s^2 + d_2^* s + d_3^*}$	$d_1^* = d_1 + d + \frac{\dot{c}_1}{c_1}$ $d_2^* = d_2 + dd_1 + d_1\frac{\dot{c}_1}{c_1} - \dot{d}_1 - \frac{\ddot{c}_1}{c_1} + \left(\frac{\dot{c}_1}{c_1}\right)^2$ $d_3^* = dd_2 + d_2\frac{\dot{c}_1}{c_1} + \frac{\dot{c}_1}{c_1}(d - d_1) + \left(\frac{\dot{c}_1}{c_1}\right)^2(d_1 - d) + \frac{\dddot{c}_1}{c_1} + \left(\frac{\dot{c}_1}{c_1}\right)^3 + \ddot{d}_1 - \dot{d}_2 - 2\frac{\dot{c}_1\ddot{c}_1}{c_1^2}$; $c_0^* = c_1^* = 0; \quad c_2^* = cc_1$;

продолжение таблицы 2

*9		
4	$\widetilde{W}_1 = \dfrac{c}{s^2 + ds + f}$ $\widetilde{W}_2 = \dfrac{c_1}{s^2 + d_1 s + d_2}$ $\widetilde{W}_3 = \dfrac{c_0^* s^3 + c_1^* s^2 + c_2^* s + c_3^*}{s^4 + d_1^* s^3 + d_2^* s^2 + d_3^* s + d_4^*}$	$c_0^* = c_1^* = c_2^* = 0; \quad c_3^* = cc_1; \quad d_1^* = d + d_1 + 2\dfrac{\dot c}{c};$ $d_2^* = d_2 + dd_1 + f + \dfrac{\dot c}{c}(2d + d_1) + 4\left(\dfrac{\dot c}{c}\right)^2 - 3\dfrac{\ddot c}{c} - 2\dot d;$ $d_3^* = dd_2 + d_1 f + \dfrac{\dot c}{c}(2d_1^2 + 3dd_1 - d^2 - 2d_2 - f) - d\dot d - 5d\dfrac{\ddot c}{c} + \dfrac{\dddot c}{c} - 2\dot f + \ddot d -$ $- 6\dot d\dfrac{\dot c}{c} - 12\dfrac{\dot c \ddot c}{c^2};$ $d_4^* = d_2 f - \dfrac{\dot c}{c}(10d^3 + 22d^2 d_1 + 6dd_1^2 + 12d_1^3 - 10df - 15d_1 f - 8dd_2 - 12d_1 d_2) - 8\dot d(d^2 +$ $+ dd_1 + d_1^2 - f - d_2) - \dfrac{\ddot c}{c}(10d^2 + 10dd_1 + 7d_1^2 - 3f - 10d_2) - 4\ddot d(d + d_1) - \left(\dfrac{\dot c}{c}\right)^2(4d^2 -$ $- 4dd_1 - 3d_1^2 - 8f + 4d_2) - d_1\dot f - 8d_1\dot d\dfrac{\dot c}{c} - 2\dfrac{\dot c \ddot c}{c^2}(4d - 3d_1);$
5	$W_1 = \dfrac{c}{s + d}$ $\widetilde{W}_2 = \dfrac{c_2}{s^3 + d_1 s^2 + d_2 s + d_3}$ $\widetilde{W}_3 = \dfrac{c_0^* s^3 + c_1^* s + c_2^* s + c_3^*}{s^4 + d_1^* s^3 + d_2^* s^2 + d_3^* s + d_4^*}$	$c_0^* = c_1^* = c_2^* = 0; \quad c_3^* = cc_2; \quad d_1^* = d + d_1 + 3\dfrac{\dot c}{c};$ $d_2^* = d_2 + dd_1 + \dfrac{\dot c}{c}(3d + 2d_1) - 3\dot d - 6\dfrac{\ddot c}{c} + 9\left(\dfrac{\dot c}{c}\right)^2;$ $d_3^* = d_3 + dd_2 - \dfrac{\dot c}{c}(8d^2 + 14dd_1 + 8d_1^2 - d_2) + 2d_1\dot d - \dfrac{\ddot c}{c}(18d + 22d_1) + 3\left(\dfrac{\dot c}{c}\right)^2(3d +$ $+ d_1) + 27\left(\dfrac{\dot c}{c}\right)^3 - 30\dot d\dfrac{\dot c}{c} - 6\ddot d - 10\dfrac{\dddot c}{c};$ $d_4^* = dd_3 - \dfrac{\dot c}{c}(12d^3 + 20d^2 d_1 + 28dd_1^2 + 16d_1^3 - 13dd_2 - 14d_1 d_2 + 6d_3) - \dot d(12d^2 + 6dd_1 +$ $+ 12d_1^2 - 11d_2) - \dfrac{\ddot c}{c}(30d + 40d_1) - \dfrac{\ddot c}{c}(48d^2 + 90dd_1 + 65d_1^2 - 27d_2) - \left(\dfrac{\dot c}{c}\right)^2 \cdot (18d^2 +$ $+ 45dd_1 + 31d_1^2 + 3d_2) - 27\left(\dfrac{\dot c}{c}\right)^3 \cdot (3d + d_1) - \dfrac{\dot c \ddot c}{c_2}(144d + 168d_1) - 10\dddot d + 81\left(\dfrac{\dot c}{c}\right)^4 -$ $- \dfrac{\dot c}{c}\dot d(66d + 74d_1) - 21\dot d^2 - 153\dot d\left(\dfrac{\dot c}{c}\right)^2 - 162\dfrac{\ddot c}{c}\left(\dfrac{\dot c}{c}\right)^2 - 30\dot d\dfrac{\dot c}{c} - \ddot d(24d + 22d_1) -$ $- 66\dfrac{\dot c}{c}\ddot d - 15\dfrac{c^{(IV)}}{c};$

Таблица 3

N n/n	Циклическое соединение	
1	$\widetilde{W}_p = \dfrac{c_0}{s + d_1}$ $\widetilde{W}_y = \dfrac{c_0^*}{s + d_1^*}$	$c_0^* = c_0$; $d_1^* = d_1 - c_0$.
2	$\widetilde{W}_p = \dfrac{c_0 s + c_1}{s^2 + d_1 s + d_2}$ $\widetilde{W}_y = \dfrac{c_0^* s + c_1^*}{s^2 + d_1^* s + d_2^*}$	$c_0^* = c_0$; $\quad c_1^* = c_1 + c_0(d_1^* - d_1 + c_0)$; $d_1^* = d_1 - c_0 + \dfrac{c_0 \dot{c}_0 (c_0 d_1 - c_0^2 - c_1) + c_0^2 (\ddot{c}_0 - \dot{c}_1 + c_0 \dot{d}_1)}{c_1^2 - c_0 c_1 d_1 + c_0^2 d_2 + c_0^2 \dot{c}_0}$; $d_2^* = d_2 - c_1 + 2\dot{c}_0 + \dfrac{c_0 (c_0 d_1 - c_0^2 - c_1)(\dot{c}_0 d_1 + \ddot{c}_0 - c_0 \dot{c}_1 + c_0^2 \dot{d}_1) + c^2 \dot{c}_0 (2c_1 - d_2 - \dot{c}_0 - c_0 d_1 + c_0^2)}{c_1^2 - c_0 c_1 d_1 + c_0^2 d_2 + c_0^2 \dot{c}_0}$.

телом ([7], стр. 262). Предметом анализа в данном примере будут нормальные ускорения управляемого тела в процессе сближения.

Запишем линеаризованное уравнение кинематики относительного движения в виде

$$D(t) \cdot \frac{d\dot{\varepsilon}}{dt} + 2\dot{D}\dot{\varepsilon} = n; \quad 0 \leq t \leq T, \tag{6.1}$$

где $D(t)$ — текущая дальность по опорной траектории; $\dot{\varepsilon}$ — угловая скорость линии визирования (прямой, соединяющей центры масс сближающихся тел); n — проекция ускорения объекта управления на нормаль к линии визирования; T — полное время сближения.

Звено, динамические свойства которого описываются уравнением (6.1), будем называть кинематическим звеном. Входной величиной кинематического звена будет n, выходной — $\dot{\varepsilon}$.

Считаем, что управление осуществляется по методу пропорциональной навигации — управляющий сигнал пропорционален $\dot{\varepsilon}_u$ — угловой скорости визирования, измеренной бортовым устройством. Этот сигнал, воздействуя на управляемый объект через бортовой контур управления, изменяет величину n. Считаем, что динамические свойства бортового контура управления характеризуются передаточной функцией колебательного звена с постоянными параметрами

$$W_1(s) = \frac{k\omega_0^2}{s^2 + 2\xi\omega_0 + \omega_0^2}. \tag{6.2}$$

Структурная схема рассматриваемой системы показана на рис. 2. Входной величиной системы является ошибка измерения бортовым измерительным устройством угловой скорости линии визирования, выходной величиной — n.

Ограничимся рассмотрением случая, когда на вход системы действует только случайное возмущение $\psi(t)$ типа белого шума.

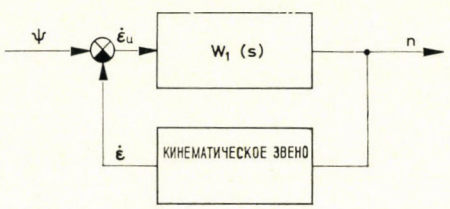


Рис. 2. Структурная схема системы управляемого сближения двух тел

Таблица 4

$\frac{N}{n/n}$	Соединение с обратной связью	
1	$\widetilde{W}_1 = \frac{c_0}{s + d_1}$	$c_0^* = c_0\,;\qquad c_1^* = c_0 d + \dot{c}_0 + c_0 \frac{\dot{c}}{c}$
	$\widetilde{W}_2 = \frac{c}{s + d}$	$d_1^* = d + d_1 + \frac{\dot{c}}{c} + \frac{\dot{c}_0}{c_0}\,;$
	$\widetilde{W}_4 = \frac{c_0^* s + c_1^*}{s^2 + d_1^* s + d_2^*}$	$d_2^* = dd_1 + cc_1 + d_1\left(\frac{\dot{c}}{c} + \frac{\dot{c}_0}{c_0}\right);$
2	$\widetilde{W}_1 = \frac{c_0 s + c_1}{s^2 + d_1 s + d_2}$	$d_1^* = d_1 + c_0 + \frac{c_0 \dot{c}_0 (c_1 - c_0 d_1 - c_0^2) + c_0^2 (\dot{c}_1 - \ddot{c}_0 - \dot{c}_0 d_1)}{c_1^2 + c_0^2 d_2 - c_0 c_1 d_1 - c_1^2 \dot{c}_0}\,;$
	$\widetilde{W}_2 = 1$	$d_2^* = d_2 + c_1 - 2\dot{c}_0 + \frac{c_0 (c_1 - c_0 d_1 - c_0^2)(\dot{c}_0 d_1 + \dot{c}_0 d_1 - \dot{c}_1 - \ddot{c}_0) - c_0^2 \dot{c}_0^2}{c_1^2 + c_0^2 d_2 - c_0 c_1 d_1 - c_0^2 \dot{c}_0}\,;$
	$\widetilde{W}_4 = \frac{c_0^* s + c_1^*}{s^2 + d_1^* s + d_2^*}$	$c_0^* = c_0\,;\qquad c_1^* = c_1 + c_0 (d_1^* - d_1 - c_0)\,;$
3	$\widetilde{W}_1 = \frac{c_1}{s^2 + d_1 s + d_2}$	$d_1^* = d_1 + d + 3\frac{\dot{c}_1}{c_1} + 2\frac{\dot{c}}{c}\,;$
	$\widetilde{W}_2 = \frac{c}{s + d}$	$d_1^* = dd_1 + d_2 + d_1\left(3\frac{\dot{c}_1}{c_1} + 2\frac{\dot{c}}{c}\right);$
	$\widetilde{W}_4 = \frac{c_0^* s^2 + c_1^* s + c_2^*}{s^3 + d_1^* s^2 + d_2^* s + d_3^*}$	$d_3^* = dd_2 + cc_1 + d_2\left(3\frac{\dot{c}_1}{c_1} + 2\frac{\dot{c}}{c}\right);$
		$c_0^* = 0\,;\qquad c_1^* = c_1\,;\qquad c_2^* = c_1 (d_1^* - d_1)$

Продолжение таблицы 4

4	$\widetilde{W}_1 = \dfrac{c}{s+d}$ $\widetilde{W}_2 = \dfrac{c_1}{s^2 + d_1 s + d_2}$ $\widetilde{W}_4 = \dfrac{c_0^* s^2 + c_1^* s + c_2^*}{s^3 + d_1^* s^2 + d_2^* s + d_3^*}$	$d_1^* = d + d_1 + 3\dfrac{\dot{c}}{c} + \dfrac{\dot{c}_1}{c_1}$; $d_2^* = dd_1 + d_2 + \dfrac{\dot{c}}{c}(5d - d_1) + \dfrac{\dot{c}_1}{c_1} d_1 + 2\dfrac{\dot{c}}{c}\dfrac{\dot{c}_1}{c_1} + 9\left(\dfrac{\dot{c}}{c}\right)^2 + \left(\dfrac{\dot{c}_1}{c_1}\right)^2 - 6\dfrac{\ddot{c}}{c} - \dfrac{\ddot{c}_1}{c_1} - \dot{d}_1$; $d_3^* = dd_2 + cc_1 + d\left[\dfrac{\dot{c}}{c}(d - 2d_1) + \dfrac{\dot{c}_1}{c_1}(d - d_1) + \dot{d}_1 + 6\dfrac{\ddot{c}}{c} + \dfrac{\ddot{c}_1}{c_1} - 2\dfrac{\dot{c}}{c}\dfrac{\dot{c}_1}{c_1} - 9\left(\dfrac{\dot{c}}{c}\right)^2 - \left(\dfrac{\dot{c}_1}{c_1}\right)^2\right]$; $c_0^* = c$; $\quad c_1^* = c(d_1^* - d_1)$; $\quad c_2^* = c(d_2^* - dd_1^* + d^2)$;
5	$\widetilde{W}_1 = \dfrac{c_2}{s^2 + d_1 s^2 + d_2 s + d_3}$ $\widetilde{W}_2 = \dfrac{c}{s+d}$ $\widetilde{W}_4 = \dfrac{c_0^* s^3 + c_1 s^2 + c_2^* s + c_3^*}{s^4 + d_1^* s^3 + d_2^* s^2 + d_3^* s + d_4}$	$c_0^* = c_1^* = 0$; $\quad d_1^* = d + d_1 + 3\dfrac{\dot{c}}{c} + 4\dfrac{\dot{c}_2}{c_2}$; $c_2^* = c_2$; $\quad d_2^* = d_2 + dd_1 + d_1\left(3\dfrac{\dot{c}}{c} + 4\dfrac{\dot{c}_2}{c_2}\right)$ $c_3^* = c_2\left(d + 3\dfrac{\dot{c}}{c} + 4\dfrac{\dot{c}_2}{c_2}\right)$; $\quad d_3^* = d_3 + dd_2 + d_2\left(3\dfrac{\dot{c}}{c} + 4\dfrac{\dot{c}_2}{c_2}\right)$; $d_4^* = dd_3 + cc_2 + d_3\left(3\dfrac{\dot{c}}{c} + 4\dfrac{\dot{c}_2}{c_2}\right)$;
6	$\widetilde{W}_1 = \dfrac{c_1 s + c_2}{s^3 + d_1 s^2 + d_2 s + d_3}$ $\widetilde{W}_2 = 1$ $\widetilde{W}_4 = \dfrac{c_0^* s^2 + c_1^* s + c_2^*}{s^3 + d_1^* s^2 + d_2^* s + d_3^*}$	$d_1^* = d_1 + \dfrac{c_1^2[\dot{c}_1(5c_2 - 7c_1 d_1) - 3c_1\ddot{c}_1 + 2c_1\dot{c}_2 - 2c_1^2 d_1]}{c_1 c_2^2 d_1 - c_1^2 c_2 d_2 + c_1^3 d_3 - c_2^3 + 2c_1^3\dot{c}_1}$; $d_2^* = d_2 + c_1 + \dfrac{c_1^2(c_2 - c_1 d_1)(2c_1\dot{d}_1 + 3\ddot{c}_1 - 2\dot{c}_2) + c_1\dot{c}_1(c_1^3 - 5c_2^2 - 9c_1^2 d_1^2 + 14c_1 c_2 d_1 + 2c_1^2 d_2)}{c_1 c_2^2 d_1 - c_1^2 c_2 d_2 + c_1^3 d_3 - c_2^3 + 2c_1^3\dot{c}_1}$; $d_3^* = d_3 + c_2 + \dfrac{\dot{c}_1(3c_2^2 - 12c_2^2 c_2^3 d_1 + 9c_1^2 c_2 d_1^2 + 7c_1^3 c_2 - 9c_1^3 d_1 d_2 + 4c_1^3 d_3 - 9c_1^4 d_1)}{c_1 c_2^2 d_1 - c_1^2 c_2 d_2 + c_1^3 d_3 - c_2^3 + 2c_1^3\dot{c}_1} +$ $+ \dfrac{+ (c_2^2 + c_1^3 - c_1 c_2 d_1 + c_1^2 d_2)(2c_1\dot{c}_1 - 2c_2^2 d_1 - 3c_1\ddot{c}_1) + 4c_2^3 c_1^2}{c_1 c_2^2 d_1 - c_1^2 c_2 d_2 + c_1^3 d_3 - c_2^3 + 2c_1^3\dot{c}_1}$ $c_0^* = 0$; $\quad c_1^* = c_1$; $\quad c_2^* = c_2 + c_1(d_1^* - d)$;

Окончание таблицы 4

N п/п	Соединение с обратной связью	
7	$\widetilde{W}_1 = \dfrac{c}{s+d}$	$d_1^* = d \qquad c_0^* = c$
	$\widetilde{W}_2 = \dfrac{c_0 s + c_1}{s^2}$	$d_2^* = cc_6$
	$\widetilde{W}_3 = \dfrac{c_0^* s^2}{s^3 + d_1^* s^2 + d_2^* s + d_3^*}$	$d_3^* = cc_1 - c\dot{c}_0 - 2\dot{c}c_0$
8	$\widetilde{W}_1 = \dfrac{c}{s^2 + d_1 s + d_2}$	$d_1^* = d_1 \qquad c_0^* = 0$
	$\widetilde{W}_2 = \dfrac{c_0 s + c_1}{s^2}$	$d_2^* = d_2 \qquad c_1^* = c$
		$d_3^* = cc_0$
	$\widetilde{W}_3 = \dfrac{c_0^* s^3 + c_1^* s^2}{s^4 + d_1^* s^3 + d_2^* s^2 + d_3^* s + d_4^*}$	$d_4^* = cc_1 - 2c\dot{c}_0 - 3\dot{c}c_0$

Для вычисления дисперсий выходного сигнала $\sigma^2(t)$ в этом случае достаточно знать спектральную плотность шума S и интегральную характеристику $\mu(t)$ — степень подвижности ([1], стр. 302), определяемую равенством

$$\mu(t) = \frac{1}{\pi} \int_0^\infty |W(i\omega, t)|^2 \, d\omega .$$

Расчетное соотношение имеет вид:

$$\sigma^2(t) = S \cdot \mu(t). \tag{6.3}$$

Так как в реальной системе в начальный момент процесса управления дисперсия выходной величины равняется нулю, то этой формулой можно пользоваться для вычисления дисперсии только после окончания переходного процесса.

Если передаточная функция системы аппроксимирована дробно-рациональной функцией вида (4.1), то степень подвижности $\mu(t)$ можно определить алгебраическим путем ([1], стр. 309). В частности, для $n = 3$

$$\mu(t) = \frac{c_0^2 d_2 + c_1^2 - 2c_0 c_2 + c_2^2 \cdot \dfrac{d_1}{d_3}}{2(d_1 d_2 - d_3)} ; \tag{6.4}$$

для $n = 4$

$$\mu(t) = \frac{1}{2H}\left[c_0^2(d_2 d_3 - d_1 d_4) + (c_1^2 - 2c_0 c_2) d_3 + (c_2^2 - 2c_1 c_3) d_1 + c_3^2 \frac{d_1 d_2 - d_3}{d_4}\right],$$

$$H = d_1(d_2 d_3 - d_1 d_4) - d_3^2. \tag{6.5}$$

Передаточную функцию системы получим с помощью формул из таблицы 3. Для пользования таблицами необходимо знать передаточные функции всех звеньев.

Уравнение кинетического звена (6.1) является линейным дифференциальным уравнением первого порядка с переменными коэффициентами и решение его может быть найдено в конечном виде. Импульсная переходная функция $w_2(t, u)$ при любом законе изменения дальности $D(t)$ имеет вид

$$w_2(t, u) = \frac{D(u)}{D^2(t)}.$$

Заменим u на $t-\tau$ и преобразуем $w_2(t, t-\tau)$ по Лапласу как функцию от τ. Тогда передаточная функция кинематического звена для случая $\dot{D} =$

const выражается следующей формулой

$$W_2(s,t) = \frac{\dfrac{1}{D(t)} \cdot \left(s - \dfrac{\dot{D}}{D(t)}\right)}{s^2}. \tag{6.6}$$

Рассматриваемой системе соответствует вариант 8 таблицы 4. По формулам из таблицы 3 вычисляем коэффициенты передаточной функции системы. Передаточная функция системы будет

$$W(s,t) = \frac{k\omega_0^2 s^2}{s^4 + 2\xi\omega_0 s^3 + \omega_0^2 s^2 + \dfrac{k}{D(t)}\omega_0^2 s + \dfrac{k\dot{D}}{D^2(t)}}. \tag{6.7}$$

Для сравнения найдем передаточную функцию системы с помощью других приближенных методов.

Применяя методы стационарных систем, по дифференциальному уравнению всей системы

$$\dddot{n}\left(2\xi\omega_0 + 2\frac{\dot{D}}{D(t)}\right)\ddot{n} + \left(\omega_0^2 + 4\xi\omega_0\frac{\dot{D}}{D(t)}\right)\dot{n} + \omega_0^2\frac{k+2\dot{D}}{D(t)}n = k\omega_0^2\left(\dot{\psi} + 2\frac{\dot{D}}{D(t)}\psi\right) \tag{6.8}$$

найдем передаточную функцию «замороженной» в момент t нестационарной системы

$$W(s,t) = \frac{k\omega_0^2\left(s + 2\dfrac{\dot{D}}{D(t)}\right)}{s^3 + \left(2\xi\omega_0 + 2\dfrac{\dot{D}}{D(t)}\right)s^2 + \left(\omega_0^2 + 4\xi\omega_0\dfrac{\dot{D}}{D(t)}\right)s + k\omega_0^2 \cdot \dfrac{1}{D(t)} + 2\omega_0^2\dfrac{\dot{D}}{D(t)}}. \tag{6.9}$$

Для получения передаточной функции можем применить способ, использующий разложение импульсной переходной функции в степенной ряд, (см. разд. 2), и найти коэффициенты передаточной функции (см. разд. 4). При этом степень полинома знаменателя положим такой же, как и в формуле (6.9). Опустив выкладки, запишем

$$W(s,t) = \frac{k\omega_0^2\left(s - \dfrac{\dot{D}}{D(t)}\right)}{s^3 + \left(2\xi\omega_0 - \dfrac{\dot{D}}{D(t)}\right)s^2 + \left(\omega_0^2 - 2\xi\omega_0\dfrac{\dot{D}}{D(t)}\right)s + \omega_0\left(\dfrac{k-\dot{D}}{D(t)}\right)}. \tag{6.10}$$

Еще один вариант передаточной функции системы получим, применив к $W_1(s)$ и $W_2(s, t)$ формулу для преобразования структурных схем стационарных систем,

$$W(s,t) = \frac{k\omega_0^2 s^2}{s^4 + 2\xi\omega_0 s^3 + \omega^2 s^2 + \frac{k}{D(t)}\omega_0^2 s + \frac{k\dot{D}}{D^2(t)}\omega_0^2}. \tag{6.11}$$

Для всех вариантов передаточной функции по формулам (6.3—6.5) находим дисперсию выходного сигнала системы

$$\sigma_1^2 = S \cdot \frac{k^2\omega_0^2}{2\left(2\xi\omega_0 - 4\xi^2\frac{\dot{D}}{D(t)} - \frac{k}{D(t)}\right)} \tag{6.12}$$

$$\sigma_2^2 = S \cdot \frac{k^2\omega_0\left(\omega_0^2 D^2(t) + 8\dot{D}^2\frac{\xi\omega_0 D(t) + D}{k + 2\dot{D}}\right)}{2\left[8\xi\dot{D}^2 + \omega_0(2\xi\omega_0 D^2(t) + 2\xi^2\dot{D}D(t) - kD(t))\right]}; \tag{6.13}$$

$$\sigma_3^2 = S\frac{k^2\omega_0\left(\omega_0^2 D^2(t) + \dot{D}^2\frac{2\xi\omega_0 D(t) - \dot{D}}{k - \dot{D}}\right)}{2\left[2\xi\dot{D}^2 + \omega_0(2\xi\omega_0 D^2(t) - 4\xi^2\dot{D}D(t) - kD(t))\right]}; \tag{6.14}$$

$$\sigma_4^2 = S\frac{k^2\omega_0^2}{2\left(2\xi\omega_0 + 4\xi^2\frac{D}{D(t)} - \frac{k}{D(t)}\right)}, \tag{6.15}$$

где σ_1 соответствует передаточной функции (6.7), σ_2 — (6.9), σ_3 — (6.10), σ_4— (6.11).

На рис. 3 приведены графики изменения σ_1, σ_2, σ_3 и σ_4 по дальности, полученные по четырем формулам (6.12), (6.13), (6.14), (6.15). Графики построены для следующих значений параметров системы: $k = 5$ с/град; $w_0 = 10$ с$^{-1}$; $\xi = 0{,}7$; $S = 0{,}004$ град^2/с. Точные значения дисперсии σ^2 вычислялись как решение системы уравнений, составленной по методу Дункана [8, стр. 223]. Решение получено на ЦВМ для тех же параметров системы и входного сигнала. График σ приведен на рис. 3. Как видно из рисунка, наиболее близко к точному решение, полученное по формуле (6.12).

Заключение

Приведенный в статье метод структурных преобразований позволяет в ряде случаев существенно упростить процесс получения передаточной функции нестационарной линейной системы и тем самым анализ этой системы. Процедура применения метода для нахождения передаточной функции сис-

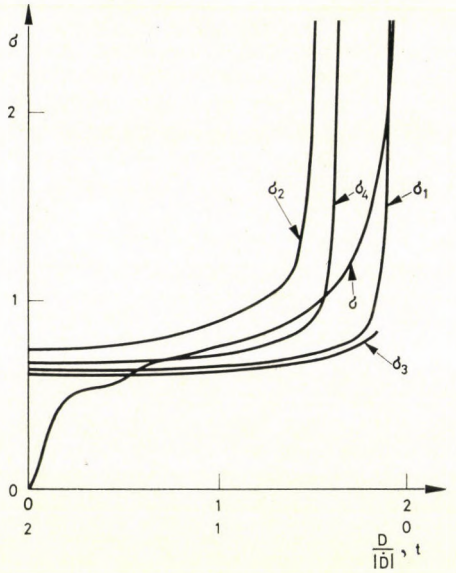


Рис. 3. Зависимость среднеквадратичного отклонения выходного сигнала от времени

темы заключается в использовании формул, приведенных в таблицах 1—4. Если в таблице не окажется подходящего варианта, то передаточную функцию соединения можно получить, находя последовательно коэффициенты степенного разложения передаточных функций звеньев, коэффициенты степенного разложения передаточной функции соединения, коэффициенты передаточной функции соединения.

Приложение

Левое преобразование Лапласа

Рассмотрим функцию $g(t)$, заданную на интервале $(-\infty, T]$, где $T < \infty$, и удовлетворяющую условиям, обеспечивающим сходимость интеграла

$$\int_{-\infty}^{T} |g(t)| \cdot \exp\,[s(t-T)] \cdot dt < \infty \qquad (1)$$

при всех значениях $t \in (-\infty, T]$ и при каком-либо вещественном значении c комплексной переменной s. Если ввести новую переменную τ, принимающую значения из интервала $[0, \infty)$, то функцию $g(t-\tau)$ при любом фиксированном значении t можно преобразовать по Лапласу. В результате этого преобразо-

вания получим функцию

$$G(s,\ t) = \int_0^\infty g(t-\tau) \cdot \exp \cdot (-s\tau)\ d\tau, \tag{2}$$

область определения которой согласно неравенству (1): $Re\ s \geq c$.

Переход от функции $g(t)$ к функции $G(s,\ t)$ назван левым преобразованием Лапласа [1, стр. 493].

Левое преобразование Лапласа ставит в соответствие каждой функции $g(t)$ с указанной выше областью определения функцию $G(s,\ t)$ комплексной переменной s и вещественной переменной t. Если функции $g(t)$ и $G(s, t)$ связаны левым преобразованием Лапласа, то функцию $g(t)$ можно назвать оригиналом, а функцию $G(s,\ t)$ — её левым изображением. Символически указанная связь функций $g(t)$ и $G(s,\ t)$ обозначается следующим образом

$$G(s,\ t) = \Lambda[g(t)]:\ \ g(t) = \Lambda^{-1}[G(s,\ t)]. \tag{3}$$

Из теории обычного преобразования Лапласа следует, что функция $G(s, t)$ — аналитическая по s при $Re\ s > c$.

На основании теоремы обращения обычного преобразования Лапласа обратное левое преобразование Лапласа определяется равенством

$$g(t) = \lim_{\tau \to 0} \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} G(s,t) \cdot \exp\ (s\tau) \cdot ds, \tag{4}$$

где $i = \sqrt{-1}$.

Литература

1. *Михайлов Ф. А., Теряев Е. Д., Булеков В. П., Саликов Л. М., Диканова Л. С.* Динамика непрерывных линейных систем с детерминированными и случайными параметрами. «Наука», М., 1971.
2. *Солодов А. В., Петров Ф. С.* Линейные автоматические системы с переменными параметрами. «Наука, М., 1971.
3. Современная теория систем управления. Под ред. *К. Т. Леондеса*, «Наука», М., 1971.
4. *Мальчиков С. В.* О синтезе линейных систем автоматического управления с переменными параметрами. «Автоматика и телемеханика», *20* (1959) 12.
5. *Деч Г.* Руководство к практическому применению переобразования Лапласа и z-преобразования. «Наука», М., 1971.
6. *Михайлов Ф. А., Тыхевич О. Ф., Хаджинов М. К.* Вычисление характеристик структурных соединений линейных нестационарных систем. Сб. «Задачи динамики управления летательных аппаратов», Труды МАИ, Вып. 240, М., 1972.
7. *Михайлов Ф. А., Теряев Е. Д., Булеков В. П., Данков Г. Ю., Саликов Л. М., Степаньянц Г. А.* Динамика нестационарных линейных систем. «Наука», М., 1967.
8. *Ленинг Д. Х., Бэттин Р. Г.* Случайные процессы в задачах автоматического управления. ИИЛ, М., 1958.
9. *Бицадзе А. В.* Основы теории аналитических функций комплексного переменного. «Наука», М., 1969.
10. *Zadeh, L. A.*, On stability of linear varying parameter system. J. Appl. Phys. **22** (April 1951).
11. *Лаврентьев М. А., Шабат Б. В.* Методы теории функций комплексного переменного. «Наука», М., 1965.

Application of structural transforms to the analysis of time-varying linear systems

F.A. Mikhailov, O.F. Tykhevich, M.K. Khadzhinov

(Moscow)

This paper deals with the time-varying linear systems. Its object is to obtain the overall Zadeh's system function in any type of interconnections provided the individual system functions are known.

Explicit equations for solving this problem are deduced. These are Eq. (1.1) for the case of parallel connection, Eq. (1.11) for the case of series connection and Eq. (1.16) for the simplest case of feedback connection. Equations (1.11) and (1.16) are complicated. It prompts to state the problem in approximate-calculus approach.

As a base of approximate approach, the authors suggest to approximate actual system functions by some functions which are rational algebraic function of complex variable s with time-varying coefficients. In order to construct these functions the series expansion (2.1) of impulsive response are used. As shown in [1], the coefficients of this expansion can be expressed in terms of coefficients of differential equation (2.2) describing the motion.

Using (2.1) and applying Laplace transformation yields the asymptotic expansion (3.1) for Zadeh's system function. On the other hand, if s is sufficiently large, one can obtain the series expansion of approximate system function (4.1). Comparing the series results in Eqs (4.3) which connect the coefficients of approximate system function with the coefficients of series expansion of impulsive response.

In order to obtain the system function in the case of series connection, we use the series expansion of impulsive response for gains, then determine the series expansion for overall approximate system function and, at last, construct the overall approximate system function. Similarly, we obtain the overall approximate system function in the case of feedback connection. The result for the case of parallel connection is trivial, since Eq. (1.1) holds. Some results for the cases of series connection and feedback connection are listed in Tables 2—4.

As a simple application of the above results a feedback system representing a mathematical model of some problem of controlled motion is considered. Under the condition that the system is excited by the input in the form of white noise process, the approximate overall system function and the dispersion of output are determined.

One can see that the described method for determining the overall approximate system function permits to simplify solving some problems of time-varying system analysis.

Ф. А. Михайлов, О. Ф. Тыхевич, М. К. Хаджинов
Московский авиационный институт им. С. Орджоникидзе
СССР, Москва А-80, ГСП,
Волоколамское ш., 4

BOOR REVIEW

RICHARD C. DORF: **Modern Control Systems.** Addison Wesley Publishing Co. Reading (Massachusetts) — Menlo Park (California)—London—Don Mills (Ontario). Second edition, 1974, 413 pages, 392 figures, 24 tables

In this book the author, who showed oneself previously as one of the pathfinders in state-space analysis and synthesis, has two purposes: (*a*) to recover balance between the time-domain and frequency-domain methods; (*b*) to give possibilities for the reader to rediscover and recreate himself the rules of control theory and practice. To fulfil both aim the topics are carefully and circumspectly selected, besides modern concepts as sensitivity analysis, performance indices, state variables and optimum control systems, a great effort is also displayed to retain the classical topics of control theory which have been proven so useful in practice.

The reviewer may completely agree with the author's assertion that we have truly learned and understood we discovered ourselves, indeed. The traditional method, confronting the student not with the problem but with the finished solution and results, means to deprive him of all excitement, to shut off all of the creative impulses, to reduce the adventure of mankind concerning and making discoveries to a dusty heap of theorems. To avoid these mistakes the book presents to the reader a lot of problems and questions and point to some of the answers that have been obtained. Furthermore, it presents also some unanswered and important problems which we contime to confront.

The text is written in a clear style. Chapter 1 serves as an introduction outlining also the basic history of control theory. Chapter 2 is concerned with the development of mathematical models of control systems. Chapter 3 is devoted to the description of feedback control system characteristics. The next two chapters examine the performance and the stability of linear control systems. Chapter 6 treats the root locus method. Chapter 7 presents the frequency response method. The next chapter shows how can the stability be investigated by using the Nyquist criterion. Chapter 9 develops the time-domain method in terms of the state variables of the control system. Finally, Chapter 10 is a good summary of the design and compensation of feedback control systems. The Appendices treat Laplace transform pairs; symbols and units; conversion factors; matrix algebra; a computional algorithm for the evaluation of the state-transition matrix, a program transition. Index and contents serve for finding the various concepts within the text.

This book can be regarded as an introductory course in linear control systems. The perspicuous figures and photos, the clear text as well as the great amount of well chosen examples are the causes why the reader will find pleasure in this book.

The problems and examples are chosen from various fields such as electrical, mechanical, chemical and industrial engineering, biology, sociology, ecology and economics intending to illustrate the unity of control system concepts, and the applicability of control theory in many fields of science and technology.

F. Cs.

W. FINDEISEN: **Multilevel Control Systems.** Panstwowe Wydawnictwo Naukowe, Warsaw. 1974, 296 pages

The book deals mainly with the control systems of technological processes. We refer here for instance to an energetical block (boiler-turbine-generator), to a technological system producing aluminous earth, etc. As the problems of control depend not only on technical viewpoints, but also on economical ones, designing a control system besides the physical and chemical relationships the economical connections have also to be taken into account.

In the examples of the book the control of technological objects is described. The investigation does not include the management and economical sphere but utilizes the prescriptions and trends determined here for optimum control.

The investigated technological processes are generally complex systems containing a lot of different elements. The structure of the control system of these processes is suitable for a multilevel or hierarchical construction. The title of the book derives from here.

In the multilevel control systems the control tasks are divided among certain levels, which differ from each other in their character, in the complexity of the algorithm, in the frequency of their action, etc. Thus the multilevel control can be determined by the structure of the control system. The book discusses methods by which the construction of the control will require less work and money efforts. Two basic ideas are taken here into consideration: the first is the division of the control tasks among the levels of the object, and the second is the parcelling of the optimization task.

The first chapter describes the mathematical models of the controlled systems and the methods of model-building.

The second chapter deals with the optimization of process technology without mentioning the way of realization. In this section continuous and periodical processes with cyclically optimum state and the dynamic optimization are treated.

The third chapter gives an introduction to the basic principles of the operation of multilevel control systems. This section discusses the basic features of such systems mostly via examples, since the technique of multilevel control is more involved task than the problems which can be handled by formal methods.

The main point of the multilevel systems is their structure and it is difficult to treat the problem of the choice of the structure only by analytical means.

The fourth chapter presents the problems in connection with the division of the control task into differeet levels.

The fifth chapter deals with the methods of optimization based on decomposition and coordination task. This part is an introduction to Chapter 6, but it can be easily understood also by those who are not perfect in the basis of nonlinear programming.

The sixth chapter — according to the purpose of the book — shows the multilevel control systems regarding the whole complex system.

The book is written not for undergraduates, but for engineers dealing with control systems design, though the major part of the book may be understandable also for those who are not perfect in the theory of optimum control.

L. K.

Printed in Hungary

AKADÉMIAI NYOMDA, BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

PROBLEMS OF CONTROL AND INFORMATION THEORY VOL, 4, NO. 1
ENGLISH TRANSLATION OF THE COMMUNICATIONS IN RUSSIAN

PROGRAMMABLE STRUCTURE AUTOMATA

A. V. KALYAEV

(Taganrog)

In the paper programmable structure automata, their properties and their distinction from programmable process automata are discussed. Theorems on behaviour of automata with various graphs and of programmable structure and programmable process automata are proved. An important class of programmable structure automata, that of programmable connection automata is considered. A notion of connection function is introduced, and a system of relations describing programmable structure automation is examined. Dependences of the transition and output functions, as well as of the state set and the output alphabet of the programmable connection automation on the connection function are elucidated. Automata are classified according to the principle of process programming, structure and connections.

Introduction

The uniform arrays which are being actively developed in the last few years and successfully implemented due to the advances of the microelectronic technology form a new class of automata whose graphs are not predetermined but are defined and changed by the control signals.

Behaviour and properties of such automata differ significantly from those of conventional rigid-graph automata.

The paper attempts to analyze peculiarities of the automata whose structure and graphs may be changed by means of a program, to discuss their properties and differences from automata having a unique graph and rigid structure, and to contemplate principles of control of the structure and graphs of such automata.

1. Programming of the process, structure and connections in automata

The majority of practical automata has rigid invariable structure. Their behaviour may be programmed only by presetting transfers from one state into another with the choice of an initial state and a particular input word. Such automata will be conditionally placed in the class of programmable process automata.

The programmable process automation (see Fig. 1) always has rigid invariable graph g, and its behaviour is determined by the operation program $\pi\{a(0)\}$ giving the initial vertex of the graph $a(0)$, and by the current information $y(t)$, $t = 1, 2, \ldots$. Hereafter the term "automaton graph" will refer to the graph formed by the state set, and transition and output functions of the automaton. Transition and output functions, input and output alphabets, state set and structure of such an automaton are established in the course of synthesis once and forever.

The choice of the initial state of the automaton $a(0) \in A$ from the state set $A = \{a_1, a_2, \ldots, a_\gamma\}$ is done in the programmable process automaton by control operator ϑ corresponding to program $\pi\{a(0)\} = \{\vartheta(A)\}$, with $\vartheta(A) \in A$. In the simplest case, control operator may be given as a set of signals or control commands $\theta = \{\vartheta_1, \vartheta_2, \ldots, \vartheta_\xi\}$, $\vartheta \in \theta$ each of them defining a specific choice of the initial vertex in the automaton.

Along with the programmable process automata we are going to discern those with programmable structure. An automaton whose behaviour is defined by its graph g is understood as programmable structure automaton. It is known that to each particular graph g_i a certain subset of various automaton structures S_i corresponds. Set of subsets of structures $S = \{S_1, S_2, \ldots, S_\gamma\}$ will correspond to the set of graphs $G = \{g_1, g_2, \ldots, g_\gamma\}$. To define automaton graph $g_i \in G$ it is necessary and sufficient to choose automaton structure S_i so that $S_{ij} \in S_i$ where $S_i \in S$.

The programmable structure automaton is, thus, an automaton whose graph of transients and outputs $g_i \in G$ may be varied by taking from the structure set S a corresponding subset S_i with subsequent arbitrary choice of a particular structure S_{ij} from this subset.

In the simplest case, one particular structure $s_i \in S_i$ may be pre-assigned to each graph $g_i \in G$, with the rest of the structures from S_i excluded from consideration. Then there will be one-to-one correspondence between elements of graph set $G = \{g_1, g_2, \ldots, g_\gamma\}$ and those of structure set $S^* = (S_1, S_2, \ldots, S_\gamma\}$. In this case programming of the structure of an automaton with given graph $g_i \in G$ reduces to the choice of a structure $s_i \in S^*$ corresponding to graph g_i.

In the programmable structure automaton (see Fig. 2) an external program $\pi\{g\}$ defines the set of states, transient and output functions, and automaton input and output alphabets as well.

The choice of graph $g \in G$ from the graph set $G = \{g_1, g_2, \ldots, g_\gamma\}$ in the programmable structure automaton is done by control operator φ defined by program $\pi\{g\} = \{\varphi(G)\}$ where $\varphi(G) \in G$. Control operator for the choice of the graph may be given by a set of signals or control commands $\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_\tau\}$, $\varphi \in \Phi$.

In the case of structural programming, the initial automaton is essentially synthesized incompletely as it were, the synthesis being completed under control of input signals depending on the problem being solved and on the environment.

Connection programming is a version of the structural programming, it consists in connecting sufficiently large functional units having predetermined graphs. The programmable connection system of automata (see Fig. 3) is a system whose behaviour is given by connection program $\pi\{k\}$ where k is a connection function, and by the current information, with graph of each subautomaton in the system being invariable under any connection.

Connection function $y(t) = k[z(t-1), x(t)]$ defines relations between switch output $y(t)$, switch input $x(t)$, and the vector of subautomata outputs $z(t-1)$, $t = 1, 2, 3, \ldots$ fed into the switch input with one clock-cycle delay (see Fig. 3).

Connection function k is chosen from connection function set $K = \{k_1, k_2, \ldots, k_\varrho\}$ by control operator ψ corresponding to program $\pi\{k\} = \{\psi(K)\}$ where $\psi(K) \in K$. Control operator intended to choose connection function may be given as a set of signals or control commands $\Psi = \{\psi, \psi_2, \ldots, \ldots, \psi_\xi\}$, $\psi \in \Psi$.

In automata it is possible, and sometimes necessary to use the processor, structural and connection programming jointly in the sense that the processor and structural programming is done separately for each sufficiently large functional unit (subautomaton) of the automaton, and the connection programming defines connections between the subautomata depending on the particular task.

Thus, in the general case, automaton behaviour program may be given jointly by the initial state, graph and connection function $\pi\{a(0), g, k\}$ (see Fig. 4). Other combinations of programmings are possible: $\pi\{a(0), g\}$, $\pi\{a(0), k\}$, $\pi\{g, k\}$ (see Table 1).

An automaton having all three kinds of programming, that is an automaton with programmable structure, connections and process (see Fig. 4) has the highest flexibility and the greatest possibilities. Such an automaton may be implemented with standard subautomata of one type having structure discretely rearrangeable in accordance with several graphs. Each standard subautomaton is connected to one or more subautomata of the system through the switch, with the connections being defined by the connection function.

Programming of such an automaton involves the choice of the subautomaton graphs $g = g[A, Y, z, \delta(a, y), \lambda(a, y)]$ from the graph set G by the control operator φ, of the connection function $y = k(z, x)$ from the connection function set K by the control operator ψ, and of the initial state $a(0)$ from the

automaton state set A by control operator ϑ. Behaviour of the automaton under consideration is, thus, defined by the program $\pi\{a(0), g, k\} = \{\vartheta(A, \varphi(G), \psi(K)\}$ where $\vartheta(A) \in A$, $\varphi(G) \in G$, $\psi(K) \in K$.

It follows that to define completely an automaton with programmable process, structure and connections, it is necessary and sufficient to give a set of the following data:

state set	$A = \{a_1, a_2, \ldots, a_k\}$
input alphabet	$X = \{x_1, x_2, \ldots, x_n\}$
intermediate alphabet	$Y = \{y_1, y_2, \ldots, y_m\}$
output alphabet	$Z = \{z_1, z_2, \ldots, z_l\}$
transient function set	$\Delta = \{\delta_1, \delta_2, \ldots, \delta_\mu\}$
output function set	$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_\nu\}$
connection function set	$K = \{k_1, k_2, \ldots, k_\varrho\}$
structure control operators	$\Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_\tau\}$
connection control operators	$\Psi = \{\psi_1, \psi_2, \ldots, \psi_\zeta\}$
process control operators	$\Theta = \{\vartheta_1, \vartheta_2, \ldots, \vartheta_\xi\}$
automaton behaviour program	$\pi\{a_i(0), g_i, k_i\} = \{\vartheta_i(A), \varphi_i(G), \psi_i(K)\}$

where

$$\vartheta_i \in \Theta,\ \varphi_i \in \Phi,\ \psi_i \in \Psi$$

$$\vartheta_i(A) \in A,\ \varphi_i(G) \in G,\ \psi_i(k) \in k.$$

In other words, to define an automaton with programmable process, structure and connections one has to define the set of graphs $G = \{g_1, g_2, \ldots, g_j\}$ the set of connection functions K, input alphabet X, sets of control operators Φ, Ψ, and θ, and program $\pi\{a(0), g, k\}$.

A particular program

$$\pi\{a_i(0), g_j, k_i\} = \{\vartheta_i(A), \varphi_i(G), \psi_i(K)\}$$

defines in the graph set a particular graph

$$g_i = g_i[A_i, Y_i, Z_i, \delta_i(a, y), \lambda_i(a, y)],\ g_i \in G$$

for which

$$A_i,\ A,\ Y_i\ Y,\ Z_i,\ Z,\ \delta_i \in \Delta,\ \lambda_i \in \Lambda$$

and in the connection function set K a particular connection function $y = k_i(z, x)$, $k_i \in K$, and gives the initial state $a_j(0) \in A_i$. As the result, a particular automaton is completely defined which at arrival of input words determines corresponding output words.

2. A difference in behaviour between programmable process and programmable structure automata

Response of the automaton to any input word $p = y(1), y(2), \ldots, y(t)$ under a given initial state $a(0)$ will be referred to as "automaton behaviour". It is a sequence of pairs of states and output letters

$$r = \{a(1), z(1)\}\ \{a(2), z(2)\} \ldots \{a(t), z(t)\}, \quad \{a(t), z(t)\} \in A \times Z,\ r \in R$$

corresponding to input letters $y(t)$ ($t = 1, 2, 3, \ldots$). $R = \{r_1, r_2, r_3, \ldots\}$ stands here for the response set.

In the general case, automaton behaviour unambiguously depends on the program and the input word

$$r_i = r[\pi\{a(0), g, k\}, p_i]$$

i.e., in the final analysis, on the graph, connections, initial state, and input word.

Consider conditions under which behaviour of two automata may be identical, and those under which their behaviours will differ. Introduce to this end additional notions of coinciding and distinguishable graphs and subgraphs.

Any graph g consists of a set of complete arcs $d = \{a_н, y, a_н, z\}$, each being characterized by the initial state $a_н$, input letter y, final state $a_k = \delta(a_н, y)$, and output letter $z = \lambda(a_н, y)$, where δ and λ are transient and output functions, respectively. The set of all complete arcs $D = \{d_1, d_2, \ldots, d_\zeta\}$ completely defines graph g.

Let there be graphs g_i and g_j characterized by sets of complete arcs $D_i = \{d_{i1}, d_{i2}, \ldots, d_{i\zeta}\}$ and $D_j = \{d_{j1}, d_{j2}, \ldots, d_{j\xi}\}$, respectively. The two graphs g_i and g_j coincide completely if for any arc of each graph, $d_i \in D_i$, $d_i = \{a_{iн}, y_i, a_{ik}, z_i\}$ or $d_j \in D_j$, $d_j = a_{jн}, y_j, a_{jk}, z_j\}$, one can select one and only one arc in the other graph such that for each of the resulting pairs of arcs (d_i, d_j) the following conditions hold:

$$\left.\begin{aligned} a_{iн} &= a_{jн} \\ y_i &= y_j \\ a_{ik} &= a_{jk} \\ z_i &= z_j \end{aligned}\right\}. \tag{1}$$

Equal graphs $g_i = g_j$ and isomorphic ones pertain to completely coinciding graphs.

If for each pair of arcs (d_i, d_j) conditions (1) hold only partially, in particular if of four equalities (1) only three, two or one hold, the corresponding graphs g_i and g_j will be referred to as "partially coinciding".

In the case where for any pair of arcs (d_i, d_j) none of the equalities (1) holds, graphs g_i and g_j are completely non-coinciding.

Non-coinciding and partially coinciding graphs will be referred to as "distinguishable" graphs. Any pair of graphs g_i and g_j may be in general decomposed into completely coinciding subgraphs $g_{i\pi} = g_{j\pi} = g_\pi = g_i \cap g_j$ for which conditions (1) hold, and into distinguishable subgraphs $g_{ip} = g_i \setminus g_j$ and $g_{jp} = g_j \setminus g_i$ for which conditions (1) do not hold either completely or partially.

Formulate now conditions for coinciding behaviour of automata. Let two automata be given defined by graphs g_i and g_j. In order that for identical input words $p_i = p_j$ behaviour r_i of automaton g_i coincides with behaviour r_j of automaton g_j, it is necessary and sufficient that pairs of states and output letters of automata, $\{a_i(t), z_i(t)\} = \{a_j(t), z_j(t)\}$, coincide at each step of their operation. It follows that for behaviours of automata g_i and g_j to coincide, it is necessary and sufficient that at each step of automata operation under coinciding input letters $y_i(t) = y_j(t)$, $t = 1, 2, 3 \ldots$ the following conditions hold

$$a_{i\text{н}}(t) = a_{j\text{н}}(t)$$
$$a_{ik}(t) = a_{jk}(t)$$
$$z_i\ (t) = z_j(t)\,.$$

The above-said enables us to formulate a theorem about behaviour of two different automata.

Theorem 1. On behaviour of automata with different graphs.

Behaviours of two completely defined automata with different graphs g_i and g_j coincide under identical input words $p_i = p_j$ and equal initial states $a_i(0) = a_j(0)$ only in the domain of completely coinciding subgraphs $g_{i\pi} = g_{j\pi} = g_\pi = g_i \cap g_j$ and do not coincide in the area of distinguishable subgraphs.

Proof. Consider coinciding subgraphs $g_{i\pi} = g_{j\pi}$ to which coinciding subsets of complete arcs $D_{i\pi} = D_{j\pi} = D_i \cap D_j$ correspond. Let initial states $a_i(0) = a_j(0)$ and input letters $y_i(1) = y_j(1)$ belonging to subgraphs $g_{j\pi}$ and $g_{j\pi}$ be given. In this case there always will be coinciding arcs $d_i(1) = \{a_i(0), y_i(1), a_i(1), z_i(1)\}$ $d_i(1) \in D_{i\pi}$ and $d_j(1) = \{a_j(0), y_j(1), a_j(1), z_j(1)\}$, $d_j(1) \in D_{j\pi}$ for which the following conditions hold:

$$y_i(1) = y_j(1)$$
$$a_i(0) = a_j(0)$$

$$a_i(1) = a_j(1)$$
$$z_i(1) = z_j(1).$$

Similarly, at the second step of automaton operation for input letters $y_i(2) = y_j(2)$ belonging to subgraphs g_{i_n} and g_{j_n} and for initial states $a_i(1) = a_j(1)$, there necessarily will be a pair of coinciding arcs $d_i(2) = \{a_i(1), y_i(2), a_i(2), z_i(2)\}$, $d_i(2) \in D_{i_n}$ and $d_j(2) = \{a_j(1), y_j(2), a_j(2), z_j(2)\}$, $d_j(2) \in D_{j_n}$ for which

$$y_i(2) = y_j(2)$$
$$a_\iota(1) = a_j(1)$$
$$a_i(2) = a_j(2)$$
$$z_i(2) = z_j(2).$$

By continuing the same reasoning for successive steps of operation of automata g_i and g_j we come to the conclusion that in the domain of coinciding subgraphs g_{i_n} and g_{j_n} conditions of behaviour coincidence

$$a_i(t-1) = a_j(t-1)$$
$$a_i(t) = a_j(t)$$
$$z_i(t) = z_j(t),\ t = 1, 2, 3 \ldots$$

will always hold for any allowable coinciding words $p_i = y_i(1), y_i(2) \ldots y_i(t)$ and $p_j = y_j(1)\, y_j(2) \ldots y_j(t)$, $p_i = p_j$, for which $y_i(t) = y_j(t)$, $t = 1, 2, 3, \ldots$.

Hence, for similar input words and initial states, behaviours of automata coincide in the domain of coinciding graphs.

Consider further distinguishable subgraphs g_{ip} and g_{jp} to which subsets of distinguishable complete arcs D_{ip} and D_{jp} correspond. The subset of distinguishable complete arcs D_{ip} consists of arcs each of which either does not coincide completely with any arc from the subset D_j of graph g_j or coincides only partially. The same holds for subset D_{jp}.

Decompose arcs of subset D_{ip} into two subsets. One of them, D^*_{ip}, will comprise all arcs which do not coincide with arcs of D_j with respect to the input alphabet letter y, and with respect to other parameters may either coincide or not. Since input letters of subset D^*_{ip} do not coincide, one cannot form input words coinciding with input word in domain D_j. Therefore behaviour of automaton g_i coinciding with that of automaton g_j under similar input words cannot arise in domain D_j.

The second subset $D^{**}_{ip} = D_{ip} D^*_{ip}$ comprises all arcs of subset D_{ip} which coincide with one or more arcs of D_j with respect to the input alphabet letter but do not with respect to coincide at least one of other parameters. In domain

D_{ip}^{**} input words p_i of automaton g_i coinciding with input words p_j of automaton g_j may be formed. However, for any arc $d_i^{**} \in D_{ip}^{**}$, $d_i^{**} = \{a_{iH}, y_i, a_{ik}, z_i\}$ any arc $d_j^{**} \in D_j$, $d_j = \{a_{jH}, y_j, a_{jk}, z_k\}$ coinciding with arc d_i^{**} with respect to the input letter $y_i = y_j$ will differ from it at least with respect to one of other parameters. This means that in domain D_{ip}^{**} the necessary and sufficient conditions of coincidence of behaviours of two automata are never satisfied.

Thus, in domain $D_{ip}^{*} \cup D_{ip}^{**} = D_{ip}$ under identical input words $p_i = p_j$ identical behaviours of automata g_i and g_j can never take place. Hence, under identical input words in the domain of distinguishable subgraphs g_{ip} and g_{pj} behaviours of automata g_i and g_j never coincide. Q.E.D.

Analyze now behaviour of automata with programmable process and those with programmable structure. Programmable process automata have a fundamental distinction from those with programmable structure that is explained below by two theorems on behaviour of these two classes of automata.

Theorem 2. On behaviour of programmable process automata.

If a programmable process automaton is given having rigid graph $g = g[A, Y, Z, \delta(a, y), \lambda(a, y)] = \text{const}$ for which under any initial state $a(0) \in A$ all states $a \in A$ are accessible, and if two distinguishable programs $\pi_i\{a(0)\} = \{\vartheta_i(A)\}$ and $\pi_j\{a(0)\} = \{\vartheta_j(A)\}$ are given, there exist for program $\pi_i\{a(0)\}$ input word p_j, and for program $\pi_j\{a(0)\}$ input word p_j such that after their completion both programs will generate under any subsequent identical words p^* identical behaviour of automaton r^*:

$$r_i r^* = r[\{\vartheta_i(A)\},\ p_i p^*]$$
$$r_j r^* = r[\{\vartheta_j(A)\},\ p_j p^*]$$

where

$$r_i = r[\{\vartheta_i(A)\},\ p_i]$$
$$r_j = r[\{\vartheta_j(A)\},\ p_j].$$

Proof. Consider a state $a^* \in A$ of the automaton. Since all automaton states are accessible, state a^* is also accessible for any $a(0) \in A$. Therefore for program $\pi_i = \{\vartheta_i(A)\}$ there will always be a finite input word p_i which will transfer the automaton from state $a_i(0) = \vartheta_i(A)$ into state $a^* = a[\{\vartheta_i(A)\}, p_i]$. Similarly, for program $\pi_j = \{\vartheta_j(A)\}$ there always will be a finite input word p_j which will transfer the automaton from initial state $a_j(0) = \vartheta_j(A)$ into this same state $a^* = a[\{\vartheta_i(A), p_j]$. The corresponding behaviour of the automaton will be defined by relations:

$$r_i = r[\{\vartheta_i(A)\}, p_i]$$
$$r_j = r[\{\vartheta_j(A)\}, p_j], r_i \neq r_j.$$

The accessed state a^* may be regarded as a new initial state $a(0) = a^*$ corresponding to a program $\pi^* = \{\vartheta^*(A)\}$ for which under arbitrary allowable word p^* automaton behaviour will be defined by $r^* = r[\{\vartheta^*(A), p^*]$.

From consecutive behaviours r_i and r^* make up behaviour $r_i r^*$. Since program $\{\vartheta^*(A)\}$ continues program $\{\vartheta_i(A)\}$ after completion of word p_i, in the case of joint behaviour $r_i r^*$ program $\{\vartheta_i(A)\}$ absorbs program $\{\vartheta^*(A)\}$:

$$r_i r^* = r[\{\vartheta_i(A)\}, p_i] r[\{\vartheta^*(A)\}, p^*] = r[\{\vartheta_i(A), p_i p^*].$$

Similarly make up behaviour $r_j r^*$ from consecutive behaviours r_j and r^*:

$$r_j r^* = r[\{\vartheta_j(A)\, p_j] r[\{\vartheta^*(A), p^*] = r[\{\vartheta_j(A), p_j p^*].$$

It follows that programs $\pi_i = \{\vartheta_i(A)\}$ and $\pi_j = \{\vartheta_j(A)\}$ after completion of, respectively, words p_i and p_j will generate under any similar input words p^* identical behaviour of automaton r^*. Q.E.D.

Theorem 3. On behaviour of programmable structure automaton.

If a programmable structure automaton is given, having rearrangeable connected graph $g_0 = g_0[A, X, Z, \delta_0(a, y), \lambda_0(a, y)] = \text{var}$, where $g_0 = g_1 \cup g_2 \cup g_3 \ldots \cup g_j$, $G = \{g_1, g_2, \ldots, g_j\}$, for which graphs $g_i \in G$ and $g_j \in G$ $(j, i = 1, 2, \ldots \gamma)$, $i \neq j$, do not pairwise coincide $g_i \setminus g_j = \varnothing$, $g_j \setminus g_i = \varnothing$, distinguishable subgraphs $g_{ip} = (g_i \setminus g_j) \subset g_i$ and $g_{jp} = (g_j \setminus g_i) \subset g_j$ are not isomorphic both among themselves and with respect to coinciding subgraphs $g_{i_\Pi} = g_{j_\Pi} = g_i \cap g_j$, and any state $a \in A_i \subset A$ is accessible under any initial state $a(0) \in A_i$ $(i = 1, 2, \ldots \gamma)$ and if two distinguishable automaton behaviour programs $\pi_i(g) = \{\varphi_i(G) \in G\}$ and $\pi_j\{g\} = \{\varphi j(G) \in G\}$ $i \neq j$ $(i, j = 1, 2, \ldots \gamma)$ are given, there is at least one input word p^* for each program such that for two programs π_i and π_j one cannot sort out pairs of words p_i and p_j after whose completion different programs π_i and π_j will generate under identical input words p^* identical automaton behaviour r^*:

$$r_i r^* = r[\{\varphi_i(G)\}, a_i(0), p_i p^*],$$

$$r_j r^* = r[\{\varphi_j(G)\}, a_j(0), p_j p^*],$$

where

$$r_i = r[\{\varphi_i(G)\}, a_i(0), p_i],$$

$$r_j = r[\{\varphi_j(G)\}, a_j(0), p_j].$$

Proof. Let programs $\pi_i = \{\varphi_i(G)\}$ and $\pi_j = \{\varphi_j(G)\}$ be given setting up programmable structure automaton to implement graphs g_i and g_j, re-

spectively. Each of the graphs may be decomposed into a distinguishable and a completely coinciding subgraph $g_i = g_{ip} \cup g_{i_\Pi}$, $g_j = g_{jp} \cup g_{j_\Pi}$ where $g_{i_\Pi} = g_{j_\Pi} = g_i \cap g_j$, $g_{ip} = g_i \setminus g_j$, $g_{jp} = g_j \setminus g_i$.

Consider behaviour of the automaton in the case of program $\pi_i = \{\varphi_i(G)\}$. No matter to which subgraph g_{ip} or g_{i_Π} the initial state $a_i(0) \in A_i$ belongs, on account of accessibility of any state $a \in A$ there is always a word p_i such that it will transfer the automaton in accordance to program π_i into state a_{pi} belonging to distinguishable subgraph g_{ip}. For the same reason there is a word p^* which after completion of word p_i will hold the automaton during a finite interval of discrete time in the domain of distinguishable subgraph g_{ip}. Therewith automaton behaviour $r_i r^* = r[\{\varphi_i(G)\}, a_i(0), p_i p^*]$ will be generated.

Referring to Theorem 1, for the same input word p^* behaviour r^* reproducible in the domain of the distinguishable subgraph g_{ip} under no condition can coincide with the behaviour in the domain of graph g_{jp}. Therefore, independently of word p_j preceeding word p^* in the domain of graph g_j, there always will be $r_j^* \neq r^*$ where $r_i r_j^* = r[\{\varphi_j(G)\}, a_j(0), p_j p^*]$. Hence, in the case of program $\pi_j = \{\varphi_j(G)\}$ behaviour r^* cannot be reproduced for any of input words $p_j \in P_j$. Q.E.D

Very important corollaries follow from these theorems.

Corollary 1. For the connected programmable process automaton finite input words exist such that after their completion two different programs are indistinguishable, i.e. behaviours of the automaton coincide for any identical succeeding input words. This means that if a connected programmable process automaton is taught to perform certain functions, recognition or decision taking ones in particular, the habits acquired during learning may be destroyed later by input information of certain types entering the automaton through its information input.

Corollary 2. For the connected programmable structure automaton there do not exist finite input words such that they lead to indistinguishability of its two distinguishable programs, i.e. to coincidence of behaviours of the automaton for any identical succeeding input words. This means that if a connected programmable structure automaton is taught to perform certain functions, recognition or decision making for instance, the habits acquired during learning cannot be henceforth destroyed by any information entering the automaton through its information input.

3. Programmable connection automata

Programmable connection automaton (see Fig. 5) is, in general, made of a set of separate automata whose inputs and outputs are connected in accordance with a given program $\pi(k\} = \{\psi(K)\}$ where $k = \psi(K)$ is connection function, $K = \{k_0, k_1, k_2 \ldots k_\varrho\}$ is connection function set, and $\psi \in \Psi$ is a connection control operator belonging to the operator set $\Psi = \{\psi_1, \psi_2, \ldots, \psi_\varrho\}$. Connection function is implemented by the switch (see Fig. 5).

The uniform set of standard programmable connection automata is of the greatest interest. Therefore we shall henceforth consider all individual standard automata from the set (Fig. 5) as belonging to one type and having identical graphs, i.e. identical sets of states, input and output alphabets, and identical transient and output functions as well:

$$
\begin{aligned}
g^* &= g^*[A^*, Y^*, Z^*, \delta^*(a^p, y^p), \lambda^*(a^p, y^p)], \\
A^* &= \{a_1^*, a_2^*, \ldots, a_b^*\}, a^p(t) \in A^*, \\
Y^* &= \{e, y_1^*, y_2^*, \ldots, y_d^*\}, y^p(t) \in Y^* \\
Z^* &= \{e, z_1^*, z_2^*, \ldots, z_l^*\}, z^\nu(t) \in Z^* \\
a^p(t) &= \delta^*[a^p(t-1), y^p(t)] \\
Z^p(t) &= \lambda^*[a^p(t-1), y^p(t)].
\end{aligned}
$$

Here p is the number of a standard individual automaton from the set of programmable connection automata, e is the empty symbol standing for the absence of the signal.

Consider a set of concurrently operating standard automata (Fig. 5). The set may be interpreted as a unique integrated automaton (see Fig. 6) representing Cartesian product of all standard automata. For this integrated automaton we have the following relations:

$$
\begin{aligned}
g_0 &= g_0[A, Y, Z, \delta_0(a, y), \lambda_0(a, y)], \\
A &= \{a_1, a_2, \ldots a_K\}, a(t) \in A, \\
Y &= \{e, y_1, y_2, \ldots y_m\}, y(t) \in Y, \\
Z &= \{e, z_1, z_2, \ldots z_l\}, z(t) \in Z \\
a(t) &= \delta_0[a(t-1), y(t)] \\
z(t) &= \lambda_0[a(t-1), y(t)],
\end{aligned}
$$

where the graph, state set and alphabets of the integrated automaton are obtained as Cartesian products of graphs, state sets and alphabets of standard automata:

$$g_0 = g^* \times g^* \times \ldots g^* = (g^*)^N$$
$$A = A^* \times A^* \times \ldots A^* = (A^*)^N$$
$$Y = Y^* \times Y^* \times \ldots Y^* = (Y^*)^N$$
$$Z = Z^* \times Z^* \times \ldots Z^* = (Z^*)^N.$$

Consider now the switch of the set of programmable connection automata (Fig. 5). Its inputs are fed with output letters $z(t-1)$ from the outputs of the integrated automaton delayed by one clock cycle, and by signals from the environment $x(t) \in X$. The outer alphabet $X = \{e, x_1, x_2, \ldots, x_n\}$ is formed evidently as a Cartesian product $X = (X^*)^M$ of the partial outer input alphabets of the switch $X^* = \{e, x_1^*, x_2^*, \ldots, x_g^*\}$. At each pth switch output connected to the input of pth standard automaton a letter of the standard automaton input alphabet $y^p(t) \in Y^*$ appears accordingly to the partial connection function K^p, which depends on the input letters of the switch $y^p(t) = k^p[z(t-1), x(t)]$.

For any pth switch output, i.e. for any standard automaton, there is a finite set of partial connection functions $K^* = \{k_0^*, k_1^*, k_2^*, \ldots, k_r^*\}$, $k^p \in K^*$. The totality of partial connection functions $k^p (p = 1, 2, \ldots, N)$ of all standard automata forms the general connection function vector $k = \{k^1, k^2, \ldots, k^N\}$, $k \in K$. The set of general or complete connection functions $K = \{k_0, k_1, k_2, \ldots, k_\varrho\}$ is a Cartesian product $K = (K^*)^N$ of partial connection function sets. Complete connection functions define relations between switch input and output $y(t) = K[z(t-1), x(t)]$.

To sum up the above-said, one can write the following relations for the switch (see Fig. 7) involved into a system of programmable structure automata (see Fig. 5):

$$y^p(t) = k^p[z(t-1), x(t)],$$
$$K^* = \{k_0^*, k_1^*, k_2^* \ldots k_r^*\}, \; k^p \in K^*,$$
$$y(t) = k[z(t-1), x(t)],$$
$$K = \{k_0, k_1, k_2 \ldots k_\varrho\}, \; k \in K,$$
$$y^p(t) \in Y^*, \; y(t) \in Y, \; z(t) \in Z, \; x(t) \in X$$
$$X^* = \{e, x_1^*, x_2^* \ldots x_g^*\}$$
$$X = \{e, x_1, x_2 \ldots x_n\}$$
$$X = X^* \times X^* \times \ldots X^* = (X^*)^M.$$

Overall, the integrated automaton of Fig. 6 and the switch (see Fig. 7) from a more complex automaton (see Fig. 8) whose characteristics and properties depend not only on the characteristics of the standard automata incorporated into the system, but largely on the particular connection functions to which the switch is specified. The connection function defines ultimately transition and output functions of the programmable connection automaton as a whole, its state set, and input and output alphabets as well.

One of the basic problems arising in the study of the programmable connection automaton is determination of alphabets, state sets and transition and output functions vs. connection function. Let us discuss it in greater detail. Behaviour of a system of automata as a whole (see Fig. 8) is completely described by the transition and output functions of the integrated automaton and by the switch connection function:

$$a(t) = \delta_0[a(t-1), y(t)],$$
$$z(t) = \lambda_0[a(t-1), y(t)],$$
$$y(t) = k[z(t-1), x(t)].$$

By means of the expressions above one can find transition and output functions of the automata system as a whole if one regards the programmable connection automaton (see Fig. 8) consisting of the integrated automaton and the switch as a single automaton which is fed with input letters $x(t) \in X$ and outputs letters $z(t) \in Z$ and passes successively through the states $a(t) \in A$. Substituting, with due regard to the above-said, letters of the intermediate alphabet $y(t) = k[z(t-1), x(t)]$ into the transition $\delta_0(a, y)$ and output $\lambda_0(a, y)$ functions, obtain:

$$a(t) = \delta_0\{a(t-1),\ k[z(t-1), x(t)]\},$$
$$z(t) = \lambda_0\{a(t-1),\ k[z(t-1), x(t)]\}.$$

Figure 9 shows a circuit implementing these expressions.

It may be seen from the obtained expressions and Fig. 9 that letters $z(t)$ coming from the automaton output to its input with one clock-cycle delay do not differ functionally from the states $a(t)$, thus enabling us to join the states $a(t)$ and the letters $z(t)$ into pairs and to describe the automaton states as a whole by a vector $v(t) = \{a(t), z(t)\} = a(t) \cup z(t)$, $v(t) \in V = A \times Z$.

As the result, the transition and output functions of the programmable connection automaton take the following form:

$$v(t) = \delta_i[v(t-1), x(t)],$$
$$z(t) = \lambda_i[v(t-1), x(t)],$$

where

$$\delta_i[v(t-1), x(t)] = \delta_0\{a(t-1), k_i[z(t-1), x(t)]\} \cup \lambda_0\{a(t-1), k_i[z(t-1), x(t)]\},$$

$$\lambda_i[v(t-1), x(t)] = \lambda_0\{a(t-1), k_i[z(t-1), x(t)]\}.$$

As it may be seen, the transition $\delta_i(v, x)$ and output $\lambda_i(v, x)$ functions greatly depend on the connection function $k_i(z, x)$. Hence, by varying the connection function one can specify the programmable connection automaton to various transition and output functions. In the general case, not only the transition and output functions depend on the connection function, but the state set V_i and output alphabet Z_i of the automaton as a whole.

Any specific connection $k_i(z, x) = \{k_i^1, k_i^2, \ldots, k_i^N\}$, where $k_i \in K$, $k_i^p \in K^*$, $p = 1, 2, \ldots, N$, leads to the fact that in each specific standard automaton only a portion of the input alphabet $Y_p^i \subset Y^*$, a portion of the output alphabet $Z_i^p \subset Z^*$ and a portion of the state set $A_i^p \subset A^*$ are used (see Fig. 10). In this connection, the state set and the alphabets of the integrated automaton (see Fig. 6) are used little more than partially:

$$A_i = \coprod_{p=1}^{N} A_i^p, \quad A_i \subset A,$$

$$Y_i = \coprod_{p=1}^{N} Y_i^p, \quad Y_i \subset Y,$$

$$Z_i = \coprod_{p=1}^{N} Z_i^p, \quad Z_i \subset Z.$$

Show that this indeed takes place, and use the following reasoning to find both partial sets A_i^p, Y_i^p, Z_i^p and sets A_i, Y_i, Z_i. The partial connection function $y^p(t) = k_i^p[z(t-1), x(t)]$ is defined over the set $(Z_i \times X)$ and defines unambiguously the set Y_i^p. This may be written symbolically as $Y_i^p = k_i^p[Z_i \times X]$, or taking into consideration that $Z_i = \coprod_{p=1}^{N} Z_i^p$, as

$$Y_i^p = k_i^p\left[\left(\coprod_{j=1}^{N} Z_i^j\right) \times X\right].$$

On the other hand, the transition $a^p(t) = \delta^*[a^p(t-1), y^p(t)]$ and output $z^p(t) = \lambda^*[a^p(t-1), y^p(t)]$ functions of the standard automata are defined over the set $(A_i^p \times y_i^p)$ and define unambiguously the corresponding sets A_i^p and Z_i^p, this may be written symbolically as

$$A_i^p = \delta^*[A_i^p \times Y_i^p],$$

$$Z_i^p = \lambda^*[A_i^p \times Y_i^p].$$

As the result, we get a system of relations defining the partial sets A_i, Y_i and Z_i vs. the partial connection functions:

$$\left.\begin{aligned} Y_i^p &= k_i^p[(\coprod_{j=1}^{N} Z_i^j)\times X], \\ A_i^p &= \delta^*[A_i^p \times Y_i^p], \\ Z_i^p &= \lambda^*[A_i^p \times Y_i^p]. \\ p &= 1, 2, \ldots N. \end{aligned}\right\}$$

Similar arguments lead to the following relations between the partial sets A_i, Y_i and Z_i, and the complete connection function $k_i(z, x)$ of the integrated automaton (see Fig. 11)

$$\left.\begin{aligned} Y_i &= k_i[Z_i \times X], \\ A_i &= \delta_0[A_i \times Y_i], \\ Z_i &= \lambda_0[A_i \times Y_i]. \end{aligned}\right\}$$

It rests to define the corresponding relations for a system of programmable connection automata as a whole (see Fig. 12). The state set of a system of automata as a whole is formed as a Cartesian product of the state set and the input alphabet of the integrated automaton $Y_i = A_i \times Z_i$. Also the transition $v(t) = \delta_i[v(t-1), x(t)]$ and output $z(t) = \lambda_i[v(t-1), x(t)]$ functions of the automaton as a whole are defined over the set $(V_i \times X)$ and define unambiguously the sets V_i and Z_i. As the result, obtain the following relations for the state set and output alphabet of the system of programmable connection automata:

$$\begin{aligned} V_i &= A_i \times Z_i \\ Y_i &= \delta_i[V \times X] \\ Z_i &= \lambda_i[V \times X], \end{aligned}$$

where $A_i \subset A$, $V_i \subset V$, $Z_i \subset Z$. Since functions $\delta_i(v, x)$ and $\lambda_i(v, x)$ depend essentially on the connection function $k_i(z, x)$. The state set $V_i$and the output alphabet Z_i depend on this function as well.

It follows from the obtained relations that in the programmable connection automaton the state set, the output alphabet, as well as the transition and output functions, on the one hand, depend on the state sets, alphabets, and transition and output functions of the standard automata and, on the other hand, largely depend on the connection function. As the result, the control of the connections becomes a powerful tool for controlling the automaton structure.

In conclusion we are going to list all relations describing the behaviour of the programmable connection automaton:

$$g_i = g_i[V_i, X, Z_i, \delta_i(v, x), \lambda_i(v, x)],$$

$$\delta_i(v, x) = \delta_0\{a(t-1), k_i[z(t-1), x(t)]\} \cup \lambda_0\{a(t-1), k_i[z(t-1), x(t)]\},$$

$$\lambda_i = \lambda_0\{a(t-1), k_i[z(t-1), x(t)]\},$$

$$Y_i = k_i[Z_i \times X_i],$$

$$A_i = \delta_0[A_i \times Y_i],$$

$$Z_i = \lambda_0[A_i \times Y_i],$$

$$V_i = A_i \times Z_i.$$

The characteristic feature of the programmable connection automaton is that its functional program $\pi\{k\} = \{\psi(K)\}$ is given in the simplest possible manner, but at the same time it enables one to vary over wide limits the graph of the automaton, to change its major parameters and, as the result, to change appreciably the properties and possibilities of the automaton as a whole.

Table 1

Classification of automata by programming

	Type of programming	$\pi = \{g, k, a(0)\}$	
		var.	conts.
Individual automaton	Programmable process	$a(0)$	g, k
	Programmable structure	g	$a(0)$, k
	Programmable structure and process	$a(0)$, g	k
System of automata	Programmable connection	k	$a(0)$, g
	Programmable connection and process	$a(0)$, k	g
	Programmable connection and structure	g, k	$a(0)$
	Programmable connection, structure and process	$a(0)$, g, k	—

Conclusion

The programmable structure automata, in particular those with programmable connections, whose graphs may be rearranged by control operators, have much wider functional possibilities than the conventional automata with fixed graphs.

The possibility to rearrange the transition graph endows the automaton with special features and flexibility in adjusting to the environment, enables it to give the most effective response to external actions.

The rearrangeable graph automaton features a potential possibility of self-adjustment, adaptation to the environment, and organizing the internal structure in the process of learning. In such automata there is possibility of random search and selection of the most viable, the most adjusted to the environment structures and graphs.

As it was shown in the present paper, the principles and methods of programming and rearrangement of graphs and structures of such automata are simple enough and may be readily implemented.

APPLICATION OF STRUCTURAL TRANSFORMS TO THE ANALYSIS OF TIME-VARYING LINEAR SYSTEMS

F. A. MIKHAILOV, O. F. TYKHEVICH, M. K. KHADZHINOV

(Moscow)

Determination of transfer functions for various combinations of time-varying linear systems with known transfer functions of components is discussed. By means of the left Laplace transform precise expressions are obtained for the functions of parallel, series and cyclic connections. The formulae of the two latter connections are shown to be too complex in practice, and therefore a rough method is proposed where the transfer function of the system and its units are approximated by a bilinear function of the complex variable s with time-dependent coefficients. Formulae are derived relating coefficients of the connection transfer function to those of the unit transfer functions. The obtained results are illustrated by an example.

Introduction

For the analysis of complicated automatic control systems, block-diagrams representing a system as a combination of simpler systems (units) are of great convenience. In the studies of stationary linear systems the units are usually described by their transfer functions which may be easily obtained from the unit differential equations. The transfer functions might be also used to characterize the time-varying linear systems; however, the problem here is complicated since not all units do have transfer functions and a simple relation between differential equations and transfer functions is lacking. As the result, in the today's methods of structural analysis of time-varying linear systems units are defined by differential equations [2, 3] or pulse transient functions [4]. At the same time, a method for the analysis of block-diagrams of linear time-varying systems based, like the analysis of stationary systems, on transfer functions is of obvious practical interest.

The present paper discusses finding the system transfer function through those of its units. In so doing the system and the time interval under consideration are assumed to be such that the transfer functions of the units and the system as a whole do exist. The presentation has the following plan:

1. Derivation of an equation relating the unit transfer functions to those of elementary connections.

2. Evaluation of complexity of these equations and formulation of the problem of finding a practicable approximation method for connection transfer functions.

3. Presentation of such a method.

4. An example illustrating application of the method for analytical purposes.

1. Structural transforms and formulae of connection transfer functions

Any complex time-varying linear system may be represented as a connection of simpler systems or units. In the general case, the system may be represented as a combination of three types of connections, parallel, series and "cyclic" (Fig. 1). Find the transfer functions of these connections.

Parallel connection is depicted in Fig. 1a. By superposition its transfer function is as follows:

$$W_n(s, t) = W_1(s, t) + W_2(s, t). \tag{1.1}$$

Series connection is shown in Fig. 1b. To find the connection transfer function $W_3(s, t)$ make use of the pulse transient function for the series connection [4]

$$w_3(t, u) = \int_u^t w_2(t, \eta) \cdot w_1(\eta, u)\, d\eta\,. \tag{1.2}$$

By substituting the argument of integration η by $\xi = t - \eta$ obtain

$$w_3(t, u) = \int_0^{t-u} w_2(t, t - \xi) \cdot w_1(t - \xi, u) d\xi\,. \tag{1.3}$$

Substitute in (1.3) instead of function $w_2(t, t - \xi)$ its expression through the transfer function of the second unit

$$w_2(t, t - \xi) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma, t) \exp(\sigma\xi) \cdot d\sigma\,, \tag{1.4}$$

where c is a positive number such that the integral converges absolutely for all Re $\sigma \geq c$. As the result find

$$w_3(t, u) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma, t) \int_0^{t-u} w_1(t - \xi, u) \exp(\sigma\xi)\, d\xi\, d\sigma\,. \tag{1.5}$$

Substituting $t - \tau$ for u and applying the Laplace transform with respect to argument τ to both parts of (1.5) obtain

$$W_3(s,t) = \frac{1}{2\pi i} \int_0^{\infty} \exp(-s\tau) \int_{c-i\infty}^{c+i\infty} W_2(\sigma,t) \int_0^{\tau} w_1(t-\xi, t=\tau) \exp(\sigma\xi)\, d\xi\, d\sigma\, d\tau . \tag{1.6}$$

Taking into consideration that $w_1(t-\xi, t-\tau) = 0$ at $\xi > \tau$ change the upper limit of integration over ξ for ∞. If we assume that all integrals in (1.6) converge uniformly, the order of integration may be changed [11]

$$W_3(s,t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma,t) \int_0^{\infty} \exp(\sigma\xi) \int_0^{\infty} w_1(t-\xi, t-\tau) \exp(-s\tau)\, d\tau\, d\xi\, d\sigma . \tag{1.7}$$

Rewrite the internal integral as

$$\int_0^{\infty} w_1[t-\xi, (t-\xi) - (\tau-\xi)] \cdot \exp(-s\tau)\, d\tau$$

and, confining ourselves to the case of $\xi \leq \tau$ (since for $\xi \leq \tau$ the integrand is zero), reduce it by the first theorem of biasing [5] to

$$\exp(-s\xi) \cdot W_1(s, t-\xi), \tag{1.8}$$

where $W_1(s, t-\xi)$ is the transfer function of the first unit with complex argument s and real one $t - \xi$.

Now the integral over ξ in (1.7) may be regarded as a biased to the complex number σ left Laplace transform [1] of function $W_1(s,t)$ regarded as function of t:

$$\int_0^{\infty} W_1(s, t-\xi) \exp[-\xi(s-\sigma)]\, d\xi = \Lambda_{t/s-\sigma}[W_1(s,t)] . \tag{1.9}$$

Here the subscripts of the left Laplace transform indicate, respectively, the real and complex arguments of the transform.

Rewrite (1.7) with due regard to (1.8) and (1.9)

$$W_3(s,t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma,t) \cdot \Lambda_{t/s-\sigma}[W_1(s,t)]\, d\sigma . \tag{1.10}$$

By applying the left Laplace transform and regarding the integrand as the left transform with complex argument σ of a function of t with a parameter s, we can write

$$W_3(s, t) = \Lambda^{-1}_{t/\sigma} \{W_2(\sigma, t) \cdot \Lambda_{t/s-\sigma} [W_1(s, t)]\} . \tag{1.11}$$

The series connection transfer function $W_3(s, t)$ is, thus, expressed in an explicit form through transfer functions $W_1(s, t)$ and $W_2(s, t)$ of its units. Note that this relation is very complex. Equation (1.11) holds also for the stationary system with unit transfer functions $W_1(s)$ and $W_2(s)$ since

$$\Lambda_{t/s-\sigma} [W_1(s)] = \frac{W_1(s)}{s - \sigma} , \tag{1.12}$$

$$W_3(s) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} W_2(\sigma) \frac{W_1(s)}{s - \sigma} d\sigma = \frac{W_1(s)}{2\pi i} \int_{c+i\infty}^{c-i\infty} \frac{W_2(\sigma)}{\sigma - s} d\sigma . \tag{1.13}$$

This equality is defined for $\operatorname{Re} \sigma \geq c$, but in this case integral (1.13) is equal to $2\pi i W_2(s)$ by the Cauchy integral formula. Hence

$$W_3(s) = W_1(s) \cdot W_2(s). \tag{1.14}$$

"*Cyclic*" *connection* is shown in Fig. 1c. To find its transfer function make use of the results for the series connection. The pulse transient function $w_y(t, u)$ may be defined by the following expression [4]:

$$w_y(t, u) = w_p(t, u) + \int_u^t w_p(t, \eta)\, w_y(\eta, u)\, d\eta , \tag{1.15}$$

By analogy with (1.11) write immediately

$$W_y(s, t) = W_p(s, t) + \Lambda^{-1}_{t/\sigma} \{W_p(\sigma, t) \cdot \Lambda_{t/s-\sigma} [W_y(s, t)]\} . \tag{1.16}$$

Equation (1.16) is a complex integral one with $W_y(s, t)$ as the unknown function. One is unable to obtain from it the explicit form of the connection transfer function $W_y(s, t)$. Like (1.11), for the stationary system (1.16) turns into the well-known relation solvable with respect to $W_y(s)$:

$$W_y(s) = W_p(s) + W_p(s) \cdot W_y(s) . \tag{1.17}$$

Thus, for all three types of connections we have formulae expressing the connection transfer functions through those of their units either explicitly, (1.1) and (1.11), or implicitly as the integral equation (1.16). However, (1.11) and (1.16) are unpractical. Therefore a problem suggests itself of finding for the series and cyclic connections approximate but applicable formulae.

2. Power expansion of the pulse transient function

Consider expansion of the pulse transient function into a power series with respect to argument $\tau = t - u$ under assumption that the expansion converges in a domain $\tau \leq Q$:

$$w(t, t-\tau) = \sum_{k=0}^{\infty} w_k(t) \cdot \frac{\tau^k}{k!} . \tag{2.1}$$

Referring to [1], coefficients $w_k(t)$ may be expressed through those of the differential equation system

$$\frac{d^n x}{dt^n} + b_1(t) \frac{d^{n-1} x}{dt^{n-1}} + \ldots + b_n(t)\, x = a_0(t) \frac{d^{n-1} y}{dt^{n-1}} + \ldots + a_{n-1}(t)\, y \tag{2.2}$$

where x is output, y is input, and $a_i(t)$ and $b_i(t)$ are analytical functions. These relations are given in Table 1.

The pulse transient function $w(t, u)$ may be found through the following formula [2]:

$$w(t, u) = \sum_{i=0}^{n-1} (-1)^{n-i-1} \cdot \frac{d^{n-i-1}}{du^{n-i-1}} [a_i(u) \cdot g(t, u)] ; \tag{2.3}$$

whose right side contains function $g(t, u)$ and coefficients of the right side of eq. (2.2) Since coefficients of (2.2) are analytical functions, $a_i(u)$ may be expanded into convergent Taylor series in a vicinity of a point $u = t$, namely

$$a_i(u) = \sum_{j=0}^{\infty} a_i^{(j)}(t) \cdot \frac{(t-u)^j}{j!} (-1)^j , \tag{2.4}$$

where $a_i^{(j)}(t)$ is derivative of order j with respect to t of function $a_i(t)$. Henceforth, function $a_i(t)$ proper will be understood as the zero-order derivative $a_i^{(0)}(t)$ of itself.

Function $g(t, u)$ is the pulse transient function of the integrating part of the system described by the right-hand side of eq. (2.2) changing $y(t)$. It is possible to use for this part the following expansion [1, p. 244]:

$$g(t, t-\tau) = \sum_{l=n-1}^{\infty} g_{n,l}(t) \frac{r}{l!}$$

where the coefficient may be expressed by $b_1(t), \ldots, b_n(t)$. This relations for $n < 4$ is given in Table 1.

Substituting (2.3) and(2.4) into (2.5) and taking into account that the product of two convergent series is also a convergent series [9], obtain

$$w(t, u) = \sum_{i=0}^{n-1} \sum_{l=n-1}^{\infty} \sum_{j=0}^{\infty} (-1)^{n-1-i+j} a_i^{(j)}(t) \cdot g_{n,l}(t) \frac{d^{n-1-i}}{dt^{n-1-i}} \left[\frac{(t-u)^{l+j}}{l!\, j!} \right] . \tag{2.5}$$

By taking $t - u = \tau$ and equating coefficients for equal powers of τ obtain

$$w_k(t) = \sum_{i=0}^{n-1} \sum_{l=n-1}^{n-1+k+i} (-1)^{n-1+k-i-l} \binom{n-1+k-i}{l} \cdot a_l^{(n-1+k-i-i)}(t)\, g_{n,l}(t), \quad (2.6)$$

where $(n_l^{(n-1+k-i)})$ are binomial coefficients.

3. Power expansion of the transfer function

By regarding in the right side of (2.1) τ as an independent variable and t as a parameter, and transforming the series termwise according to Laplace, obtain a formal expansion into a series for the transfer function

$$W(s, t) \sim \sum_{k=0}^{\infty} \frac{w_k(t)}{s^{k+1}} \quad (3.1)$$

Referring to [5], series (3.1) may not always be convergent but it always is an asymptotic expansion of function $W(s, t)$ for $s \to \infty$. And inversely, by (3.1) one can easily restore (2.1), i.e. values of $w(t, t - \tau)$ for small τ.

4. Approximation of the transfer function by bilinear one

One cannot find the analytical expression of the time-varying system transfer function for the general case. To approximate it the following bilinear function is the most appropriate:

$$\tilde{W}(s, t) = \frac{c^0(t) \cdot s^{q-1} + \ldots + c_{q-2}(t) \cdot s + c_{q-1}(t)}{s^q + d_1(t)\, s^{q-1} + \ldots + d_q(t)}, \quad (4.1)$$

where $c_i(t)$, $i = 0, 1, \ldots, q - 1$, and $d_i(t)$, $i = 1, 2, \ldots q$, are real continuous functions of time.

Transfer function $W(s, t)$ of type (4.1) is single-valued and analytical for all sufficiently large $|s|$, therefore it may be expanded in this vicinity into Laurent series

$$\tilde{W}(s, t) = \sum_{k=0}^{\infty} \frac{\tilde{w}_k(t)}{s^{k+1}}.$$

Coefficients $\tilde{w}_k(t)$ of this series are defined through the coefficients of the transfer function (4.1) by means of recurrent formula

$$\tilde{w}_k(t) = c_k(t) - \sum_{i=1}^{q} d_i(t) \cdot \tilde{w}_{k-i}(t) \quad k = 0, 1, \ldots, \infty \quad (4.2)$$

where $c_k = 0$ at $k > q-1$; $\tilde{w}_k(t) = 0$ at $i < k$.

By (4.2), all coefficients of the transfer function (4.1) may be determined if $2q$ first coefficients of $\tilde{w}_k(t)$ are known. Therefore the approximation should be based upon the following requirement:

$$\tilde{w}_k(t) = w_k(t), \quad k = 0, 1, \ldots, 2q - 1, \tag{4.3}$$

where $w_k(t)$ are coefficients of expansion (3.1).

This approximation of the transfer function is, thus, ambiguous being dependent on the power q of the bilinear function polynomial denominator. Choice of q will be discussed below, and for the time being assumed that it is chosen already. Then coefficients of the transfer function (4.1) in the bilinear form are related to the first $2q$ coefficients of the approximating series by a set of algebraic equations obtained from (4.2) ($k = 0, 1, \ldots, 2q - 1$).

The set is solved as follows. Having separated a set of equations with respect to coefficients $d_i(t)$ ($k = q, q + 1, \ldots, 2q - 1$) obtain from it all $d_i(t)$, and by substituting them into the rest of equations ($k = 0, 1, \ldots, q - 1$) obtain formulae for all coefficients $c_k(t)$.

It may happen that the set of algebraic equations has no solution with respect to coefficients $d_i(t)$ because its determinant is identically equal to zero. In this case one has to increase q and construct a new set of equations.

5. Approximate transfer functions of connections

Represent pulse transient functions of units and their connections as series (2.1). By making use of this formula one can obtain formulae for function $W_3(s, t)$ approximating the transfer function of the series connection whose coefficients are expressed through those of unit transfer functions. To this end, one can use expressions of [6] relating connection pulse transient functions to those of units for the values of the argument τ taken from the common to both connection and units convergence domain.

Coefficients $w_{3,\tau}(t)$, $r = 0, 1, \ldots$ (the first index defines unit or connection) of the expansion of the pulse transfer functions $w_3(t, u)$ of the series connection of two units (Fig. 1b) are determined as follows [6]:

$$w_{3,r}(t) = \sum_{l=0}^{r-1} \sum_{k=0}^{r-1-l} (-1)^{r-1-l-k} \cdot \binom{r-1-l}{k} \cdot w_{2,k}(t) \cdot w_{1,l}^{(r-1-l-k)}(t) \tag{5.1}$$

This expression enables one to synthesise and approximate transfer function of the series connection by means of the given approximate transfer functions of the units with allowance for (4.3).

The order of derivation of connection transfer function coefficients is as follows:

A) Find coefficients $\tilde{w}_{i,j}(t)$ of the power expansion of the given approximate transfer function of each unit expressed through coefficients $c_i(t)$ and $d_i(t)$.

B) Given power q of the approximate transfer function denominator, find by (5.1) the first $2q$ coefficients of its power expansion.

C) The obtained coefficients of the connection transfer function power expansion are substituted into the set (4.2) of equations with respect to coefficients $c_i(t)$ and $d_i(t)$ of the desired transfer function. Having solved this system obtain the coefficients of the connection transfer function expressed through the unit transfer functions.

For the cyclic connection (Fig. 1c), coefficients $W_{y,r}(t) = \tilde{w}_{y,r}(t)$ of the expansion of the connection pulse transient function are expressed by a recurrent formula through the coefficients of the pulse transient function expansion of an open-loop system [6]:

$$w_{y,r}(t) = w_{p,r}(t) + \sum_{l=0}^{r-1} \sum_{k=0}^{r-1-l} (-1)^{r-l-1-k} \cdot \binom{r-1-l}{k} w_{p,k}(t) \cdot w_{y.l}^{(r-1-l-k)}(t)$$

$$w_{y,0}(t) = w_{p,0}(t), \quad r = 1, 2, \ldots \tag{5.2}$$

The sequence of finding the coefficients of the cyclic connection is the same as with the series connection.

The procedure above is also applicable if the approximate transfer functions of connection units are unknown and the units are described by differential equations defined over interval $(-\infty, T]$ and with coefficients being analytical functions. But in this case coefficients of the power expansion for the units are determined by (2.8) through the coefficients of their equations. All following operations are the same as above.

The procedure above was used to derive the transfer function ocefficients for series and cyclic connections in the case where the power of the denominator polynomial of the approximate transfer functions is taken equal to the sum of the denominator power of the unit transfer functions. The formulae are tabulated. In Table 2 $W_1(s, t)$ and $W_2(s, t)$ stand for the unit transfer function and $W_3(s, t)$ is that of their series connection (Fig. 1b). In Table 3 the transfer function of the open circuit is denoted with $W_p(s, t)$, and that of cyclic connection with $W_y(s, t)$ (Fig. 1c). Table 4 gives formulae for connections with nonunit negative feedback. Here $W_1(s,t)$, $W_2(s,t)$ and $W_3(s,t)$ stand for the transfer function, respectively, the forward circuit, the feedback and the connection as a whole (Fig. 1d).

6. Applicability of the method and an example

The method is based upon the power series expansion of the pulse transient function $w(t, u)$ in a vicinity of point $t = u$ with respect to powers $\tau = t - u$. The convergence rate within the convergence circle decreases with the distance form the expansion point thus deteriorating precision of (5.1) and (5.2) with increase of $\tau = t - u$. Yet the approximation assumes that a certain number of first approximation terms coincide, that is why the method works well with systems whose output depends basically on inputs at moments sufficiently close to those of response observations, i.e. with systems having pulse transient function quickly attenuating with respect to argument τ. For stationary systems the obtained formulae give precise results. Therefore the proposed method of structural transformations is effective when applied to systems with slowly changing parameters.

Example. Make use of the above results to obtain the transfer function and output variance of a time-varying system which is the simplest mathematical model of a movement control system.

Consider rendez-vous in a plane of two bodies of which one is uncontrolled and moves uniformly along a rectilinear path, and another has a constant speed and is controlled so as to approach the first body [7]. We are going to analyse normal accelerations of the controlled body during approach.

Write the linearized equation of the relative movement kinematics as

$$D(t) \cdot \frac{d\dot{\varepsilon}}{dt} + 2\dot{D}\dot{\varepsilon} = n\,; \quad 0 \leq t \leq T, \tag{6.1}$$

where $D(t)$ is the current distance relative to the reference trajectory; $\dot{\varepsilon}$ is the angular speed of the line of sight (i.e. the line connecting the centers of masses of the two bodies); n is the projection of the control plant on the normal to the line of sight, and T is the total rendez-vous time.

The unit whose dynamics is described by (6.1) will be termed as "kinematic". n and $\dot{\varepsilon}$ are, respectively, its input and output.

Control is assumed to be exercised according to the proportional navigation method where the control signal is in proportion to the angular sighting speed $\dot{\varepsilon}_4$ as measured by an on-board device. This signal acting upon the control plant through the on-board control circuit, changes the value of n. The on-board control circuit dynamics is described by the transfer function of a constant parameter oscillating unit:

$$W_1(s) = \frac{k\omega_0^2}{s^2 + 2\xi\,\omega_0\, s + \omega_0^2}\,. \tag{6.2}$$

The block-diagram of the system under consideration is depicted in Fig. 2. The error of sighting line angular speed measurement by the on-board devices is the system input, and n is its output.

Confine ourselves to the case when there is only random disturbance $\psi(t)$ of the white noise type at the system input.

To compute the output variance $\sigma^2(t)$ it suffices in this case to know the noise spectral density S and the integral characteristic $\mu(t)$, mobility degree defined as follows:

$$\mu(t) = \frac{1}{\pi} \int_0^\infty |W(i\omega, t)| \, d\omega \, .$$

The mathematical formula is as follows:

$$\sigma^2(t) = S \cdot \mu(t). \tag{6.3}$$

Since in the physical system the initial output variance is zero, this formula may be used to calculate the variance only after the completion of the transient.

Approximate the system transfer function by a bilinear function of (4.1) kind, the mobility degree $\mu(t)$ may be then determined algebraically [1]. In particular, for $n = 3$

$$\mu(t) = \frac{c_0^2 d_2 + c_1^2 - 2c_0 c_2 + c_2^2 \dfrac{d_1}{d_3}}{2(d_1 d_2 - d_3)} \, ; \tag{6.4}$$

for $n = 4$

$$\mu(t) = \frac{1}{2H}\left[c_0^2(d_2 d_3 - d_1 d_4) + (c_1^2 - 2c_0 c_2) d_3 + (c_2^2 - 2c_1 c_3) d_1 + c_3^2 \frac{d_1 d_2 - d_3}{d_4}\right],$$

$$H = d_1(d_2 d_3 - d_1 d_4) - d_3^2 \, . \tag{6.5}$$

The system transfer function is obtained by means of formulae of Table 3. To make use of the tables one should know the transfer functions of all units.

The kinematic unit equation (6.1) is linear, first-order differential equation with variable coefficients and has a finite solution. Under any distance variation profile $D(t)$ the pulse transient function $w_2(t, u)$ is as follows:

$$w_2(t, u) = \frac{D(u)}{D^2(t)} \, .$$

Substituting $t - \tau$ for u and performing Laplace transform of $w_2(t, t - \tau)$ with respect to τ obtain the kinematic unit transfer function for $\dot{D} = \text{const}$:

$$W_2(s, t) = \frac{\dfrac{1}{D(t)} \cdot \left(s - \dfrac{\dot{D}}{D(t)}\right)}{s^2} . \tag{6.6}$$

Entry 8 in Table 4 corresponds to the system under consideration. Using these formulae, compute the system transfer function coefficients. The transfer function is

$$W(s, t) = \frac{k\omega_0^2 s^2}{s^4 + 2\xi\omega_0 s^3 + \omega_0^2 s^2 + \dfrac{k}{D(t)}\omega_0^2 s + \dfrac{k\dot{D}}{D^2(t)}} . \tag{6.7}$$

For the sake of comparison determine the system transfer function by means of some other approximation techniques.

Applying stationary system methods and using the differential equation of the system as a whole

$$\dddot{n} + \left(2\xi\omega_0 + 2\frac{\dot{D}}{D(t)}\right)\ddot{n} + \left(\omega_0^2 + 4\xi\omega_0\frac{\dot{D}}{D(t)}\right)\dot{n} + \omega_0^2\frac{k + 2\dot{D}}{D(t)}n =$$

$$= k\omega_0^2\left(\dot{\psi} + 2\frac{\dot{D}}{D(t)}\psi\right) \tag{6.8}$$

find the transfer function of the time-varying system "frozen" at time t:

$$W(s, t) = \frac{k\omega_0^2\left(s + 2\dfrac{\dot{D}}{D(t)}\right)}{s^3 + \left(2\xi\omega_0 + 2\dfrac{\dot{D}}{D(t)}\right)s^2 + \left(\omega_0^2 + 4\xi\omega_0\dfrac{\dot{D}}{D(t)}\right)s + k\omega_0^2 \cdot \dfrac{1}{D(t)} + 2\omega_0^2\dfrac{\dot{D}}{D(t)}} . \tag{6.9}$$

The transfer function may be obtained by the expansion of the pulse transient function into a power series (see Section 2) and succeeding determination of the transfer function coefficients (see Section 4). In so doing, take the same power of the denominator polynomial as in (6.9), Skipping calcula-

tions, write down the result:

$$W(s,t) = \frac{k\omega_0^2\left(s - \dfrac{\dot{D}}{D(t)}\right)}{s^3 + \left(2\xi\omega_0 - \dfrac{\dot{D}}{D(t)}\right)s^2 + \left(\omega_0^2 - 2\xi\omega_0\dfrac{\dot{D}}{D(t)}\right)s + \omega_0\left(\dfrac{k-\dot{D}}{D(t)}\right)}. \tag{6.10}$$

Another version of the transfer function may be obtained by applying the formula for tranformation of stationary system block-diagrams to $W_1(s)$ and $W_2(s, t)$

$$W(s,t) = \frac{k\omega_0^2 s^2}{s^4 + 2\xi\omega_0 s^3 + \omega_0^2 s^2 + \dfrac{k}{D(t)}\omega_0^3 s + \dfrac{k\dot{D}}{D^2(t)}\omega_0^2}. \tag{6.11}$$

For all the versions of the transfer function (6.3), (6.4), (6.5) find system output variance:

$$\sigma_1^2 = S \cdot \frac{k^2\omega_0^2}{2\left(2\xi\omega_0 - 4\xi^2\dfrac{\dot{D}}{D(t)} - \dfrac{k}{D(t)}\right)} \tag{6.12}$$

$$\sigma_2^2 = S \cdot \frac{k^2\omega_0\left(\omega_0^2 D^2(t) + 8\dot{D}^2\dfrac{\xi\omega_0 D(t) + \dot{D}}{k + 2\dot{D}}\right)}{2\left[8\xi\dot{D}^2 + \omega_0\left(2\xi\omega_0 D^2(t) + 8\xi^2\dot{D}D(t) - kD(t)\right)\right]} \tag{6.13}$$

$$\sigma_3^2 = S\frac{k^2\omega_0\left(\omega_0^2 D^2(t) + \dot{D}^2\dfrac{2\xi\omega_0 D(t) - \dot{D}}{k - \dot{D}}\right)}{2\left[2\xi\dot{D}^2 + \omega_0\left(2\xi\omega_0 D^2(t) - 4\xi^2\dot{D}D(t) - kD(t)\right)\right]} \tag{6.14}$$

$$\sigma_4^2 = S\frac{k^2\omega_0^2}{2\left(2\xi\omega_0 + 4\xi^2\dfrac{\dot{D}}{D(t)} - \dfrac{k}{D(t)}\right)}, \tag{6.15}$$

where σ_1, σ_2, σ_3, σ_4 correspond to transfer functions (6.7), (6.9), (6.10), (6.11), respectively.

Plots showing σ_1, σ_2, σ_3, σ_4 *vs.* distance and computed through (6.12), (6.13), (6.14), (6.15) are depicted in Fig. 3 for the following system parameters: $k = 5$ sec/deg; $w_0 = 10\,\text{sec}^{-1}$; $\xi = 0.7$; $S = 0.004$ deg^2/sec. The precise values of σ^2

were found by means of a digital computer as the solution to a set of equations constructed by the Duncan method [8] for the same system parameters and input signal. The plot for σ is shown in Fig. 3. As illustrated in the figure, the solution through (6.12) is the closest to the precise one.

Conclusion

The structural transformation as described here enables one in some cases to simplify appreciably derivation of the time-varying linear system transfer function. The procedure of finding the system transfer function makes use of the formulae in Tables 1—4. If the tables do not contain a suitable entry, the transfer function may be obtained by finding successively coefficients of the unit transfer function power expansions, coefficients of the connection transfer function power expansion, coefficients of the connection transfer function.

Appendix

Left Laplace transform

Consider a function $g(t)$ defined over the interval $(-\infty, T]$ where $T < \infty$ and satisfying convergence conditions of the integral

$$\int_{-\infty}^{T} |g(t)| \exp[s(t-T)]\,dt < \infty \tag{1}$$

at all values of $t \in (-\infty, T]$ and at some real value c of the complex variable s. If a new variable τ is introduced taking values from the interval $[0, \infty)$, function $g(t-\tau)$ may be Laplace transform at any fixed t resulting in

$$G(s, t) = \int_{0}^{\infty} g(t-\tau) \cdot \exp(-s\tau)\,d\tau \tag{2}$$

having, with due regard to inequality (1), the domain of convergence $\mathrm{Re}\ s \geq c$.

Passage from $g(t)$ to $G(s, t)$ is referred to as the "left Laplace transform" ([1], p. 493).

The left Laplace transform assigns to each $g(t)$ with the domain of definition as above function $G(s, t)$ of complex variable s and real one t. If functions $g(t)$ and $G(s, t)$ are related by the left Laplace transform, $g(t)$ may be termed as the "pre-image", and $G(s, t)$ as its "left image". Symbolics of this relation is as follows:

$$G(s, t) = \Lambda[g(t)];\ g(t) = \Lambda^{-1}[G(s, t)]. \tag{3}$$

From the theory of the conventional Laplace transform it follows that $G(s, t)$ is analytical over s for $\mathrm{Re}\ s > c$. With due regard to the conventional Laplace reconversion, the left Laplace reconversion is as follows

$$g(t) = \lim_{\tau \to 0} \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} G(s, t) \cdot \exp(s\tau)\,ds, \tag{4}$$

where $i = \sqrt{-1}$.

References

1. *Mikhailov, F. A., Teryaev, E. D., Bulekov, V. P., Salikov, L. M., Dikanova, L. S.*, "Dynamics of continuous linear systems with determined and random parameters", "Nauka", Moscow, 1971 (in Russian).
2. *Solodov, A. V., Petrov, F. S.*, "Variable parameter linear automatic systems", "Nauka" Moscow, 1971 (in Russian).
3. *Leondes, C. T.* (ed.), Modern Control Systems Theory, McGraw-Hill, Inc.
4. *Malchikov, S. V.*, Synthesis of linear automatic control systems with varying parameters, "Automatics and Remote Control", **20** (1959) 12.
5. *Deutsch, G.*, Anleitung zum Praktischen Gebrauch der Laplace-Transformation, R. Oldenburg, München, Wien, 1967.
6. *Mikhailov, F. A., Tykhevich, O. F., Khadzhinov, M. K.*, Calculation of characteristics of structural connections of linear non-stationary systems, in "Dynamics of flight vehicle control", Proceedings of Moscow Aviation Inst., **240** Moscow, 1972, (in Russian).
7. *Mikhailov, F. A., Teryaev, E. D., Bulekov, V. P., Dankov, G. Yu., Salikov, L. M., Stephan'yants, G. A.*, "Dynamics of non-stationary linear systems", "Nauka", Moscow, 1967 (in Russian).
8. *Laning, J., Batin, R.*, Random processes in automatic control, New York, a. o., McGraw-Hill, 1956.
9. *Bitsadze, A. V.*, "Fundamentals of the theory of analytical functions of the complex variable", "Nauka", Moscow, 1969 (in Russian).
10. *Zadeh, L. A.*, On stability of linear varying parameter system, J. Appl. Ph., **22** (April 1951).
11. *Lavrent'ev, M. A., Shabat, B. V.*, "Methods in the theory of functions of the complex variable", "Nauka", Moscow, 1965 (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 которы 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 гранок.

Статье должна предшествовать аннотация объемом до 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 высылаются гранки статьи, которые они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Csibi, S.: Learning under computational constraints from weakly dependent samples (*Чиби Ш.* Обучение при вычислительных ограничениях на основе статистически зависимых данных) 3

Bársony, A.: A linear learning algorithm for the purpose of identification (*Баршонь А.:* Применение линейного алгоритма обучения для идентификации динамических характеристик) 23

Каляев А. В.: Автоматы с программируемой структурой (*Kalyaev, A. V.:* Programmable structure automata) 35

Almásy, G. A., Sztanó, T.: Checking and correction of measurements on the basis of linear system model (*Алмаши Г. А., Стано Т.* Проверка и поправка результатов измерений на основе линейной модели системы) 57

Михаилов Ф. А., Тыхевич О. Ф., Хаджинов М. К. Применение структурных преобразований к анализу нестационарных линейных систем (*Mikhailov, F. A., Tykhevich, O. F., Khadzhinov, M. K.:* Application of structural transforms to the analysis of time-varying linear systems) 71

Book Review 95

316.920

VOL. 4 * NUMBER 2
ТОМ 4 * НОМЕР 2

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

1975

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 four issues published per year make up a volume of some 320 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.

The quarterly is published by

AKADÉMIAI KIADÓ
Publishing House of the Hungarian Academy of Sciences
H-1363 Budapest, P.O. Box 24, Hungary

Президиумами Академии Наук СССР и Академии Наук ВНР было принято решение о совместном издании журнала

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал создан для обеспечения быстрейшей публикации материалов о новейших результатах научных исследований, разработок и кратких сообщений 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 издается четыре раза в год общим объёмом около 320 печатных страниц.

Журнал в первую очередь будет публиковать научные достижения обеих стран, а также статьи из других стран, представляющие международный интерес.

Издатель журнала:

AKADÉMIAI KIADÓ
Издательство Академии Наук Венгрии
H-1363 Будапешт, п. о. 24, Венгрия

Distributor for:

Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Roumania, USSR, People's Republic of Viet-Nam, Yugoslavia

KULTURA
Hungarian Trading Co. for Books and Newspapers
H-1389 Budapest 62, P.O.B. 149, Hungary

Distributor for all other countries:

PLENUM PUBLISHING CO. LTD
4a Lower John Street
London. W1R 3PD
ENGLAND

Annual subscription price: US $ 35,– plus postage of US $ 1,80

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. EMELIANOV (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. 4 (2), pp. 97—102 (1975)

НЕВЕРОЯТНОСТНАЯ ВЗАИМНАЯ ИНФОРМАЦИЯ БЕЗ ПАМЯТИ*

В. Д. ГОППА

(Москва)

(Поступила в редакцию 12 февраля 1974 г.)

Вводится невероятностный аналог взаимной информации. Предлагается универсальный алгоритм декодирования для дискретных каналов без памяти.

1. Для любых двух слов $x \in X^n$, $x = (x_1, \ldots, x_n)$, $|X| = L_1$, $y \in Y^n$, $y = (y_1, \ldots, y_n)$, $|Y| = L_2$ образуем новое слово $x \otimes y \in (X \otimes Y)^n$, $x \otimes y = ((x_1, y_1), (x_2y_2), \ldots, (x_ny_n))$. Пусть $\{m_i\}$, $\{n_j\}$, $\{m_{ij}\}$ — композиции слов x, y и $x \otimes y$ соответственно, $i = 1, 2, \ldots, L_1$, $j = 1, 2, \ldots, L_2$.

Очевидно, $\{m_{ij}/n\}$ является распределением вероятностей на $X \otimes Y$, а $\{m_i/_n\}$ и $\{n_j/n\}$ — его частными распределениями на X и Y соответственно, поскольку $\sum\limits_{i,j} m_{ij} = n$, $\sum\limits_{j} m_{ij} = m_i$, $\sum\limits_{i} m_{ij} = n_j$. Отсюда следует, что $\Phi(x \otimes y) \leq \leq \Phi(x) + \Phi(y)$, где $\Phi(x)$ — вес Фитингофа слова x [1].

Введем следующие величины:

$$\Phi(x : y) = \Phi(x) + \Phi(y) - \Phi(x \otimes y),$$

$$\Phi(y/x) = \Phi(x \otimes y) - \Phi(x).$$

Легко проверяются следующие свойства этих величин:

1) $\Phi(x : y) = \Phi(y : x)$;

2) $\Phi(x : x) = \Phi(x)$ (поскольку $\Phi(x \otimes x) = \Phi(x)$);

3) $\Phi(x : y) = \Phi(x) - \Phi(x/y) = \Phi(y) - \Phi(y/x)$.

Лемма 1. Пусть Q — распределение вероятностей на $X \otimes Y$, а Q^n — распределение на $X^n \otimes Y^n$, являющееся произведением мер на $X \otimes Y$ (дискретный канал без памяти), $R = H(X : -H(X/Y)$ — шенноновская взаимная информация. Тогда

$$\Phi(x: y) \xrightarrow[n.B.]{} R, \quad \Phi(y/x) \xrightarrow[n.B.]{} H(Y/X).$$

Доказательство следует из определений и закона больших чисел. Введем теперь расстояние между словами x и y следующим образом:

$$\varrho(x, y) = [\Phi(x/y) + \Phi(y/x)]/2.$$

* Основные результаты данной работы были доложены на III-ем Международном симпозиуме по теории информации (Таллин, 1973).

Заметим, что x и y являются словами одной и той же длины, но, быть может, из разных алфавитов.

Выясним условия, при которых $\Phi(y/x) = 0$. Согласно определению, это означает, что $\Phi(x \otimes y) = \Phi(x)$. Последнее равенство будет справедливо при выполнении следующего условия: если в слове x буква $a_i \in X$ занимает места $j_1, j_2, \ldots, j_k$, то на этих местах в слове y будет стоять одна и та же буква $b_i \in Y$, безразлично какая. Если же равенство $y_{j_1} = y_{j_2} = \ldots = y_{jk}$ не будет выполняться, это приводит к тому, что $\Phi(x \otimes y) > \Phi(x)$.Например, для пары слов

$$\begin{matrix} x = (11121233) \\ y = (77777011) \end{matrix} \qquad \Phi(y/x) \neq 0, \qquad \Phi(x/y) \neq 0,$$

для слов

$$\begin{matrix} x = (11121233) \\ y' = (77777711) \end{matrix} \qquad \Phi(y'/x) = 0, \qquad \Phi(x/y') \neq 0,$$

а для слов

$$\begin{matrix} x' = (11111133) \\ y' = (77777711) \end{matrix} \qquad \Phi(y'/x') = 0, \qquad \Phi(x'/y') = 0.$$

Таким образом, равенство $\varrho(x, y) = 0$ означает, что слово y получено из слова x просто переименованием букв, причем разные буквы слова x имеют различные новые названия. Ясно, что отношение $x \sim y \Leftrightarrow \varrho(x, y) = 0$ является отношением эквивалентности и мы будем отождествлять все слова одного и того же класса эквивалентности.

Лемма 2. $\varrho(x, y)$ *является метрикой на множестве всех слов длины* n.

Доказательство. Достаточно проверить неравенство треугольника. Шенноном была введена метрика $\eta(X, Y) = H(X/Y) + H(Y/X)$ на пространстве всех случайных переменных с конечной энтропией. Введенное нами расстояние $\varrho(x, y)$ между словами индуцировано метрикой Шеннона: каждой паре слов x, y мы ставим в соответствие распределение вероятностей $\{m_{ij}/n\}$ на $X \otimes Y$ с частными распределениями $\{m_i/n\}$ и $\{n_j/n\}$ на X и Y и метрика Шеннона $d(X, Y)$ для этих распределений как раз совпадает с расстоянием $\varrho(x, y)$.

Приведем для полноты доказательство справедливости неравенства треугольника для метрики Шеннона, следуя Y. Хорибе [2].

$$H(X/Y) + H(Y/Z) \geq H(X/Y, Z) + H(Y/Z) =$$
$$= H(X, Y/Z) \geq H(X/Z), \qquad \text{таким образом,}$$
$$H(X/Y) + H(Y/Z) \geq H(X/Z) \qquad \text{и, заменяя } X \text{ на } Z,$$
$$H(Y/X) + H(Z/Y) \geq H(Z/X).$$

Складывая эти два неравенства, получаем:

$$d(X, Y) + d(Y, Z) \geq d(X, Z).$$

Лемма 3.

$$\varrho(x, y) = \Phi(x \otimes y) - [\Phi(x) + \Phi(y)]/2;$$

$$\varrho(x, y) = [\Phi(x) + \Phi(y)]/2 - \Phi(x : y).$$

Лемма 4. Пусть $x_0 \in X^n$, $V_{x_0}^{d,w} = \{y \in Y^n : \varrho(x_0, y) \leq d, \Phi(y) \leq w\}$ *(обобщенный шар с центром в точке* x_0*). Тогда*

$$\log |V_{x_0}^{d,w}| \leq n\big(d + (w - \Phi(x_0))/2 + 0(1)\big).$$

Доказательство. Согласно лемме 3, для любого $y \in V_{x_0}^{d,w}$ выполняется неравенство

$$\Phi(x_0 \otimes y) \leq d + \big(w + \Phi(x_0)\big)/2. \tag{1}$$

Число пар (x, y), удовлетворяющих неравенству (1), оценивается с помощью леммы 1 статьи [1]:

$$\log N^{d,w} \leq n\big(d + (\Phi(x) + w)/2 + 0(1)\big).$$

Если x и x' имеют одну и ту же композицию, то x' получается из x в результате перестановки позиций букв, и если обозначить через B_x множество слов $y \in Y^n$, таких, что пара (x, y) удовлетворяет неравенству (1), то $B_x = B_{x'}$, поскольку $B_{x'}$ получается из B_x также в результате перестановки позиций. Число слов с одной и той же композицией, как известно, оценивается следующим образом [3]:

$$\log N_x \geq n\big(\Phi(x) - 0(1)\big).$$

Окончательно получаем:

$$\log |V_x^{d,w}| \leq \log N^{d,w} - \log N_x \leq n\big(d + (w - \Phi(x))/2 + 0(1)\big),$$

2. Под *кодом* будем понимать подмножество U пространства X^n. Введем следующие характеристики кода:

1) скорость передачи кода $R(U) = (\log |U|)/n$;

2) кодовое расстояние $d(U) = \min\limits_{x, y \in U} \varrho(x, y)$, $x \neq y$;

3) пропускная способность кода $C(U) = \max\limits_{x, y \in U} \Phi(x : y)$, $x \neq y$;

4) вес кода $w(U) = \operatorname{Max}\limits_{x \in U} \Phi(x)$.

Код, все слова которого имеют один и тот же вес Фитингофа, будем называть *кодом постоянного веса.*

Теорема 1. Существует код постоянного веса, для которого

$$R(U) \geq C(U) - 0(1) = w(U) - d(U) - 0(1)$$

для любого $C(U) > 0$.

Доказательство. Из определения $\varrho(x, y)$ следует, что для любого кода

$$d(U) \leq w(U),$$

а из леммы 3 следует, что для кода постоянного веса

$$d(U) = w(U) - C(U).$$

Для фиксированного $d > 0$ выберем произвольное $w \geq d$ и будем строить код постоянного веса с параметрами $d(U) = d$, $w(U) = w$, используя стандартный «метод выбрасывания» Гилберта. Число всех слов веса w не меньше величины

$$2^{n(w-0(1))},$$

а число выбрасываемых точек шара радиуса d не превышает величины

$$2^{n(d+0(1))}. \qquad \text{(лемма 4).}$$

Таким образом, существует код со скоростью

$$R(U) \geq w - d - 0(1) = C(U) - 0(1).$$

3. Введем *Φ-декодирование* для произвольного дискретного канала без памяти следующим образом: слово $z \in Y^n$ декодируется в кодовое слово x тогда и только тогда, когда $\Phi(x : z) > \Phi(x' : z)$ для любого другого кодового слова x'. Если не окажется кодового слова, удовлетворяющего этому условию, то скажем, что слово z приводит к отказу от декодирования, и такой отказ будем включать в ошибку декодирования.

Теорема 2. Для любого дискретного канала без памяти найдется последовательность кодов постоянного веса со скоростью передачи, как угодно близкой к пропускной способности канала и как угодно малой вероятностью ошибки Φ-декодирования.

Доказательство. Применим «классический» метод выбрасывания» Файнстейна. Пусть C — пропускная способность канала, $H(X)$ — энтропия максимизирующего распределения, $H(Y)$ — энтропия распределения на выходе, R — скорость передачи кода, $R < C$. Пусть $\alpha = (R + C)/2$ $M = \{(x, y) : x \in X^n, y \in Y^n, \Phi(x : y) > \alpha, \Phi(y) \in H(Y), \Phi(x) = H(X) + 0(1)\}$,

M_y — проекция M на Y^n,

M_x — проекция M на X^n.

Если $P(M_y) \geq 1 - \varepsilon$, $0 < \varepsilon < 1$, то, поскольку

$$P(M_y) = \sum_{x \in M_x} P(x)\, P(M_y/x),$$

найдется слово $x_1 \in M_x$ такое, что $p(M_y/x_1) \geq 1 - \varepsilon$. Выберем это слово в качестве первого кодового слова и выбрасываем множество $B_1 = \{y \in M_y : \Phi(x_y : y) > \alpha\}$ из M_y. Если теперь $p(M_y - B_1) \geq 1 - \varepsilon$, то найдется другое слово x_2, удовлетворяющее условию $p(M_y/x_2) \geq 1 - \varepsilon$. Процесс выбрасывания заканчивается тогда, когда

$$P\Big(M_y - M_y \cap \Big(\bigcup_{i=1}^{N} B_i\Big)\Big) < 1 - \varepsilon.$$

При этом

$$P(M_y) = P\Big(M_y \cap \Big(\bigcup_{i=1}^{N} B_i\Big)\Big) + P\Big(M_y - M_y \cap \Big(\bigcup_{i=1}^{N} B_i\Big)\Big) \leq \sum_{i=1}^{N} P(B_i) + 1 - \varepsilon. \quad (2)$$

Чтобы оценить $P(B_i)$, заметим, что, согласно лемме 3, для любой пары $(x, y) \in M$ выполняется неравенство

$$\varrho(x, y) < \big(H(X) + H(Y)\big)/2 - \alpha + 0(1),$$

а из леммы 4 следует, что

$$|B_i| \leq 2^{n(H(Y) - \alpha + 0(1))}.$$

Далее, при всех достаточно больших n все слова $y \in Y^n$, для которых $\Phi(y) \leq \leq H(Y)$, становятся равнораспределенными, а так как число этих слов не меньше величины $2^{n(H(Y) - 0(1))}$, то

$$P(y) \lesssim 2^{-n(H(Y) + 0(1))},$$

Поэтому

$$P(B_i) \lesssim P(y)\,|B_i| \lesssim 2^{-n(\alpha + 0(1))}. \quad (3)$$

Подставляя (3) в (2), получаем:

$$P(M_y) \leq N \cdot 2^{-n(\alpha + 0(1))} + 1 - \varepsilon.$$

Таким образом, для вероятности ошибки при всех достаточно больших n выполняется неравенство

$$P_e \leq N \cdot 2^{-n(\alpha + 0(1))} + P[\Phi(x : y) \leq \alpha],$$

и так как $\alpha = (R + C)/2$, а $N = 2^{nR}$, то

$$P_e \leq 2^{-n((C - R)/2 + 0(1))} + P[|\Phi(x : y| - C| \geq (C - R)/2]. \quad (4)$$

Поскольку $\Phi(x : y) \xrightarrow[n.B.]{} C$, $p_e \to 0$ с ростом n.

4. Величина $\Phi(x:y)$ в этой статье выступает как невероятностный аналог «взаимной информации без памяти». Эта величина обладает всеми свойствами шенноновской взаимной информации и оказывается полезной для декодирования. Введенное здесь Φ-декодирование не зависит от конкретного канала, т.е. является универсальным декодированием. Как уже отмечалось в работе [1], существование таких методов было известно давно, но конкретный универсальный алгоритм для дискретных каналов без памяти приводится, повидимому, впервые. Общее определение информации по Фитингофу будет проанализировано в дальнейшем.

Литература

1. *Гоппа В. Д.* Универсальное декодирование для симметричных каналов. Проблемы передачи информации. 1975, 11, 1.
2. *Horibe, Y.*, A Note on Entropy Metrics. Information and Control 22, 4, May 1973.
3. *Фано Р.* Передача информации. «Мир», М., 1964.

Nonprobabilistic mutual information without memory

V. D. GOPPA

(Moscow)

Let X, Y be finite alphabets, $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$, $x_i \in X$, $y_i \in Y$, $i = 1, 2, \ldots, n$. Construct the word $x \otimes y \in (X \otimes Y)^n$ as follows: $x \otimes y = ((x_1y_1), (x_2y_2), \ldots, (x_ny_n))$.

Define
$$\Phi(x:y) = \Phi(x) + \Phi(y) - \Phi(x \otimes y),$$
$$\Phi(y/x) = \Phi(x \otimes y) - \Phi(x),$$

where $\Phi(x)$ is the Fitingoff weight of x. The following of their properties are readily verified: (a) $\Phi(x:y) = \Phi(y:x)$; (b) $\Phi(x:x) = \Phi(x)$; (c) $\Phi(x:y) = \Phi(x) - \Phi(x/y) = \Phi(y) - \Phi(y/x)$.

The subset U of X^n will be called a code. Let us introduce now the two parameters of a code: 1) code transmission rate $R(U) = (\log|U|)/n$; 2) code capacity $C(U) = \max \Phi(x:y)$, $x \neq y$, $x, y \in U$.

Theorem 1. There exists a code, such that $R(U) > C(U) - 0(1)$ for any $C(U) > 0$. Define now Φ-decoding as follows: any output word z is decoded into the code word x if and only if $\Phi(x:z) > \Phi(x':z)$ for all code words $x' \neq x$.

Theorem 2. For any discrete channel without memory there exists a sequence of the codes, the rate of a code being close to the channel capacity and probability of error of Φ-decoding being vanishingly small.

Notice, that Φ-decoding does not depend on the channel (universal decoding).

В. Д. Гоппа
Институт проблем передачи информации
СССР, Москва Е-24,
Авиамоторная ул., 8а, корп. 2

Problems of Control and Information Theory, Vol. 4 (2), pp. 103—120 (1975)

METHOD FOR DESIGN OF LINEAR OPTIMAL SYSTEMS IN FREQUENCY DOMAIN

F. CSÁKI, EL-SAID HAMADA

(Budapest)

It is shown by an extension of optimality condition studied extensively in the literature [1] (Kalman 1964) that there are certain relationships between the response of optimal system, the elements of the weighting vector of the reduced performance index and the optimal feedback matrix. Two algorithms are given on the basis of these relationships. The first calculates the weighting vector of the reduced performance index from the desired optimal response, the other calculates the corresponding optimal feedback vector.

The designer procedure specifies the performance index for the optimal response and the appropriate law of optimality can be calculated as a routine work. An example is given to illustrate the design technique.

1. Introduction

The theory of optimal linear multivariable systems has been well developed and several methods exist for calculating optimal feedback in order to minimize chosen performance index. These methods are based on the solution of the Riccati matrix equation. One of the difficultes is, however, the choice of such a performance index which ensures the desired system response. Conventionally, an initial criterion is chosen, the optimal feedback is calculated, then the system response is obtained by simulation. Usually, the response obtained in this way is unsatisfactory and so the performance index has to be modified by trial and error or experience and this procedure must be repeated until a satisfactory response is obtained. The present paper contributes to this design aspect by using the relationship between the performance index, the weighting matrices and the system response. Thus the weighting matrices become design parameters specified by the desired system response.

The procedure contains two algorithms working in the frequency domain for computing the performance index and the optimal feedback vector.

This solution requires shorter computer time, less memory and gives better results. Presently our procedure is elaborated only for single variable systems, but its generalization for multivariable systems is possible.

2. Preliminaries in the theory

2.1. *System equations in the frequency domain*

In our paper linear, time invariant, finite dimensional systems are considered which can be described by state equations

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}; \tag{1}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x}; \tag{2}$$

where

$\mathbf{x}$ is the $(n \times 1)$ state vector,

$\mathbf{u}$ is the $(r \times 1)$ control vector,

$\mathbf{y}$ is the $(m \times 1)$ output vector;

$\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ are constant matrices of appropriate dimensions.

The open-loop system described in Eqs. (1) and (2) is shown in Fig. 1.

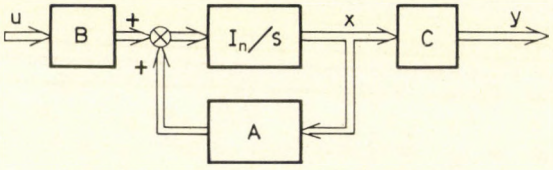


Fig. 1. Open-loop system

The transfer-function matrix of the open-loop system is given by

$$\mathbf{W}_0(s) = \mathbf{C}[s\mathbf{I}_n - \mathbf{A}]^{-1}\mathbf{B}, \tag{3}$$

where $\mathbf{I}_n$ is the $(n \times n)$ identity matrix.

In this equation $[s\mathbf{I} - \mathbf{A}]^{-1}$ can be expressed as

$$\frac{\operatorname{adj}[s\mathbf{I}_n - \mathbf{E}]}{\det[s\mathbf{I}_n - \mathbf{A}]} = \boldsymbol{\Phi}_0(s). \tag{4}$$

Then the transfer-function matrix becomes

$$\mathbf{W}_0(s) = \mathbf{C}\,\boldsymbol{\Phi}_0(s)\,\mathbf{B}. \tag{5}$$

If state-variable feedback is applied now then the input $\mathbf{u}$ of the system may be expressed in the form

$$\mathbf{u} = \mathbf{r} - \mathbf{K}^T\mathbf{x}, \tag{6}$$

where $\mathbf{K}^T$ is a constant $(r \times n)$ matrix.

The state equations of the closed-loop system become

$$\dot{\mathbf{x}} = [\mathbf{A} - \mathbf{B}\mathbf{K}^T]\,\mathbf{x} + \mathbf{B}\mathbf{r}; \tag{7}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x}. \tag{8}$$

The configuration of the resulting system is shown in Fig. 2.

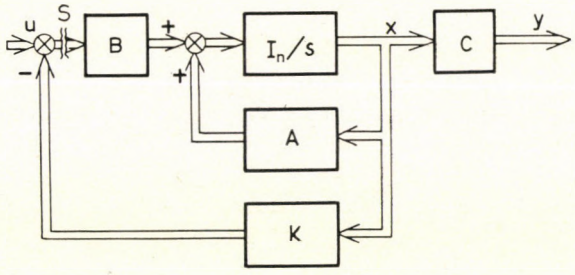


Fig. 2. System with state-variable feedback

Using the Laplace transformation, Eqs. (7) and (8) can be written as

$$s\mathbf{x}(s) = [\mathbf{A} - \mathbf{B}\mathbf{K}^T]\,\mathbf{x}(s) + \mathbf{B}\mathbf{r}(s) \tag{9}$$

$$\mathbf{y}(s) = \mathbf{C}\mathbf{x}(s) \tag{10}$$

where $\mathbf{x}(s)$, $\mathbf{r}(s)$ and $\mathbf{y}(s)$ are Laplace transformed vectors of $\mathbf{x}$, $\mathbf{r}$ and $\mathbf{y}$, respectively.

Equations (9) and (10) may be written in the form:

$$\mathbf{x}(s) = [s\mathbf{I}_n - \mathbf{A} + \mathbf{B}\mathbf{K}^T]^{-1}\,\mathbf{B}\mathbf{r}(s); \tag{11}$$

$$\mathbf{y}(s) = \mathbf{C}[s\mathbf{I}_n - \mathbf{A} + \mathbf{B}\mathbf{K}^T]^{-1}\,\mathbf{B}\mathbf{r}(s). \tag{12}$$

The transfer-function matrix of closed-loop system relating $\mathbf{r}(s)$ to $\mathbf{y}(s)$ is given by

$$\mathbf{W}_c(s) = \mathbf{C}\boldsymbol{\Phi}_c(s)\,\mathbf{B} \tag{13}$$

where

$$\boldsymbol{\Phi}_c(s) = [s\mathbf{I}_n - \mathbf{A} + \mathbf{B}\mathbf{K}^T]^{-1}.$$

The closed-loop transfer-function matrix (13) and the open-loop transfer-function matrix (5) are related by the return difference matrix $\mathbf{W}(s)$ through

$$\mathbf{W}_c(s) = \mathbf{W}_0(s)\;\mathbf{W}^{-1}(s) = \mathbf{C}(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}$$
$$[\mathbf{I}_r + \mathbf{K}^T(s\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{B}]^{-1} \tag{14}$$

where

$$\mathbf{W}(s) = \mathbf{I}_r + \mathbf{K}^T\boldsymbol{\Phi}_0(s)\,\mathbf{B}. \tag{15}$$

The matrix $\mathbf{I}_r + \mathbf{K}^T \mathbf{\Phi}_0(s)\, \mathbf{B}$ is a return difference matrix with the loop broken at point S of Fig. 2.

It can be shown [1, 2, 3, 4] that

$$\det [\mathbf{I}_r + \mathbf{K}^T \mathbf{\Phi}_0(s)\, \mathbf{B}] = \frac{\det [s\mathbf{I}_n - \mathbf{A} + \mathbf{B}\mathbf{K}^T]}{\det [s\mathbf{I} - \mathbf{A}]} =$$

$$= \frac{\text{characteristic polynomial of closed loop system}}{\text{characteristic polynomial of open-loop system}}. \tag{16}$$

2.2. *Optimality condition in frequency domain*

The problem of optimality may simply be formulated in the following way: find such a vector $\mathbf{u}$ which minimizes the quadratic performance criterion

$$PI = \int_0^\infty (\mathbf{x}^T \mathbf{Q}\, \mathbf{x} + \mathbf{u}^T \mathbf{P}\mathbf{u})\, dt\,. \tag{17}$$

where $\mathbf{Q}$ and $\mathbf{P}$ are symmetric constant weighting matrices. In order to get a well-defined solution, the integrand in (17) must be strictly concave regarding $\mathbf{u}$ and this requires $\mathbf{Q}$ to be positive semidefinite and $\mathbf{P}$ to be positive definite. The well-known advantages of such a quadratic criterion are that the computation are relative easy to perform, the obtained optimal system is stable, and the feedback matrix $\mathbf{K}$ is time invariant. As is well known, the problem of optimality can be solved by time-domain analysis, i.e. as a solution of the so-called reduced Riccati equation

$$\mathbf{A}^T\mathbf{R} + \mathbf{R}\mathbf{A} - \mathbf{R}\mathbf{B}\mathbf{P}^{-1}\mathbf{B}^T\mathbf{R} + \mathbf{Q} = \mathbf{0}. \tag{18}$$

Having the symmetric solution matrix $\mathbf{R}$ the optimal feedback control is then

$$\mathbf{u}^0(\mathbf{x}) = \mathbf{r} - \mathbf{P}^{-1}\mathbf{B}^T\, \mathbf{R}\mathbf{x} = \mathbf{r} - \mathbf{K}^T\mathbf{x} \tag{19}$$

where

$$\mathbf{K} = \mathbf{R}\mathbf{B}\mathbf{P}^{-1}. \tag{20}$$

(The problem can also be solved in the time domain by using dynamic programming or other methods.)

Recently great interest has arised in frequency-domain methods [4–10].

Using the frequency-domain approach, the condition of optimality regarding the minimization of a quadratic performance index (17) for linear, time-invariant systems (1), (2) can be expressed in the following way.

The state control law is optimal if, and only if, the product of the conjugate transpose of the corresponding return-difference matrix and itself minus the identity matrix is a positive definite matrix for all frequencies.

More precisely, as it was shown in [5, 7, 8] the necessary and sufficient condition of the optimality in frequencydomain can be expressed as [4, 5]:

$$\mathbf{W}^T(-s)\,\mathbf{W}(s) = \mathbf{I}_r + \mathbf{D}^{-1}\mathbf{B}_r\,\mathbf{\Phi}_0^T(-s)\,\mathbf{G}\mathbf{G}^T\mathbf{\Phi}_0(s)\,\mathbf{B}\mathbf{D}^{-1}; \tag{21}$$

where $\mathbf{W}(s)$ have been defined in (15) whereas $\mathbf{G}\mathbf{G}^T = \mathbf{Q}$ and $\mathbf{D}\mathbf{D}^T = \mathbf{P}$. The matrices $\mathbf{P}$ and $\mathbf{Q}$ can be always written in these forms, since $\mathbf{P}$ and $\mathbf{Q}$ are positive definite and semidefinite, respectively. In addition, since $\mathbf{P}$ is always positive definite, $\mathbf{D}$ can always be chosen to be symmetric and positive definite. Therefore $\mathbf{P}$ may also be written as: $\mathbf{P} = \mathbf{D}\mathbf{D}$ and $\mathbf{P}^{-1}$ given by $\mathbf{P}^{-1} = \mathbf{D}^{-1}\mathbf{D}^{-1}$ where $\mathbf{D}^{-1}$ exists, since $\mathbf{D}$ is positive definite.

2.3. *Single-variable case*

For the special case of single-variable systems certain simplifications can be made regarding the condition of optimality, i.e. the feedback is optimal if and only if the absolute value of the return difference is at least unity for all frequencies.

In this particular case of single-variable systems, the necessary and sufficient condition of optimality, that is, Eq. (21) can be written in scalar equation form as [7]:

$$\frac{\Delta_c(-s)\,\Delta_c(s)}{\Delta_0(-s)\,\Delta_0(s)} = 1 + \frac{1}{\varrho^2}\,[\mathbf{b}^T\,\mathbf{\Phi}_0^T(-s)\,\mathbf{g}]\cdot[\mathbf{g}^T\,\mathbf{\Phi}_0(s)\,\mathbf{b}]\,, \tag{22}$$

where $\Delta_c(s) = \det[W_c(s)] = \det[s\mathbf{I}_n - \mathbf{A} + \mathbf{b}\mathbf{k}^T]$ is the characteristic polynomial of the closed-loop optimal system, $\Delta_0(s) = \det[s\mathbf{I}_n - A]$ is the characteristic polynomial of the open-loop system, $\mathbf{g}^T$ is the weighting vector and ϱ is a weighting scalar. It has been proved by Wonham and Johnson [7, 8] that the minimization of the performance index in Eq. (17) is equivalent with the minimization of the reduced performance index

$$PI = \int_0^\infty [(\mathbf{g}^T\mathbf{x})^2 + \varrho u^2]\,dt. \tag{23}$$

It is clear now that there is no loss of generality for single-variable case, in writing a column matrix $\mathbf{g}$ instead of matrix $\mathbf{G}$ in both the condition of optimality (22) and the performance index (23).

2.4. *The relationship between the system response and the performance index*

The response of any dynamic system is determined by the location of its poles and also by the location of its zeros. Using time-domain methods there is no explicit relationship between the poles of the system and the

elements of **Q** and **P**. If determinants are first taken of both sides of Eq. (21) and using the fact that the determinant of the product is the product of the determinants, then we get

$$\frac{\det\left[-s\mathbf{I}_n - \mathbf{A} + \mathbf{B}\mathbf{K}^T\right]^T}{\det\left[-s\mathbf{I}_n - A^T\right]} \cdot \frac{\det\left[s\mathbf{I}_n - \mathbf{A} + \mathbf{B}\mathbf{K}^T\right]}{\det\left[s\mathbf{I}_n - \mathbf{A}\right]} =$$

$$= \det\left[\mathbf{I}_r + \mathbf{D}^{-1}\mathbf{B}^T\mathbf{\Phi}^T(-s)\mathbf{G}\mathbf{G}^T\mathbf{\Phi}_0(s)\,\mathbf{B}\mathbf{D}^{-1}\right] \tag{24}$$

where $\det\left[s\mathbf{I}_n - \mathbf{A} + \mathbf{B}\mathbf{K}^T\right]$ and $\det\left[s\mathbf{I}_n - \mathbf{A}\right]$ are the closed-loop and open-loop characteristic polynomials respectively.

Expression (24) clearly gives the characteristic equation of the optimal system as an explicit function of the elements of $\mathbf{G}\mathbf{G}^T$ and $\mathbf{D}\mathbf{D}$. For the single-variable case, the relationship between the system response and the performance index, as given in Eq. (24) reduces to Eq. (22).

3. Design technique for single-variable systems

Equation (22) gives an explicit and computable relationship between the characteristic polynomial of the optimal system and its performance index. Assuming in the continuation $s = j\omega$, Eq. (22) can be written in the following form:

$$\left|\frac{\Delta_c(s)}{\Delta_0(s)}\right|^2 = 1 + \frac{1}{\varrho^2}\left|\mathbf{g}^T\mathbf{\Phi}_0(s)\,\mathbf{b}\right|^2. \tag{25}$$

From Eq. (15) we obtain

$$\left|\frac{\Delta_c(s)}{\Delta_0(s)}\right|^2 = \left|1 + \mathbf{k}^T\mathbf{\Phi}_0(s)\,\mathbf{b}\right|^2 \tag{26}$$

that is,

$$\left|1 + \mathbf{k}^T\mathbf{\Phi}_0(s)\,\mathbf{b}\right|^2 = 1 + \frac{1}{\varrho^2}\left|\mathbf{g}^T\mathbf{\Phi}_0(s)\,\mathbf{b}\right|^2. \tag{27}$$

If the closed-loop transfer function is specified then $\Delta_c(s)$ is known, the weighting vector $\mathbf{g}^T$ can be calculated by using Eq. (25) and taking $\varrho = 1$, and hence PI is specified. On the other hand, if the weighting vector $\mathbf{g}^T$ and the weighting scalar ϱ are known and using Eq. (27), the optimal feedback vector $\mathbf{k}^0$ and hence from Eqs. (14) and (15) the optimal closed loop transfer function may be determined:

$$W_c(s) = \frac{\mathbf{c}^T\mathbf{\Phi}_0(s)\,\mathbf{b}}{1 + \mathbf{k}^T\mathbf{\Phi}_0(s)\,\mathbf{b}}. \tag{28}$$

3.1. *An algorithm for computation of vector* $\mathbf{g}^T$

By choice $\varrho = 1$ and substitution $s = j\omega$ from Eq. (25) we get

$$\left|\frac{\Delta_c(j\omega)}{\Delta_0(j\omega)}\right|^2 = 1 + |\,\mathbf{g}^T\,\mathbf{\Phi}_0\,(j\omega)\mathbf{b}\,|^2 \tag{29}$$

or

$$\left|\frac{\Delta_c(j\omega)}{\Delta_0(j\omega)}\right|^2 = 1 + \left|\mathbf{g}^T\,\frac{L(j\omega)}{\Delta_0(j\omega)}\right|^2, \tag{30}$$

where

$$L(jw) = \text{adj}\,[jw\mathbf{I} - \mathbf{A}]\,\mathbf{b}. \tag{31}$$

Taking the real and imaginary parts in both sides of Eq. (30), by writing $\mathbf{\Phi}_0(j\omega)\,\mathbf{b} = \mathbf{h}(\omega) + j\mathbf{m}(\omega)$; we get (see the Appendix):

$$|\,e(\omega) + jf(\omega)\,|^2 = 1 + |\,\mathbf{g}^T[\mathbf{h}(\omega) + j\mathbf{m}(\omega)]\,|^2. \tag{32}$$

Hence

$$e^2(\omega) + f^2(\omega) = 1 + [\mathbf{g}^T\mathbf{h}(\omega)]^2 + [\mathbf{g}^T\mathbf{m}(\omega)]^2 \tag{33}$$

taking $T(\omega) = e^2(\omega) + f^2(\omega)$, and defining

$$T_1(\omega) = T(\omega) - 1 \tag{34}$$

then

$$T_1(\omega) = \mathbf{g}^T\,[\mathbf{h}(\omega)\,\mathbf{h}^T(\omega) + \mathbf{m}(\omega)\,\mathbf{m}^T(\omega)]\,\mathbf{g} \tag{35}$$

or

$$T_1(\omega) = \mathbf{g}^T\mathbf{H}(\omega)\,\mathbf{g} \tag{36}$$

where

$$\mathbf{H}(\omega) = \mathbf{h}(\omega)\,\mathbf{h}^T(\omega) + \mathbf{m}(\omega)\,\mathbf{m}^T(\omega). \tag{37}$$

By choice of a suitable sequence $\{\omega_i\}_{i=1}^n$, Eq. (36) can be written as:

$$T_1(\omega_i) = \mathbf{g}^T\,\mathbf{H}(\omega_i)\,\mathbf{g} \quad (i = 1, 2, \ldots, n). \tag{38}$$

Introducing the vector

$$\mathbf{V}(\omega) = \begin{bmatrix} T_1(\omega_1) \\ T_1(\omega_2) \\ \cdot \\ \cdot \\ \cdot \\ T_1(\omega_n) \end{bmatrix}; \tag{39}$$

and the matrix

$$\mathbf{D}_1(\mathbf{g}^T) = \begin{bmatrix} \mathbf{g}^T\,\mathbf{H}(\omega_1) \\ \mathbf{g}^T\,\mathbf{H}(\omega_2) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{g}^T\,\mathbf{H}(\omega_n) \end{bmatrix} \tag{40}$$

we get

$$\mathbf{V}(\omega) = \mathbf{D}_1(\mathbf{g}^T)\,\mathbf{g}$$

which can be solved by successive approximation method. Taking $\mathbf{g}_0$ as an initial vector the iteration equation

$$\mathbf{g}_{i+1} = [\mathbf{D}_1(\mathbf{g}_i^T)]^{-1}\,\mathbf{V}(\omega)$$

is used until a given accuracy.

3.2. *An algorithm for the calculation of the optimal feedback vector* $\mathbf{k}^0$

If the aforementioned performance index or more precisely if vector $\mathbf{g}$ and scalar ϱ are used in Eq. (27), then the optimal feedback vector $\mathbf{k}^0$ and hence the optimal law $u^0 = -\mathbf{k}^{0T}\mathbf{x}$ can be determined. Equation (27) with $\varrho = 1$ is repeated here for convenience:

$$|\,1 + \mathbf{k}^T\,\boldsymbol{\Phi}_0(j\omega)\,\mathbf{b}\,|^2 = 1 + |\,\mathbf{g}^T\boldsymbol{\Phi}_0(j\omega)\,\mathbf{b}\,|^2. \tag{41}$$

The first step is to rewrite Eq. (41) in the following form:

$$|\,1 + \mathbf{k}^T[\mathbf{h}(\omega) + j\mathbf{m}(\omega)]\,|^2 = 1 + |\,\mathbf{g}^T[\mathbf{h}(\omega) + j\mathbf{m}(\omega)]\,|^2 \tag{42}$$

where

$$\boldsymbol{\Phi}(j\omega)\,\mathbf{b} = \mathbf{h}(\omega) + j\mathbf{m}(\omega)$$

and then

$$|\,1 + \mathbf{k}^T\mathbf{h}(\omega) + j\mathbf{k}^T\mathbf{m}(\omega)\,|^2 = 1 + |\,\mathbf{g}^T\mathbf{h}(\omega) + j\mathbf{g}^T\,\mathbf{m}(\omega)\,|^2 \tag{43}$$

from which

$$1 + 2\mathbf{k}^T\,\mathbf{h}(\omega) + \mathbf{k}^T\,\mathbf{h}(\omega)\,\mathbf{h}^T(\omega)\,\mathbf{k} + \mathbf{k}^T\,\mathbf{m}(\omega)\,\mathbf{m}^T(\omega)\,\mathbf{k} =$$

$$= 1 + \mathbf{g}^T\,\mathbf{h}(\omega)\,\mathbf{h}^T(\omega)\,\mathbf{g} + \mathbf{g}^T\,\mathbf{m}(\omega)\,\mathbf{m}^T(\omega)\,\mathbf{g}; \tag{44}$$

and

$$2\mathbf{k}^T\mathbf{h}(\omega) + \mathbf{k}^T\,[\mathbf{h}(\omega)\,\mathbf{h}^T(\omega) + \mathbf{m}(\omega)\,\mathbf{m}^T(\omega)]\,\mathbf{k} =$$

$$= \mathbf{g}^T\,[\mathbf{h}(\omega)\,\mathbf{h}^T(\omega) + \mathbf{m}(\omega)\,\mathbf{m}^T(\omega)]\,\mathbf{g}. \tag{45}$$

By defining $\mathbf{F}(\omega) = \mathbf{h}(\omega)\,\mathbf{h}^T(\omega) + \mathbf{m}(\omega)\,\mathbf{m}^T(\omega)$, Eq. (45) will be

$$2\mathbf{k}^T\mathbf{h}(\omega) + \mathbf{k}^T\mathbf{F}(\omega)\,\mathbf{k} = \mathbf{g}^T\mathbf{F}(\omega)\,\mathbf{g}. \tag{46}$$

If in Eq. (46) a suitable sequence $\{\omega_i\}_{i=1}^n$ is chosen with not too high positive values, i.e. a sequence $(\omega_1, \omega_2, \ldots, \omega_n)$ which satisfies (46), then for each ω_i $(i = 1, 2, \ldots, n)$, $\mathbf{h}(\omega_i)$ and $\mathbf{F}(\omega_i)$ may be computed. From Eq. (46) the following simultaneous equations are obtained:

$$2\mathbf{k}^T\mathbf{h}(\omega_i) + \mathbf{k}^T\mathbf{F}(\omega_i)\,\mathbf{k} = \mathbf{g}^T\mathbf{F}(\omega_i)\,\mathbf{g}; \tag{47}$$
$$(i = 1, 2, \ldots, n)$$

where $\mathbf{F}(\omega)$ is a symmetrical matrix, i.e. $\mathbf{F}^T(\omega) = \mathbf{F}(\omega)$. After some algebraic manipulation

$$[2\mathbf{h}^T(\omega_i) + \mathbf{k}^T\mathbf{F}(\omega_i)]\,\mathbf{k} = \mathbf{g}^T\mathbf{F}(\omega_i)\,\mathbf{g}; \tag{48}$$
$$(i = 1, 2, \ldots, n).$$

These latter equations may also be written as:

$$[2\hat{\mathbf{H}} + \hat{\mathbf{Q}}(\mathbf{k}^T)]\,\mathbf{k} = \hat{\mathbf{g}} \tag{49}$$

where

$$\hat{\mathbf{H}} = [\mathbf{h}(\omega_1), \mathbf{h}(\omega_2), \ldots, \mathbf{h}(\omega_n)]^T;$$

$$\hat{\mathbf{Q}} = [\mathbf{F}^T(\omega_1)\,\mathbf{k}, \mathbf{F}^T(\omega_2)\,\mathbf{k}, \ldots, \mathbf{F}^T(\omega_n)\,\mathbf{k}]^T; \tag{50}$$

$$\hat{\mathbf{g}} = [\mathbf{g}^T\mathbf{F}(\omega_1)\,\mathbf{g}, \mathbf{g}^T\mathbf{F}(\omega_2)\,\mathbf{g}, \ldots, \mathbf{g}^T\mathbf{F}(\omega_n)\,\mathbf{g}]^T. \tag{51}$$

Defining $\hat{\mathbf{D}}(\mathbf{k}^T) = 2\hat{\mathbf{H}} + \hat{\mathbf{Q}}(\mathbf{k}^T)$; then (49) becomes

$$\hat{\mathbf{D}}(\mathbf{k}^T)\,\mathbf{k} = \hat{\mathbf{g}}. \tag{52}$$

This equation may be written as:

$$\mathbf{k}_{j+1} = [\hat{\mathbf{D}}(\mathbf{k}_j^T)]^{-1}\hat{\mathbf{g}} \qquad (j = 1, 2, \ldots) \tag{53}$$

and is solved by successive approximation method, i.e. starting by a suitable initial vector $\mathbf{k}_0$, after some iterations we get the final optimal feedback vector $\mathbf{k}^0$.

4. Sufficient condition for the convergency of iteration

When the system of equation

$$\mathbf{k}_{j+1} = [\hat{\mathbf{D}}(\mathbf{k}_j^T)]^{-1}\hat{\mathbf{g}}; \qquad (j = 1, 2, \ldots) \tag{54}$$

is solved by the successive approximation method then the sufficient condition of the convergency is ensured by the following two theorems [11]: Con-

sider a sequence of $\mathbf{k}_j$ ($j = 0, 1, 2, \ldots$) where $\mathbf{k}_0 \in G_0$ and G_0 is the domain of definition for the vector function $[\hat{\mathbf{D}}(\mathbf{k}_j^T)]^{-1}\hat{\mathbf{g}}$.

Theorem 1. If iteration (54) remains in the closed domain G_0 and for any two points condition

$$\| [\hat{\mathbf{D}}(\mathbf{k}_1)]^{-1}\hat{\mathbf{g}} - [\hat{\mathbf{D}}(\mathbf{k}_2)]^{-1}\hat{\mathbf{g}} \| \leq \varrho_1 \| \mathbf{k}_1 - \mathbf{k}_2 \|,$$
$$0 \leq \varrho_1 < 1 \tag{55}$$

s satisfied, then for any initial vector $\mathbf{k}_0$

(1) the sequence converges, that is, solution $\mathbf{k}^0 = \operatorname{Lim}_{j\to\infty} \mathbf{k}_j$ exists,

(2) solution $\mathbf{k}^0$ is unique in the domain G_0,

(3)
$$\left\| \mathbf{k}^0 - \mathbf{k}_j \right\| \leq \frac{\varrho_1^j}{1 - \varrho_1} \left\| \mathbf{k}_1 - \mathbf{k}_0 \right\|. \tag{56}$$

Theorem 2. If $[\hat{\mathbf{D}}(\mathbf{k})]^{-1}\hat{\mathbf{g}}$ and $\{[\hat{\mathbf{D}}(\mathbf{k})]^{-1}\hat{\mathbf{g}}\}'$ are continuous in the closed domain G_0, where prime means derivation with respect to $\mathbf{k}$, and $\mathbf{k}_{j+1} = [\hat{\mathbf{D}}(\mathbf{k}_j)]^{-1}\hat{\mathbf{g}}$ falls into the domain G_0, then the sufficient condition of the convergency is

$$\| \{[\hat{\mathbf{D}}(\mathbf{k})]^{-1}\hat{\mathbf{g}}\}' \| \leq \varrho_1 < 1, \tag{57}$$

where ϱ_1 is an arbitrary constant. In this case the unequality gives the unique solution of the system of equations in domain G_0. The system of equations (54) may be written as

$$\hat{\mathbf{D}}(\mathbf{k})\,\mathbf{k} - \hat{\mathbf{g}} = 0; \tag{58}$$

from which

$$\mathbf{k} = \mathbf{k} + \mathbf{M}[\hat{\mathbf{D}}(\mathbf{k})\,\mathbf{k} - \hat{\mathbf{g}}] \tag{59}$$

where $\mathbf{M}$ is an arbitrary matrix.

Then

$$[\hat{\mathbf{D}}(\mathbf{k})]^{-1}\hat{\mathbf{g}} = \mathbf{k} + \mathbf{M}[\hat{\mathbf{D}}(\mathbf{k})\,\mathbf{k} - \hat{\mathbf{g}}] \tag{60}$$

and

$$\{[\hat{\mathbf{D}}(\mathbf{k})]^{-1}\mathbf{g}\}' = I + \mathbf{M}[\hat{\mathbf{D}}(\mathbf{k})]. \tag{61}$$

Applying Theorem 2 matrix $\mathbf{M}$ must be chosen in such a way that equation

$$\| \{[\hat{\mathbf{D}}(\mathbf{k}_0)]^{-1}\hat{\mathbf{g}}\}' \| = \mathbf{0} \tag{62}$$

should be satisfied.

This means that $\mathbf{I} + \mathbf{M}[\hat{\mathbf{D}}(\mathbf{k}_0)]$ have to be equal to the zero matrix.

If $\{\hat{\mathbf{D}}(\mathbf{k}_0)\,\mathbf{k}_0 - \hat{\mathbf{g}}\}'$ is nonsingular by choice $\mathbf{k}_0$, then taking $\mathbf{M} = -[\hat{\mathbf{D}}(\mathbf{k}_0)]^{-1}$. We can write

$$\mathbf{k} = \mathbf{k}_0 - [\hat{\mathbf{D}}\mathbf{k}_0)]^{-1}\,\hat{\mathbf{D}}(\mathbf{k}_0)\,\mathbf{k}_0 + \hat{\mathbf{D}}^{-1}(\mathbf{k}_0)\,\hat{\mathbf{g}} \tag{63}$$

from which

$$\mathbf{k}_{j+1} = [\hat{\mathbf{D}}(\mathbf{k}_j)]^{-1}\,\hat{\mathbf{g}} \tag{64}$$

which meets the convergency condition.

If $\{\hat{\mathbf{D}}(\mathbf{k}_0)\,\mathbf{k}_0 - \hat{\mathbf{g}}\}'$ would be singular, then other vector $\mathbf{k}_0$ has to be chosen.

5. Examples

In order to demonstrate this technique, two examples are considered.

5.1. Let us consider that the desired optimal transfer function of the Ward–Leonard speed control system shown in Fig. 3 is the following

$$\frac{Y(s)}{R(s)} = \frac{1798.47}{s^4 + 13.66\,s^3 + 143.82\,s^2 + 742.93\,s + 1798.47} \tag{E 1}.$$

Field resistance $R_f = 75$ ohms, field inductance $L_f = 9$ henries, rotational inductance $G_{fa} = 0.6$ henry, armature resistance $R_a = 0.15$ ohm, armature inductance is negligible, inertia of motor plus load $J_m = 5.0$ kg m^2, winding coefficient of motor plus load $D_m = 0.07$ newton ms/rad.

The amplidyne parameters: control field resistance $R_{af} = 1200$ ohms, control field inductance $L_{af} = 200$ henries, control field rotational inductance G_a^{qf} *a* 1.5 henries, quadrative winding resistance $R_a^q = 1.0$ ohm, quadrature winding inductance $L_a^q = 0.6$ henry, quadrature winding rotational

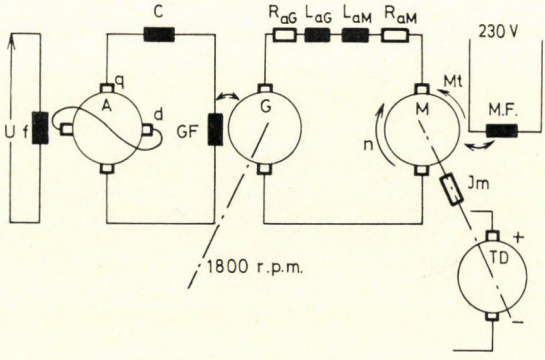


Fig. 3. Ward—Leonard speed control

inductance $G_a^{dq} = 0.26$ henry, total armature resistance $R_a^d + R_a^c = 2.0$ ohms, assume unity coupling between the C and d windings.

Tachometer: Gain $K_{T0} = 1$ volt sec/rad, the transfer function at the open-loop system is:

$$\frac{\Omega(S)}{U(S)} = \frac{1798.47}{s^4 + 18.467\,s^3 + 112.2\,s^2 + 256.6\,s + 193.8}. \tag{E 2}$$

The phase-variable equation of the system is:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -193.8 & -256.6 & -112.2 & -18.467 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1798.47 \end{bmatrix} u; \tag{E 3}$$

$$\mathbf{y} = [1, 0, 0, 0]^T \mathbf{x}.$$

According to point (3.1) and by writing the data for REPCO program, which is prepared to the computation of the weighting vector $\mathbf{g}$ of the reduced performance index from the optimal transfer function Eq. (E 1) and the phase-variable representation Eq. (E 3).

Taking the initial value of vector $\mathbf{g}$ that is, $\mathbf{g}_0 = [1, 1.75, 2.15, 1]^T$; and sequence in Eq. (38) as $\omega = \{0.8, 1, 1.5, 2\}$, furthermore, $\varrho = 1$, the result of the first stage is:

$$\mathbf{g}^0 = [0.99418, 1.032, 0.1757, 0.02]^T.$$

This result was obtained by eight steps of iteration. According to point (3.2) and by introducing the data for LISIKFID program which is written to compute the constant optimal feedback vector $\mathbf{k}^0$ from the phase-variable equations of the system in Eq. (E 3) and by knowing vector $\mathbf{g}^0$, the vector $\mathbf{k}^0$

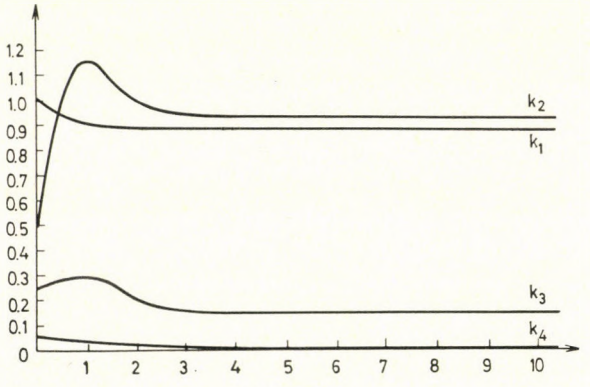


Fig. 4

is obtained. Taking the initial value of vector $\mathbf{k}$ that is, $\mathbf{k}_0 = [1, 0.5, 0.25, 0.05]^T$; and the same values of sequence $\{\omega\}$, we get the following results:

$$\mathbf{k}^0 = [0.892245, 0.928466, 0.149994, 0.0156585]^T.$$

This result is obtained by six steps of iteration (see Fig. 4).

The response of the system is shown in Fig. 5 with $\varrho = 1$. In order to compare the results of the frequency-domain direct method with those obtained

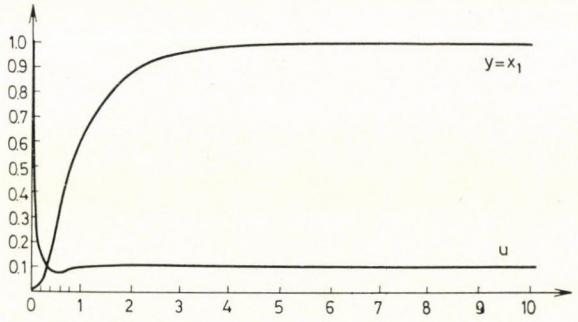


Fig. 5a

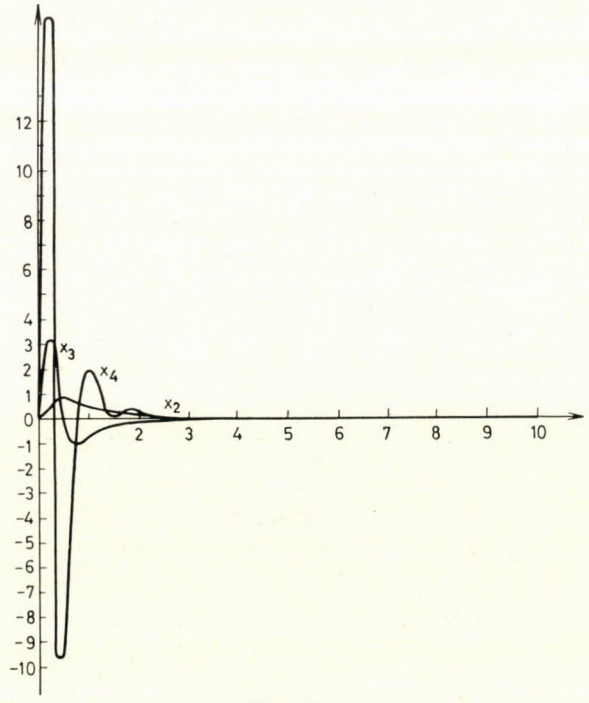


Fig. 5b

by the previous methods, the feedback vector $\mathbf{k}^0$ is calculated by the solution of Riccati matrix equation (18). The Riccati equation is solved by considering the control period T infinite, $\mathbf{R}(\tau)$ constant for finite values of t and is found by integration with respect to τ until the constant $\mathbf{R}(\infty)$ is reached. It is known that the plant input is given by the equation:

$$U(t) = \frac{1}{\varrho} \mathbf{b}^T \mathbf{R}(\tau)\, \mathbf{x}(t) = \mathbf{k}^T(\tau)\, \mathbf{x}(t)$$

where

$$\tau = T - t.$$

Equation (18) can be written in the form:

$$\dot{\mathbf{R}}(\tau) = \mathbf{R}(\tau)\, \mathbf{A} + \mathbf{A}^T \mathbf{R}(\tau) - \frac{1}{\varrho} \mathbf{R}(\tau)\, \mathbf{b}\mathbf{b}^T \mathbf{R}(\tau) + \mathbf{g}\mathbf{g}^T$$

and with the boundary condition $\mathbf{R}(0) = \mathbf{0}$.

By using the weighting vector $\mathbf{g}$ of the reduced *PI* which was obtained in the first stage of the frequency-domain technique, approximately the same results are obtained using the SRESP program. The obtained optimal vector is:

$$\mathbf{k}^0 = [0.8922,\ 0.9382,\ 0.1483,\ 0.01720]^T.$$

The system response was simulated with both results. The results of simulation are given in Fig. 5. At the 10th second the output is 0.99999 with the results obtained by the direct method and is 0.99998 with the result obtained by the solution of Riccati equation.

The calculations are carried out by RAZDAN computer. The simulation is carried out by ODRA computer.

5.2. Considering that the desired optimal transfer function is:

$$\Delta_c(s) = \frac{Y(s)}{R(s)} = \frac{80}{s^3 + 14\, s^2 + 48\, s + 80};$$

and the open-loop system in the state-variable form is:

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 5 \\ 0 & 0 & -5 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 16 \end{bmatrix} u; \qquad \text{(E 4)}$$

$$\mathbf{y} = [1,\ 0,\ 0]^T \mathbf{x}$$

taking the initial value of vector $\mathbf{g}$, $\mathbf{g}_0 = [0.5, 0.5, 0.5]$, and the values of sequence ω $\{1, 2, 3\}$, and $\varrho = 1$, the result of the first stage of design is:

$$\mathbf{g}^0 = [1, 0.362735, 0.537645]^T.$$

Using this result in the second stage, taking the data of the first stage and initial value of vector $\mathbf{k}$, $\mathbf{k}_0 = [0.1, 0.1, 0.1]^T$, the following results are obtained:

$$\mathbf{k}^0 = [1, 0.4375, 0.5000]^T.$$

6. Appendix

In a special case of single-variable systems, in the knowledge of system matrices $\mathbf{A}$, $\mathbf{b}$ and $\mathbf{c}$, the following computation technique may be applied to get $\mathbf{h}(\omega) + j\mathbf{m}(\omega)$ from $\mathbf{\Phi}_0(j\omega)\mathbf{b}$. The fundamental matrix $\mathbf{\Phi}_0(s) = [s\mathbf{I} - \mathbf{A}]^{-1}$ is computed according to the procedure

$$\mathbf{\Phi}_0(s) = \frac{\operatorname{adj}[s\mathbf{I} - \mathbf{A}]}{\det[s\mathbf{I} - \mathbf{A}]} = \frac{\mathbf{N}^*(s)}{\Delta_0(s)} \tag{65}$$

where

$$\Delta_0(s) = \sum_{i=0}^{n} d_i s^{n-i} \tag{66}$$

is the characteristic equation of the open-loop system of nth degree and

$$\mathbf{N}^*(s) = \sum_{i=0}^{n-1} \mathbf{N}_i^* s^{n-i-1} \tag{67}$$

is a matrix whose elements are polynomials in s of degree $(n - 1)$ at most, with $\mathbf{N}_0^*, \mathbf{N}_1^*, \mathbf{N}_2^*, \ldots, \mathbf{N}_{n-1}^*$ are $(n \times n)$ matrices with constant elements.

The technique for computing the coefficient matrices $\mathbf{N}_i^*$ of $\mathbf{\Phi}_0(s)$ is given in the following form: Given an $(n \times n)$ constant system matrix $\mathbf{A}$, then the coefficients d_i of the polynomial $\Delta_0(s)$ defined in (66) and the matrix coefficients $\mathbf{N}_i^*$ of the matrix polynomial $\mathbf{N}^*(s)$ defined in (67) are given by the following pair of equations [12, pp. 301–303]:

$$d_i = \frac{1}{i} \operatorname{tr}(\mathbf{N}_{i-1}^* \mathbf{A}); \qquad \mathbf{N}_i^* = \mathbf{N}_{i-1}^* \mathbf{A} + d_i \mathbf{I} \tag{68}$$

with initial $\mathbf{N}_0^* = \mathbf{I}$ and $\mathbf{A}$ is the system matrix. This iterative technique for successively computing the d_i's and the $\mathbf{N}_i^*$'s determines the characteristic polynomial by (66) and $\mathbf{N}^*(s)$ by (67) step by step.

Formulas (68) are very suitable for computation by digital computer.

Using Eq. (67)

$$\boldsymbol{\Phi}_0(s)\,\mathbf{b} = \frac{1}{\Delta_0(s)} \sum_{i=0}^{n-1} \mathbf{N}_i^*\,\mathbf{b}\, s^{n-i-1} = \frac{1}{\Delta_0(s)} \sum_{i=0}^{n-1} \mathbf{a}_i\, s^{n-i-1} \tag{69}$$

where $\mathbf{a}_i = \mathbf{N}_i^*\mathbf{b}$ is ith coefficient vector of the $\boldsymbol{\Phi}_0(s)\mathbf{b}$ vector polynomial, defining the mth element of $\sum_{i=0}^{n-1} \mathbf{a}_i\, s^{n-i-1}$ polynomial vector by $a_{i_m}(s)$, we obtain

$$a_{i_m}(s)\Big|_{s=j\omega} = \sum_{i=0}^{n-1} a_{i_m}\, s^{n-i-1}\Big|_{s=j\omega}. \tag{70}$$

This may be expressed simply by the Horner form:

$$a_{i_m}(j\omega) = \big(\ldots (a_{1_m} j\omega + a_{2_m})\, j\omega + \ldots + a_{(n-2)m}\big) j\omega + a_{(n-1)_m} \tag{71}$$

by means of which each element of the polynomial vector and the characteristic polynomial may be computed.

Considering

$$a_{i_m}(j\omega) = \hat{U}_m + j\hat{V}_m; \tag{72}$$

and

$$\Delta_0(j\omega) = \hat{e} + jf. \tag{73}$$

From Eq. (69) we get

$$\boldsymbol{\Phi}_0(j\omega)\,\mathbf{b} = \frac{1}{\hat{e}^2 + j\hat{F}} \begin{bmatrix} \hat{U}_1 + j\hat{V}_1 \\ \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \\ \hat{U}_n + j\hat{V}_n \end{bmatrix} = \mathbf{h}(\omega) + j\mathbf{m}(\omega), \tag{74}$$

where h_i and m_i are the real and the imaginary elements respectively, i.e.

$$h_i = \frac{\hat{U}_i\,\hat{e} + \hat{V}_i\,\hat{f}}{\hat{e}^2 + \hat{f}^2} \qquad (i = 1, 2, \ldots, n); \tag{75}$$

and

$$m_i = \frac{\hat{V}_i\,\hat{e} + \hat{U}_i\,\hat{f}}{\hat{e}^2 + \hat{f}^2}.$$

These values may be used in Eqs. (32) and (42).

7. Conclusion

Procedures for computing a weighting vector $\mathbf{g}^0$ for the reduced performance index and its optimal $\mathbf{k}^0$ feedback coefficient vector in order to satisfy a desired optimal transfer function have been presented, using frequency-domain method. Depending upon the relationships between the response of an optimal system, the elements of the weighting vector of the reduced performance index and the optimal feedback vector, a design technique has been put forward.

This allows the design of an optimal system for a desired response and removes the difficulty in choosing the elements of the performance index met with in conventional optimization techniques. This method in its presented form is confined to single-variable systems. Further work is being carried out to generalize the method for multivariable systems.

References

1. *Kalman, R. E.*, When is a linear control system optimal? Trans. ASME, J. Basic Eng., ser. D. **86** (1964) 51—60.
2. *Sandberg, W.*, On the theory of linear multiloop feedback systems. Bell Syst. Techn. J. **42** (1963) 355—382.
3. *Perkins, W. R., Cruz, J.*, Feedback properties of linear regulators. IEEE Trans. on Automatic Control **16** (1971), 6, 659—664.
4. *Heym, K.*, Die Synthese des optimalen Zustandsreglers im Frequenzbereich. Regelungstechnik und Prozess-Datenverarbeitung **20** (1972) 9, 388—393.
5. *Fallside, F., Seraji, H.*, Design of optimal systems by a frequency-domain technique. Proc. IEE **117** (1970) Oct. 2017—2024.
6. *Fallside, F., Seraji, H.*, Optimality and closed-loop pole positions for single-input systems. Electronics Letters **6** (1970) 22, 721—723.
7. *Schultz, D. G., Melsa, J. L.*, State functions and linear control systems. McGraw-Hill, 1967.
8. *Csáki, F.*, Fejezetek a szabályozástechnikából. Állapotegyenletek (Chapters from control technique. State equations). Műszaki Könyvkiadó, Budapest, 1973.
9. *Kaczorek, T.*, Synteza liniowych ukladow stacjonarnych za pomoca czlonow proporcjonalnych. Archiwum Automatyki i Telemechaniki **17** (1972) 3 247—256.
10. *Amborski, K.*, Organiczenia w syntezie ukladow regulacji przy niepelnej dostepnosci zmiennych stanu. Archiwum Automatyki i Telemechaniki **17** (1972) 3 257—267.
11. *Kis, Ottó, Kovács, M.*, Numerikus módszerek (Numerical methods). Műszaki Könyvkiadó, Budapest, 1973.
12. *Zadeh, L. A., Desoer, Ch. A.*, Linear System Theory. The state space approach. McGraw-Hill, 1963.

Метод конструкции линейных оптимальных систем в области частот

Ф. ЧАКИ, ЭЛ-САИД ХАМАДА

(Будапешт)

В этой статье показан двухстепенчатый метод, состоящий из двух алгоритмов, для оптимального синтеза линейных систем с квадратным критерием. Если существует минимизируемая сметная функция, то первая степень определяется алгоритмом на основе конфигурации или спецификации характеристического уравнения оптимальной замкнутой системы.

С другой стороны, если дана сметная функция или уже определена применением первого алгоритма, то оптимальная матрица коэффициентов обратной связи определяется применением второго алгоритма. Оба алгоритма могут быть применены отдельно но и соединено.

Причём, если приходится определить только оптимальную матрицу коэффициентов обратной связи, то её можно вычислить с помощью непосредственного но подобного алгоритма на основе конфигурации замкнутой системы без определения местной функции. Метод даёт возможность ознакомиться с соотношением значений и матрицы веса замкнутой системы. С помощью матрицы веса можно непосредственно доказать определение стратегии регулирования. Причём метод делает возможным изменение матрицы веса в системе, уже конструированной с данным квадратным критерием, чтобы полюсы, находящиеся в нежелательной окружаемости перенесились в желательную окружаемость. Соединение двух алгоритмов показано вычислением одного примера Вард—Леонард привода.

F. Csáki and
El-Said Hamada
Department of Automation
Technical University of Budapest
H–1111, Budapest, XI., Goldmann Gy. tér 3
Hungary

Problems of Control and Information Theory, Vol. 4 (2), pp. 121—130 (1975)

ОБ ОДНОМ ПОДХОДЕ К РЕШЕНИЮ ОЦЕНОЧНЫХ ЗАДАЧ

В. Н. ЗАХАРОВ, Д. А. ПОСПЕЛОВ

(Москва)

(Поступила в редакцию 27 мая 1974 г.)

В статье рассматривается использование аппарата многозначной логики для построения оценочной функции, позволяющей отражать психологию принятия той или иной оценки ситуации человеком в условиях неполного знания значений параметров, определяющих ситуацию. На примере задачи коллективного поведения автоматов показывается, что подход, предлагаемый в данной работе, оказывается полезным в моделях децентрализованного управления.

1. Постановка задачи

В повседневной практике мы часто сталкиваемся с необходимостью оценки различных явлений, событий, поступков. Мы говорим, например, «неинтересная книга», «хорошая погода», «полезная лекция», формулируя тем самым свою оценку этих явлений. Эти оценки оказывают существенное влияние на формирование линии нашего поведения. В ряде случаев мы не можем решиться на то или иное действие, не будучи в состоянии оценить ситуацию, в которой это действие нужно принять.

Стремление к автоматизации сложных технических, экономических и организационных систем, характерное для последнего десятилетия развития теории управления, привело к необходимости создания систем управления, работающих в непредсказуемо меняющихся условиях, в той или иной степени автономно. Роботы, служащие для исследований и работ в космосе или океане, лишь — один из немногих примеров систем такого типа. Но любое автономное автоматическое устройство должно либо иметь специальный механизм оценки ситуаций, возникающих в управляемом объекте и среде, в которой он находится, либо пользоваться подобной информацией от других устройств или человека. При этом эффективность функционирования таких систем управления оказывается тесно связанной с эффективностью нахождения оценок.

Будем предполагать, что в некоторые дискретные моменты времени на вход системы управления поступает информация,подлежащая оценке. Эта информация даёт описание некоторой ситуации S, характеризуемой набором значений n параметров, важных с точки зрения оценки S. Таким образом, ситуация в любой момент времени t характеризуется набором $\langle \pi_1^t, \pi_2^t, \ldots, \pi_n^t \rangle$,

в котором π_i^t означает некоторое фиксированное значение параметра π_i в момент времени t. Оценки для S берутся из некоторого множества Ξ. В простейшем случае Ξ может быть множеством, состоящим только из двух оценок, которые условно можно обозначить как 0 и 1. Интерпретация этих оценок может быть различной. Примерами интерпретации могут служить пары: хорошо — плохо, полезно — бесполезно, полезно — вредно и т. п. В более сложных случаях Ξ может состоять из m элементов с соответствующей интерпретацией. Примерами могут служить: полезно — безразлично — вредно или очень приятно — приятно — безразлично — неприятно — очень неприятно и т. п. Задача оценки ситуации S состоит в нахождении такого отображения φ, которое по каждому конкретному набору $\langle \pi_1^t, \pi_2^t, \ldots, \pi_n^t \rangle$ указывало бы однозначно некоторый элемент из множества Ξ.

В человеческой практике наиболее часто встречается двухэтапная задача оценки. На первом этапе происходит индивидуальная оценка значений параметров π_i. Другими словами, реализуются n отображений вида $\pi_i^t \to \Xi_i$, где Ξ_i — множество значений оценок для параметра π_i. После этого для оценки ситуации S используется не набор $\langle \pi_1^t, \pi_2^t, \ldots, \pi_n^t \rangle$, а набор оценок параметров ситуации $\langle \xi_1(\pi_1^t), \xi_2(\pi_2^t), \ldots, \xi_n(\pi_n^t) \rangle$, где $\xi_i(\pi_i^t)$ — элементы из Ξ_i. Заметим, что параметры π_i могут принимать не только числовые значения а оценки этих параметров всегда могут быть закодированы натуральными числами. Выберем среди множеств Ξ и Ξ_i $i = 1, 2, \ldots, n$ множество наибольшей мощности (случай бесконечного множества оценок здесь не рассматривается так как для практики он представляет незначительный интерес). Пусть это множество есть Ξ^*, а его мощность равна l. Тогда пополним все остальные множества до этой мощности и будем рассматривать одно общее множество значений оценок для параметров и ситуации. При n параметрах и l значениях оценки будем иметь l^n различных наборов $\langle \xi_1(\pi_1), \xi_2(\pi_2), \ldots \xi_n(\pi_n) \rangle$. Оценка S будет сводиться к построению отображения вида $(\Xi^*) \to \Xi^*$. Это отображение естественным образом описывается таблицей, состоящей из l^n строк и $(n+1)$-го столбца. В первых n столбцах записываются оценки параметров, а в $(n+1)$-ом столбце — значения оценки S. Как известно, такая таблица задаёт функцию l-значной логики.

С помощью описанной процедуры задача оценки ситуации S по набору оценок параметров, её определяющих, сводится к задаче нахождения подходящей функции l-значной логики.

2. Двоичный случай

Рассмотрим наиболее простой случай, когда мощность множества Ξ^* равна двум. В этом случае оценочная функция является обычной функцией алгебры логики. Для иллюстрации применения аппарата двоичной логики

рассмотрим следующую содержательную постановку задачи. В качестве ситуации, которую необходимо оценивать, будем рассматривать погоду. Параметры, имеющие значение для погоды, — это температура, осадки, состояние неба и состояние почвы. Обозначим эти параметры соответственно через π_1 π_2, π_3 и π_4. В таблице 1 показано, как осуществляется переход к оценкам этих параметров.

Таблица 1

Параметр	Наименование	Порог оценки	Оценка
π_1	Температура	не ниже 20° ниже 20°	1 0
π_2	Осадки	нет есть	1 0
π_3	Состояние неба	безоблачно есть облака	1 0
π_4	Состояние почвы	сухая влажная	1 0

Наша задача состоит в оценке погоды. Эта оценка имеет два значения: 1 — хорошая погода и 0 — плохая погода. По оценкам, приведённым в таблице 1, нетрудно составить новую таблицу, содержащую 16 строк. Каждый из наборов будет характеризовать совокупность значений параметров. Например, набор ⟨0110⟩ означает, что погода безоблачная, без осадков, почва влажная, а температура воздуха ниже 20°. Не все возможные комбинации оценок параметров будут иметь смысл. Например, невозможна ситуация, при которой «дождь среди ясного неба». Поэтому на практике некоторые строки не будут порождать никакой определённой оценки для функции, оценивающей S. Функция алгебры логики получится недоопределённой.

Отметим также, что оценка ситуации в целом весьма тесно связана с целью этой оценки. От цели зависит как вид оценочной функции, так и оценки параметров. Если, например, мы собираемся совершить прогулку, то ситуация, описываемая набором ⟨1 1 1 1⟩, несомненно должна оцениваться нами как «хорошая погода» и оценочная функция на этом наборе обращается в единицу. Однако, если проводить оценку погоды с точки зрения проведения посевной кампании, то набор ⟨1 1 1 1⟩ может быть оценён как

«плохая погода» и оценочная функция на этом наборе должна обратиться в нуль. Во втором случае оценки индивидуального типа для параметров оказываются несогласованными с оценкой для всей ситуации. Эта несогласованность проявляется в том, что интерпретация 1 для параметров и ситуации в целом различна. Оценочные функции для указанных двух целей приведены в таблице 2. В этой таблице φ_1 означает функцию для цели «прогулка», а φ_2 — для цели «посев». Прочерки в столбцах значений функций соответствуют тем наборам, которые не имеют физического смысла. На этих наборах оценочные функции можно доопределять произвольным образом, подставляя вместо прочерка 0 или 1.

Таблица 2

Оценки параметров				φ_1	φ_2
$\xi_1(\pi_1)$	$\xi_2(\pi_2)$	$\xi_3(\pi_3)$	$\xi_4(\pi_4)$		
0	0	0	0	0	0
0	0	0	1	0	0
0	0	1	0	—	—
0	0	1	1	—	—
0	1	0	0	1	1
0	1	0	1	1	0
0	1	1	0	1	1
0	1	1	1	1	0
1	0	0	0	0	1
1	0	0	1	0	0
1	0	1	0	—	—
1	0	1	1	—	—
1	1	0	0	1	1
1	1	0	1	1	0
1	1	1	0	1	1
1	1	1	1	1	0

От подобной таблицы нетрудно перейти к аналитическому представлению оценочных функций в виде стандартных дизъюнктивных или конъюнктивных нормальных форм или к какому-либо другому аналитическому представлению.

К сожалению, двоичный случай представляет мало возможностей для исследования реальных процессов оценки ситуаций и принятых человеком решений. Поэтому более интересно и поучительно рассматривать случай, когда множество Ξ^* имеет не менее трёх элементов.

3. Троичный случай

Будем считать, что множество Ξ^* содержит три элемента 0, 1/2 и 1. Будем интерпретировать значения оценок как «неблагоприятно», «информация отсутствует» и «благоприятно». Таким образом, будем допускать наличие неопределённости в оценках параметров и ситуации. Такое допущение делает модель оценки интересной и позволяет ввести функции оценки, в какой-то мере отражающие психологические особенности людей, дающих оценку. Эти психологические особенности будут выявляться при оценке ситуаций, в которых имеются параметры с оценкой 1/2. При выбранной нами интерпретации этой оценки возможно её истолкование как в неблагоприятном, так и в благоприятном для оценивающего индивидуума смысле. Индивидуумы, склонные к истолкованию 1/2 в неблагоприятном смысле, образуют класс пессимистов. а склонные к истолкованию 1/2 в благоприятном смысле — класс оптимистов Индивидуумы, которые за 1/2 не видят ни 0, ни 1, образуют класс объективистов.

Для более точного определения состава этих классов несколько сузим класс возможных оценочных функций. Как показывают многочисленные психологические исследования (например, [1]), человек при принятии решений использует в подавляющем большинстве случаев функции конъюнкции, дизъюнкции, отрицания и дизъюнктивные и конъюнктивные формы с весьма небольшим числом членов и с небольшой длиной каждого из члена формы. Это позволяет нам рассматривать психологические особенности индивидуумов на примере функций дизъюнкции и конъюнкции. Особенность оценки будет проявляться в том, что вместо объективных конъюнкций и дизъюнкций индивидуум будет использовать «похожие» на эти функции пессимистические и оптимистические конъюнкции и дизъюнкции. Поскольку в строгом смысле эти смещённые функции не являются конъюнкциями и дизъюнкциями, то будем называть их квазиконъюнкциями и квазидизъюнкциями. Из квазиконъюнкций и квазидизъюнкций могут конструироваться аналоги дизъюнктивных и конъюнктивных нормальных форм.

Рассмотрим вначале квазидизъюнкции. Поставим в соответствие каждой оценке некоторый числовой вес, зависящий от числа параметров, определяющих ситуацию. А именно: оценке 0 будет соответствовать нулевой вес, оценке 1 — вес, равный единице, а оценке 1/2 — вес, равный $1/(n+1)$. Под весом набора $\langle \xi_1(\pi_1), \xi_2(\pi_2), \ldots, \xi_n(\pi_n) \rangle$ будем понимать сумму весов оценок, образующих этот набор. Вес для оценки 1/2 выбран так, чтобы вес набора, состоящего только из оценок 1/2, был бы меньше единицы. Зададим теперь два числа α и β ($0 \leq \alpha, \beta \leq 1$), которые будем называть соответственно нижним и верхним порогами. Определим оценочную функцию для ситуации следую-

щим образом. Она принимает значение 0 на наборах, вес которых меньше α, принимает значение 1 на наборах, вес которых не меньше β и принимает значение 1/2 на остальных наборах. Изменяя значения нижнего и верхнего порога, можно породить множество оценочных функций, обладающих подобно дизъюнкции свойством монотонности относительно значений аргументов, упорядоченных естественным образом ($0 < 1/2 < 1$). Функции, получающиеся таким способом, и будем называть квазидизъюнкциями. Обычная дизъюнкция получается, если положить $\alpha = 1/(n+1)$, а $\beta = 1$. Если теперь зафиксировать нижний порог таким, какой он есть у дизъюнкции, и понижать значение β, то часть наборов, в которые будут входить лишь оценки, равные 0 и 1/2, будут давать в качестве значений оценочной функции значение, равное 1. Такие квазидизъюнкции естественно назвать оптимистическими.

Если же зафиксировать верхний порог таким, какой он есть у дизъюнкции, и повышать значение α, то часть наборов, в которые будут входить оценки, равные 0 и 1/2, будут давать в качестве значений оценочной функции значение, равное 0. Такие квазидизъюнкции естественно назвать пессимистическими. Таким образом, оптимисты склонны часть неопределённых оценок параметров считать благоприятными, а пессимисты всегда подозревают, что среди тех параметров, оценки которых являются неопределёнными, всегда имеются неблагоприятные оценки.

Рассмотрим теперь квазиконъюнкции. Для этих функций будем использовать такие же веса оценок параметров, как и для случая квазидизъюнкций. Однако вес набора теперь будем определять не суммой весов, составляющих его оценок, а произведением этих весов. Как и для квазидизъюнкций введём нижний γ и верхний δ пороги. Для обычной конъюнкции $\gamma = 1/(n+1)^n$, а $\delta = 1$. Изменяя соответствующим образом γ или δ получим всё множество оптимистических и пессимистических квазиконъюнкций. В таблице 3 приведены для случая двух параметров, определяющих ситуацию, значения всех квазидизъюнкций и квазиконъюнкций. В двух левых столбцах показаны веса наборов соответственно для квазидизъюнкций и квазиконъюнкций. Для оценочных функций нижние индексы определяют тип функции (индекс o соответствует оптимистическим функциям, индекс p — пессимистическим, отсутствие индекса на втором месте внизу — объективным). Верхние индексы показывают степень оптимизма или пессимизма. Чем выше эта степень, тем более оптимистической или пессимистической является оценочная функция. Например, функция $\varphi^1_{\&_o}$ устроена так, что при наличии в наборе оценок параметров одной оценки 1/2 предполагается, что истинная оценка этого параметра благоприятна. Если же в наборе имеется две оценки 1/2, то оценка ситуации в целом остаётся неопределённой. Функция же $\varphi^2_{\&_o}$ является функцией, характеризующей «оголтелого оптимиста», который даже на наборе

$\langle 1/2\ 1/2\rangle$ склонен считать, что все параметры имеют благоприятные для него значения.

Таблица 3

Σ	П	$\xi_1(\pi_1)$	$\xi_2(\pi_2)$	Квазидизъюнкции					Квазиконъюнкции				
				$\varphi^2_{\vee p}$	$\varphi^1_{\vee p}$	$\varphi_\vee$	$\varphi^1_{\vee 0}$	$\varphi^2_{\vee 0}$	$\varphi^2_{\& p}$	$\varphi^1_{\& p}$	$\varphi_\&$	$\varphi^1_{\& 0}$	$\varphi^2_{\& 0}$
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1/3	0	0	1/2	0	0	1/2	1/2	1	0	0	0	0	0
1	0	0	1	1	1	1	1	1	0	0	0	0	0
1/3	0	1/2	0	0	0	1/2	1/2	1	0	0	0	0	0
2/3	1/9	1/2	1/2	0	1/2	1/2	1	1	0	0	1/2	1/2	1
4/3	1/3	1/2	1	1	1	1	1	1	0	1/2	1/2	1	1
1	0	1	0	1	1	1	1	1	0	0	0	0	0
4/3	1/3	1	1/2	1	1	1	1	1	0	1/2	1/2	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1

Для иллюстрации того, как используются введённые нами функции, рассмотрим следующий шуточный пример. Некоторый индивидуум А решил, наконец, после долгих колебаний жениться. При этом для него принятие решения о женитьбе связано с оценкой двух важных с его точки зрения параметров: характера девушки и характера её матери, с которой ему придётся, повидимому, жить вместе. Однако в период знакомства А не сумел полностью выяснить оценки этих параметров. Решение о женитьбе ему надо принять в условиях набора $\langle 1/2\ 1/2\rangle$. Вид оценочной функции — квазиконъюнкция. Если А является «оголтелым оптимистом», то он считает, что в жизни ему, безусловно, должно везти, и характер будущей жены и её матери является для него благоприятным. Такой индивидуум примет решение о женитьбе. Если же А — «осторожный пессимист», т. е. пессимист первого ранга, то он не примет решения о женитьбе, так как не надеется, что при отсутствии информации оба параметра окажутся одновременно благоприятными для него. «Оголтелый пессимист» не будет жениться даже в условиях наборов $\langle 1/2\ 1\rangle$ и $\langle 1\ 1/2\rangle$, считая, что «мир устроен плохо». «Осторожный пессимист» в этой же ситуации займёт выжидательную позицию, не принимая окончательного решения о женитьбе.

4. Некоторые обобщения

То, что здесь рассматривался троичный случай, а не любой l-ичный, не должно смущать читателя. Переход от троичного случая к l-ичному совершается безболезненно и не даёт никаких принципиально новых результатов.

Увеличится лишь число различных рангов для пессимистов и оптимистов и число порогов, которое надо будет вводить для получения значений оценочной функции.

До сих пор рассматривался случай, когда все параметры в оценке ситуации имели одинаковое значение. Однако на практике часто встречаются случаи, в которых одни параметры более важны для оценки ситуации, а другие — менее важны. Можно показать, что такое неравноправие параметров устраняется введением индивидуальных весов для них, учитываемых при формировании веса набора.

Более сложную проблему представляет случай коррелированных значений параметров. В этом случае требуется, повидимому, несколько иной подход к решению задачи определения вида оценочной функции.

5. Пример

В заключение статьи нам представляется интересным привести один пример использования аппарата, развитого в предшествующих разделах. В теории коллективного поведения автоматов, тесно связанной с созданием систем управления децентрализованного типа, неоднократно рассматривалась задача распределения ограниченного множества ресурсов [2]. Эта задача интересна для практики, так как она легко интерпретируется в виде задач теории массового обслуживания или сетевого планирования, распределения промышленного оборудования или проектирования хороших операционных систем для вычислительных комплексов. Однако во всех ранее рассмотренных моделях коллектив автоматов предполагался однородным. С помощью моделирования на ЭВМ был проанализирован случай решения задачи распределения ограниченных ресурсов в неоднородном коллективе автоматов. Каждый автомат имел фиксированную пессимистическую или оптимистическую оценочную функцию и принимал решение о том или ином выборе на основе значений этой функции. Задача ставилась следующим образом. Имелось q кормушек, каждая из которых характеризовалась параметром $P_i(t)$, который интерпретировался как вероятность того, что в момент времени t в i-ой кормушке будет пища. Значения $P_i(t)$ автоматам априорно известны не были. Если в момент времени t в данной кормушке была пища, то её количество всегда было равно некоторой фиксированной и одинаковой для всех кормушек величине a. Если в этот момент времени в кормушке было r автоматов, то каждый из них получал a/r единиц пищи. Фиксировался период времени (число тактов распределения автоматов по кормушкам) T и указывалось число b, одинаковое для всех автоматов. Если за T тактов работы автомат накапливал суммарный запас пищи меньший чем b, то он

получал штраф пропорциональный разности T и накопленной им пищи. При превышении суммарной накопленной пищи над b он получал пропорциональный разности выигрыш. Задача коллективного поведения состояла в максимизации суммарного выигрыша по коллективу автоматов.

На основании накапливаемого опыта автоматы формировали два параметра: среднюю частоту появления пищи в каждой из кормушек и среднее число автоматов, посещающих каждую из кормушек. Значения этих параметров для различных автоматов различны, так как они формируются только на основе личных посещений кормушек автоматами. Для оценки значений этих параметров вводились соответствующие пороги. Например, если вероятность появления пищи в данной кормушке, оцениваемая средней частотой её появления, найденной данным автоматом, была меньше 1/4, то параметр $P_i(t)$ оценивался как 0, если эта частота была больше 3/4, то параметр $P_i(t)$ оценивался как 1. В остальных случаях в качестве оценки $P_i(t)$ использовалось значение 1/2. Оценка второго параметра была устроена аналогичным образом в зависимости от того, сколько автоматов посещало данную кормушку.

В каждый такт времени все автоматы индивидуально оценивали все кормушки с помощью тех оценочных функций, которые им были присвоены. После оценки они выбирали кормушку с максимальной оценкой, а если таких кормушек было несколько, то равновероятно любую из этих кормушек.

Был также подсчитан суммарный выигрыш коллектива автоматов, состоящих только из объективистов, и суммарный выигрыш коллектива автоматов, в котором были представлены равномерно все оптимисты, пессимисты и объективисты, оценочные функции которых даны в таблице 3. Оказалось, что смешанный коллектив даёт более высокий суммарный выигрыш по сравнению с однородным коллективом автоматов.

Литература

1. *Пиаже Ж.* Речь и мышление ребёнка. М.—Л., 1932.
2. *Варшавский В. И.* Коллективное поведение автоматов. М., 1973.

An approach to estimation problems

V. N. ZAKHAROV, D. A. POSPELOV

(Moscow)

The paper considers the possibility of application of apparatus of multivalued logic to the problem of evaluation of situations. This approach should correspond to the psychology of human decision being taken in the face of incomplete knowledge of data, necessary for the situation description. The functions of optimistic and pessimistic valuation analogous to ordinary conjunction and disjunction are introduced. An example of resource allocation between the group of finite automata in condition of incomplete information shows the advantages of the proposed approach. It is shown that an existance of "pessimistic" and "optimistic" automata among the group leads to the amelioration of the results in comparison with that of the homogeneous group.

В. Н. Захаров, Д. А. Поспелов
Вычислительный центр АН СССР
СССР, Москва В-333,
ул. Вавилова, д. 40

Problems of Control and Information Theory, Vol. 4 (2), pp. 131—138 (1975)

ERGODICITY RESULTS FOR SAMPLE COVARIANCES

T. SÖDERSTRÖM

(Uppsala)

(Received December 6, 1974)

Ergodicity results concerning convergence of sample covariances are established. Sufficient conditions on the actual signals are either that they are deterministic and periodic or that they are stationary stochastic processes obtained by sending white noise through a stable filter.

1. Introduction

In several applications of control and information theory ergodicity theorems are valuable tools for analysis. One example is the analysis of different identification methods (see e.g. [1], [4], [8]).

In this paper we will analyse the convergence of the sample covariance

$$\frac{1}{N} \sum_{t=1}^{N} z_1(t)\, z_2(t) \tag{1.1}$$

where $z_1(t)$ and $z_2(t)$ are either stationary stochastic processes or deterministic and periodic signals. Similar results for finite order systems have been obtained [2], where also the asymptotic distribution of (1.1) is considered. An analysis for some non-stationary stochastic processes can be found in [7].

Sufficient conditions for convergence of (1.1) with probability one (which often is the most desirable convergence concept) and in mean square will be given. It will turn out that when both $z_1(t)$ and $z_2(t)$ are stationary stochastic processes the conditions will also imply

$$\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} z_1(t)\, z_2(t) = E z_1(t)\, z_2(t)\,. \tag{1.2}$$

2. Preliminaires

The analysis made in the paper is a straightforward application of two well-known ergodicity theorems. They are restated here as Lemmas for clearness.

Lemma 2.1. Assume that $x(t)$ is a stationary stochastic process with discrete time and finite variance. If the covariance function fulfils

$$r_x(\tau) \to 0 \qquad \tau \to \mathbf{W} \tag{2.1}$$

then

$$\frac{1}{N} \sum_{t=1}^{N} x(t) \to Ex(t) \qquad N \to \infty \tag{2.2}$$

with probability one and in mean square.

Proof. See [5].

Lemma 2.2. Assume that $x(t)$ is a stochastic process with zero mean. If the covariance function fulfils

$$|r(t, s)| \leq K \frac{t^{\alpha} + s^{\alpha}}{1 + |t - s|^{\beta}} \tag{2.3}$$

with

$$K > 0, \quad 0 \leq 2\alpha < \beta < 1 \tag{2.4}$$

then

$$\frac{1}{N} \sum_{t=1}^{N} x(t) \to 0 \qquad N \to \infty \tag{2.5}$$

with probability one and in mean square.

Proof. See [3].

3. Main results

In the analysis of (1.1) the signals $z_i(t)$ will be one of the following forms

(i) $$z_i(t) = G(q^{-1})e(t) = \sum_{j=0}^{\infty} g_j q^{-j} e(t) = g_j e(t - j) \tag{3.1}$$

where $G(q^{-1})$ is a stable filter ($\sum_{i=0}^{\infty} g_i^2 < \infty$) and $e(t)$ denotes white noise (a sequence of equally distributed, independent random variables with finite moments). The stability requirement will imply that $z_i(t)$ is a stationary stochastic process with finite variance. The requirement is in particular fulfilled for a finite order, asymptotically stable filter.

(ii) $z_i(t)$ is a deteministic and periodic signal.

It is no restriction to assume that the white noise has zero mean. If not, rewrite (3.1) as

$$z_i(t) = \sum_{j=0}^{\infty} g_j(e(t-j) - Ee(t-j)) + Ee(t) \sum_{j=0}^{\infty} g_j \tag{3.2}$$

so $z_i(t)$ will be the sum of a stochastic part and a constant (a particular form of case (ii)). By superposition this case is thus easily handled.

The following three cases will then have to be considered

(i) $z_1(t)$ and $z_2(t)$ both stochastic
(ii) $z_1(t)$ stochastic and $z_2(t)$ deterministic
(iii) $z_1(t)$ and $z_2(t)$ both deterministic.

These three cases will now be treated. The results are given as Lemmas 3.1–3.3.

Lemma 3.1. Consider the stationary stochastic processes $z_1(t)$ and $z_2(t)$ given by

$$z_1(t) = G(q^{-1})e_1(t) = \sum_{i=0}^{\infty} g_i e_1(t-i) \tag{3.3}$$

$$z_2(t) = H(q^{-1})e_2(t) = \sum_{i=0}^{\infty} h_i e_2(t-i) \tag{3.4}$$

where $e_1(t)$ and $e_2(t)$ are white noise signals with zero mean.

Assume that

$$\sum_{i=0}^{\infty} g_i^2 < \infty \quad \sum_{i=0}^{\infty} h_i^2 < \infty \tag{3.5}$$

and that either $e_1(t) = e_2(t)$ or that the noise signals $\{e_1(t)\}$ and $\{e_2(t)\}$ are mutually independent.

Then

$$\frac{1}{N} \sum_{t=1}^{N} z_1(t)z_2(t) \to Ez_1(t)z_2(t) = \sum_{i=0}^{\infty} g_i h_i Ee_1(t)e_2(t), \quad N \to \infty \tag{3.6}$$

with probability one and in mean square.

Proof. The equality in (3.6) follows by a simple calculation due to the assumption on the noise signals. Introduce

$$x(t) = z_1(t)z_2(t) \tag{3.7}$$

which is a stationary stochastic process. Through some technical but straightforward calculations it will be shown that Lemma 2.1 can be applied.

Introduce

$$m_{ij} = Ee_i(t)\, e_j(t) \tag{3.8}$$

$$p_{12} = Ee_1(t)^2\, e_2(t)^2. \tag{3.9}$$

For the case $e_1(t) = e_2(t) = e(t)$ we have $m_{11} = m_{12} = m_{22} = Ee(t)^2$, $p_{12} = Ee(t)^4$, and for the case of mutual independence we have $m_{12} = 0$, $P_{12} = Ee_1(t)^2 \cdot Ee_2(t)^2$.

The covariance function of x is given by

$$r_x(\tau) = E\Big(z_1(t+\tau)\, z_2(t+\tau) - m_{12} \sum_{i=0}^{\infty} g_i h_i\Big)\Big(z_1(t)\, z_2(t) - m_{12} \sum_{i=0}^{\infty} g_i h_i\Big) =$$

$$= E \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} g_i h_j g_k h_n e_1(t-i) e_2(t-j) e_1(t+\tau-k) e_2(t+\tau-n) -$$

$$- (m_{12} \sum_{i=0}^{\infty} g_i h_i)^2. \tag{3.10}$$

In (3.10) the expectation operator commutates with the summations. Consider therefore

$$Ee_1(t-i)\, e_2(t-j)\, e_1(t+\tau-k)\, e_2(t+\tau-n). \tag{3.11}$$

If a time argument occurs only once, the value of the expression must be zero. Using Kronecker's δ-symbol we have

$$\begin{aligned} &Ee_1(t-i)\, e_2(t-j)\, e_1(t+\tau-k)\, e_2(t+\tau-n) = \\ &= m_{12}^2\, \delta_{ij}\, \delta_{kn} + m_{11}\, m_{22}\, \delta_{k,i+\tau}\, \delta_{n,j+\tau} + m_{12}^2\, \delta_{n,i+\tau}\, \delta_{k,j+\tau} + \\ &+ (p_{12} - 2m_{12}^2 - m_{11}^2\, m_{22})\, \delta_{ij}\, \delta_{kn}\, \delta_{k,i+\tau}\,. \end{aligned} \tag{3.12}$$

The first three terms describe the cases of pairwise coincident time argument. The last terms describe the effect when all time arguments are equal.

When (3.12) is used in (3.10) we obtain

$$r_x(\tau) = m_{12}^2 \sum_{i=0}^{\infty} g_i h_i \sum_{k=0}^{\infty} g_k h_k + m_{11} m_{22} \sum_{i=0}^{\infty} g_i g_{i+\tau} \sum_{j=0}^{\infty} h_j h_{j+\tau} +$$

$$+ m_{12}^2 \sum_{i=0}^{\infty} g_i h_{i+\tau} \sum_{j=}^{\infty} g_{j+\tau} h_j + (p_{12} - 2m_{12}^2 - m_{11} m_{22}) \sum_{i=0}^{\infty} g_i g_{i+\tau} h_i h_{i+\tau} -$$

$$- (m_{12} \sum_{i=0}^{\infty} g_i h_1)^2. \tag{3.13}$$

Using the Cauchy–Schwarz lemma

$$|r_x(\tau)| \leq m_{11}\, m_{22}\left[\sum_{i=0}^{\infty} g_i^2 \sum_{j=0}^{\infty} g_{j+\tau}^2 \sum_{k=0}^{\infty} h_k^2 \sum_{n=0}^{\infty} h_{n+\tau}^2\right]^{1/2} +$$

$$+ m_{12}^2\left[\sum_{i=0}^{\infty} g_i^2 \sum_{j=0}^{\infty} h_{j+1}^2 \sum_{k=0}^{\infty} g_{k+\tau}^2 \sum_{n=0}^{\infty} h_n^2\right]^{1/2} +$$

$$+ (p_{12} - 2m_{12}^2 - m_{12}\, m_{22})\left[\sum_{i=0}^{\infty} g_i^2\, h_i^3 \sum_{j=0}^{\infty} g_{i+\tau}^2\, h_{i+\tau}^2\right]^{1/2}. \tag{3.14}$$

But

$$\sum_{i=0}^{\infty} g_{i+\tau}^2 = \sum_{i=0}^{\infty} g_i^2 - \sum_{i=0}^{\tau-1} g_i^2 \to 0, \quad \tau \to \infty \tag{3.15}$$

which with (3.14) implies that

$$r_x(\tau) \to 0, \qquad \tau \to \infty. \tag{3.16}$$

Invoking Lemma 2.1 the proof is finished.

Remark. The special case of $G(q^{-1}) = H(q^{-1})$, $e_1(t) \equiv e_2(t)$ can be found e.g. in [6].

Lemma 3.2. Consider the stochastic process

$$z_1(t) = G(q^{-1})\, e(t) = \sum_{i=0}^{\infty} g_i\, e(t-i) \tag{3.17}$$

where $e(t)$ is white noise with zero mean. Let $z_2(t)$ be a deterministic (uniformly) bounded sequence. Assume that the filter $G(q^{-1})$ satisfies

$$\left|\sum_{i=0}^{\infty} g_i\, g_{i+\tau}\right| \leq C_\tau^{-\gamma} \tag{3.18}$$

for some C, some γ, $0 < \gamma < 1$ and all $\tau \geq 1$.

Then

$$\frac{1}{N}\sum_{t=1}^{N} z_1(t)\, z_2(t) \to 0, \qquad N \to \infty \tag{3.19}$$

with probability one and in mean square.

Remark. The condition (3.18) is always fulfilled for a finite order, asymptotically stable filter.

Proof. Introduce

$$x(t) = z_1(t)\, z_2(t). \tag{3.20}$$

By the assumption

$$| z_2(t) | < D \text{ all } t \tag{3.21}$$

for some D. The stochastic process $x(t)$ has zero mean and is not necessarily stationary. Its covariance function satisfies

$$| r_x(t, s) | = | Ez_1(t)\, z_2(t)\, z_1(s)\, z_2(s) | \leq D^2 | r_{z_1}(t - s) | , \tag{3.22}$$

but from (3.18)

$$| r_{z_1}(\tau) | = Ee(t)^2 \left| \sum_{i=0}^{\infty} g_i g_{i+\tau} \right| \leq Ee(t)^2 C_\tau^{-\varphi} \qquad \tau \geq 1. \tag{3.23}$$

Thus

$$| r_x(t, s) | \leq Ee(t)^2 CD^2 | t - s |^{-\gamma} \qquad | t - s | \geq 1. \tag{3.24}$$

The Lemma is now proved by invoking Lemma 2.2 with $\alpha = 0$, $\beta = \gamma$ and

$$K = Ee(t)^2 D^2 \max \left(\sum_{i=0}^{\infty} q_i^2, C \right). \tag{3.25}$$

Lemma 3.3. Let $z_1(t)$ and $z_2(t)$ be two bounded periodic sequences. Then

$$\frac{1}{N} \sum_{t=1}^{N} z_1(t)\, z_2(t) \tag{3.26}$$

converges as $N \to \infty$.

Proof. Clearly also $x(t) = z_1(t)\, z_2(t)$ is periodic. Denote its period with T. For given N, let $n = n(N)$ denote the largest integer such that $nT < N$.

Due to the assumption

$$| x(t) | \leq C \quad \text{some} \quad C, \text{ all } t. \tag{3.27}$$

Thus

$$\begin{aligned} \left| \frac{1}{N} \sum_{t=1}^{N} z_1(t)\, z_2(t) - \frac{1}{T} \sum_{t=1}^{T} x(t) \right| &= \left| \frac{1}{N} \sum_{t=1}^{nT} x(t) - \frac{1}{nT} \sum_{t=1}^{nT} x(t) + \frac{1}{N} \sum_{t=nT+1}^{N} x(t) \right| \leq \\ &\leq \frac{N - nT}{N\, nT} \left| \sum_{t=1}^{nT} x(t) \right| + \frac{1}{N} \left| \sum_{t=1}^{N-nT} x(t) \right| \leq \\ &\leq 2C \frac{N - nT}{N} \leq 2C \frac{T}{N} \to 0, \quad N \to \infty . \end{aligned} \tag{3.28}$$

Thus the convergence of (3.26) to $\frac{1}{T} \sum_{t=1}^{T} x(t)$ is proved.

The following example shows that $x(t)$ bounded does not imply convergence of

$$\frac{1}{N}\sum_{t=1}^{N} x(t)\,.$$

This means especially that $z_i(t)$ bounded is a too weak condition in Lemma 3.3.

Example. Define $x(t)$ by

$$\begin{aligned} x(t) &= 1 \quad t = 1 \\ &= -1 \quad t = 2 \ldots 4 \\ &= 1 \quad t = 5 \ldots 12 \\ &= -1 \quad t = 13 \ldots 36 \end{aligned}$$

and generally

$$x(t) = (-1)^{m-1}, \quad 4 \cdot 3^{m-1} + 1 \leq t \leq 4 \cdot 3^{m}.$$

Put

$$s_N = \frac{1}{N}\sum_{t=1}^{N} x(t)\,,$$

then $s_N = 1/2$ if $N = 4 \cdot 3^m$ m odd

and $s_N = -1/2$ if $N = 4 \cdot 3^m$ m even.

From this it follows that $\liminf s_N < \limsup s_N$ and thus $\lim s_N$ does not exist.

4. Conclusions

Some ergodicity results for sample covariances has been established. Sufficient conditions for convergence with probability one and in mean square have been derived. The results are useful e.g. in the analysis of consistency of different methods for identification of dynamic systems.

References

1. *Åström, K. J., Bohlin, T., Wensmark, S.*, Automatic construction of linear stochastic dynamic models for stationary industrial processes with random disturbances using operating records. IBM Nordic Laboratory, Sweden, Report TP18. 150 (1965).
2. *Åström, K. J.*, Div. of Automatic Control. Lund Inst. of Tech., unpublished results.
3. *Cramér, I. H., Leadbetter, M. R.*, Stationary and related stochastic processes. John Wiley & Sons, New York (1967).
4. *Finigan, B., Rowe, I. H.*, On the identification of linear discrete time system models using the instrumental variable method. Paper TM-5, 3rd IFAC Symposium on Identification and System Parameter Estimation, the Hague/Delft (1973).
5. *Gnedenko, B. V.*, The theory of probability. Chelsea Publishing Company, New York (1963).
6. *Hannan, E. J.*, Time series analysis. Meuthen and Co., London (1960).
7. *Ljung, L.*, On consistency for prediction error identification methods. Report 7405, Div. of Automatic Control, Lund Inst. of Tech. (1974).
8. *Söderström, T.*, Convergence properties of the generalized least squares identification method. Automatica **10,** Nov. (1974).

Эргодические результаты для ковариациям, полученные на основе реализащий

Т. СОДЕРСТРОМ

(Уппсала)

В статье описаны эргодические результаты сходимости по отношению к ковариации сигналов дискретных систем. Даны удовлетворительные условия в случае детерминированных и периодических сигналов как и в случае если сигналы являются стационарными стохастическими процессами. Стационарные стохастические процессы вызваны белым шумом, пропущенным через устойчивый фильтр. Описанные в статье результаты могут быть очень полезными для появляющейся оценки ошибки при проблемах идентификации.

T. Söderström
Institute of Technology
Uppsala University
Sweden

Problems of Control and Information Theory, Vol. 4 (3), pp. 139—146 (1975)

THE ANALYSIS OF THE ACCURACY OF DYNAMIC SYSTEMS CHANGING THEIR STRUCTURE IN THE RANDOM TIME MOMENT

V. A. BUKHALEV

(Moscow)

(Received April 25, 1974)

The system under discussion is described by nonlinear stochastic differential equations. The structure of the system is changed by jumping in the random time moment with known probability density. Equations for a characteristic function and the moments of the vector of phase coordinates of the system are obtained.

1. Problem formulation

Many systems during operation can change their structure by jumping in the random time moment due to a sudden damage of some elements, operation duty change and ambient conditions. Equations for the mathematical expectations and dispersions of the vectors of phase coordinates of the linear system, the structure of which is changed in the random time moment, are obtained in [1–3]. In the present paper the problem is formulated for defining the distribution law of the phase coordinates vector of the nonlinear system.

The class of the nonlinear dynamic systems the structure of which has N possible states, is considered. In state k the system is described by the equation

$$\mathbf{X}^{(k)} = \boldsymbol{\varphi}^{(k)}(\mathbf{X}^{(k)}, t) + \boldsymbol{\sigma}^{(k)}(\mathbf{X}^{(k)}, t)\,\mathbf{V}(t) \quad (k = 1, 2, \ldots, N) \quad (1)$$

where $\mathbf{X}^{(k)}(t)$ — n-dimensional phase coordinates vector, $\mathbf{V}(t)$ — n-dimensional vector of white noises with matrix of intensities equal to 1 and mathematical expectation equal to zero, $\boldsymbol{\varphi}^{(k)}(\mathbf{X}^{(k)}, t)$ — n-dimensional vector function, $\boldsymbol{\sigma}^{(k)}(\mathbf{X}^{(k)}, t)$ — $n \times n$ matrix, the elements of which are differentiated along all the coordinates of vector $\mathbf{X}^{(k)}$.

In the initial time moment $t_0 = 0$ vector $\mathbf{X}^{(k)}(0)$ has a conventional characteristic function $g^{(k)}(\boldsymbol{\lambda}, 0)$. The structure of the system transits into state k in the random time moment τ with the conventional probability density $p^{(kl)}(\tau)$, provided that in the moment t_0 the structure had state l. The probability of state l in the time moment t_0 is equal to $P_0^{(l)}$. Let us define statistical characteristics of vector $\mathbf{X}$.

2. Solution

On the basis of the conditions the characteristic function $g(\boldsymbol{\lambda}, t)$ of the stochastic process $\mathbf{X}(t)$ is defined by

$$g(\boldsymbol{\lambda}, t) = \sum_{k=1}^{N} g_*^{(k)}(\boldsymbol{\lambda}, t), \tag{2}$$

$$g_*^{(k)}(\boldsymbol{\lambda}, t) = \sum_{\substack{l=1 \\ (l \neq k)}}^{N} P_0^{(l)} \int_0^t g^{(kl)}(\boldsymbol{\lambda}, t, \boldsymbol{\tau})\, p^{(kl)}(\tau)\, d\tau +$$

$$+ P_0^{(k)} g^{(kk)}(\boldsymbol{\lambda}, t) \Big(1 - \sum_{\substack{l=1 \\ (l \neq k)}}^{N} \int_0^t p^{(lk)}(\tau)\, d\tau\Big) \tag{3}$$

where $g^{(kl)}(\boldsymbol{\lambda}, t, \tau) = M[e^{i\boldsymbol{\lambda}^T \mathbf{X}} \mid l_{t_0}, k_\tau, \tau]$ — conventional characteristic function provided that the structure, which had state l in the time moment t_0, transited in state k in the time moment τ, $g^{(kk)}(\boldsymbol{\lambda}, t) = M[e^{i\boldsymbol{\lambda}^T \mathbf{X}} \mid k_{t_0}, k_t]$ — conventional characteristic function provided that the structure, which had state k in the time moment t_0, preserved state k in the time moment t.

Conventional characteristic functions $g^{(kl)}(\boldsymbol{\lambda}, t, \tau)$ and $g^{(kk)}(\boldsymbol{\lambda}, t)$ are defined by the equation of V. S. Pugachev [4]

$$\frac{\partial g^{(kl)}(\boldsymbol{\lambda}, t)}{\partial t} = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} e^{i(\boldsymbol{\lambda} - \boldsymbol{\omega})^T \mathbf{x}} \boldsymbol{\Phi}^{(k)}(\boldsymbol{\lambda}, \mathbf{x}, t)\, g^{(kl)}(\boldsymbol{\mu}, t)\, d\boldsymbol{\mu}\,, \tag{4}$$

$$\boldsymbol{\Phi}^{(k)}(\boldsymbol{\lambda}, \mathbf{x}, t) = i\boldsymbol{\lambda}^T \mathbf{A}^{(k)}(\mathbf{x}, t) - \frac{1}{2} \boldsymbol{\lambda}^T \mathbf{B}^{(k)}(\mathbf{x}, t)\, \boldsymbol{\lambda}$$

($k, l = 1, 2, \ldots, N$; argument τ of the characteristic functions is deleted for the sake of simplification, superscript T means transposition), where $\boldsymbol{\lambda}$ — n-dimensional vector, $\mathbf{A}^{(k)}(\mathbf{x}, t)$ — n-dimensional vector of transfer factors, $\mathbf{B}^{(k)}(\mathbf{x}, t)$ — $n \times n$ matrix of diffusion factors, which are defined by the following formulae (5)

$$A_\nu^{(k)}(\mathbf{x}, t) = \varphi_\nu^{(k)}(\mathbf{x}, t) + \frac{1}{2} \sum_{p,q=1}^{n} \frac{\partial \sigma_{\nu p}^{(k)}(\mathbf{x}, t)}{\partial x_q}\, \sigma_{qp}^{(k)}(\mathbf{x}, t)\,,$$

$$B_{\nu\mu}^{(k)}(\mathbf{x}, t) = \sum_{p=1}^{n} \sigma_{\nu p}^{(k)}(\mathbf{x}, t)\, \sigma_{\mu p}^{(k)}(\mathbf{x}, t) \tag{5}$$

$$(\nu, \mu = 1, 2, \ldots, n; \quad k = 1, 2, \ldots, N).$$

Equations (4) are solved under the initial conditions

$$[g^{(kl)}(\boldsymbol{\lambda}, t, \tau)]_{t=\tau} = g^{(ll)}(\boldsymbol{\lambda}, \tau) \qquad (l \neq k),$$

$$[g^{(kk)}(\boldsymbol{\lambda}, t)]_{t=t_0} = g^{(k)}(\boldsymbol{\lambda}, 0)\,. \tag{6}$$

The values of function $g^{(ll)}(\boldsymbol{\lambda}, \tau)$ are random, because τ is a random value. That is why we use the first initial condition (6) as an intermediate result, without integrating Eq. (4) by $g^{(kl)}(\boldsymbol{\lambda}, t, \tau)$.

Differentiating Eq. (3) by t and using expressions (4) and (6), we get

$$\frac{\partial g_*^{(k)}(\boldsymbol{\lambda}, t)}{\partial t} = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} d\mathbf{x} \int_{-\infty}^{\infty} e^{i(\lambda-\mu)^T \mathbf{x}}\, \boldsymbol{\Phi}^{(k)}(\boldsymbol{\lambda}, \mathbf{x}, t) \times$$

$$\times\, g_*^{(k)}(\boldsymbol{\mu}, t)\, d\boldsymbol{\mu} - P_0^{(k)}\, g^{(kk)}(\boldsymbol{\lambda}, t) \sum_{\substack{l=1 \\ (l\neq k)}}^{N} p^{(lk)}(t) + \sum_{\substack{l=1 \\ (l\neq k)}}^{N} P_0^{(l)}\, g^{(ll)}(\boldsymbol{\lambda}, t)\, p^{(kl)}(t) \tag{7}$$

$$(k = 1, 2, \ldots, N)\,.$$

Equation (7) together with Eqs. (2), (4), (5) define characteristic function $g(\boldsymbol{\lambda}, t)$ and are integrated under the initial conditions

$$g^{(kk)}(\boldsymbol{\lambda}, 0) = g^{(k)}(\boldsymbol{\lambda}, 0), \quad g_*^{(k)}(\boldsymbol{\lambda}, 0) = g^{(k)}(\boldsymbol{\lambda}, 0)\, P_0^{(k)}. \tag{8}$$

It is possible to define also the probability of state k and conventional characteristic function $g^{(k)}(\boldsymbol{\lambda}, t)$ provided that in the time moment t the structure had state k:

$$g^{(k)}(\boldsymbol{\lambda}, t) = \frac{g_*^{(k)}(\boldsymbol{\lambda}, t)}{P^{(k)}(t)}\,; \tag{9}$$

$$\dot{P}^{(k)}(t) = \frac{\partial g_*^{(k)}(\mathbf{0}, t)}{\partial t} = -P_0^{(k)} \sum_{\substack{l=1 \\ (l\neq k)}}^{N} p^{(lk)}(t) + \sum_{l=1(l\neq k)}^{N} P_0^{(l)}\, p^{(kl)}(t)\,. \tag{10}$$

Using the formula

$$\alpha_{r_1 \ldots r_n} = \frac{1}{i^R} \left[\frac{\partial^R g(\boldsymbol{\lambda}, t)}{\partial \lambda_1^{r_1}\, \partial \lambda_2^{r_2} \ldots \partial \lambda_n^{r_n}} \right]_{\lambda=0}$$

where $\alpha_{r_1 \ldots r_n}(t) = M\left[\prod_{\nu=1}^{n} X_\nu^{r_\nu}\right]$ – mixed moment of Rth order $(r_1, r_2, \ldots, r_n = 0, 1, 2, \ldots, R;\ \sum_{\nu=1}^{n} r_\nu = R)$, the following equations are obtained for the

moments, from (2), (4), (7)

$$\alpha_{r_1 \dots r_n} = \sum_{k=1}^{N} \alpha^{(k)}_{* r_1 \dots r_n},$$

$$\dot{\alpha}^{(k)}_{* r_1 \dots r_n} = Q\{f^{(k)}_*(\mathbf{x}, t)\} - P_0^{(k)} \alpha^{(k)}_{r_1 \dots r_n} \sum_{\substack{l=1 \\ (l \neq k)}}^{N} p^{(lk)} + \sum_{\substack{l=1 \\ (l \neq k)}}^{N} P_0^{(l)} \alpha^{(ll)}_{r_1 \dots r_n} p^{(kl)}, \quad (11)$$

$$\dot{\alpha}^{(kk)}_{r_1 \dots r_n} = Q\{f^{(kk)}(\mathbf{x}, t)\}.$$

where

$$Q\{f^{(\xi)}(\mathbf{x}, t)\} = \int_{-\infty}^{\infty} \prod_{q=1}^{n} x_q^{r_q} \Big[\sum_{\nu=1}^{n} r_\nu A^{(k)}_\nu(\mathbf{x}, t)\, x_\nu^{-1} + \sum_{\nu, \mu=1}^{n} r_\nu (r_\nu - 1) B^{(k)}_{\nu\mu}(\mathbf{x}, t)\, x_\nu^{-2} +$$

$$+ \sum_{\substack{\nu, \mu=1 \\ (\nu \neq \mu)}}^{n} r_\nu r_\mu B^{(k)}_{\nu\mu}(\mathbf{x}, t)\, x_\nu^{-1} x_\mu^{-1} \Big] f^{(\xi)}(\mathbf{x}, t)\, dx.$$

If we expand probability densities $f^{(k)}(\mathbf{x}, t)$ and $f^{(kk)}(\mathbf{x}, t)$ in Eq. (11) into Ageworth type series [6], preserving necessary number of members in them, we get the closed system of equations for the moments. These equations are approximate, as they comprise approximate expressions for probability densities $f^{(k)}(\mathbf{x}, t)$ and $f^{(kk)}(\mathbf{x}, t)$ which are composed of several members of Ageworth type series. In particular preserving only the first member in each series, we get equations for the first two moments from (11) with the account of (10).

$$\mathbf{m} = \sum_{k=1}^{N} \mathbf{m}^{(k)} P^{(k)}, \qquad \theta = \sum_{k=1}^{N} \theta^{(k)} P^{(k)},$$

$$\dot{\mathbf{m}}^{(k)} = \int_{-\infty}^{\infty} \mathbf{A}^{(k)}(\mathbf{x}, t)\, f_r^{(k)}(\mathbf{x}, t)\, d\mathbf{x} +$$

$$+ \frac{1}{P^{(k)}} \Big[\mathbf{m}^{(k)} (\gamma^{(k)} - \beta^{(k)}) - \mathbf{m}^{(kk)} \gamma^{(k)} + \sum_{\substack{l=1 \\ (l \neq k)}}^{N} \mathbf{m}^{(ll)} P_0^{(l)} p^{(kl)} \Big],$$

$$\dot{\mathbf{m}}^{(kk)} = \int_{-\infty}^{\infty} \mathbf{A}^{(k)}(\mathbf{x}, t)\, f_r^{(kk)}(\mathbf{x}, t)\, d\mathbf{x},$$

$$\dot{\theta}^{(k)} = \int_{-\infty}^{\infty} \mathbf{A}^{(k)}(\mathbf{x}, t)\, \mathbf{x}^T f_r^{(k)}(\mathbf{x}, t)\, d\mathbf{x} + \int_{-\infty}^{\infty} \mathbf{x}\mathbf{A}^{(k)T}(\mathbf{x}, t)\, f_r^{(k)}(\mathbf{x}, t)\, d\mathbf{x} +$$

$$+ \int_{-\infty}^{\infty} \mathbf{B}^{(k)}(\mathbf{x}, t)\, f_r^{(k)}(\mathbf{x}, t)\, d\mathbf{x} + \frac{1}{P^{(k)}} \Big[\theta^{(k)} (\gamma^{(k)} - \beta^{(k)}) - \tag{12}$$

$$- \theta^{(kk)} \gamma^{(k)} + \sum_{\substack{l=1 \\ (l \neq k)}}^{N} \theta^{(ll)} P_0^{(l)} p^{(kl)} \Big],$$

$$\dot{\theta}^{(kk)} = \int_{-\infty}^{\infty} \mathbf{A}^{(k)}(\mathbf{x}, t)\, \mathbf{x}^T f_r^{(kk)}(\mathbf{x}, t)\, d\mathbf{x} +$$

$$+ \int_{-\infty}^{\infty} \mathbf{x}\mathbf{A}^{(k)T}(\mathbf{x}, t)\, f_r^{(kk)}(\mathbf{x}, t)\, d\mathbf{x} + \int_{-\infty}^{\infty} \mathbf{B}^{(k)}(\mathbf{x}, t)\, f_r^{(kk)}(\mathbf{x}, t)\, d\mathbf{x},$$

where

$$\gamma^{(k)} = P_0^{(k)} \sum_{\substack{l=1 \\ (l \neq k)}}^{N} p^{(lk)}, \qquad \beta^{(k)} = \sum_{\substack{l=1 \\ (l \neq k)}}^{N} P_0^{(l)} p^{(kl)},$$

$f_r^{(k)}(\mathbf{x}, t)$, $f_r^{(kk}(\mathbf{x}, t)$ — Gaussian probability densities with moments $\mathbf{m}^{(k)}$, $\theta^{(kk)}$ and $\mathbf{m}^{(kk)}$, $\theta^{(kk)}$ respectively;

$$\mathbf{m} = M[\mathbf{X}], \qquad \theta = M[\mathbf{X}\mathbf{X}^T].$$

Probabilities $P^{(k)}$ are defined from Eq. (10). Equations (10), (12) are integrated under initial conditions

$$m^{(k)}(0) = m^{(kk)}(0) = \left[\frac{\partial g^{(k)}(\boldsymbol{\lambda}, 0)}{i\, \partial \boldsymbol{\lambda}} \right]_{\lambda=0}^{T},$$

$$\theta^{(k)}(0) = \theta^{(kk)}(0) = \left\{ \frac{\partial}{i\, \partial \boldsymbol{\lambda}} \left[\frac{\partial g^{(k)}(\boldsymbol{\lambda}, 0)}{i\, \partial \boldsymbol{\lambda}} \right]^T \right\}_{\lambda=0}, \tag{13}$$

$$P^{(k)}(0) = P_0^{(k)}.$$

In particular case of the linear system with two possible states of the structure at $P_0^{(2)} = 0$ Eqs. (12) coincide with those obtained in [1].

In the particular case when the system, which had in the initial time moment state j with probability $P_0^{(j)} = 1$, transits into state i in the known time moment t_n at $\sigma^{(l)}(\mathbf{X}^{(l)}, t)$ independent of $\mathbf{X}^{(l)}$ $(l = i, j)$, Eqs. (12) coincide with those obtained in [7] by the statistical linearization method [8], [9].

3. Example

The structure of the dynamic system has two possible states. In the first state the system is described by

$$\dot{X}^{(1)} = -a^{(1)} \operatorname{sgn} X^{(1)} + \sqrt{G^{(1)}}\, V(t), \tag{14}$$

and in the second state by

$$\dot{X}^{(2)} = -a^{(2)} X^{(2)} + \sqrt{G^{(2)}}\, V(t), \tag{15}$$

where $a^{(k)}$, $G^{(k)}$ $(k = 1, 2)$ — deterministic values, $V(t)$ — Gaussian white noise with the intensity equal to 1 and mathematical expectation equal to zero.

The system structure transits into state k in the random time moment τ, distributed according to Reley law with the conventional probability density

$$p^{(kl)}(\tau) = 2h^{(l)} \tau\, e^{-h^{(l)}\tau^2}$$

provided that in the initial time moment $t_0 = 0$ the structure had state l $(l, k = 1, 2; l \neq k)$. Initial state probabilities equal to $P_0^{(1)}$ and $P_0^{(2)}$ respectively.

Let us define statistical characteristics $X(t)$ in the steady state duty. Equations (2), (4), (5), (7) for the given example in the steady state duty yield $(t \to \infty)$

$$g(\lambda, \infty) = \sum_{k=1}^{2} g_*^{(k)}(\lambda, \infty), \quad g_*^{(k)}(\lambda, \infty) = g^{(k)}(\lambda, \infty)\, P^{(k)}(\infty),$$

$$g_*^{(1)}(\lambda, \infty) = \frac{\left(\dfrac{a^{(1)}}{G^{(1)}}\right)^2 P_0^{(2)}}{\left(\dfrac{a^{(1)}}{G^{(1)}}\right)^2 + \left(\dfrac{\lambda}{2}\right)^2}, \quad g_*^{(2)}(\lambda, \infty) = P_0^{(1)}\, e^{-\frac{\lambda^2 G^{(2)}}{4a^{(2)}}} \tag{16}$$

$$P^{(1)}(\infty) = P_0^{(2)}, \quad P^{(2)}(\infty) = P_0^{(1)},$$

which yields

$$g(\lambda, \infty) = \frac{\left(\dfrac{a^{(1)}}{G^{(1)}}\right)^2 P_0^{(2)}}{\left(\dfrac{a^{(1)}}{G^{(1)}}\right)^2 + \left(\dfrac{\lambda}{2}\right)^2} + P_0^{(1)}\, e^{-\frac{\lambda^2 G^{(2)}}{4a^{(2)}}}, \tag{17}$$

$$m(\infty) = 0, \quad \text{dispersion } D(\infty) = \frac{1}{2}\left[\left(\frac{G^{(1)}}{a^{(1)}}\right)^2 P_0^{(2)} + \frac{G^{(2)}}{a^{(2)}} P_0^{(1)}\right].$$

In the case of approximate solution of the problem we get from Eqs. (12) for the second moments

$$\theta = \sum_{k=1}^{2} \theta^{(k)} P^{(k)},$$

$$\dot{\theta}^{(1)} = -\frac{2a^{(1)}}{\sqrt{2\pi D^{(1)}}} \int_{-\infty}^{\infty} x \operatorname{sgn} x \exp\left[-\frac{(x - m^{(1)})^2}{2D^{(1)}}\right] dx + \\ + \frac{1}{P^{(1)}} [\theta^{(1)}(\gamma - \beta) - \theta^{(11)}\gamma + \theta^{(22)}\beta] + G^{(1)}, \tag{18}$$

$$\dot{\theta}^{(2)} = -2a^{(2)}\theta^{(2)} + \frac{1}{P^{(2)}} [\theta^{(2)}(\beta - \gamma) + \theta^{(11)}\gamma - \theta^{(22)}\beta] + G^{(2)},$$

$$\dot{\theta}^{(11)} = -\frac{2a^{(1)}}{\sqrt{2\pi D^{(11)}}} \int_{-\infty}^{\infty} x \operatorname{sgn} x \exp\left[-\frac{(x - m^{(11)})^2}{2D^{(11)}}\right] dx + G^{(1)},$$

$$\dot{\theta}^{(22)} = -2a^{(2)}\theta^{(22)} + G^{(2)},$$

$$D^{(1)} = \theta^{(1)} - (m^{(1)})^2, \quad D^{(11)} = \theta^{(11)} - (m^{(11)})^2,$$

$$\dot{P}^{(1)}(t) = P_0^{(1)} e^{-h^{(1)}t^2} + P_0^{(2)}(1 - e^{-h^{(2)}t^2}),$$

$$P^{(2)}(t) = 1 - P^{(1)}(t),$$

where $\gamma(t) = P_0^{(1)} p^{(21)}(t)$, $\beta(t) = P_0^{(2)} p^{(12)}(t)$.

In the steady state duty

$$m(\infty) = 0,$$

$$D(\infty) = \theta(\infty) = \frac{1}{2}\left[\frac{\pi}{4}\left(\frac{G^{(1)}}{a^{(1)}}\right)^2 P_0^{(2)} + \frac{G^{(2)}}{a^{(2)}} P_0^{(1)}\right].$$

The comparison of the exact solution (17) with the approximate one (19) shows that when $G^{(2)} = 0$ or $P_0^{(1)} = 0$ the relative error in the definition of the root-mean-square deviation $\sqrt{D(\infty)}$ for the given example reaches its maximum which is approximately equal to 11 %.

References

1. *Mishulina, O. A.*, The analysis of the linear control systems with random change of parameters with the use of computers. Izv. AN SSSR, Tekhnicheskaya kibernetika (1968) **3.**
2. *Sklarevich, A. N.*, Statistical characteristics of the output signal of the dynamic system with possible damage. Avtomatika i telemekhanika (1970) **4.**

3. *Sklarevich, A. N.*, Introduction into statistical dynamics of systems with possible damages. „Zinatne", Riga, 1973.
4. *Pugachev, V. S.*, The use of Markov processes theory for the analysis of the automatic systems accuracy. Izv. AN SSSR, OTN, Energetika i avtomatika (1961) **3.**
5. *Khazen, A. M.*, Methods of optimal statistical solution and problems of optimal equations. Sov. Radio, 1968.
6. *Geidarov, T. G.*, Expansion of n-dimensional laws of distribution in Ageworth series. Izv. AN Az. SSR, Series of physical, technical and mathematical sciences (1966) **4.**
7. *Aleksandrov, V. M., Batkov, A. M., Staroverov, A. N., Stchukin, B. A.*, Analysis of the accuracy of the nonlinear non-stationary systems by means of statistical linearization. Avtomatika i telemekhanika **26** (1965) 3.
8. *Kazakov, I. Ye.*, Statistical analysis of systems with multidimensional nonlinearities. Avtomatika i telemechanika **26** (1965) 3.
9. *Sinitsin, I. N.*, Methods of statistical linearization (Review). Avtomatika i telemekhanika (1974) **5.**

Анализ точности динамических систем, изменяющих структуру в случайный момент времени

В. А. БУХАЛЕВ

(Москва)

Многие системы в процессе работы могут скачкообразно изменять свою структуру в случайный момент времени вследствие внезапного отказа некоторых элементов, изменения режима работы или внешних условий. В [1—3] получены уравнения для определения математических ожиданий и корреляционных моментов фазовых координат линейных систем, структура которых изменяется скачком в случайный момент времени. В настоящей работе рассматривается нелинейная динамическая система, структура которой имеет N возможных состояний. В k-ом состоянии система описывается стохастическим дифференциальным уравнением (1) ($k = 1, 2, \ldots, N$). Структура системы, имеющая в начальный момент времени состояние l переходит в случайный момент времени в состояние $k(l, k = 1, 2, \ldots, N; l \neq k)$. Ставится задача нахождения закона распределения фазовых координат системы при известных законах распределения фазовых координат и состояния структуры в начальный момент времени и известной плотности вероятности случайного момента времени перехода.

Условная характеристическая функция вектора фазовых координат при известном моменте времени перехода определяется уравнением (4) В. С. Пугачева [4]. Использование уравнения (4) приводит к уравнению (7), которое совместно с выражениями (2), (4), (5) определяет характеристическую функцию вектора фазовых координат. Из уравнений для характеристической функции получены уравнения для моментов (11), позволяющие путем разложения плотностей вероятностей в ряд Эджворта получить систему приближенных уравнений (12) для первых двух моментов вектора фазовых координат.

Приведенный пример показывает, что для некоторых динамических систем первого порядка, которые могут содержать нелинейности разрывного типа, уравнение для характеристической функции позволяет находить закон распределения в установившемся режиме. Сравнение точного решения (17) с приближенным (19) показывает, что относительная ошибка в определении среднеквадратического отклонения не превышает в данном примере 11%.

В. А. Бухалев
Институт проблем управления
СССР, Москва В-485,
Профсоюзная ул., 81

Problems of Control and Information Theory, Vol. 4 (2), pp. 147–154 (1975)

A SEPARATING TRANSFORMATION TO THE MINIMUM VARIANCE CONTROL

J. HETTHÉSSY, R. HABER, L. KEVICZKY

(Budapest)

(Received May 1, 1974)

In this paper a useful dead time transformation is derived for single input–single output, discrete-time linear systems. The suggested method can be used for the determination of the minimum variance control strategy of systems, which are described by an Åström model.

Introduction

Considering the optimal control of single input—single output, discrete-time linear systems the optimal control signal $u(t)$ should always be determined on the basis of the observations available at the time t. If the system is described by an Åström model [1]

$$A(z^{-1})\, y(t+d) = B(z^{-1})\, u(t) + \lambda C(z^{-1})\, e(t+d),$$

where

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + \ldots + a_n z^{-n} \\ B(z^{-1}) &= b_0 + b_1 z^{-1} + \ldots + b_m z^{-m} \\ C(z^{-1}) &= 1 + c_1 z^{-1} + \ldots + c_k z^{-k}, \end{aligned}$$

then the expression of $y(t+d)$ involves $y(t+d-1)$, $y(t+d-2)$, $\ldots$, $y(t+1)$. If the purpose of the control is to minimize the loss function

$$V = E\{y^2(t+d)\}$$

then in order to separate the values before and after t a transformation must be introduced [2].

This paper presents a simple method, which is called dead time transformation [3], [4].

Main result

The dead time transformation is a required method to constitute the difference equation

$$y(t+d) + \sum_{j=0}^{n-1} \alpha_j y(t-j) = \sum_{j=0}^{m+d-1} \beta_j u(t-j) + \lambda \sum_{j=0}^{k+d-1} \gamma_j e(t+d-j) \quad (1)$$

from the system equation

$$A(z^{-1})\,y(t+d) = B(z^{-1})\,u(t) + \lambda C(z^{-1})\,e(t+d)\,. \tag{2}$$

To determine the coefficients α_j $(j = 0, 1, \ldots, n-1)$, β_j $(j = 0, 1, \ldots, m+d-1)$ and γ_j $(j = 0, 1, \ldots, k+d-1)$ let us introduce the following vectors:

$$\begin{aligned}
\mathbf{y}^T(t+d) &= [y(t+d),\, y(t+d-1),\, \ldots,\, y(t-n+1)] \\
\mathbf{e}^T(t+d) &= [e(t+d),\, e(t+d-1),\, \ldots,\, e(t-k+1)] \\
\mathbf{u}^T(t) &= [u(t),\, u(t-1),\, \ldots,\, u(t-d-m+1)] \\
\mathbf{a}^T &= [1,\, a_1,\, \ldots,\, a_n,\, 0,\, 0,\, \ldots,\, 0] \\
\mathbf{b}^T &= [b_0,\, b_1,\, \ldots,\, b_m,\, 0,\, 0,\, \ldots,\, 0] \\
\mathbf{c}^T &= [1,\, c_1,\, \ldots,\, c_k,\, 0,\, 0,\, \ldots,\, 0] \\
\boldsymbol{\alpha}^T &= [1,\, 0,\, 0,\, \ldots,\, 0,\, \alpha_1,\, \alpha_2,\, \ldots,\, \alpha_n] \\
\boldsymbol{\beta}^T &= [\beta_0,\, \beta_1,\, \ldots,\, \beta_{m+d-1}] \\
\boldsymbol{\lambda}^T &= [1,\, \gamma_1,\, \ldots,\, \gamma_{k+d-1}],
\end{aligned}$$

where $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\boldsymbol{\alpha}$ contain $(d-1)$ zeros.

Determine the matrices $\mathbf{H}_A$, $\mathbf{H}_B$ and $\mathbf{H}_C$ in such a way that the following equations should hold

$$\mathbf{H}_A\,\mathbf{a} = \boldsymbol{\alpha} \qquad \mathbf{H}_B\,\mathbf{b} = \boldsymbol{\beta} \qquad \mathbf{H}_C\,\mathbf{c} = \boldsymbol{\lambda}. \tag{3}$$

Using the introduced notations Eq. (2) can be written as follows

$$\mathbf{y}^T(t+d)\,\mathbf{a} = \mathbf{u}^T(t)\,\mathbf{b} + \lambda\mathbf{e}^T(t+d)\,\mathbf{c}\,. \tag{4}$$

On the other hand Eq. (2) can be rewritten for any time t, e.g. for $t = t-1$:

$$\mathbf{y}^T(t+d-1)\,\mathbf{a} = \mathbf{u}^T(t-1)\,\mathbf{b} + \lambda\mathbf{e}^T(t+d-1)\,\mathbf{c}\,. \tag{5}$$

The construction of the parameter vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ makes it possible to formulate Eq. (5) by shifting the parameter vectors instead of $\mathbf{y}$, $\mathbf{u}$ and $\mathbf{e}$. Using the shift or Toeplitz matrix

$$\mathbf{S}_1 = \begin{bmatrix} 0 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 0 & 0 & \ldots\ 1 & 0 \end{bmatrix}$$

of the proper order for the parameter vectors, Eq. (5) has the following form:

$$\mathbf{y}^T(t+d)\,\mathbf{S}_1^A\,\mathbf{a} = \mathbf{u}^T(t)\,\mathbf{S}_1^B\,\mathbf{b} + \lambda\mathbf{e}^T(t+d)\,\mathbf{S}_1^C\,\mathbf{c}\,. \tag{6}$$

In order to ensure that $\boldsymbol{\alpha}$ should have $(d-1)$ zeros the shifted equations must be reduced $(d-1)$ times from Eq. (4) considering proper weighting factors h_j:

$$\begin{aligned}\mathbf{y}^T(t+d)\,[\mathbf{a} + h_1\mathbf{S}_1^A\mathbf{a} + h_2\mathbf{S}_2^A\mathbf{a} + \ldots + h_{d-1}\mathbf{S}_{d-1}^A\mathbf{a}] = \\ = \mathbf{u}^T(t)\,[\mathbf{b} + h_1\mathbf{S}_1^B\mathbf{b} + h_2\mathbf{S}_2^B\mathbf{b} + \ldots + h_{d-1}\mathbf{S}_{d-1}^B\mathbf{b}] + \\ + \lambda\mathbf{e}^T(t+d)\,[\mathbf{c} + h_1\mathbf{S}_1^C\mathbf{c} + h_2\mathbf{S}_2^C\mathbf{c} + \ldots + h_{d-1}\mathbf{S}_{d-1}^C\mathbf{c}],\end{aligned} \tag{7}$$

where the shifted equations are obtained by means of the shift matrices

$$\mathbf{S}_2 = \mathbf{S}_1 \cdot \mathbf{S}_1 = \begin{bmatrix} 0 & . & . & \ldots & . & 0 \\ 0 & 0 & . & \ldots & . & 0 \\ 1 & 0 & 0 & \ldots & . & 0 \\ 0 & 1 & 0 & \ldots & . & . \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ . & & & & 0 & . \\ 0 & 0 & . & .1 & 0 & 0 \end{bmatrix};$$

$$\mathbf{S}_3 = \mathbf{S}_2\,\mathbf{S}_1;\ \ldots;\ \mathbf{S}_{d-1} = \mathbf{S}_{d-2}\,\mathbf{S}_1\,.$$

Writing (7) in a matrix form we get

$$\mathbf{y}^T(t+d)\,\mathbf{S}_A\,\mathbf{a} = \mathbf{u}^T(t)\,\mathbf{S}_B\,\mathbf{b} + \lambda\mathbf{e}^T(t+d)\,\mathbf{S}_C\mathbf{c}, \tag{8}$$

where

$$\mathbf{S}_A = \mathbf{I}^A + \sum_{j=1}^{d-1} h_j\,\mathbf{S}_j^A$$

$$\mathbf{S}_B = \mathbf{I}^B + \sum_{j=1}^{d-1} h_j\,\mathbf{S}_j^B$$

$$\mathbf{S}_C = \mathbf{I}^C + \sum_{j=1}^{d-1} h_j\,\mathbf{S}_j^C$$

and $\mathbf{I}$ is the identity matrix of the proper order. In order to get the matrices $\mathbf{H}_A$, $\mathbf{H}_B$ and $\mathbf{H}_C$ let us write Eq. (1) by means of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$:

$$\mathbf{y}^T(t+d)\,\boldsymbol{\alpha} = \mathbf{u}^T(t)\,\boldsymbol{\beta} + \lambda\,\mathbf{e}^T(t+d)\,\boldsymbol{\gamma}. \tag{9}$$

Comparing the equation above with Eq. (8) the following equations must be hold:

$$\mathbf{S}_A\mathbf{a} = \boldsymbol{\alpha} \tag{10}$$

$$\mathbf{S}_B\mathbf{b} = \boldsymbol{\beta} \tag{11}$$

$$\mathbf{S}_C\mathbf{c} = \boldsymbol{\gamma}, \tag{12}$$

whereof it is seen the $\mathbf{S}_A$, $\mathbf{S}_B$ and $\mathbf{S}_C$ gives the required matrices $\mathbf{H}_A$, $\mathbf{H}_B$ and $\mathbf{H}_C$.

Entering into the details Eq. (10) gives

$$\begin{bmatrix} 1 & 0 & . & \dots & . & . & 0 \\ h_1 & 1 & 0 & \dots & . & . & 0 \\ h_2 & h_1 & 1 & \dots & . & . & . \\ . & h_2 & h_1 & \dots & . & . & . \\ . & . & . & \dots & . & . & . \\ . & . & . & \dots & . & . & . \\ h_{d-1} & . & . & \dots & 1 & . & . \\ 0 & h_{d-1} & . & \dots & h_1 & 1 & 0 \\ 0 & 0 & h_{d-1} & \dots & h_2 & h_1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ . \\ . \\ . \\ a_n \\ 0 \\ . \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ . \\ 0 \\ \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_n \end{bmatrix}. \tag{13}$$

Using the first d equations from this matrix equation we obtain the weighting factors h_j $(j = 1, 2, \dots, d-1)$ in a recursive manner:

$$h_j = -\sum_{i=1}^{j} a_i h_{j-i} \qquad (h_0 = 1; \quad j = 1, 2, \dots, d-1).$$

In the knowledge of the weighting factors h_j, α_j $(j = 0, 1, \dots, n-1)$, β_j $(j = 0, 1, \dots, m+d-1)$ and γ_j $(j = 0, 1, \dots, k+d-1)$ can be determined by means of Eqs. (10)–(12).

To summarize the main property of the described transformation we can say that the system equation

$$A(z^{-1})\, y(t+d) = B(z^{-1})\, u(t) + \lambda C(z^{-1})\, e(t+d)$$

can be written in the following form:

$$y(t+d) = T(z^{-1})\, y(t) + S(z^{-1})\, u(t) + \lambda R(z^{-1})\, e(t+d),$$

where

$$T(z^{-1}) = t_0 + t_1 z^{-1} + \dots + t_{n-1} z^{1-n}$$

$$S(z^{-1}) = s_0 + s_1 z^{-1} + \dots + s_{m+d-1} z^{1-m-d}$$

$$R(z^{-1}) = 1 + r_1 z^{-1} + \dots + r_{k+d-1} z^{1-k-d}$$

and further

$$t_j = -\alpha_j \quad (j = 0, 1, \ldots, n-1)$$
$$s_j = \beta_j \quad (j = 0, 1, \ldots, m+d-1)$$
$$r_j = \gamma_j \quad (j = 0, 1, \ldots, k+d-1).$$

Examples

The properties of the derived transformation will be shown by two simple examples.

First a second-order system with $d = 2$ will be considered

$$y(t) + a_1 y(t-1) + a_2 y(t-2) = b_0 u(t-2) + b_1 u(t-3) + \\ + \lambda[e(t) + c_1 e(t-1) + c_2 e(t-2)], \tag{14}$$

and the desired form by Eq. (1) will be determined by means of a straightforward computation and using the suggested method, too. Equation (14) has the following form for $t = t+1$ and $t = t+2$:

$$y(t+2) + a_1 y(t+1) + a_2 y(t) = b_0 u(t) + b_1 u(t-1) + \\ + \lambda[e(t+2) + c_1 e(t+1) + c_2 e(t)], \tag{15}$$

as well as

$$a_1 y(t+1) + a_1^2 y(t) + a_1 a_2 y(t-1) = a_1 b_0 u(\ -1) + a_1 b_1 u(t-2) + \\ + a_1 \lambda[e(t+1) + c_1 e(t) + c_2 e(t-1)]. \tag{16}$$

Notice that Eq. (16) can be obtained multiplying both sides by a_1. Substracting Eq. (16) from Eq. (15) and rearranging we get the desired form

$$y(t+2) + (a_2 - a_1^2)\, y(t) - a_1 a_2 y(t-1) = \\ = b_0 u(t) + (b_1 - a_1 b_0)\, u(t-1) - a_1 b_1 u(t-2) + \\ + \lambda[e(t+2) + (c_1 - a_1)\, e(t+1) + (c_2 - a_1 c_1)\, e(t) - a_1 c_2 e(t-1)]. \tag{17}$$

Using the equations of the dead time transformation, Eq. (13) gives

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ h_1 & 1 & 0 & 0 \\ 0 & h_1 & 1 & 0 \\ 0 & 0 & h_1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} \tag{18}$$

whence

$$\begin{aligned} h_1 &= -a_1 \\ \alpha_1 &= \quad a_2 - a_1^2 \\ a_2 &= -a_1 a_2 \,. \end{aligned}$$

Further by Eqs. (11) and (12) we get

$$\begin{bmatrix} 1 & 0 & 0 \\ h_1 & 1 & 0 \\ 0 & h_1 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \tag{19}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ h_1 & 1 & 0 & 0 \\ 0 & h_1 & 1 & 0 \\ 0 & 0 & h_1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ c_1 \\ c_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix}, \tag{20}$$

and thus

$$\begin{aligned} \beta_0 &= b_0 \\ \beta_1 &= b_1 - a_1 b_0 \\ \beta_2 &= -a_1 b_1 \\ \gamma_1 &= c_1 - a_1 \\ \gamma_2 &= c_2 - a_1 c_1 \\ \gamma_3 &= -a_1 c_2 \,. \end{aligned}$$

In the following example a second-order system with $d = 3$ will be investigated:

$$\begin{aligned} y(t+3) + a_1 y(t+2) + a_2 y(t+1) = b_0 u(t) + b_1 u(t-1) + \\ + \lambda[e(t+3) + c_1 e(t+2) + c_2 e(t+1)]. \end{aligned} \tag{21}$$

Constituting $a_1 y(t+2)$ and substituting it into Eq. (21) we obtain

$$\begin{aligned} y(t+3) + (a_2 - a_1^2) y(t+1) - a_1 a_2 y(t) = b_0 u(t) + (b_1 - a_1 b_0) u(t-1) - \\ - a_1 b_1 u(t-2) + \lambda[e(t+3) + (c_1 - a_1) e(t+2) + \\ + (c_2 - a_1 c_1) e(t+1) - a_1 c_2 e(t)]. \end{aligned} \tag{22}$$

Taking again Eq. (21) for $t = t + 1$ and multiplying it by $(a_2 - a_1^2)$:

$$\begin{aligned} (a_2 - a_1^2) y(t+1) + (a_2 - a_1^2) a_1 y(t) + (a_2 - a_1^2) a_2 y(t-1) = \\ = (a_2 - a_1^2) b_0 u(t-2) + (a_2 - a_1^2) b_1 u(t-3) + \\ + \lambda[(a_2 - a_1^2) e(t+1) + (a_2 - a_1^2) c_1 e(t) + (a_2 - a_1^2) c_2 e(t-1)]. \end{aligned} \tag{23}$$

Substracting Eq. (23) from Eq. (22) and rearranging we get:

$$\begin{aligned} y(t+3) + (a_1^3 - 2a_1 a_2)\, y(t) + (a_1^2 a_2 - a_2^2)\, y(t-1) = \\ = b_0\, u(t) + (b_1 - a_1 b_0)\, u(t-1) + (a_1^2 b_0 - a_1 b_1 - a_2 b_0)\, u(t-2) + \\ + (a_1^2 b_1 - a_2 b_1)\, u(t-3) + \\ + \lambda[e(t+3) + (c_1 - a_1)\, e(t+2) + (c_2 - a_1 c_1 - a_2 + a_1^2)\, e(t+1) + \\ + (a_1^2 c_1 - a_2 c_1 - a_1 c_2)\, e(t) + (a_1^2 c_2 - a_2 c_2)\, e(t-1)]. \end{aligned} \tag{24}$$

Instead of this tiresome, step-by-step procedure the suggested method can be used, by which from Eq. (13):

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ h_1 & 1 & 0 & 0 & 0 \\ h_2 & h_1 & 1 & 0 & 0 \\ 0 & h_2 & h_1 & 1 & 0 \\ 0 & 0 & h_2 & h_1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ a_1 \\ a_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

and thus

$$\begin{aligned} h_1 &= -a_1 \\ h_2 &= a_1^2 - a_2 \\ \alpha_1 &= a_1^3 - 2a_1 a_2 \\ \alpha_2 &= a_1^2 a_2 - a_2^2 . \end{aligned}$$

Using Eqs. (11) and (12) we get

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ h_1 & 1 & 0 & 0 \\ h_2 & h_1 & 1 & 0 \\ 0 & h_2 & h_1 & 1 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ h_1 & 1 & 0 & 0 & 0 \\ h_2 & h_1 & 1 & 0 & 0 \\ 0 & h_2 & h_1 & 1 & 0 \\ 0 & 0 & h_2 & h_1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ c_1 \\ c_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{bmatrix},$$

whence

$$\begin{aligned} \beta_0 &= b_0 \\ \beta_1 &= b_1 - a_1 b_0 \\ \beta_2 &= a_1^2 b_0 - a_2 b_0 - a_1 b_1 \\ \beta_3 &= a_1^2 b_1 - a_2 b_1 \end{aligned}$$

$$\gamma_1 = c_1 - a_1$$
$$\gamma_2 = a_1^2 - a_2 - a_1 c_1 + c_2$$
$$\gamma_3 = a_1^2 c_1 - a_2 c_1 - a_1 c_2$$
$$\gamma_4 = a_1^2 c_2 - a_2 c_2 ;$$

It can be seen that this method gives the desired results by means of a simpler procedure.

References

1. *Åström, K. J.*, Introduction to Stochastic Control Theory. Academic Press, New York, 1970.
2. *Åström, K. J., Wittenmark, B.*, On Self-Tuning Regulators. Automatica **9** (1973) 185–199.
3. *Hetthéssy, J., Keviczky, L.*, On the Minimum Variance Control Strategies. Periodica Polytechnica **18** (1974) 291–308.
4. *Hetthéssy, J., Keviczky, L.*, Some Innovations to the Minimum Variance Control. IFAC Symposium on Stochastic Control, Budapest, 1974

О сепарабельном преобразовании в задачах управления с минимальной среднеквадратической ошибкой

Й. ХЕТЕШИ, Р. ХАБЕР, Л. КЕВИЦКИ

(Будапешт)

В статье излагают трансформацию запаздывания для одной входной и одной выходной, линейных, дискретных (квантованных по времени) систем с постоянными параметрами при управлении с минимальным разбросом для разъединения, появляющегося проблемой. Представленный метод основывается на использовании шифта (сдвига) или Тоеплиц матрицы и даёт желательные результаты довольно простым и наглядным образом. Кроме того, что в статье детально отмечены общие связи трансформации, с помощью двух численных примеров наглядно показаны и оказывающиеся преимущества изложенного метода перед методом постепенного вычисления («шаг за шагом»).

R. Haber and
L. Keviczky
Department of Automation
Technical University of Budapest
H-1111 Budapest, Goldmann György tér 3.
Hungary

J. Hetthéssy
Institute for Electric Power Research
H-1051 Budapest, Zrinyi u. 1.
Hungary

Problems of Control and Information Theory, Vol. 4 (2), pp. 155--163 (1975)

ESTIMATING THE DYNAMICS OF MODIFIED SYSTEMS WITH VARIABLE STRUCTURE

S. M. TSONKOV

(Sofia)

(Received June 3, 1974)

The high dynamical properties of the systems with variable structure (SVS) elaborated in the USSR under the guidance of S. V. Emelianov, Corresponding Member of the Academy of Sciences of the USSR, are widely known today. In our country experiments with a view to the creation of constant time elements for the SVS for converting normally closed systems to SVS-class local contour automation in the slipping mode of operation were carried out. The high dynamical characteristics of the SVS are attained by slight modifications and the utilization of existing elements.

The paper gives the evaluation of the dynamics of the modified SVS and compares them with the characteristics of the classical automatic control systems; it proves the justification of this new approach to the control of non-minimum phase objects of the SVS-class with improved characteristics.

The performed experiments proved the great advantage of the suggested modified SVS over the linear systems.

The introduction of more productive technological processes, the application of better machines and apparatuses, the efforts aimed at attaining maximum efficiency of the aggregates used in production, increased productivity, improved production quality, economy in the consumption of electric power and the costly raw materials led to the creation of new type systems. Various modifications of the linear automatic control systems (ACS) of variable structure, whose theoretical foundations were elaborated in the USSR at the Institute of Control Problems, awarded by the Lenin Prize, by a research group under the guidance of S. V. Emelianov, Corresponding Member of the Academy of Sciences of the USSR, were successfully introduced recently in several industrial branches (energetics, metallurgy, petrochemistry). The performance of these automatic control systems compared to the ACS of constant structure is sufficiently convincing and is treated widely in literature [1, 2].

For realizing the ACS of variable structure serious attention must be devoted to the simplicity and reliability of the technical means, their interconnections and the technical instrumentation of the objects to be controlled. The ACS of variable structure is designed advisedly for actual industrial objects equipped with industrial control, by completing the latter with nonlinear functional commutators.

Be the object to be controlled O and be the control P the invariable part (IP) of the given ACS, which is then completed by the control device CD (Fig. 1) for varying the structure of the ACS according to the previously selected control algorithm.

The possibility of introducing the linear ACS into the class of the systems of variable structure (SVS) has been proved by several studies [3, 4, etc.].

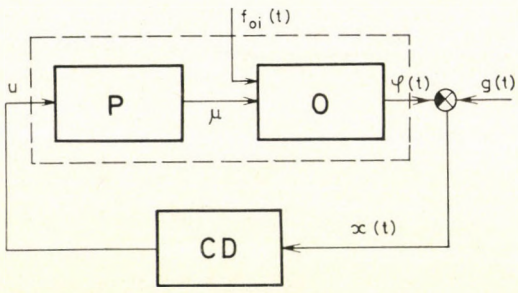


Fig. 1

In this case a modified SVS (MSVS) with a defined constant kinetic moment in the slipping mode of operation is created in a way that the MSVS preserves all the positive properties of the SVS. The structure and the parameters of CD for controlling first and second order objects by PI and PID controllers respectively are defined in the above-mentioned papers.

If the transfer function of the MSVS is

$$W_0(p) = \frac{K_0}{T_0 p + 1} \tag{1}$$

and that of the PI-controller is

$$W_{\mathrm{PI}}(p) = K_c \frac{T_i p + 1}{T_i p} \tag{2}$$

then the structural scheme of CD designed for controlling the free movement of the MSVS will be as shown in Fig. 2, where K_0, T_0 are the 0 parameters and K_c, T_i — the P parameters.

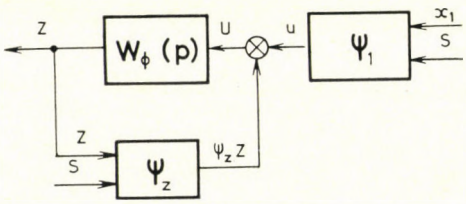


Fig. 2

The differential equation of the MSVS is obtained as:

$$\frac{dx_1}{dt} = x_2$$

$$\frac{dx_2}{dt} = -\frac{1}{T_0} x_2 + \left(\frac{K_0 K_c}{T_0} D + \frac{K_c K_0}{T_i T_0} \right) u ; \quad \left(x \equiv x_1 ; \frac{d}{dt} \equiv D \right) . \tag{3}$$

By (3) it is seen that the input effects of the MSVS are differentiated and as these are jump-type time functions, some of the phase coordinates of the system are cut off, when the describing points (DP) fall on the shifting hyposurface S, while the quasi-invariance of the external and internal non-stationary states in the MSVS is reached, if the constant filters m are introduced into the scheme of the control, with 1.5-fold backfeed. Here m is the degree of the differentiation of the control function U. The unequivocal state of the MSVS is characterized by the coordinate to be controlled x_i and the central coordinate z_i. Similar problems arise in objects of non-minimum phase of the SVS-class, so the relationships obtained in this case may be accepted also here [1], with the control U given as the linear combination of the error x_i and its $(n-1)$ differential quotients with the coordinates z_i $(i = 1, 2, \ldots, m)$, with the coefficients varying in the jumping mode.

In the considered case

$$U = \psi_1 \cdot x_1 + \psi_z \cdot z , \tag{4}$$

where the transfer function of the filter is

$$W_f(p) = \frac{1}{T_f p + 1} \tag{5}$$

and the commutation coefficients are

$$\psi_1 = \begin{cases} \alpha \text{ for } Sx_1 > 0 \\ \beta \text{ for } Sx_1 < 0 \end{cases} \tag{6a}$$

and

$$\psi_z = \begin{cases} \gamma \text{ for } S \cdot z > 0 \\ \delta \text{ for } S \cdot z < 0. \end{cases} \tag{6b}$$

The conditions of attaining the constant slipping movement over the shifting line S are given by the equation

$$S = cx_1 + x_2 . \tag{7}$$

According to [1, 3] we obtain:

$$\frac{K_c \alpha}{T_i} > \sup_{T,a} (ac - Tc^2)$$

$$\frac{K_c \beta}{T_i} < \inf_{T,a} (ac - Tc^2)$$

$$\gamma > 1 - \frac{T_f}{T_i}$$

$$\delta < 1 - \frac{T_f}{T_i}; \quad \left(a = \frac{1}{K_0}; \quad T = \frac{T_0}{K_0}\right). \tag{8}$$

Consequently, values for the commutational coefficients ensuring the constant slipping movement in the MSVS may be selected. This movement is determined by the single equation of the shifting surface and does not depend on the variable parameters of the object.

The transfer process of the MSVS in this phase may be written up in the form of the exponential equation:

$$x_1(t) = Ae^{-ct} \qquad (A = \text{const}). \tag{9}$$

By varying the coordinate z, we obtain from (5)

$$z(t) = e^{-\int \frac{1}{T_i} dt} \int e^{\int \frac{1}{T_i} dt} \cdot \frac{B(t)}{K_c T_i} e^{-ct} + z(0)\, e^{-\int \frac{1}{T_i} dt} \tag{10}$$

where

$$B(t) = B(K_c, T_i, K_0, T_0); \; z(0) = z(t = 0).$$

From (10) follows $\lim_{t \to \infty} z \to 0$ for $T_i > 0$, showing that the proposed MSVS is realizable.

From Eq. (8) follows that if $T_\Phi = T_i$ is assumed in the MSVS, then the control U may be formed without the z coordinates. Also the ratio of the parameters of the PI-controller may be given as $\frac{K_c}{T_i} \to \max$, for attaining optimum filtration relative to the disturbed $f_{0i}(t)$ and the U control effects.

Similar results are obtained for the linear controls with independent variables. The author [4] proves that first-order objects may be controlled by a commutation of only K_c with the help of the PI-controller, and second-order objects by the K_c and T_i commutations with the help of PID-controllers.

If the preliminary control of the dynamic characteristics of the MSVS is absent, the control period is t_c and the maximum deviation of the controlled quantity is X_M, whose variation with the varying parameters of the object is of importance.

Let us assume that the MSVS consists of first-order objects $W_0(p) = \left(W_0(p) = \frac{\varepsilon_0}{p}\right)$ and a PI-controller. In this case the differential equation of the MSVS at $T_f = T_{i\,\min}$ is obtained as

$$\frac{dx_2}{dt} + \frac{\varepsilon_0 K_c}{T_i} i = \dot{F}(t) \tag{11}$$

where $g(t)$ is given by the system and $F(t) = \dot{g}(t) + \varepsilon_c f_{0i}(t)$.

From (8) ($a = 0$), the commutation coefficients α and β are obtained as [3]:

$$\alpha = \frac{c_2}{\varepsilon_{0\,\min}} - \left(\frac{T_i}{K_c}\right)_{\min} \quad \text{and} \quad \beta = -\alpha. \tag{12}$$

The slope of the shifting line is obtained from (11) as

$$c = \omega_{0\,\min} = \frac{\varepsilon_{0\,\min} K_c}{T_i} \beta.$$

In this case S coincides with the asymptotic family of the hyperbolas corresponding to the second structure. This is the limit C-value determining the position of S, when α and β are given.

The jumplike effects $f_{0i}(t) = 1(t)$ may be considered as nonzero initial conditions:

$$x_{10} = 0 \quad \text{and} \quad x_{20} = \varepsilon_0 f_0$$

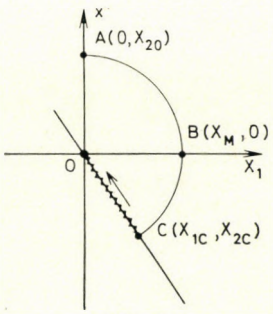


Fig. 3

The period of the transfer process may be obtained by integrating the movement phase trajectories of the individual phases:

a) t_{ac} — movement over the ellyptic trajectory;

b) t_{c0} — movement over S up to $x_1 = 0.01\, X_M$.

The coordinate points of C are obtained by

$$x_{1c} = 0 = \frac{X_M \omega_1}{\sqrt{\omega_1^2 + \omega_{1\min}^2}}. \tag{13}$$

With $\varepsilon_0 = \varepsilon_{0\min}$

$$x_{1c} = X_M \frac{\sqrt{2}}{2} \tag{13a}$$

and with $\varepsilon_0 = \varepsilon_{0\max}$

$$x_{1c} = X_M \frac{1}{\sqrt{1 + \left(\frac{\omega_{1\min}}{\omega_{1\max}}\right)^2}}. \tag{13b}$$

From (11) we obtain:

$$\omega_1^2 = \varepsilon_0 \frac{K_c}{T_i} \alpha. \tag{13c}$$

In this case the period of the movement of the MSVS is [5]:

a) in the first phase:

$$t_{ac} = \frac{1}{\omega_1}\left[\int_0^{X_M} \frac{dx_1}{\sqrt{X_M - x_1^2}} - \int_{X_M}^{X_M \frac{\sqrt{2}}{2}} \cdot \frac{dx_1}{\sqrt{X_M - x_1^2}}\right]; \tag{14a}$$

b) in the second phase:

$$t_{c0} = \int_{x_{1c}}^{0.05 X_M} \frac{dx_1}{c x_1} = -\frac{1}{\omega_1} \ln 0.05 \sqrt{1 - \left(\frac{\omega_{1\min}}{\omega_1}\right)^2}. \tag{14b}$$

The maximum dynamic deviations is given by

$$X_M = \frac{\varepsilon_{0\max}}{\omega_1} f_0. \tag{15}$$

If this must be limited ($X_M \leq X_{M\,\text{permitted}}$, then from (13) and (15) we obtain:

$$\frac{K_c}{T_i} \alpha \geq \varepsilon_0 \frac{f_0^2}{X_{M\,\text{permitted}}^2}. \tag{16}$$

With $X_M \leq X_{M\,\text{permitted}}$ we obtain for $\varepsilon_0 = \varepsilon_{0\,\min}$ from (13), (14) and (16):

$$t_{p\max} = \frac{3\pi}{4\omega_{1\min}} - \frac{1}{\omega_{1\min}} \ln \frac{X_{M\,\text{permitted}}}{f_0 \sqrt{\frac{\varepsilon_{0\min}}{\alpha} \cdot \frac{T_i}{K_c}}}, \tag{17}$$

and for $\varepsilon_0 = \varepsilon_{0\max} \gg \varepsilon_{0\min}$

$$t_{c\min} = \sqrt{\frac{T_i}{\varepsilon_{0\max} \cdot K_c \alpha}} \left(\frac{\pi}{2} - \ln \frac{0.05\, X_{M\,\text{permitted}}}{f_0 \sqrt{\frac{\varepsilon_{0\max}}{\alpha} \cdot \frac{T_i}{K_c}}} \right). \tag{18}$$

Figure 4 shows the calculated dependence of $t_c = f\left(\varepsilon_0, \frac{K_c}{T_i}\right)$ for $\frac{X_{M\,\text{permitted}}}{f_0} = 1$ and Fig. 5 shows $\frac{X_M}{f_0} = f(\varepsilon_0)$ for $\frac{K_c}{T_i} \cdot \alpha = \text{const}$. For comparison also $\frac{X_M}{f_0} = f(\varepsilon_0)$ for $\frac{K_c}{T_i} = 2$ and $t_p = f\left(\varepsilon_0, \frac{K_c}{T_i}\right)$ for $\varepsilon_0 \frac{K_c}{T_i} = 8$ are shown, as well as $K_c = \sqrt{2}$ for the linear ACS with the same parameters.

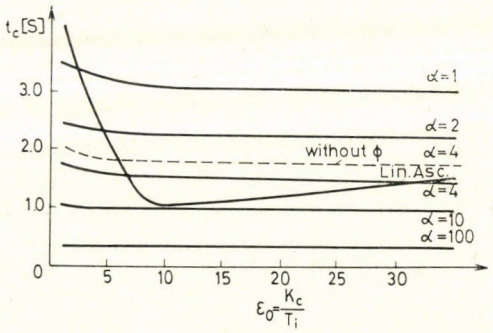


Fig. 4

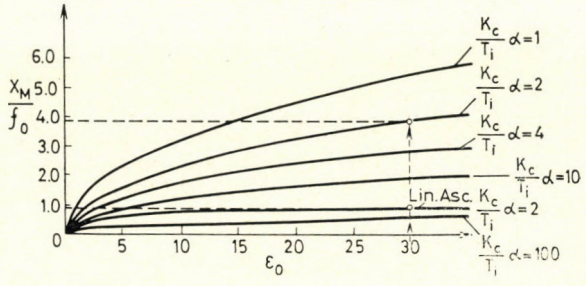


Fig. 5

By this comparison it is seen that the positive qualities of the MSVS as to the control period t_c (independence of the variation of the parameter $\varepsilon_0 = \text{var}$ and the possibility of obtaining $t_c \leq t_{c_{\text{given}}}$), the dynamical error exceeds that of the linear ACS. This deficiency appears also in the classical SVS with non-minimum phase objects and is due to the constantly connected damping filters reducing — in the first structure, where the maximum deviation is formed — the effect of the jump-type control effects (with the PI-controller this is second order).

In this case the formation of control functions of the SVS-class for controlling non-minimum phase objects (or MSVS with PI- and PID-controllers) including damping filters proves to be efficient for transferring the describing points of the process to the shifting surface S. As the suggested synthesis method of CD for the SVS with zero transfer functions of the invariable parts of the system permits to combine fully the quality reducing the maximum dynamical error, supplied by the PI-controller with the properties of the SVS in the slipping mode of operation, the invariance in the system is realized in the closing phase of the movement.

If we assume $\alpha = 1$ for $\frac{K_c}{T_i} = 2$ (Fig. 5), then the dynamical error in such a MSVS coincides with that of the ACS. It is seen that it decreases by several times (for $\varepsilon_0 = 30$ by 4-times) proving the advantage of the suggested method for the synthesis of SVS of this type. The control period in this case (without Φ) differs only insignificantly for $\alpha = 4$, as seen in Fig. 4.

The evaluation of the dynamics of the MSVS based on the obtained results shows the advantage of the MSVS as compared with the linear ACS. In this case the modification is gained with insignificant variations and easily realizable complementary control devices. With the introduction of the MSVS in the slipping mode of operation the external quasi-variance, the instability of the parameters and the maximum error may be reduced within the given range. The maximum filtration in the system is realized by ensuring the ratio $\frac{K_c}{T_i} \rightarrow \max$ of the parameters of the controller.

The discussions support the applicability of the modified systems with variable structure for the high-quality control of technological processes.

References

1. *Emelianov, S. V.*, The theory of the systems of variable structure. Nauka, M., 1970.
2. *Utkin, V. I.*, Slipping modes of operation in the system of variable structure. Nauka, M., 1974.
3. *Tsonkov, S. M.*, Reduction of linear automatic control systems in the class of systems with variable structures. Bulletin de l'Institute de Mécanique Appliquée et de l'Électrotechnique „Lenine", **XXVIII,** *4* (1973).
4. *Tsonkov, S. M.*, The synthesis of modified systems with variable structure, with linear controllers of independent parameters for forced movement control. Tekhnicheskaya Misl, Sofia, 1 (1973).
5. *Eidinov, R. M.*, On estimating the period of transfer processes in the automatic control system with variable structure. Bulletin of the Academy of Sciences of the USSR, TK, *5* (1965).

Оценка динамики модифицированных систем с переменной структурой

С. М. ЦОНКОВ

(София)

В настоящее время хорошо известны высокие динамические свойства систем с переменной структурой (СПС), созданных в СССР под руководством чл.-корр. АН СССР С. В. Емельянова. Ввиду имеющихся трудностей создания элементов СПС, в нашей стране были сделаны попытки преобразований классических систем для автоматизации локальных контуров в классе СПС в режиме скольжения. При этом незначительная модификация и использование существующих элементов позволяют получить высокие динамические показатели.

В статье дана оценка динамики таких модифицированных систем и сравнение их с классическими системами автоматического регулирования. Также обоснован новый подход к управлению неминимальнофазовыми объектами в классе СПС с улучшенными показателями.

Сделанные исследования выявили большие преимущества предложенных модифицированных систем с присущими СПС свойствами по сравнению с линейными системами.

С. М. Цонков
Высший электротехнический институт им. Ленина
София
Болгария

Problems of Control and Information Theory, Vol. 4 (2), pp. 165—181 (1975)

THE REDUCTION OF THE DIMENSION OF INPUT PATTERN RECOGNITION VECTORS BY SEARCHING THE OPTIMAL PARAMETER SET FOR LINEAR REGRESSION

GY. MUSZÉLY

(Budapest)

(Received May 15, 1974)

A method is presented for reducing the number of input parameters of pattern recognition objects. It proposes to select the most informative measurements after linear features have been found in the space of all original input parameters. The selection is solved by transferring the problem to the reduction of the number of parameters for linear regression.

This more general problem is considered in the case of more than one dependent variables. A criterion is suggested as a measure of the goodness of fit of regression. The paper presents a procedure which gives the best subset of independent variables without enumerating all the subsets which are subject to the constraint of limitation.

As an example of illustration an experiment connected with the recognition of intonation of a spoken Russian sentence was performed.

Introduction

The input parameters are the different measurements taken under the given object of recognition. Selection of input parameters is one of the basic phases of designing a pattern recognition procedure. Unfortunately this step can be algorithmized least of all. The original parameters are, for the most part, chosen on certain physical or heuristic principles and constitute a large and redundant set. The reduction of the size of this set may be performed in order to minimize the cost of measuring the parameters or for other purpose. Owing to the fact that the input parameter set is gathered in a heuristic way it seems proper to minimize its dimension after features have been extracted in the space of all original input parameters. The same approach is suggested by Turbovich, Gitis and Maslov [1]. In this case the process of choosing the minimal set among the input parameters can be carried out after the irrelevant information has been removed.

By the criterion presented in the paper those parameter sets are considered to be efficient which have linear combinations close to the features. In this manner the problem can be solved by searching the optimal set of independent variables for linear regression. In this context the case of simultaneous regression of more than one dependent variables must be examined. A criterion is proposed to measure the goodness of fit of the regression function.

The paper presents a procedure for finding the optimal subset of parameters subject to certain constraint.

It gives the exact solution without enumerating all the subsets satisfying the limiting condition.

As an example of illustration an experiment connected with the recognition of a spoken Russian sentence was performed.

1. The basic approach to the reduction of the number of input parameters for pattern recognition

1.1. *The first step: Extraction of linear features in the space of all parameters*

Linear combinations of input parameters are called linear features if (1) they form a set which is substantially smaller than that of the initial parameters and (2) in the space defined by them the pattern classes can be separated from each other not substantially worse than in the space of input parameters. Many techniques have been developed for extracting linear features. The most simple approach is the Karhunen–Loeve expansion applied to pattern recognition by Watanabe [2] and Chien and Fu [3]. Methods of discriminant analysis are more adequate to this problem (see [1]). In [1] an appropriate weighting of difference vectors between samples from different classes ensures the discriminant analysis methods to separate the closest pattern classes.

1.2. *Reduction of the input parameter set*

The process of choosing the minimal subset of the original set of input parameters can be carried out after linear features have been extracted. The space of the reduced parameter set also has to contain a subspace of linear features where the discrimination of classes is guaranteed.

For each subset of parameters, however, one can rarely calculate the linear features and test their discriminating ability because of the large amount of computations. Therefore we try to replace the requirements by more rigorous ones which can be checked easier. We require the reduced parameter set to have linear functions which are close to the linear features determined by the whole set of input parameters. For a fixed subset of parameters the coefficients of the linear functions can be obtained from the sample by least squares estimation. Depending on the given application, both the measurement vectors of individual patterns and the difference vectors between patterns from different classes can be considered as sample vectors. In the latter case an appropriate weighting of the small difference vectors results in the

existence of a subspace of reduced parameter sets where the structure of close pattern classes is similar to the structure of the same pattern classes in the space of the features of the full parameter set.

The above approach suggests to treat the problem as a reduction of the set of independent variables for linear regression. We note that in our case the features are taken for dependent variables and as a rule their number is more than one. For this generalized case a criterion of the goodness of fit of the regression will be chosen in the next section.

Suppose we know the costs of measuring the individual input parameters. We may formulate two different problems of optimization:

1°: To find the subset of parameters maximizing the measure of the goodness of fit among the subsets for which the cost of measuring is less than a given level.

2°: To find the subset of parameters minimizing the cost of measuring among the subsets for which the measure of the goodness of fit exceeds a given level.

Both problems will be treated in a generalized form in the next section.

2. Selection of regression variables for a set of dependent variables

2.1. *The criterion of the goodness of fit*

Suppose we have an $m \times N$ matrix $\mathbf{Y}$ of deviations of the N measurements of m random variables from the sample means of each variables. We try to estimate $\mathbf{Y}$ by $\hat{\mathbf{Y}} = \hat{\mathbf{A}}\mathbf{X}$ where $\mathbf{X}$ is the $n \times N$ matrix of deviations of the independent variables. The rows of $\hat{\mathbf{A}}$ are obtained by least squares estimation.

Suppose that $\mathbf{X}$ is of rank n, $\mathbf{Y}$ is of rank m and $m + 1 < n < N$. Then by the method of least squares we have

$$\hat{\mathbf{A}} = \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1} . \tag{1}$$

In the case of $m = 1$ the usual measure of the goodness of fit of the regression function $\hat{\mathbf{Y}}$ is the square of the multiple correlation coefficient:

$$R^2(\mathbf{Y};\mathbf{X}) = 1 = \frac{\sum_{j=1}^{N} (y_j - \hat{y}_j)^2}{\sum_{j=1}^{N} y_j^2} , \tag{2}$$

where y_j and $\hat{y}_j$ are the jth elements of row matrices $\mathbf{Y}$ and $\hat{\mathbf{Y}}$.

For $m > 1$ there exist simple generalizations of (2), but they do not fulfill some elementary requirements. For example, consider the two formulas:

$$\sum_{i=1}^{m}\left(1-\frac{\sum_{j=1}^{N}(y_{ij}-\hat{y}_{ij})^2}{\sum_{j=1}^{N}y_{ij}^2}\right), \qquad 1-\frac{\sum_{i=1}^{m}\sum_{j=1}^{N}(y_{ij}-\hat{y}_{ij})^2}{\sum_{i=1}^{m}\sum_{j=1}^{N}y_{ij}^2}.$$

The first of them does not remain invariant under the orthogonal, the second one under the diagonal transformation of matrix $\mathbf{Y}$.

The criterion of the goodness of fit should be determined by the deviation which occurs in the course of the practical application when the dependent variables are replaced by their least square estimates. In most applications, however, either the variables are used in different situations or their role cannott be exactly formulated. We assume that we are applying several linear functions of the dependent variables,

$$L(y_j, b) = b' y_j,$$

where b is any $k \times 1$ vector and y_j is the jth column of $\mathbf{Y}$. The criterion of the goodness of fit can be defined as

$$M = \inf_{b \neq 0} R^2(b'\mathbf{Y}; \mathbf{X}). \tag{3}$$

The following lemma relates to the calculation of M.

Lemma 1. M is the minimal root in λ^2 of the equation:

$$|\lambda^2 \cdot \mathbf{YY}' - \mathbf{YX}'(\mathbf{XX}')^{-1}\mathbf{XY}'| = 0. \tag{4}$$

Proof. By using (1) $R^2(b'\mathbf{Y}; \mathbf{X})$ can be written as

$$R^2(b'\mathbf{Y}; \mathbf{X}) = \frac{b'\hat{\mathbf{Y}}\hat{\mathbf{Y}}'b}{b'\mathbf{YY}'b} = \frac{b'\mathbf{YX}'(\mathbf{XX}')^{-1}\mathbf{XY}'b}{b'\mathbf{YY}'b}.$$

We note that $R^2(b'\mathbf{Y}; \mathbf{X})$ remains invariant if we multiply b by a scalar, therefore it has a minimum value among nonzero b.

For the numerator of the derivative of the last term with respect to the vector b we have

$$\mathbf{YX}'(\mathbf{XX}')^{-1}\mathbf{XY}'b(b'\mathbf{YY}'b) - \mathbf{YY}'b(b'\mathbf{YX}(\mathbf{XX}')^{-1}\mathbf{XY}'b) = \mathbf{0},$$

i.e.

$$\mathbf{YX}'(\mathbf{XX}')^{-1}\mathbf{XY}'b - \mathbf{YY}'b \cdot R^2(b'\mathbf{Y}; \mathbf{X}) = \mathbf{0}. \tag{5}$$

There exists a nontrivial b fulfilling (5) if and only if $R^2(b'\mathbf{Y}; \mathbf{X})$ satisfies (4). Consequently the minimal $R^2(b'\mathbf{Y}; \mathbf{X})$ is equal to the smallest root of the equation.

Remark 1. It follows from the lemma that M remains invariant under any linear transformation of $\mathbf{Y}$.

Remark 2. The roots of (4) are called the squares of the canonical correlation coefficients between the dependent and the independent variables. Thus M is the square of the smallest canonical correlation coefficient.

2.2. *Formulation of the optimization problem*

Let $E = \{1, \ldots n\}$ the set of indices of all the independent variables. Suppose we know the costs s_i $(i \in E)$ of measuring the individual independent variables. Then the cost of measuring a subset given by $G \subset E$ is

$$s_G = \sum_{i \in G} s_i \,.$$

In the most simple case, when $s_i \equiv 1$, the cost is equal to the number of elements in G. We may formulate two different problems of optimization:

Problem A Find G which maximizes M subject to the constraint $s_G \leq c$.

Problem B Find G which minimizes s_G subject to the constraint $M \geq c$.

2,3. *The basic theorem for reduced enumeration*

2.3.1. Both problems given in the previous section can be solved by enumerating all the subsets which satisfy the constraint inequality. However, for even a moderate number of independent variables the amount of computations would be quite large. The next section gives an algorithm for finding the exact solution without enumerating all the subsets satisfying the limiting constraint. Before describing the algorithm we present a theorem which enables it to be exact. Let the subset $G \subset E$ have k elements. Let $C_k^h(G)$ denote the class of those subsets of G all of which have h elements. Let x_i denote the ith row vector of measurements of the ith independent variable, i.e. x_i denotes the ith row of $\mathbf{X}$. Let $\mathbf{X}_H$ and $\mathbf{X}_G$ be the matrices of rows of $\mathbf{X}$ defined by the subsets H, $G \subset E$. Suppose the rank of $[\mathbf{X}_G', \mathbf{Y}']$ is $k + m$.

Theorem. With the above notations and assumptions

$$\max_{H \in C_k^h(G)} \min_{i \in G-H} R^2(x_i\,; [\mathbf{Y}'\,\mathbf{X}_H']') \geq 1 - \frac{(h+1)^2}{\left(\sum_{j=1}^{h+1} \psi_j\right)^2},$$

where

$$\psi_j = \begin{cases} 1 & \text{if} \quad j \leq k - m, \\ (1 - \lambda^2_{j-k+m})^{-1/2} & \text{if} \quad k - m < j \leq k, \end{cases}$$

and $\lambda_1^2 \leq \lambda_2^2 \leq \ldots \leq \lambda_m^2 < 1$ are the roots of the equation

$$|\lambda^2 \mathbf{YY}' - \mathbf{YX}'_G(\mathbf{X}_G\mathbf{X}'_G)^{-1}\mathbf{X}_G\mathbf{Y}'| = 0.$$

For any $\lambda_1, \ldots, \lambda_m$, h and $\mathbf{Y}$ there exists an $\mathbf{X}_G$ such that the inequality becomes an equality.

Remark 1. Let $M(G)$ denote the measure of the goodness of fit when the variables corresponding to G are in the regression. By Lemma 1 we get $\lambda_1^2 = M(G)$.

Remark 2. For $h \leq k - m - 1$ the right side of the inequality is equal to 0.

2.3.2. *Proof.* The theorem follows from the forthcoming Lemmas 2—4.

Lemma 2. Let $\mathbf{\Phi}$ be a diagonal matrix with elements $0 \leq \varphi_1 \leq \varphi_2 \leq \ldots \leq \varphi_m$. Let $v_1, \ldots, v_g$ an orthonormal system of m dimensional vectors. With these notations

$$\sum_{j=1}^{g} (v'_j \mathbf{\Phi} v_j)^{1/2} \geq \sum_{j=1}^{g} \varphi_j^{1/2}.$$

The lemma is a particular case of the theorem of 4.1.4 in [4].

2.3.3. Let $\mathbf{D}$ be an $m \times m$ matrix of full rank. Let $\mathbf{K}$ denote a symmetric positive definite $m \times m$ matrix with eigenvalues μ_i $(i = 1, \ldots, m)$. We have the following lemma:

Lemma 3. If for each $j = 1, \ldots m$

$$\{\mathbf{DKD}'\}_{jj} = 1 \tag{6}$$

then

$$\max_j \{(\mathbf{DD}')^{-1}\}_{jj} \geq \Big(\sum_{i=1}^{m} \mu_i^{1/2}\Big)^2/m^2.$$

For any $\mathbf{K}$ there exists a $\mathbf{D}$ for which the inequality becomes an equality.

Proof. Any nonsignificant $\mathbf{D}$ can be written in the form of $\mathbf{D} = \mathbf{P\Psi Q}$, where $\mathbf{P}$ and $\mathbf{Q}$ are orthogonal matrices and $\mathbf{\Psi}$ is a diagonal matrix with non-zero elements $\psi_1, \ldots, \psi_m$. With these notations we can write

$$\max_j \{(\mathbf{DD}')^{-1}\}_{jj} \geq \frac{1}{m} \operatorname{tr} (\mathbf{DD}')^{-1} = \frac{1}{m} \operatorname{tr} (\mathbf{P\Psi}^{-2}\mathbf{P}') = \frac{1}{m} \sum_{i=1}^{m} \psi_i^{-2}, \tag{7}$$

where tr denotes the sum of the diagonal elements of the matrix in the argument. Let Γ_1 denote the set of matrices $\mathbf{D}$ satisfying (6). If $\Gamma_2 = \{\mathbf{D} : \operatorname{tr} \mathbf{DKD}' = m\}$, then $\Gamma_1 \subset \Gamma_2$, and

$$\min_{\mathbf{D} \subset \Gamma_1} \sum_{i=1}^{m} \psi_i^{-2} \geq \min_{\mathbf{D} \subset \Gamma_2} \sum_{i=1}^{m} \psi_i^{-2}. \tag{8}$$

Further, we have

$$\operatorname{tr} \mathbf{DKD}' = \operatorname{tr} \mathbf{KD}'\mathbf{D} = \operatorname{tr} \mathbf{K}\,\mathbf{Q}'\boldsymbol{\Psi}^2\mathbf{Q} = \operatorname{tr} \mathbf{QKQ}'\boldsymbol{\Psi}^2 = \sum_{i=1}^{m} \psi_i^2 q_i,$$

where

$$q_i = \{\mathbf{QKQ}'\}_{ii}. \tag{9}$$

At the $\mathbf{D}$, where the minimum of

$$\sum_{i=1}^{m} \psi_i^{-2} \tag{10}$$

subject to the constraint

$$\sum_{i=1}^{m} \psi_i^2 q_i \tag{11}$$

occurs, the partial derivatives of the function

$$\sum_{i=1}^{m} \psi_i^{-2} - \xi \left(\sum_{i=1}^{m} \psi_i^2 q_i - m \right)$$

with respect to $\psi_1, \ldots, \psi_m$ vanish:

$$-2\psi_i^{-3} + 2\xi q_i \psi_i = 0 \quad (i = 1, \ldots, m) \tag{12}$$

Substituting the solutions for ψ_i into (11) we get

$$\xi^{-1/2} \sum_{i=1}^{m} q_i^{1/2} = m\,. \tag{13}$$

(12) and (13) give

$$\psi_i = \frac{1}{m} \left(\sum_{j=1}^{m} q_i^{1/2} \right)^{-1/2} q_i^{-1/4}, \tag{14}$$

i.e.

$$\sum_{i=1}^{m} \psi_i^{-2} = \frac{1}{m} \left(\sum_{i=1}^{m} q_i^{1/2} \right)^2. \tag{15}$$

By Lemma 2 we have

$$\sum_{i=1}^{m} q_i^{1/2} \geq \sum_{i=1}^{m} \mu_i^{1/2}. \tag{16}$$

(7), (8), (15) and (16) give the required inequality.

The second statement of the lemma can be proved as follows: By properly chosen $\mathbf{Q}$ we can get $q_i = \mu_i$ which give the equality in (16). For such $\mathbf{Q}$ and for $\mathbf{\Psi}$ defined by (14) we have

$$(\mathbf{DD}')^{-1} = \mathbf{P}\mathbf{\Psi}^{-2}\mathbf{P}' = m\left(\sum_{i=1}^{m} \mu_i^{1/2}\right)^2 \mathbf{DKD}'. \tag{17}$$

Choosing $\mathbf{P}$ in such a way that the diagonal elements of $\mathbf{P}\mathbf{\Psi}^{-2}\mathbf{P}'$ are equal, (17) gives the same for the diagonal elements of $\mathbf{DKD}'$. Then on the one hand the equality holds in (7), on the other hand $\mathbf{D}$ minimizes (10) in both Γ_1 and Γ_2.

2.3.4. Let F be any subset of $h+1$ elements out of G. Denote the set $\{j : j \in F,\ j \neq i\}$ by $F - i$. We have the following lemma.

Lemma 4. With the notations of the Theorem

$$\max_{i \in F} R^2(x_i\,;\,[\mathbf{Y}_1'\,\mathbf{X}_{F-i}']') \geq 1 - \frac{(h+1)^2}{\left(\sum_{j=1}^{h+1} \psi_j\right)^2}.$$

For any $\lambda_1, \ldots, \lambda_m$, h, and $\mathbf{Y}$ there exists an $\mathbf{X}_G$ such that for a certain F the inequality becomes an equality.

Proof. Without loss of generality we may assume

$$x_i x_i' = 1\,. \tag{18}$$

We introduce the symbol $d(v;\,\mathbf{Z})$ to denote the Euclidean distance of vector v from the space defined by the rows of the matrix $\mathbf{Z}$. Let $\mathbf{Q}_F$ denote the matrix of the difference vectors between the rows of $\mathbf{X}_F$ and their projections on the space of the rows of $\mathbf{Y}$. $\mathbf{Q}_{F-i}$ and q_i can be obtained from $\mathbf{X}_{F-1}$ and x_i in the same way. By this notation we may write

$$\begin{aligned} R^2(x_i;\,[\mathbf{Y}',\,\mathbf{X}_{F-i}']') &= 1 - d^2(x_i;\,[\mathbf{Y}',\,\mathbf{X}_{F-i}']') = \\ &= 1 - d^2(q_i\,;\,[\mathbf{Y}',\,\mathbf{Q}_{F-i}']') = 1 - d^2(q_i\,;\,\mathbf{Q}_{F-i})\,. \end{aligned} \tag{19}$$

Let the rows of the matrix $\mathbf{U}$ form an orthonormal basis of the space defined by the rows of $\mathbf{Y}$. Let $\hat{\mathbf{U}}$ be its least square estimator when $\mathbf{X}_G$ is in the regression. The rows of $\hat{\mathbf{U}}$ are the projections of rows of $\mathbf{U}$ on the space spanned by the rows of $\mathbf{X}_G$. Let the rows of $\mathbf{W}$ form an orthonormal basis in the space that completes the space of rows of $\hat{\mathbf{U}}$ to the space of rows of $\mathbf{X}_G$. We have $\mathbf{WU}' = 0$ and $\mathbf{W}\hat{\mathbf{U}}' = 0$. $\mathbf{X}_F$ can be written as

$$\mathbf{X}_F = \mathbf{A}\hat{\mathbf{U}} + \mathbf{BW}. \tag{20}$$

We shall try to determine $\mathbf{Q}_F$ in the form of

$$\mathbf{Q}_F = \mathbf{X}_F - \mathbf{CU}. \tag{21}$$

Multiplying (21) by $\mathbf{U}'$, in view of (20) we obtain

$$\mathbf{0} = \mathbf{A}\hat{\mathbf{U}}\mathbf{U}' - \mathbf{C}.$$

Thus (21) can be written as

$$\mathbf{Q}_F = \mathbf{A}\hat{\mathbf{U}} + \mathbf{BW} - \mathbf{A}\hat{\mathbf{U}}\mathbf{U}'\mathbf{U}.$$

We note that

$$\hat{\mathbf{U}}\hat{\mathbf{U}}' = \hat{\mathbf{U}}\mathbf{U}' = \mathbf{U}\hat{\mathbf{U}}'.$$

Denote this symmetric positive definite matrix by $\mathbf{M}$. It can be written as

$$\mathbf{M} = \mathbf{R}\boldsymbol{\Lambda}^2\mathbf{R}',$$

where $\mathbf{R}$ is an orthogonal, $\boldsymbol{\Lambda}$ a diagonal matrix with positive diagonal elements $\lambda_1, \ldots, \lambda_m$ defined in the Theorem. By these notations we may write

$$\mathbf{Q}_F\mathbf{Q}_F' = \mathbf{A}(\mathbf{M} - \mathbf{M}^2)\mathbf{A}' + \mathbf{BB}' = \mathbf{AR}(\boldsymbol{\Lambda}^2 - \boldsymbol{\Lambda}^4)\mathbf{R}'\mathbf{A}' + \mathbf{BB}'$$

and

$$\mathbf{X}_F\mathbf{X}_F' = \mathbf{AMA}' + \mathbf{BB}' = \mathbf{AR}\boldsymbol{\Lambda}^2\mathbf{R}'\mathbf{A}' + \mathbf{BB}'.$$

Let $\mathbf{D} = [\mathbf{AR}\boldsymbol{\Lambda}(\mathbf{I} - \boldsymbol{\Lambda}^2)^{1/2}, \mathbf{B}]$. Then

$$\mathbf{Q}_F\mathbf{Q}_F' = \mathbf{DD}'. \tag{22}$$

Since the λ_is are less than 1, we can write

$$\mathbf{X}_F\mathbf{X}_F' = \mathbf{D}\boldsymbol{\Phi}\mathbf{D}', \tag{23}$$

where

$$\boldsymbol{\Phi} = \left[\begin{array}{c|c} (\mathbf{I} - \boldsymbol{\Lambda}^2)^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{I} \end{array}\right].$$

It follows from the definition of the matrix $[\mathbf{A}, \mathbf{B}]$ that $\mathbf{DD}'$ is nonsignificant. Let the rows of the matrix $\mathbf{T}$ form the orthonormal basis of the space spanned by the rows of $\mathbf{D}$ for which $\mathbf{T}\boldsymbol{\Phi}\mathbf{T}' = \boldsymbol{\Phi}_1$ is a diagonal matrix with positive diagonal elements. $\mathbf{D}$ can be written as

$$\mathbf{D} = \mathbf{D}_1\mathbf{T},$$

where $\mathbf{D}_1$ is quadratic. By (22) and (23) we get

$$\mathbf{Q}_F\mathbf{Q}_F' = \mathbf{D}_1\mathbf{D}_1'$$

and

$$\mathbf{X}_F\mathbf{X}_F' = \mathbf{D}_1\boldsymbol{\Phi}_1\mathbf{D}_1'.$$

In view of (18) Lemma 3 can be applied to $\mathbf{D}_1 \mathbf{\Phi}_1 \mathbf{D}_1'$ and $(\mathbf{D}_1 \mathbf{D}_1')^{-1}$:

$$\max_j \{(\mathbf{D}_1 \mathbf{D}_1')^{-1}\}_{jj} \geq \Big(\sum_{i=1}^{h+1} \varphi_{1,i}^{1/2}\Big)^2 / (h+1)^2, \tag{24}$$

where $\varphi_{1,i}, \ldots, \varphi_{1,h+1}$ are the diagonal elements of $\mathbf{\Phi}_1 = \mathbf{T} \mathbf{\Phi} \mathbf{T}'$. By Lemma 2 we get

$$\max_j \{(\mathbf{D}_1 \mathbf{D}_1')^{-1}\}_{jj} \geq \Big(\sum_{j=1}^{h+1} \psi_i\Big)^2 / (h+1)^2. \tag{25}$$

Substituting

$$d^2(q_i ; \mathbf{Q}_{F-i}) = \frac{|\mathbf{Q}_F \mathbf{Q}_F'|}{|\mathbf{Q}_{F-i} \mathbf{Q}_{F-i}'|} = \{(\mathbf{Q}_F \mathbf{Q}_F^1)^{-1}\}_{ii}^{-1} = \{(\mathbf{D}_1 \mathbf{D}_1')^{-1}\}_{ii}^{-1} \tag{26}$$

into (19) and applying (25) we get the inequality of the lemma. Choosing the $\mathbf{A}$ and $\mathbf{B}$ such that for the corresponding $\mathbf{D}$ in (24) and (25) the equality holds, (20) gives the $\mathbf{X}_F$ for which the inequality of the lemma becomes an equality.

2.3.5. To prove the theorem we shall consider the set H_0 which maximizes $|\mathbf{Q}_H \mathbf{Q}_H'|$ over all $H \subset C_k^h(G)$. For any $j \notin H_0$ let $F_j = j \cup H_0$. From (19) and (26) it follows that

$$\max_{i \in F_j} R^2(x_i; [\mathbf{Y}', \mathbf{X}_{F_j - i}']') = R^2(x_j; [\mathbf{Y}', \mathbf{X}_{H_0}']') .$$

Then Lemma 4 gives

$$R^2(x_j ; [\mathbf{Y}', \mathbf{X}_{H_0}']') \geq 1 - \frac{(h+1)^2}{\Big(\sum_{j=1}^{h+1} \psi_j\Big)^2}$$

for each $j \notin H_0$. This result is an alternative formulation of the first part of the theorem.

Let the $k \times k$ matrix $\mathbf{D}_G$ be given as follows:

$$\mathbf{D}_G = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{O} \\ \hline \mathbf{O} & \mathbf{D}_1^* \end{array}\right],$$

where $\mathbf{D}_1^*$ denotes the choice of $\mathbf{D}_1$ for which in (25) the equality holds. If $\mathbf{A}_G$ and $\mathbf{B}_G$ are defined by

$$\mathbf{D}_G = [\mathbf{A}_G \mathbf{R} \mathbf{\Lambda} (\mathbf{I} - \mathbf{\Lambda}^2)^{1/2}, \mathbf{B}_G]$$

then for $\mathbf{X}_G = \mathbf{A}_G \hat{\mathbf{U}} + \mathbf{B}_G \mathbf{W}$ the inequality in the theorem becomes an equality.

2.4. *Procedures of reduced enumeration for various cases*

2.4.1. All the algorithms for various cases coincide in restriction of the enumeration only to a certain class of subsets containing a small number of independent variables. The numbers of elements of these subsets are on an average m less than the number of elements of subsets examined by full enumeration. The subsets will be completed to form real candidates for the set of regression variables only in some cases. The value given by the right side of the inequality of the theorem for $h = k - m$ will be used as a level for selecting additional variables. It can be written as

$$\beta = 1 \Big| - \frac{(k - m + 1)^2}{[k - m + (1 - M_0)^{-1/2}]^2}, \tag{27}$$

where M_0 is the current maximal value of the criterion of goodness.

2.4.2. Procedure for Problem A in case of $s_i \equiv 1$. The flow chart of the algorithm appears in Fig. 1. The level c can be defined as the integer k. Then the optimal subset will have k elements. To compare the procedure to that of full enumeration of all combinations of k out of n, we mention that the presented procedure needs the enumeration of the combinations of $k - m$ out of n only. Calculation of the criterion of goodness has to be performed only for certain k-combinations of the independent variables.

Explanation of the boxes. Box 1. In M_0 and in β there will be stored the current maximal value of the criterion of goodness, and the level given by (27).

Box 2. If Box 2 is called, a new combination will be stored in H. If the enumeration is over, the program will be terminated.

Box 3. The dependent variables are added to the set of independent variables defined by H. The set of indices i out of $E - H$ for which

$$R^2(x_i; [\mathbf{X}'_H, \mathbf{Y}']') \geq \beta$$

will be stored in L. If the number of its elements $r \geq m$ then Box 4 will be called otherwise the control is returned to Box 2.

Box 4. If Box 4 is called, a new combination of m out of L will be stored in $H1$. If all combinations are generated the control is transferred to Box 2.

Box 5. The criterion of goodness is calculated for $G = H \cup H1$. If it does not exceed the best value found until that time then the control will be transferred to Box 4, else Box 6 will be performed.

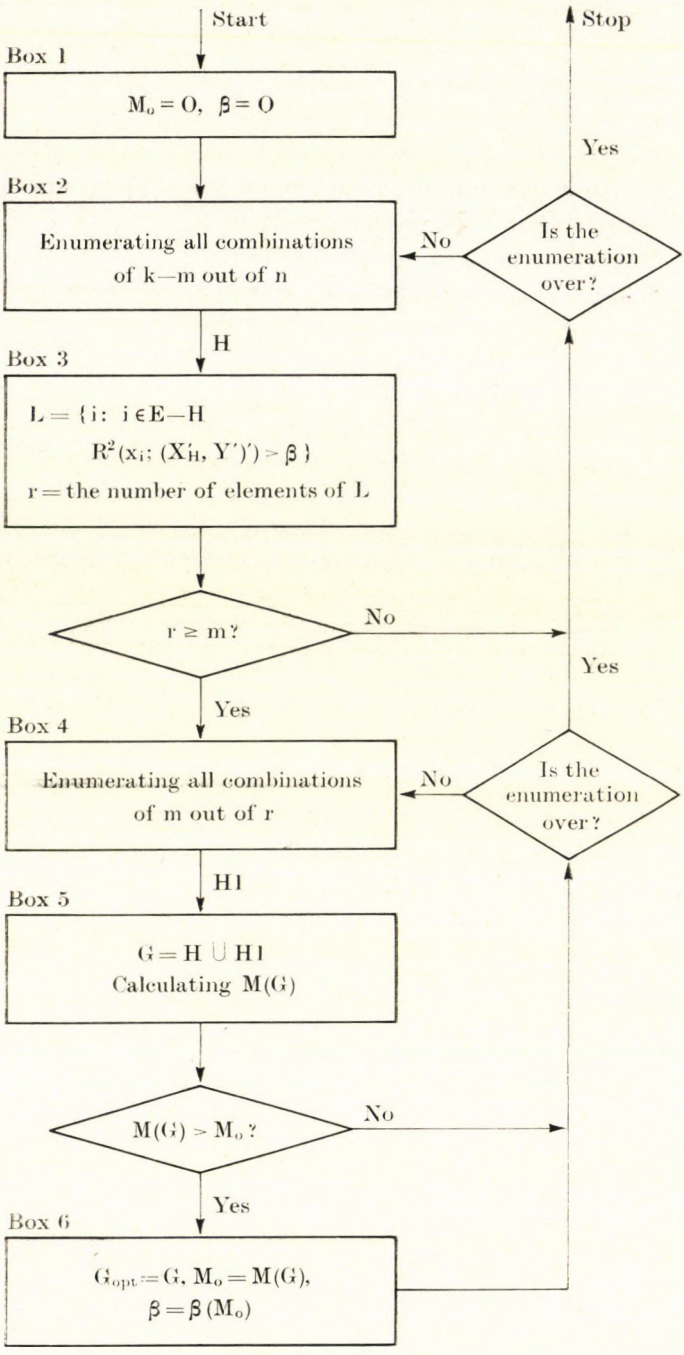

Start
Box 1
$M_o = 0,\ \beta = 0$
Box 2
Enumerating all combinations of k—m out of n
No
Is the enumeration over?
Yes
Stop
H
Box 3
$L = \{i:\ i \in E - H$
$R^2(x_i; (X'_H, Y')') > \beta\}$
r = the number of elements of L
$r \geq m$?
No
Yes
Box 4
Enumerating all combinations of m out of r
No
Is the enumeration over?
Yes
H1
Box 5
$G = H \cup H1$
Calculating M(G)
$M(G) > M_o$?
No
Yes
Box 6
$G_{opt} = G,\ M_o = M(G),$
$\beta = \beta(M_o)$

Fig. 1

Box 6. The current optimal subset and the optimal value of the criterion are stored in G_{opt} and M_0, respectively. The constant β is modified by (27).

2.4.3. Procedure for Problem B in case of $s_i \equiv 1$. The procedure presented in Fig. 2 is very similar to the previous one. The algorithm described above is performed for $k = m + 1, m + 2, \ldots$ etc, till the control is transferred to Box 6. When it is the first time it happens the program goes to stop.

2.4.4. Procedure for both problems in the general case. A very brief flow chart is given in Fig. 3.

3. Recognition of intonation

As an example of illustration an experiment connected with the recognition of intonation of a spoken Russian sentence was performed. A sentence was pronounced by 15 speakers with (a) affirmative, (b) interrogative intonations and (c) as the first part of a compound sentence. We should remark that the grammar of the Russial language permits the order of words to be the same in these sentences. 40 different measurements were made on each of the sentences.

By a method of feature extracting [1] two linear functions of all measurement parameters were found as features. In order to reduce the dimensionality of pattern recognition vectors, the subsets of measurement parameters were tested regarding the ability of their linear functions to approximate the features. The difference vectors between measurements from different classes were weighted in inverse proportion to the fourth power of their lengths and were taken as sample vectors. The cost of measuring was assumed to be equal to the number of parameters being in the regression.

There were solved Problem A in three different cases $k = 2, 3, 4$, and Problem B in cases $c = 0.93, 0.96, 0.97$. As a rough approximation to determine computational efficiency, both the method suggested in the present paper and that of full enumeration were compared by counting the number of times, the calculation of M was performed. Results and comparisons of these approximative amounts of operations are presented in Tables 1 and 2. With respect to pattern classification it should be noted that for $k = 4$, the space of approximates of the features assures an acceptable discrimination of the three classes of patterns. The figure shows the shift of the sample in the plane of the two features after the number of parameters have been reduced to four.

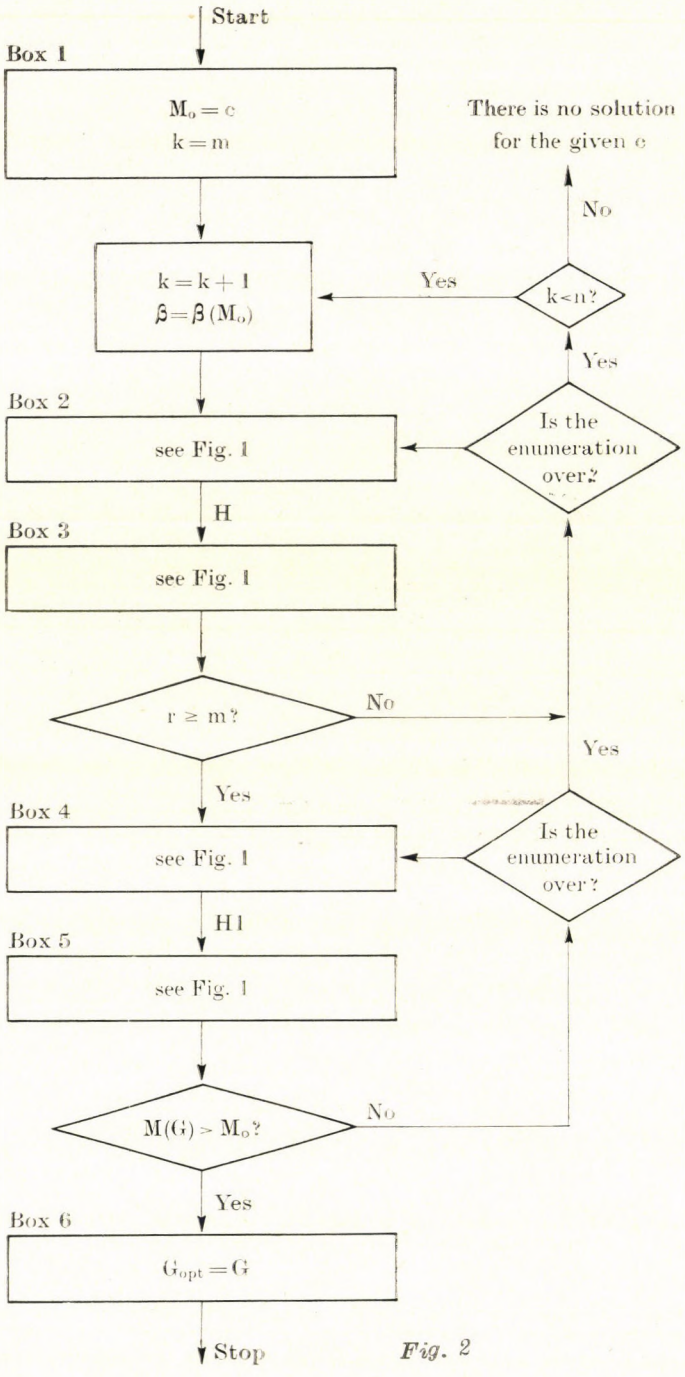

Start
Box 1
$M_o = c$
$k = m$
There is no solution for the given c
No
$k = k + 1$
$\beta = \beta(M_o)$
Yes
k<n?
Yes
Box 2
see Fig. 1
Is the enumeration over?
H
Box 3
see Fig. 1
r ≥ m?
No
Yes
Yes
Box 4
see Fig. 1
Is the enumeration over?
H1
Box 5
see Fig. 1
M(G) > M_o?
No
Yes
Box 6
$G_{opt} = G$
Stop

Fig. 2

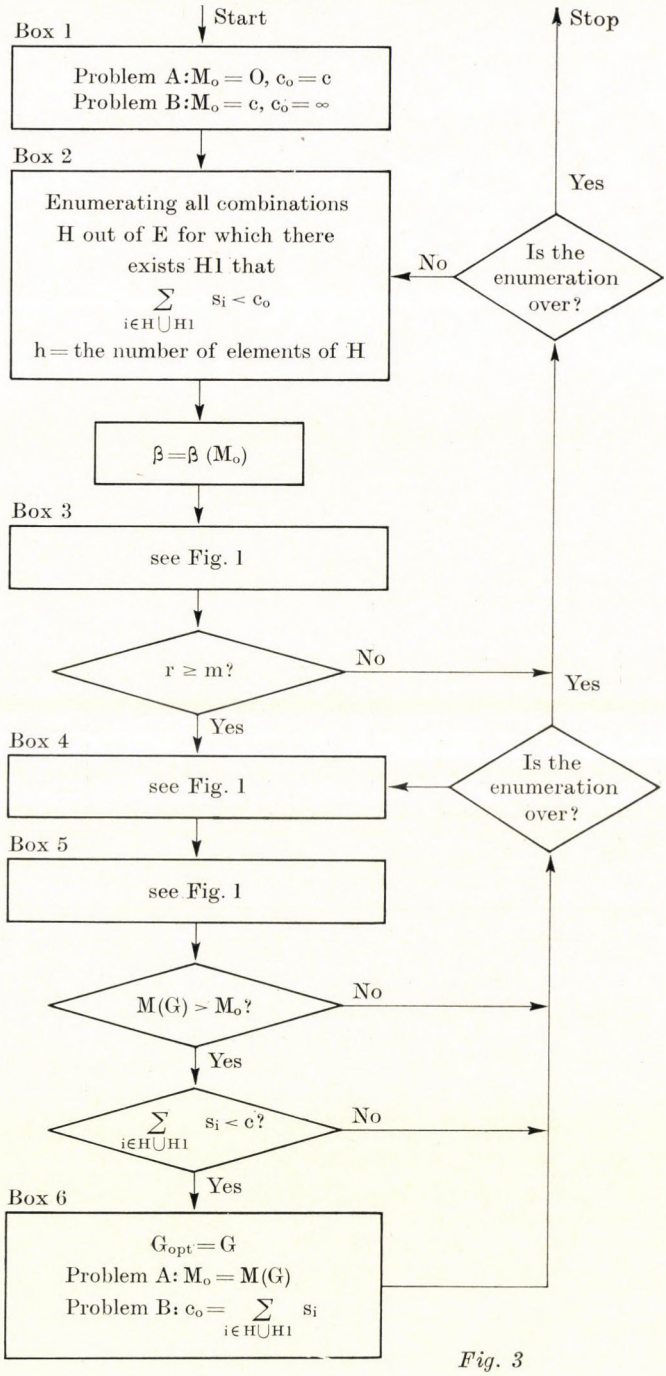

Start
Stop
Box 1
Problem A: $M_o = O$, $c_o = c$
Problem B: $M_o = c$, $c_o = \infty$
Box 2
Enumerating all combinations H out of E for which there exists H1 that $\sum_{i \in H \cup H1} s_i < c_o$
h = the number of elements of H
No
Is the enumeration over?
Yes
$\beta = \beta(M_o)$
Box 3
see Fig. 1
$r \geq m$?
No
Yes
Yes
Box 4
see Fig. 1
Is the enumeration over?
Box 5
see Fig. 1
$M(G) > M_o$?
No
Yes
$\sum_{i \in H \cup H1} s_i < c$?
No
Yes
Box 6
$G_{opt} = G$
Problem A: $M_o = M(G)$
Problem B: $c_o = \sum_{i \in H \cup H1} s_i$

Fig. 3

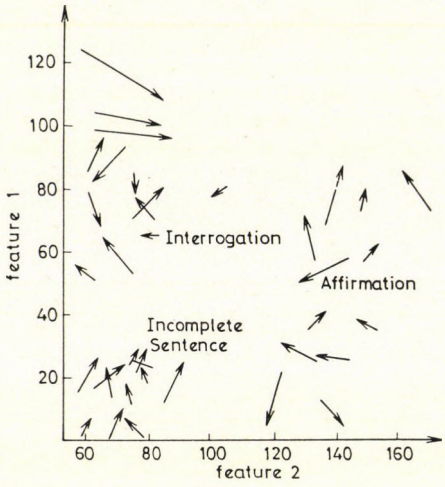


Fig. 4

Table 1

k	The optimal parameter set	$M(G)$	The number of times the calculation of $M(G)$ was performed	
			Full enumeration	Shortened enumeration
2	36 40	0.9258	780	—
3	19 36 40	0.9518	9 880	253
4	5 19 36 40	0.9707	91 390	6320

Table 2

c	The optimal parameter set	$M(G)$	The number of times the calculation of $M(G)$ was performed	
			Full enumeration	Shortened enumeration
0.93	18 36 40	0.9370	8 710	890
0.96	2 19 36 40	0.9609	24 610	1462
0.97	5 19 36 40	0.9707	46 060	2062

Conclusions

The procedure for reducing the number of input pattern recognition parameters presented in the paper has necessitated dealing with the problem of reducing the number of regression variables in case of more than on

dependent variables. A criterion of the goodness of fit of the regression and a procedure for searching the optimal set of variables have been proposed. Further experimental or perhaps theoretical investigations are necessary before definite conclusions regarding the saving of computer time would be arrived at.

References

1. *Turbovich, I. T., Gitis, V. G., Maslov, V. K.*, Opoznanie obrazov. Nauka, Moscow 1971.
2. *Watanabe, S.*, "Karhunen-Loeve Expansion and Factor Analysis — Theoretical Remarks and Applications." Proc. Fourth Prague Conference on Information Theory, 1965.
3. *Chien, Y. T., Fu, K. S.*, Selection and Ordering of Feature Observations in a Pattern Recognition System. Information and Control **12** (1968) 395—414.
4. *Marcus, M., Minc, H.*, A Survey of Matrix Theory and Matrix Inequalities. Allyn and Bacon, Inc., Boston, 1964.

Сокращение исходного описания в задаче распознавания образов с помощью метода поиска оптимального подмножества параметров линейной регрессии

Д. МУСЕЛИ

(Будапешт)

По предложенному методу сокращение исходного описания осуществляется после нахождения линейных признаков в пространстве полного описания. Предлагается выбирать такие параметры, линейные комбинации которых хорошо аппроксимируют признаки. Таким образом, решение задачи сводится к нахождению оптимального подмножества параметров для линейной регрессии. Эта более общая задача рассматривается в случае нескольких зависимых переменных. Дается критерий для измерения эффективности аппроксимации.

Формулируются две задачи: (а) найти подмножества с максимальной мерой эффективности среди тех подмножеств параметров, для которых стоимость измерений меньше некоторого порога, (б) найти подмножества с минимальной стоимостью измерений среди тех подмножеств, для которых мера эффективности больше некоторого порога.

В статье предлагается алгоритм сокращенного перебора, позволяющий получить решение поставленных выше задач. В качестве иллюстрации предлагаемой методики рассматривалась задача распознавания интонации.

Gy. Muszély
Computer and Automation Institute
H-1502 Budapest, XI. Kende u. 13—17
Hungary

Problems of Control and Information Theory, Vol. 4 (2), pp. 183—184 (1975)

INTERNATIONAL IFAC SYMPOSIUM ON DISCRETE SYSTEMS

International IFAC Symposium on Discrete Systems held in Riga (USSR) from September 30 to October 4, 1974, was attended by 262 persons representing 72 organizations from twelve countries (Belgium, Bulgary, Hungary, German Democratic Republic, Italy, Poland, USSR, USA, Federal Republic of Germany, France, Czechoslovakia, Japan). At its 6 sessions, 89 papers were presented dealing with synthesis of discrete systems, automation of their design, reliability and diagnostics, probabilistic automata and collective behaviour of automata, uniform arrays and other aspects of the discrete system theory. In addition, two round-tables were devoted to synthesis and technical diagnostics of discrete systems.

Sessions were preceded by a plenary session at which 5 survey papers were read discussing most important study areas related to discrete systems.

A new way to studying discrete system design automation based on methods of the system analysis was outlined in a paper by M. A. Gavrilov (USSR). The same matter was discussed by H.-J. Zander (GDR) who gave characterization of an automated system RENDIS, under development in GDR.

Organization and implementation of uniform arrays were discussed in papers by V. G. Lazarev (USSR) and J. Florine (Belgium). The former deals with implementation principles and peculiarities of automaton synthesis in uniform arrays, and the latter presents results of studies on construction of "macrocell" computers. Each macrocell is an involved device consisting of several registers, adders, and other computer components. All cells have identical structure and are interconnected by two closed information transmission paths, horizontal and vertical.

Paper by V. I. Varshavskii was devoted to the theory of aperiodical automata with self-synchronization whose organization differs from both classical synchronous, and asynchronous automata.

Discrete system synthesis session had 22 papers discussing new, nontraditional aspects of synthesis such as synthesis of multifunctional logic

elements, design of multifunctional computer modules, and synthesis procedures for discrete devices consisting of elements with given arbitrary structure, as well as solutions of classical problems.

Majority of papers in the discrete system diagnosis and reliability session were devoted to detection of contact-type failures in combinatorial circuits and to methods of sequential discrete system checking. Most interesting papers were from France, USSR, Bulgaria and Hungary.

Reliability of combinatorial and sequential discrete devices was discussed in two papers from USSR and Bulgaria.

Papers at the session on discrete device design automation presented results of studies carried out in USSR in three fields microprogram automaton automation synthesis, computer structure synthesis, and discrete device design automation. In addition, two papers from France dealt with simulation of logic device operation and modular design of industrial logic devices.

At the session on probabilistic automata, papers by Soviet authors discussed transformation of stochastic sequences by means of deterministic and probabilistic automata. The same topic was touched upon in a paper from DDR which considered transformation of various types of stochastic sequences by various types of automata.

At the session devoted to uniform arrays papers were presented by scientists from Belgium, Bulgaria, USSR and FRG. Several papers dealt with general questions of the uniform array theory concerning their structure, behaviour and density.

Another group of papers discussed applied aspects of the uniform array theory, namely, those concerning particular methods for implementation of automata in uniform arrays, signal transmission in these arrays, provision of reliability of automata implemented in uniform arrays, etc.

Diverse problems were discussed in papers read at the session on general problems of discrete automata. They dealt with problems of great theoretical importance, as well as with specific problems of applied nature.

On the whole, the papers reflected the today state-of-art and trends in the development of discrete system theory.

Discussion and exchange of views between symposium participants will prove very useful in further studies in the field of the discrete system theory.

V. M. Ostianu

Problems of Control and Information Theory, Vol. 4 (2), pp. 185—188 (1975)

SIXTH ALL-UNION CONFERENCE ON CONTROL PROBLEMS

The regular All-Union Conference on Control Problems was held in Moscow for the sixth time between *October 7 and 11, 1974.* It was sponsored by the Academy of Sciences of the USSR and the Soviet National Committee on Automatic Control. The conference was attended by about one thousand persons representing ministries, departments, scientific organizations, industrial enterprises and educational institutions. About 300 papers and communications covering most interesting results obtained in the field of theory, hardware and control systems were presented and discussed.

The plenary sessions were opened by a short introductory speach by academician V. A. Trapeznikov, Chairman of the Conference Organizing Committee, followed by invited papers of A. I. Milikovskii "Problems involved into the development of the capitalist economics", Yu. M. Popov "Application of lasers to optical information processing" and B. V. Raushenbakh "Attitude control of orbital stations".

The conference involved 20 sessions:

1. Theory of control systems;
2. Adaptation and recognition in automatic systems;
3. Optimal systems;
4. Discrete automata;
5. Stochastic systems;
6. Design automation;
7. Automation of scientific researches;
8. Control of economic systems;
9. Control of management systems;
10. Artificial intelligence and robots;
11. Information systems in sensor systems;
12. Control in biology and medicine;
13. Discrete and hybrid devices for automatics and computers;
14. Analog devices for automatic control;

15. Technical diagnostics and reliability of automatic control systems;
16. Automatic control systems;
17. Automatic control systems for technological processes;
18. Control of hierarchical and complex systems;
19. Control of moving objects;
20. Education, technical means and equipment for management.

At meetings between sessions the following invited papers were presented and discussed: V. M. Glushkov "A sequential technique for plan optimization at the macro-economic level"; G. S. Pospelov, D. A. Pospelov "Basic directions of research in the field of artificial intelligence"; A. M. Petrovskii "Control theory and medicine"; M. A. Aizerman "The man and the collective as elements of control systems"; O. M. Belotserkovskii "Technological progress and training of personnel"; M. A. Gavrilov, V. N. Mostachenko, O. N. Tatur, E. A. Trakhtengerts, V. V. Boiko "Present-day problems in design control automation".

The conference resolution says:

"Presently, theoretical and applied studies and developments in the field of control are being carried out with success in areas involving management-administrative, management-technical, economical, social, biological and other complex systems as well as in the traditional technical area.

"Further progress was made in the fields of optimal and finite control, and numerical methods for determining optimal controls. Significant advances were made in the optimization of high-dimensionality systems, the nonlinear invariant systems, the variable structure systems, the stochastic and terminal systems. New identification and control methods were developed and limits of adaptation alogrithms were estimated for the adaptive and self-aligning system theory. The theory of finite systems, finite automata involving human collectives, relay circuits, sequential machines, the theory of control under uncertain conditions were further explored.

"Progress was made in the study and development of applied problems of economic control. Significant results were obtained in generation of effective numerical methods for economical and mathematical problems of high dimensionality, in computer simulation of block diagrams for control of economic systems, in design of information systems for generation of rapid information and preparation of arrays of economical data for planning. Formalized description of socio-economic plants is being developed and mechanisms of its functioning are being studied. An Automated Planning System for the State Planning Committee, Planning Committees of Union republics, and also for appropriate ministries and departments has been developed and accepted.

"Studies of theoretical and practical aspects of management systems proved successful. Important results were obtained in analysis and synthesis of management structures, decision making, and planning of scientific studies.

"Work was carried out on study and modelling of control in living organisms. New models of information perception and processing in living organisms have been constructed which are of importance in terms of technical system design. Studies on artificial intelligence were continued further. Works are being carried out on construction of robots of 1st, 2nd and next generations. In cooperation with health service workers, studies were continued on disease diagnostics and medical treatment control organization.

"Intensive work was carried out on automatic control systems for technological processes, and administration and management at the levels of separate enterprises, sectors of the national economy, and the national economy as a whole. Reasonable ways of modelling and generation of hierarchical control structures, methods for reduction of information redundancy and correct organization of information flows between system elements, and effective ways to increase dependability of processed information were established.

"Design automation has been developing successfully for recent 5—6 years. Important studies are being carried out on automation of scientific studies. General-purpose systems for communication between measuring instrumentation and computers, and a number of experiment planning methods have been developed.

"Works on technical diagnostics and reliability met with success. Principles were worked out for organization of systems determining technical condition of complex plants during their assembly and use. Computer systems were developed for generation of tests for modern standard computer elements.

"In the field of reliability, methods for study and analytical description of complex control system behaviour allowing for particular conditions of their operation, and approaches to construction of reliable systems from non-reliable elements were developed. Normative documentation was prepared giving regulations for reliability of various classes of technical units and systems in terms of their importance.

"Works in the field of control hardware were going on intensively. New principles were worked out for construction of hardware for automation and computers based on advances of the quantum electronics, on magnetic domains, resonance phenomena, etc. Intensive work was carried out on enhancement of physical reliability and precision of the hardware and reduction of its size. The State System of Instrumentation was further developed; its

branches are being developed involving new areas of hardware such as vibrometry, analytical and electrical measuring instruments, etc.

"Significant results were obtained on the theoretical foundations and principles of construction of high-throughout reliable multi-processor systems for computers. Further advances were made in the field of rearrangeable structure systems based on uniform microelectronic arrays and oriented to effective real-time solution of control problems.

"Great attention was paid to the development and analysis of the principles of increasing effectiveness of the computer technology: consideration was given to numerical networks, fixed digital analog computers, multiple use of analog computing elements, automatic setting of a problem and checking of its execution on analog computers. Intensive works were carried out on the development of effective methods for application of hybrid computing systems to optimization, simulation of stochastic systems and lumped parameter ones.

"Studies of the man-automaton problem were carried out to advantage. Techniques were developed for the design of man-computer systems. Work was commenced on design of universal sets of means for simulating such systems.

"The conference believes that extensive introduction into the national economy of present advances of the control sciences, as well as of those which will result from further studies will contribute to the significant progress of the national economy of this country".

I. N. Sinitsin—M. E. Shaikin

Problems of Control and Information Theory, Vol. 4 (2), pp. 189—191 (1975)

AICA SYMPOSIUM ON HYBRID COMPUTATION IN DYNAMIC SYSTEMS DESIGN

ROME, NOVEMBER 11—14, 1974

The international symposium on Hybrid Computation in Dynamic Systems Design, organized by AICA (Assotiation Internationale pour le Calcul Analogique) in cooperation with IFAC—IFIP—SCS—AEI—ANIPLA under the sponsorship of

— Consiglio Nazionale delle Ricerche
— Istituto di Automatica — Università di Roma
— Comitato Nazionale per l'Energia Nucleare
— Ente Nazionale per l'Energia Elettrica,

was arranged for the seventh time.

The symposium was prepared by an Italian Organizing Comittee headed by Professor Alessandro De Carli (Istituto di Automatica, Rome).

The symposium was held in the Conference Hall of the "Consiglio Nazionale delle Ricerche", Piazzale delle Scienze.

The 66 accepted papers were selected from 102 submitted papers and divided according to the following topics:

(A) COMPUTING METHODS AND STRUCTURES
(B) METHODS OF HYBRID COMPUTATION FOR

1. Identification
2. Optimization
3. Analysis of Distributed Systems

(C) APPLICATIONS OF HYBRID COMPUTATION TO INDUSTRIAL SYSTEMS:

1. Chemical
2. Iron and Steel
3. Mechanical
4. Aerospace
5. Electrical and Electronic
6. Nuclear and Power

(D) APPLICATIONS OF HYBRID COMPUTATION TO ENVIRONMENTAL SYSTEMS:

1. Ecological
2. Social

(E) APPLICATIONS OF HYBRID COMPUTATION TO BIOMEDICAL SYSTEMS

The symposium started November 11, 1974 with a survey paper presented by R. Vichnevetsky, President of AICA. Then the presentation of the technical papers followed. Two round table sessions were organized, the first one was devoted to the education in hybrid computation, the second one treated the actual and future role of hybrid computation in dynamic systems design. A technical visit to the Hybrid Computation Center of the National Nuclear Energy Committee was organized as well.

It is worth mentioning how the accepted papers distributed among the countries:

Country	Number of accepted papers	Country	Number of accepted papers
Austria	1	Holland	3
Belgium	6	Hungary	2
Bulgaria	1	Iran	1
Canada	1	Israel	1
Czechoslovakia	4	Italy	10
Egypt	1	Poland	11
England	2	Spain	1
Finland	1	USA	3
France	6	USSR	6
FRG	5		

As it is shown in the table the 66 technical papers were accepted from 19 countries which also indicates the really international character of the symposium.

The 103 registered participants coming from 17 European, American, African and Asiatic countries also emphasized the international importance of the subject of the symposium.

Because of the limited length of time, the lectures were delivered in parallel sections. Time for presentation and discussion was 20 minutes. More serious discussion did not developed.

It is interesting to mention how important role played the practical application in the presented papers. Almost every lecture has dealt with problems arising in the course of industrial or development works.

Studying the subject of the symposium one can observe the comprehensive possibility of application of hybrid computers on the one hand and a very definite progress of possibilities given by hybrid computers for users on the other hand.

The above facts and the decreasing importance of analog computers in itself propose to change the first letter of AICA from A into H.

A cocktail party held on the first day and a dinner party on the third day of the symposium afforded possibility to enlarge personal connections.

The official language of the symposium was English.

The Tamburini Editor of Milan published the Proceedings of Symposium approximately in 400 pages.

M. Habermayer

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY VOL. 4, NO. 2
ENGLISH TRANSLATION OF THE COMMUNICATIONS IN RUSSIAN

NONPROBABILISTIC MUTUAL INFORMATION WITHOUT MEMORY*

V. D. GOPPA

(Moscow)

A nonprobabilistic analog of the mutual information is introduced. A universal decoding algorithm for discrete memoryless channels is proposed.

1. For any two words $x \in X^n$, $x = (x_1, \ldots, x_n)$, $|X| = L_1$; $y \in Y^n$, $y = (y_1, \ldots, y_n)$, $|Y| = L_2$ form a new word $x \otimes y \in (X \otimes Y)^n$, $x \otimes y = ((x_1y_1), (x_2y_2), \ldots, (x_ny_n))$. Let $\{m_i\}$, $\{n_i\}$, $\{m_{i,j}\}$ be compositions of the words x, y and $x \otimes y$, respectively, $i = 1, 2, \ldots, L_1$; $j = 1, 2, \ldots, L_2$.

$\{m_{ij}/n\}$ is evidently the probability distribution on $X \otimes Y$, and $\{m_i/n\}$ and $\{n_j/n\{$ are its partial distributions on X and Y, respectively, because $\sum\limits_{i,j} m_{ij} = n$, $\sum\limits_{j} m_{ij} = m_i$, $\sum\limits_{i} m_{ij} = n_j$. It follows from the above that $\varphi(x \otimes y) < < \varphi(x) + \varphi(y)$ where $\varphi(x)$ is the Fitingof weight of x [1].

Introduce the following notations:

$$\varphi(x : y) = \varphi(x) + \varphi(y) - \varphi(x \otimes y)$$

$$\varphi(y/x) = \varphi(x \otimes y) - \varphi(x).$$

The following of their properties are readily verified:

1. $\varphi(x : y) = \varphi(y : x)$;
2. $\varphi(x : x) = \varphi(x)$ because $\varphi(x \otimes x) = \varphi(x)$;
3. $\varphi(x : y) = \varphi(x) - \varphi(x/y) = \varphi(y) - \varphi(y/x)$.

Lemma 1. Let Q be a probability distribution on $X \otimes Y$, Q^n be the distribution on $X^n \otimes Y^n$ which is the product of measures on $X \otimes Y$ (discrete memoryless channel), $R = H(x) - H(X/Y)$ be Shannon's mutual information. Then $\varphi(x : y) \xrightarrow{p} R$, $\varphi(y/x) \xrightarrow{p} H(Y/X)$.

The proof follows from the definitions and the law of the large numbers.

* The main results of this paper were discussed on the 3rd International Symposium on Information Theory (Tallinn, 1973).

Now introduce the distance between words x and y as follows:

$$\varrho(x, y) = [\varphi(x/y) + \varphi(y/x)]/2.$$

Note that the words x and y have the same length but may belong to different alphabets.

Find conditions under which $\varphi(y/x) = 0$. According to the definition this means that $\varphi(x \otimes y) = \varphi(x)$. The latter equality holds if the following condition is satisfied: if in the word x the letter $a_i \in X$ occupies locations $j_1, j_2, \ldots, j_k$, the same locations in the word y will be occupied by some letter $b_i \in Y$, no matter which. If the equality $y_{j1} = y_{j2} = \ldots = y_{jk}$ does not hold, this leads to $\varphi(x \otimes y) > \varphi(x)$.

For instance, for the pair of words

$x = (11121233)$

$$\varphi(y/x) \neq 0, \ \varphi(x/y) \neq 0,$$

$y = (77777011)$

for the words

$x = (11121233)$

$$\varphi(y'/x) = 0, \ \varphi(x/y') \neq 0,$$

$y' = (77777711)$

and for the words

$x' = (11111133)$

$$\varphi(y'/x') = 0, \ \varphi(x'/y') = 0.$$

$y' = (77777711)$

Thus, the equality $\varrho(x, y) = 0$ means that the word y was obtained from the word x by just renaming the letters, various letters of the word receiving various new names. It is clear that the relation $x \sim y \Leftrightarrow \varrho(x, y) = 0$ is that of equivalence, and we are going to regard all the words belonging to the same equivalence class as identical.

Lemma 2. $\varrho(x, y)$ is metrics on the totality of all words of length n.

Proof. It is sufficient to verify the triangle inequality. Shannon has introduced the metrics $d(X, Y) = H(X/Y) + H(Y/X)$ in the space of all stochastic variables with finite entropy. Our distance between words $\varrho(x, y)$ is induced by Shannon's metrics: to each pair of words x, y we assign a probability distribution $\{m_{ij}/n\}$ over $X \otimes Y$ with partial distributions $\{m_i/n\}$ and $\{n_j/n\}$ over X and Y, and for these distributions Shannon's metrics $d(X, Y)$ exactly coincides with the distance $\varrho(x, y)$.

For the sake of completeness we are going to present the proof of the triangle inequality for the Shannon metrics according to Y. Horibe [2].

$$H(X/Y) + H(Y/Z) \geq H(X/Y, Z) + H(Y/Z) = H(X, Y/Z) \geq H(X/Z)$$

thus

$$H(Y/X) + H(Y/Z) \geq H(X/Z)$$

and substituting Z for X we obtain $H(Y/X) + H(Z/Y) \geq H(Z/X)$. By summing the two inequalities

$$d(X, Y) + d(Y, Z) \geq d(X, Z).$$

Lemma 3.

$$\varrho(x, y) = \varphi(x \otimes y) - [\varphi(x) + \varphi(y)]/2;$$

$$\varrho(x, y) = [\varphi(x) + \varphi(y)]/2 - \varphi(x : y).$$

Lemma 4. Let $x_0 \in X^n$; $V_{x_0}^{d,w} = \{y \in Y^n : \varrho(x_0, y) \leq d,\ \varphi(y) \leq w\}$ (a generalized sphere with the center in the point x_0). Then

$$\log |V_{x_0}^{d,w}| \leq n\big(d + (w - \varphi(x_0))/2 + 0(1)\big).$$

Proof. According to Lemma 3 for any $y \in V_{x_0}^{d,w}$ the inequality

$$\varphi(x_0 \otimes y) \leq d + \big(w + \varphi(x_0)\big)/2 \tag{1}$$

holds. The number of pairs satisfying inequality (1) is estimated through Lemma 1 of [1]: $\log N^{d,w} \leq n(d + (\varphi(x) + w)/2 + 0\,(1))$.

If x and x' have the same composition, x' may be obtained from x by permutation of letter locations, and if one denotes by Bx a set of words $y \in Y^n$ such that the pair (x, y) satisfies (1), $Bx = Bx'$ because Bx' is also obtained from Bx by location permutations. The number of words with the same composition is known to have the following [3] estimate:

$$\log N_x \geq n\big(\varphi(x) - 0(1)\big).$$

Finally obtain:

$$\log |V_{x_0}^{d,w}| < \log N^{d,w} - \log N_x < n\big(d + (w - \varphi(x))/2 + 0(1)\big).$$

2. By code we understand a subset U of the space X^n. Introduce the following characteristics of the code:

1. transmission rate $R(U) = (\log |U|)/n$;

2. distance $d(U) = \min\limits_{x,y \in U} \varrho(x, y),\ x \neq y$;

3. capability $C(U) = \underset{x,y \in U}{\text{Max}}\, \varphi(x : y), \quad x \neq y;$

4. weight $W(U) = \underset{x \in U}{\text{Max}}\, \varphi(x).$

The code whose words have the same Fitingof weight will be referred to as constant weight code.

Theorem 1. There exists a constant weight code for which

$$R(U) \geq C(U) - 0(1) = w(U) - d(U) - 0(1)$$

for any $C(U) > 0$.

Proof. It follows from the definition of $\varrho(x, y)$ that for any code

$$d(U) \leq w(U),$$

and Lemma 3 implies that for the constant weight code

$$d(U) = w(U) - C(U).$$

For a fixed $d > 0$ choose an arbitrary $w \geq d$ and construct a constant weight code with parameters $d(U) = d$, $w(U) = w$ by using the standard "rejection method" by Gilbert. The number of all words of weight w is at least and the number of rejected points of the sphere of radius d is at most

$$2^{n(w-0(1))}$$

$$2^{n(d+0(1))}$$

(Lemma 4).

Thus, there exists a code with the rate

$$R(U) \geq w - d - 0(1) = C\,(U) - 0(1).$$

3. Introduce φ-decoding for an arbitrary discrete memoryless channel as follows: a word $z \in Y^n$ is decoded into the code word x if and only if $\varphi(x : z) > \varphi(x' : z)$ for any other code word x'. If there is no code word satisfying this condition, the word z will be said to lead to decoding failure included into the decoding error.

Theorem 2. For any discrete memoryless channel there will be a sequence of constant weight codes with a rate arbitrarily close to the channel capacity and with arbitrarily small error probability of φ-decoding.

Proof. Apply the classical "rejection method" by Feinstein. Let C be the channel capacity, $H(X)$ be the entropy of the maximizing distribution,

$H(Y)$ be the entropy of input distribution, R be the code transmission rate $R < C$.

Let $d = (R + C)/2$, $M = \{(x, y) : x \in X^n, y \in Y^n, \varphi(x : y) > \alpha, \varphi(y) \leq \leq H(Y), \varphi(x) = H(X) + 0(1)\}$,

M_y be projection of M on Y^n,

M_x be projection of M on X^n.

If $P(M_y) \geq 1 - \varepsilon$, $0 < \varepsilon < 1$, then, since $P(M_y) = \sum_{x \in M_x} p(x) \cdot p(M_y/x)$, there will be a word $x_1 \in M_x$ such that $P(M_y/x_1) < 1 - \varepsilon$. Take this word as the first code word and reject the set $\beta_1 = \{y \in M_y : \varphi(x_1 : y) > \alpha\}$ from M_y Now if $P(M_y - \beta_1) < 1 - \varepsilon$, there will be another word x_2 satisfying $P(Mr/x_2) < 1 - \varepsilon$. The process of rejection ends when

$$P\big(M_y - M_y \cap (\bigcup_{i=1}^{N} \beta_i)\big) < 1 - \varepsilon .$$

In this case

$$P(M_y) = P\big(M_y \cap (\bigcup_{i=1}^{N} \beta_i)\big) + P\big(M_y - M_y \cap (\bigcup_{i=1}^{N} \beta_i)\big) \leq \sum_{i=1}^{N} P(\beta_i) + 1 - \varepsilon . \tag{2}$$

To evaluate $P(\beta_i)$ note that according to Lemma 3 for any pair $(x, y) \in M$ the inequality

$$\varrho(x, y) < \big(H(X) + H(Y)\big)/2 - \alpha + 0\,(1)$$

holds, and, by Lemma 4

$$|\beta_i| \leq 2^{n(H(Y) - \alpha + 0(1))} .$$

Further, for all sufficiently large n all words $y \in Y^n$, for which $\varphi(y) \leq \leq H(Y)$, become equally probable, and since their number is at least $2^{n(H(Y) - 0(1))}$ it follows that:

$$P(\beta_i) \lesssim P(y)\,|M_y| \lesssim 2^{-n(\alpha + 0(1))} . \tag{3}$$

Substituting (3) into (2) obtain

$$P(\alpha) \leq N \cdot 2^{n(\gamma + 0(1))} + 1 - \varepsilon .$$

Thus, for all sufficiently large n the following inequality for error probability holds: $P_e < N \cdot 2^{-n(\alpha + 0(1))} + P[\varphi(x : y) \leq \alpha]$, and since $d = (R + C)/2$ and $N = 2^{nR}$, we have

$$P_e \leq 2^{-n((C - R)/2 + 0(1))}\, P[|\varphi(x : y) - C| \geq (C - R) . \tag{4}$$

Since $\varphi(x : y)$ n.b. C, $P_e \to 0$ with increase of n.

4. $\varphi(x : y)$ is used in the present paper as nonprobabilistic analog of the "mutual memoryless information". It has all properties of Shannon's mutual information and is very useful in decoding. The φ-decoding introduced here is independent of a particular channel, i.e. is a universal decoding. As it is mentioned in [1], such methods have long been known, but a specific universal algorithm for discrete memoryless channels seems to be described for the first time. The general definition of information based on Fitingof ideas will be analyzed in the future.

References

1. *Goppa, V. D.*, Universal Decoding for Symmetric Channels. "Problemi Peredachi Informatsii" 1975, 11, 1.
2. *Horibe, Y.*, A Note on Entropy Metrics. Inf. and Control **22** (May 1973).
3. *Fano, R. M.*, Transmission of Information. M.I.T. Press and John Wiley & Sons Inc., New York—London, 1961.

AN APPROACH TO ESTIMATION PROBLEMS

V. N. ZAKHAROV, D. A. POSPELOV

(Moscow)

The paper discusses application of the multivalued logic to construct an estimate function allowing for the psychology of estimating a situation at incomplete knowledge of the situation parameters. An example of collective behaviour of automata demonstrates the usefulness of the proposed approach in models of the decentralized control.

1. Problem formulation

In everyday life we often have to evaluate various phenomena, events or actions. We say, for instance, "a boring book", "fine weather", "a useful lecture" thus formulating our estimates of them. These estimates affect essentially our behaviour. In some cases we cannot decide for taking one or another action being unable to estimate the situation.

The present need for control systems working, in a manner, autonomously in unpredictably changing environment stems from the trend to automation of complicated technological, economical and management systems that characterizes the last decade of the control theory's history. Robots for research and work in the outer space and oceans serve as only one of possible examples of such systems. Every autonomous device must either have a special estimating mechanism for situations within the control plant and in its environment, or make use of appropriate information from human or mechanical sources. Effectiveness of these control systems happens to be closely related to that of making estimates.

Assume that information to be estimated is fed into a system at discrete time intervals. The information describes a situation S characterized by a set of n parameters necessary for estimating S. Thus, at any time t the situation is described by set $\langle \pi_1^t, \pi_2^t, \ldots, \pi_n^t \rangle$ where π_i^t is fixed value of the parameter at time t. Estimates for S are taken from set Ξ. In the simplest case Ξ may consist only of two estimates conditionally denoted by 0 and 1 and having various interpretations, for instance, "good-bad", "useful-unuseful", "useful-harmful", etc. In a more complicated case, set Ξ may involve m elements having appropriate interpretation, for example, "useful-indifferent-harmful"

or "very pleasant-pleasant-don't care-unpleasant-very unpleasant", etc. The problem of estimating situation S is to find a mapping φ which for each specific set $\langle \pi_1^t, \pi_2^t, \ldots, \pi_n^t \rangle$ would unambiguously indicate an element in Ξ.

The two-valued estimate is the most common in the human practice. At the first step values of parameters π_i are estimated individually, that is to say n mappings of the kind $\pi_i^t \to \Xi_i$ are performed where Ξ_i is the set of estimates for parameter π_i. Then in order to estimate situation S, a set of estimates of the situation parameters $\langle \xi_1(\pi_1^t), \xi_2(\pi_2^t), \ldots, \xi_n(\pi_n^t) \rangle$ is used rather than set $\langle \pi_1^t, \pi_2^t, \ldots, \pi_n^t \rangle$ where $\xi_i(\pi_i^t)$ are elements of Ξ_i. Note that parameters π_i may take not only numerical values and that their estimates may be always coded with natural numbers. Take among Ξ and Ξ_i $i = 1, 2, \ldots, n$, the set of the highest cardinality (the case of infinite set of estimates is not considered as being of nearly no practical importance). Let this set be Ξ^* with cardinality l. Then increase the cardinalities of other sets up to this value and consider one common set of estimates for the parameters and the situation. With n parameters and l values of the estimate there are l^n different sets $\langle \xi_1(\pi_1), \xi_2(\pi_2), \ldots, \xi_n(\pi_n) \rangle$. Finding estimate of S will reduce to a mapping $(\Xi^*)^n \to \Xi^*$ which may be naturally represented as a table with l^n rows and $(n + 1)$ columns. The first n columns contain estimates of parameters, and the $(n + 1)$-st one gives values of the estimate of S. Such a table is known to define a function of the l-valued logic.

Thus, the procedure above reduces the estimation of S by a set of estimates of its parameters to finding an appropriate function of the l-valued logic.

2. The binary case

Consider the simplest case where cardinality of set Ξ^* is two. Here the estimate function is a conventional Boolean one. To illustrate application of the binary logic consider the following example. Take weather as a situation to be estimated. The relevant parameters are temperature, precipitation, state of the sky and state of the soil. Denote them, respectively, by π_1, π_2, π_3 and π_4. Table 1 gives estimates of these parameters.

Our task is to estimate the weather. The estimate has two values — 1 for good, and 0 for bad weather. By means of estimates in Table 1 a new table having 16 rows may be readily constructed where each set characterizes a combination of parameters, e.g. $\langle 0110 \rangle$ means that the weather is cloudless, without precipitation with humid soil and temperature below 20°C. However, not all possible combinations of parameter estimates will be meaningful. For instance, "a bolt from the blue" is impossible. Therefore, some rows will

Table 1

Parameter	Name	Estimate threshold	Estimate
π_1	Temperature	at least 20°C below 20°C	1 0
π_2	Precipitation	no yes	1 0
π_3	State of the sky	cloudless cloudy	1 0
π_4	State of the soil	dry humid	1 0

actually generate no definite estimate for the logic function estimating S, thus making it subdefinite.

Note also that not the least of the factors that the situation estimate is closely related to the aim of estimating which defines both the kind of the estimate function and the parameter estimates. If, for example, we are going to take a stroll, the situation defined by set $\langle 1111 \rangle$, should be doubtless estimated as "good weather", and the estimate function gives at this set 1. However, if we estimate weather in the terms of sowing campaign, set $\langle 1111 \rangle$ should be estimated as "bad weather", and the estimating function at this set should be 0. In the second case the individual estimates of the parameters turn out to be in discord with the estimate for the situation as a whole. This disagreement manifests itself in the fact that interpretations of 1 for the parameters and the situation *in toto* differ. Table 2 gives estimate functions for these two cases. Here φ_1 denotes the "stroll" function, and φ_2 denotes the "sowing" one. Empty entries in the columns of function values correspond to physically meaningless sets where the estimate functions may be arbitrarily defined by 0's or 1's.

From such a table one can easily pass to the analytical representation of the estimate functions in the form of standard sum-of-products or product-of-sums, or in some other form.

Unfortunately, the binary case offers poor possibilities of studying the real situation estimating and human decisions. It would be more interesting and instructive to discuss a case where set Ξ^* has at least three elements.

Table 2

Parameter estimates				φ_1	φ_2
$\xi_1(\pi_1)$	$\xi_2(\pi_2)$	$\xi_3(\pi_3)$	$\xi_4(\pi_4)$		
0	0	0	0	0	0
0	0	0	1	0	0
0	0	1	0		
0	0	1	1		
0	1	0	0	1	1
0	1	0	1	1	0
0	1	1	0	1	1
0	1	1	1	1	0
1	0	0	0	0	1
1	0	0	1	0	0
1	0	1	0		
1	0	1	1		
1	1	0	0	1	1
1	1	0	1	1	0
1	1	1	0	1	1
1	1	1	1	1	0

3. The ternary case

Assume that set Ξ^* has three elements 0, 1/2 and 1 interpreted as "unfavourable", "lack of information" and "favourable". We thus allow for indeterminacy in the estimates of parameters and situations thereby making the estimation model more interesting and enabling us to more psychologically realistic estimate functions. The psychological features of an individual who makes estimates will manifest themselves in situations having parameters with the estimate 1/2. With our interpretation of the estimate, it might be seen by this individual in both favourable or unfavourable light. The individuals disposed to the unfavourable interpretation of 1/2 form a class of pessimists, and those who are inclined to the favourable interpretation of 1/2 form that of optimists. The individuals who do not see behind 1/2 neither 0, nor 1 form the class of objectivists.

To define these classes more precisely narrow the class of estimate functions. As it was shown by numerous psychological studies, see e.g. [1], in the majority of cases the man makes use of the conjunction, disjunction and negation functions, and of disjunctive and conjunctive forms with small

number of short terms. This enables us to discuss psychological features of individuals by means of disjunction and conjunction functions. The singularity of estimation will manifest itself in that instead of using objective conjunctions and disjunctions the individual will employ "similar" pessimistic or optimistic conjunctions and disjunctions. Since these biased functions are not, strictly speaking, conjunctions and disjunctions, we shall refer to them as quasi-conjunctions and quasi-disjunctions respectively. From quasi-conjunctions and quasi-disjunctions one constructs analogs of disjunctive and conjunctive normal forms.

Consider first quasi-disjunctions. Assign to each estimate a numerical weight depending on the number of parameters defining a situation. Namely, weight 0 will correspond to estimate 0, weight 1 to estimate 1 and weight $1/n + 1$ — to estimate 1/2. The sum of estimate weights will be referred to as the weight of set $\langle \xi_1(\pi_1), \xi_2(\pi_2), \ldots, \xi_n(\pi_n) \rangle$. The weight of estimate 1/2 is chosen so that the weight of a set consisting only of estimates 1/2 is less than 1. Take now two numbers α and β $(0 \leq \alpha, \beta \leq 1)$ which will be referred to, respectively, as the lower and upper thresholds, and define the estimate function as assuming 0 at sets with weight less than α, 1 at sets with weight at least equal to β, and 1/2 at all other sets. By varying the values of the lower and upper thresholds one can generate a set of estimate functions which are, like disjunction, monotonous with respect to the values of naturally ordered arguments $(0 < 1/2 < 1)$. Functions obtained in this manner will be termed quasi-disjunctions. If one takes α $1/(n + 1)$ and $\beta = 1$, the conventional disjunction results. If now one fixes the lower threshold such as in disjunction and decreases the value of β, a portion of sets consisting any of estimates 0 and 1/2 will have 1 as the value of the estimate function. Such quasi-disjunctions may be naturally called optimistic. If now one fixes the upper threshold such as in disjunction and increases the value of α, a portion of sets involving estimates 0 and 1/2 will have 0 as the value of the estimate function. Such quasi-disjunctions may be naturally termed pessimistic. Thus, the optimists are inclined to regard a part of indeterminate estimates as favourable, but pessimists always suspect that among parameters with indeterminate estimates there always are unfavourable estimates.

Consider now quasi-conjunctions. We are going to use for these functions the same parameter estimate weights as for the case of quasi-disjunctions. The weight of a set, however, will be determined now by multiplying the weights rather than summing them. As with the quasi-disjunctions, introduce lower γ and upper δ thresholds. For the conventional conjunction $\gamma = 1/(n + 1)^n$ and $\delta = 1$. By varying correspondingly γ or δ we obtain the complete set of optimistic and pessimistic quasi-conjunctions. For the case

of two parameters defining a situation, values of all quasi-disjunctions and quasi-conjunctions are tabulated in Table 3. The two left columns show the weights of sets for quasi-disjunctions and quasi-conjunctions, respectively. For the estimate functions, the subscripts define the type of function (0 means optimistic, p — pessimistic, and absence of the second subscript means objective). The superscripts indicate the degree of optimism and pessimism. The higher is this degree, the more optimistic/pessimistic the estimate function is. For instance, in function $\varphi^1_{\delta 0}$, under presence of one 1/2 in the set of parameter estimates the true estimate of the parameter is assumed to be favourable. If the set has two estimates 1/2, the situation is, on the whole, indeterminate. Function $\varphi^2_{\delta 0}$ characterizes the "frantic optimist" who even with set $\langle 1/2, 1/2\rangle$ is inclined to believe that all parameters are favourable to him.

Table 3

Σ	π	$\xi_1(\pi_1)$	$\xi_2(\pi_2)$	Quasi-disjunction					Quasi-conjunction				
				$\varphi^2_{\vee p}$	$\varphi^1_{\vee p}$	$\varphi_{\vee}$	$\varphi^1_{\vee 0}$	$\varphi^2_{\vee 0}$	$\varphi^2_{\& p}$	$\varphi^1_{\& p}$	$\varphi_{\&}$	$\varphi^1_{\& 0}$	$\varphi^2_{\& 0}$
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1/3	0	0	1/2	0	0	1/2	1/2	1	0	0	0	0	0
1	0	0	1	1	1	1	1	1	0	0	0	0	0
1/3	0	1/2	0	0	0	1/2	1/2	1	0	0	0	0	0
2/3	1/9	1/2	1/2	0	1/2	1/2	1	1	0	0	1/2	1/2	1
4/3	1/3	1/2	1	1	1	1	1	1	0	1/2	1/2	1	1
1	0	1	0	1	1	1	1	1	0	0	0	0	0
4/3	1/3	1	1/2	1	1	1	1	1	0	1/2	1/2	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1

The following facetious example illustrates the functions we have introduced. An individual A after long hesitations has made up his mind to marry. In his opinion two parameters are of importance for taking decision: the character of his fiancée and that of her mother whom he seemingly shall have to live with. But during his period of courtship A has been unable to establish positive estimates of these parameters and has to decide upon uniting himself in holy matrimony on the basis of set $\langle 1/2, 1/2\rangle$. The estimate function is quasi-conjunction. If A is "frantic optimist", he believes that he should be lucky all his life and that characters of his wife-to-be and her mother are favourable to him. Such a person will decide to marry. If A is a "careful pessimist" (i.e. pessimist of the first rank), he will not decide upon a marriage because he does not expect that, information lacking, both parameters will

be favourable. The "frantic pessimist" will not marry even with sets $\langle 1/2, 1\rangle$ and $\langle 1, 1/2\rangle$ being convinced that the world is badly organized. The "careful pessimist" in this situation will temporize without taking the final decision.

4. Some generalization

The reader should not be disturbed by the fact that we studied the ternary rather than the l-valued case. The passage from the ternary case to the l-valued one is done without difficulty and does not lead to any principally new results. Only the number of ranks for pessimists and optimists increases, as well as the number of thresholds which should be fixed to find the values of the estimate function.

So far we regarded all parameters as equally important. In practice, one often meets with cases where some parameters are more important for estimating a situation, than some others. It may be shown that this inequality is corrected by introducing individual weights of parameters taken into consideration when determining the weight of a set.

The case of correlated parameters presents a more difficult problem where approach to determination of the estimate function seems to be somewhat different.

5. An example

It seems to us interesting to conclude the paper with an example illustrating application of the apparatus developed above. The theory of collective behaviour of automata which is closely related to the development of decentralized control systems, have given much consideration to allocation of a limited set of resources [2]. This problem is of practical interest because it may be readily interpreted as a problem in the theory of bulk service or network planning, in allocation of industrial equipment design of good operation systems for computer networks. In all models considered thus far the collective of automata was assumed to be homogeneous. We have carried out computer simulation of limited resource allocation in a nonhomogeneous collective of automata. Each automaton had a fixed pessimistic or optimistic estimate function and took decisions depending on its value. The problem was posed as follows: there are q cribs characterized by parameter $p_i(t)$ each interpreted as the probability that at time t the ith crib would contain food. Values of $p_i(t)$ were not known by the automata *a priori*. If at time t a crib contained food, its value was always a, a fixed value which was the same for all cribs. If at that time there were r automata at the crib, they got a/r units of food each. A time interval T (number of cycles in allocation of automata among

cribs) and a number b, the same for all automata, were fixed. If during T cycles of operation an automaton had total reserve of food less than b, it was fixed in proportion to the difference between T and its food reserve. If the total reserve was greater than b, it got a prize proportional to the difference. The problem of collective behaviour consisted in maximization of the total collective gain. On the basis of their experience the automata formed two parameters — mean frequency of appearance of food in each of the cribs and mean number of automata coming to each crib. Values of these parameters for different automata differed because they were formed on the basis of personal experience of each automaton. If, for instance, probability of food in a given crib as evaluated by a given automaton through the mean frequency of appearance of food was less than 1/4, parameter $p_i(t)$ was evaluated as 0, if this frequency was greater than 3/4, parameter $p_i(t)$ was evaluated as 1. In all other cases the value of $p_i(t)$ was 1/2. The second parameter was estimated in a similar manner depending on the number of automata visiting a given crib.

At each cycle all automata were evaluating individually all cribs by means of the estimate functions assigned to them. After this they choose the crib with maximal estimate and if there were several such estimates, they choose any of them with equal probability.

The total gain was also computed for collectives consisting of objectivists only, and of a uniform mix of all optimists, pessimists and objectivists whose estimate functions are tabulated in Table 3. It turned out that the mixed collective has greater gain as compared to that of the uniform one.

References

1. *Piaget, J.*, Le langage et la pensée chez l'enfant. Neuchatel, Delachaux et Nieftl, 1923.
2. *Varshavskii, V. I.*, Collective behaviour of automata. Moscow, 1973 (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 которы 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 гранок.

Статье должна предшествовать аннотация объемом до 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 высылаются гранки статьи, которые 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. Невероятностная взаимная информация без памяти (*Goppa, V. D.:* Nonprobabilistic mutual information without memory) 97

Csáki, F., El-Said Hamada: Method for design of linear optimal systems in frequency domain (*Чаки Ф., Эл-Саид Хамада* Метод конструкции линейных оптимальных систем в области частот) 103

Захаров В. Н., Поспелов Д. А. Об одном подходе к решению оценочных задач (*Zakharov, V. N., Pospelov, D. A.:* An approach to estimation problems) 121

Söderström, T.: Ergodicity results for sample covariances (*Содерстром Т.* Эргодические результаты для коварианции, полученные на основе реализаций) 131

Bukhalev, V. A.: The analysis of the accuracy of dynamic systems changing their structure in the random time moment (*Бухалев В. А.* Анализ точности динамических систем, изменяющих структуру в случайный момент времени) 139

Hetthéssy, J., Haber, R., Keviczky, L.: A separating transformation to the minimum variance control (*Хетеши Й., Хабер Р., Кевицки Л.* О сепарабельном преобразовании в задачах управления с минимальной среднеквадратической ошибкой) 147

Tsonkov, S. M.: Estimating the dynamics of modified systems with variable structure (*Цонков С. М.* Оценка динамики модифицированных систем с переменной структурой) 155

Muszély, Gy.: The reduction of the dimension of input pattern recognition vectors by searching the optimal parameter set for linear regression (*Мусели Д.* Сокращение исходного описания в задаче распознавания образов с помощью метода поиска оптимального подмножества параметров линейной регрессии) 165

Ostianu, V. M.: International IFAC Symposium on Discrete Systems 183

Sinitsin, I. N., Shaikin, M. E.: Sixth All-Union Conference on Control Problems 185

Habermayer, M.: AICA Symposium on Hybrid Computation Dynamic Systems Design 189

316.920

VOL. 4 * NUMBER 3
ТОМ 4 * НОМЕР 3

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

1975

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 four issues published per year make up a volume of some 320 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

Президиумами Академии Наук СССР и Академии Наук ВНР было принято решение о совместном издании журнала

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал создан для обеспечения быстрейшей публикации материалов о новейших результатах научных исследований, разработок и кратких сообщений 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 издается четыре раза в год общим объёмом около 320 печатных страниц.

Журнал в первую очередь будет публиковать научные достижения 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. EMELIANOV (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. 4 (3), pp. 193–204 (1975)

OPTIMAL CONTROL OF A STORAGE LEVEL USING STOCHASTIC PROGRAMMING

A. PRÉKOPA

(Budapest)

(Received October 14, 1974)

A stochastic programming model system is formulated for the control of a discrete time parameter Gaussian process. The problems representing the model system are solved by nonlinear programming procedures. The method can be applied e.g. for the dynamic regulation of the water level of the Lake Balaton.

1. Introduction

In the present paper we deal with the following problem: a discrete time parameter stochastic process $\xi_1, \xi_2, \ldots$ is given; a sequence of numbers $z_1, z_2, \ldots$ is to be determined such that the probability of the event

$$a_k \leqq \zeta_k - Z_k = b_k \qquad k = 1, 2, \ldots \tag{1.1}$$

be maximal subject to the constraints

$$0 \leqq z_k \leqq K, \qquad k = 1, 2, \ldots \tag{1.2}$$

Here $a_1, a_2, \ldots, b_1, b_2, \ldots$ are fixed numbers, K is a fixed positive number and for the sake of simplicity let ζ_0, the initial storage level, be also a fixed number. The decision variables are $z_1, z_2, \ldots$.

We assume that $z_1, z_2, \ldots$ can be determined sequentially so that the value of z_k is to be determined only after the realization of the random variables $\xi_1, \ldots, \xi_{k-1}$.

In our investigation we assume that $\xi_1, \xi_2, \ldots$ is a Gaussian stochastic process. It follows from this that the probability distribution of the process is uniquely determined by the expectations

$$\mu_k = E(\xi_k), \qquad k = 1, 2, \ldots,$$

the dispersions

$$\sigma_k = \sqrt{E\left[(\xi_k - \mu_k)^2\right]}, \qquad = 1, 2, \ldots$$

and the correlation coefficients belonging to the pairs ξ_i, ξ_k equal to

$$\frac{E[(\xi_i - \mu_i)(\xi_k - \mu_k)]}{\sigma_i \sigma_k}$$

if $\sigma_i > 0$, $\sigma_k > 0$ and by definition equal to zero if $\sigma_i \sigma_k = 0$.

Not only the Gaussian process can be used in the method described below. We restrict ourselves, however, to the case of the Gaussian process because in the special practical problem for the solution of which we essentially developed the method, this type of stochastic process occurs. The mentioned problem is the regulation of the water level of the Lake Balaton. Regarding the Gaussian nature of the relevant stochastic process see [1, pp. 4—42].

Time is subdivided into periods, where one period may mean e.g. 1 month. The random variable ξ_k means the inflow in the kth period minus the loss of evaporation, z_k the amount of water to be released through the Sió channel into the River Danube, K is the capacity of the Sió channel, i.e. the maximal amount of water what is possible to release in one period, finally ζ_0 is the initial water content of the Lake Balaton. Our purpose is to formulate a release policy which ensures that the water level is always between prescribed levels, more exactly which maximizes the probability that the water level is between prescribed levels. In the sequel we shall describe what we mean under the above probability maximization. Various further problems can be formulated in connection with the input and release process. Stochastic programming is reach in such models which may be applicable here. One further example will be mentioned at the end of the paper.

We do not assume that the process $\xi_1, \xi_2, \ldots$ has independent values or that it is a Markov process. There are conditions which usually occur in stochastic control problems. Our model system takes into account the general nature of the stochastic process within the frame that it is Gaussian. It will turn out from the discussion what other stochastic processes allow the essentially same treatment.

2. The dynamic model system formulated for the regulation problem

For the determination of the value of the first decision variable we formulate the following stochastic programming problem

$$0 \leq z_k \leq K_k, \qquad k = 1, \ldots, N, \tag{2.1}$$

$$\max P(a_k \leq \zeta_k - Z_k \leq b_k, \qquad k = 1, \ldots, N).$$

Here $z_1, \ldots, z_n$ are the decision variables. After the solution of problem (2.1) we accept only the first component of the optimal solution as a final value. In order to avoid sophisticated notation we do not introduce another symbol but denote this optimal value also by z_1. We define final z values by induction. Suppose that the final values $z_1, \ldots, z_n$ are already determined. In order to determine z_{n+1} we consider the following stochastic programming problem:

$$0 \leqq z_k \leqq K, \qquad k = n+1, \ldots, n+N, \tag{2.2}$$

$$\max P\,(a_k \leqq \zeta_k - Z_k \leqq b_k, \qquad k = n+1, \ldots, n+N \mid \xi_1, \ldots, \xi_n).$$

As a solution we obtain a number system $z_{n+1}, \ldots z_{n+N}$ but we only accept z_{n+1} as a final value. We remark that the N appearing in problem (2.2) may depend on n. It is advisable to choose such N that the dependence between the two groups of random variables $\xi_1, \ldots, \xi_n$ and $\xi_{n+1}, \xi_{n+2}, \ldots$ be loose, if it is possible to do. If N is very large, then computational difficulty arises. In fact problem (2.2) is a nonlinear programming problem which we have to solve numerically. When applying any nonlinear programming procedure, we frequently have to determine probabilities standing in the objective function which is very time consuming since we have to integrate the N-variate probability density on a subset of the space R^n.

3. Mathematical properties of the dynamic model system

First we mention some general notions and theorems. A *point function* $f \geqq 0$ defined on R^m is said to be *logarithmic concave* if for every pair $x_1, x_2 \in R^m$ and every $0 < \lambda < 1$ we have

$$f(\lambda \underline{x}_1 + (1-\lambda)\,\underline{x}_2) \geqq [f(\underline{x}_1)^{\lambda}[f(\underline{x}_2)]^{1-\lambda}. \tag{3.1}$$

The *measure* P defined on the measurable subsets of R^m is said to be logarithmic concave if for every convex subsets A, B of R^m and every $0 < \lambda < 1$ we have

$$P(\lambda A + (1-\lambda)\,B) \geqq [\mathrm{P}(A)]^{\lambda} \quad [P(B)]^{1-\lambda}. \tag{3.2}$$

In general the Minkowski sum $G + D$ of the sets G, D is the set $\{\underline{g} + \underline{d} \mid \underline{g} \in G, \underline{d} \in D\}$ and the λ multiple λG of the set G is the set $\{\lambda \underline{g} \mid \underline{g} \in G\}$. In [9] and [10] we proved the following

Theorem 1. If f is a logarithmic concave function defined on R^m and the measure P is generated by the function f, i.e.

$$P(G) = \int_G f(\underline{x})\, d\underline{x}, \tag{3.3}$$

where G is a measurable subset of the space R^m, then P is a logarithmic concave measure.

Theorem 1 implies that for every convex subset A of the space R^m, the integral

$$\int_{A+x} f(\underline{t})\, d\underline{t} \tag{3.4}$$

is a logarithmic concave point function of the variable $\underline{x}$ in the entire space R^m.

Now we are dealing with functions f which are probability densities. The probability density function of a nondegenerated normal distribution is the following:

$$f(\underline{x}) = \left(\frac{\det \mathbf{C}^{-1}}{(2\pi)^n}\right)^{\frac{1}{2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})' C^{-1}(\underline{x}-\underline{\mu})}. \tag{3.5}$$

where μ is the vector of expectations and $\mathbf{C}$ is the covariance matrix. Since the distribution is nondegenerated, $\mathbf{C}$ is a positive definite matrix hence C^{-1} is also positive definite. This implies that the function

$$(\underline{x} - \underline{\mu})' C^{-1}(\underline{x} - \underline{\mu}) \tag{3.6}$$

is convex and that the function f is logarithmic concave. Thus by Theorem 1 the probability distribution generated by the density function (3.5) is logarithmic concave.

Let us consider a nondegenerated normal distribution in the $m = N + n$-dimensional space, where N and n are positive integers and suppose that it is the joint distribution of the N component $\underline{\xi}$ and the n component $\underline{\eta}$ random vectors. Let us partition the corresponding $(N + n)\times(N + n)$ covariance matrix $\mathbf{C}$ according to this in the following manner

$$\mathbf{C} = \begin{pmatrix} S & U \\ U' & T \end{pmatrix}, \tag{3.7}$$

$$\underline{\mu} = \begin{pmatrix} \underline{v} \\ \underline{\tau} \end{pmatrix}. \tag{3.8}$$

It is well known that the conditional distribution of $\underline{\xi}$ gives that $\underline{\eta} = \underline{y}$ is also a normal distribution in R^N with expectation

$$\underline{v} + UT^{-1}(\underline{y} - \tau) \tag{3.9}$$

and covariance

$$S - UT^{-1}U'. \tag{3.10}$$

Denoting by $f(\underline{x} \mid \underline{y})$ the conditional probability density, by Theorem 1 we conclude that the measure

$$\int_G f(\underline{x} \mid \underline{y})\, d\underline{x} \tag{3.11}$$

defined on the measurable subsets of R^N, is logarithmic concave for every fixed $\underline{y}$. Now we prove two general theorems.

Theorem 2. The function $f(\underline{x} \mid \underline{y})$ is a logarithmic concave function of the variables in $\underline{x}$ and $\underline{y}$.

Proof. The detailed expression of the function $f(\underline{x} \mid \underline{y})$ is the following

$$f(\underline{x} \mid \underline{y}) = K \exp\left\{-\frac{1}{2}[\underline{x} - \underline{v} - UT^{-1}(\underline{y} - \underline{\tau})]'\right.$$
$$\left.[S - UT^{-1}U']^{-1}[\underline{x} - \underline{v} - UT^{-1}(\underline{y} - \underline{\tau})]\right\}, \tag{3.12}$$

where

$$K = \left[\frac{\det(S - UT^{-1}U')}{(2\pi)^n}\right]^{\frac{1}{2}}. \tag{3.13}$$

Since the function of the variables $\underline{x}$, $\underline{y}$

$$\underline{x} - \underline{v} - UT^{-1}(\underline{y} - \underline{\tau}) \tag{3.14}$$

is linear, our assertion follows immediately from the logarithmic concavity of the probability density function of the normal distribution.

Theorem 3. If G is an arbitrary but fixed convex subset of the space R^N, then the function of the variable $\underline{y}$

$$\int_G f(\underline{x} \mid \underline{y})\, d\underline{x} \tag{3.15}$$

is logarithmically concave in the space R^n.

Proof. Theorem 6 in [10] states that if $h(\underline{x}, \underline{y})$ is a logarithmic concave function of all variables in $\underline{x}$ and $\underline{y}$ and we integrate it with respect to $\underline{x}$ over a fixed convex set of $\underline{x}$ vectors, then this integral is a logarithmic concave function of $\underline{y}$. By Theorem 2 this implies our assertion.

Let us return to the dynamic model system formulated in Section 2. Assume that all finite-dimensional distributions of the stochastic process $\xi_1, \xi_2, \ldots$ are nondegenerated. The basic properties of the model system are expressed in the following theorems.

Theorem 4. The objective function of problem (2.1) is a logarithmic concave function of the variables $z_1, \ldots, z_N$.

Proof. Let

$$\begin{matrix} z_1^{(1)}, \ldots, z_N^{(1)}. \\ z_1^{(2(}, \ldots, z_N^{(2)} \end{matrix} \tag{3.16}$$

be two number system and $0 < \lambda < 1$. Define the following number system

$$z_1^{(3)} = \lambda z_1^{(1)} + (1 - \lambda) z_1^{(2)}, \ldots, z_N^{(3)} = \lambda z_N^{(1)} + (1 - \lambda) z_N^{(2)}. \tag{3.17}$$

The corresponding partial sums will be denoted by

$$Z_k^{(1)}, Z_k^{(2)}, Z_k^{(3)}, \qquad k = 1, \ldots, n.$$

Introduce further the notation

$$X_k = \zeta_0 + x_1 + \ldots + x_k, \quad k = 1, \ldots, N. \tag{3.18}$$

It is easy to see that for the vectors consisting of components $x_1, \ldots, x_N$, the following equality holds:

$$\begin{aligned} \lambda\{\underline{x} \mid a_k \leqq X_k - Z_k^{(1)} \leqq b_k,\ k = 1, \ldots, N\} + \\ + (1 - \lambda)\ \{\underline{x} \mid a_k \leqq X_k - Z_k^{(2)} \leqq b_k, k = 1, \ldots, N\} = \\ = \{\underline{x} \mid a_k \leqq X_k - Z_k^{(3)} \leqq b_k,\ k = 1, \ldots, N\}. \end{aligned} \tag{3.19}$$

The probability that the random vector having components $\xi_1, \ldots, \xi_N$ falls into the set

$$\{\underline{x} \mid a_k \leqq X_k - Z_k^{(i)} \leqq b_k, \quad k = 1, \ldots, N\} \tag{3.20}$$

is equal to

$$P(a_k \leqq \zeta_k - Z_k^{(i)} \leqq b_k, \quad k = 1, \ldots, N) \tag{3.21}$$

for $i = 1, 2, 3$ and this random vector has logarithmic concave density, our assertion follows from Theorem 1.

Before formulating the next theorem, we rewrite the objective function of problem (2.2) in the following form

$$P(a_k \leqq \zeta_k - Z_k \leqq b_k,\ k = n + 1, \ldots, n + N \mid \xi_1, = y_1, \ldots, \xi_n = y_n). \tag{3.22}$$

Theorem 5. For any fixed $y_1, \ldots, y_n$, the function (3.22) is logarithmic concave in the variables $z_{n+1}, \ldots, z_{n+N}$.

Proof. For fixed $y_1, \ldots, y_n$ the joint distribution of $\xi_{n+1}, \ldots, \xi_{n+N}$ is a normal distribution, hence the proof of Theorem 5 can be carried out similarly to the proof of Theorem 4 with a trivial modification.

4. Algorithms for the solution of problems (2.1) and (2.2)

Without loss of generality we can restrict ourselves to the algorithmic solution of problem (2.1). In fact we mentioned in Section 3 that if a collection of random variables has a joint normal distribution, the conditional distribution of some of them given the others is also a normal distribution. We also gave the expressions for its expectation vector and covariance matrix. Thus any method proposed for the solution of problem (2.1) is also a proposed method for the solution of problem (2.2). Two algorithms will be described for the solution of problem (2.1).

Algorithm I. This is the gradient method where we take into account that the constraints are linear and the objective function is logarithmic concave. We introduce new variables which are $y_1, \ldots, y_N$ so that we set

$$y_k = z_1 + \ldots + z_k, \qquad k = 1, \ldots, N. \tag{4.1}$$

Obviously $y_k = Z_k$. The reason of the introduction of the new symbols $y_1, \ldots, y_N$ is that we would like to denote variables by small latin letters. With these variables problem (2.1) has the following form

$$\begin{aligned} &0 \leqq y_1 \leqq K, \\ &0 \leqq y_2 - y_1 \leqq K, \\ &\quad\vdots \\ &0 \leqq y_N - y_{N-1} \leqq K, \\ &\max P(a_k \leqq \zeta_k - y_k \leqq b_k,\ k = 1, \ldots, N). \end{aligned} \tag{4.2}$$

The expectations of the random variables are given by

$$E(\zeta_k) = \zeta_0 + \mu_1 + \ldots + \mu_k, \qquad k = 1, \ldots, N. \tag{4.3}$$

The covariance matrix $\mathbf{V}$ of the random variables $\zeta_1, \ldots, \zeta_N$ has the form

$$\mathbf{V} = \mathbf{DSD}', \tag{4.4}$$

where $\mathbf{S}$ is the covariance matrix of the random variables $\xi_1, \ldots, \xi_N$ and $\mathbf{D}$ is the following $N + N$ matrix

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \ldots 0 \\ 1 & 1 & 0 \ldots 0 \\ - & - & - \quad - \quad - \\ 1 & 1 & 1 \ldots 1 \end{pmatrix}. \tag{4.5}$$

Each iteration of the gradient method consists of two parts. In the first one we find a direction and in the second one we determine a step length. In the begin-

ning of the method we start from an arbitrary vector

$$\underline{y} = \underline{y}^{(1)}$$

which statisfies the constraints of the problem. Suppose that $\underline{y}^{(1)}, \ldots, \underline{y}^{(j)}$ are already determined which all satisfy the constraints. For the determination of the vectory $y^{(j+1)}$ first we solve the linear programming problem

$$\begin{aligned} &0 \leqq y_1 \leqq K, \\ &0 \leqq y_2 - y_1 \leqq K, \\ &\quad \vdots \\ &0 \leqq y_N - y_{N-1} \leqq K, \\ &\max \nabla P(a_k \leqq \zeta_k - y_k^{(j)} \leqq b_k, \quad k = 1, \ldots, N) \\ &\qquad (\underline{y} - \underline{y}^{(j)}), \end{aligned} \tag{4.6}$$

where $y_1^{(j)}, \ldots, y_N^{(j)}$ are the components of $\underline{y}^{(j)}$. The gradient of a function is considered as a row vector while other vectors are column vectors. Thus in the objective function of problem (4.6) there stands a scalar product. Problem (4.6) has a finite optimum since the convex polyhedron defined by the constraints is bounded. If the optimum value is equal to 0, then $\underline{y}^{(j)}$ is an optimal solution of problem (4.2). If this is not the case and $\underline{y}_{\text{opt}}^{(j)}$ is an optimal solution of problem (4.2), the we maximize the objective function of problem (4.2) on the intersection of the ray

$$\{\underline{y}^{(j)} + \lambda(\underline{y}_{\text{opt}}^{(j)} - \underline{y}^{(j)}) \mid \lambda \geq 0\} \tag{4.7}$$

and the set of feasible solutions of problem (4.2). This fixes the step length. Choosing an arbitrary maximizing vector, this will be $\underline{y}^{(j+1)}$.

If the procedure does not terminate in a finite number of iterations, then the sequence of values of the objective function of problem (4.2) belonging to the vectors $\underline{y}^{(1)}, \underline{y}^{(2)}, \ldots$ converges to the optimum value of problem (4.2).

The above-mentioned assertions hold true in general concerning bounded constraining convex polyhedron and in R^N everywhere defined objective function having continuous gradient in the entire space (see e.g. [5], [8], [12]).

Still we have to give the expression of the gradient of the objective function of problem (4.2). First we rewrite the function in the following form

$$P\left(\frac{a_k + y_k - \xi_0 - \mu_1 - \ldots - \mu_k}{\sqrt{v_{kk}}} \leqq \frac{\xi_k - \xi_0 - \mu_1 - \ldots - \mu_k}{\sqrt{v_{kk}}} \leqq \right.$$

$$\left. \leqq \frac{b_k + y_k - \xi_0 - \mu_1 - \ldots - \mu_k}{\sqrt{v_{kk}}}, k = 1, \ldots, N\right). \tag{4.8}$$

where $v_{11}, \ldots, v_{NN}$ are the diagonal elements of the matrix $\mathbf{V}$. The random variables standing in the middle within the parentheses in (4.8) have $N(0, 1)$ distribution and their correlation matrix is the following

$$\mathbf{R} = \begin{pmatrix} \frac{1}{\sqrt{v_{11}}} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{v_{NN}}} \end{pmatrix} \quad \mathbf{V} = \begin{pmatrix} \frac{1}{\sqrt{v_{11}}} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \frac{1}{\sqrt{v_{NN}}} \end{pmatrix}. \tag{4.9}$$

where on the right-hand side the two detailed matrices are diagonal matrices. Thus the objective function (4.8) can be expressed in a well-known manner with the aid of the Nth differences of the function

$$\begin{aligned} \Phi(y; R) &= \int_{\underline{t} \leq \underline{y}} \varphi(\underline{t}; R)\, d\underline{t}, \\ \varphi(\underline{t}; R) &= \left(\frac{\det R^{-1}}{(2\pi)^n}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\underline{y}' R^{-1} \underline{y}}. \end{aligned} \tag{4.10}$$

For example, in case of $N = 2$ the function (4.8) is equal to

$$\begin{aligned} &\Phi\left(\frac{b_1 + y_1 - \xi_0 - \mu_1}{\sqrt{v_{11}}}, \frac{b_2 + y_2 - \xi_0 - \mu_1 - \mu_2}{\sqrt{v_{22}}}; R\right) - \\ &- \Phi\left(\frac{b_1 + y_1 - \xi_0 - \mu_1}{\sqrt{v_{11}}}, \frac{a_2 + y_2 - \xi_0 - \mu_1 - \mu_2}{\sqrt{v_{22}}}; R\right) - \\ &- \Phi\left(\frac{a_1 + y_1 - \xi_0 - \mu_1}{\sqrt{v_{11}}}, \frac{b_2 + y_2 - \xi_0 - \mu_1 - \mu_2}{\sqrt{v_{22}}}; R\right) + \\ &+ \Phi\left(\frac{a_1 + y_1 - \xi_0 - \mu_1}{\sqrt{v_{11}}}, \frac{a_2 + y_2 - \xi_0 - \mu_1 - \mu_2}{\sqrt{v_{22}}}; R\right). \end{aligned} \tag{4.11}$$

The gradient of the objective function can be derived from the mentioned Nth difference and the following well-known formula [2]:

$$\frac{\partial \Phi(\underline{y}; R)}{\partial y_i} = \Phi\left(\frac{y_j - r_{ji} y_i}{\sqrt{1 - r_{ji}^2}};\ j = 1, \ldots, i-1, i+1, \ldots, N; R_1\right) \varphi(y_i), \tag{4.12}$$

where $\varphi(y)$ is the density function of the $N(0, 1)$ distribution and R_1 is the correlation matrix with entries

$$\frac{r_{jk} - r_{ji} r_{ki}}{\sqrt{1 - r_{ji}^2}\sqrt{1 - r_{ki}^2}},\ j, k = 1, \ldots, i-1, i+1, \ldots, N. \tag{4.13}$$

When applying the gradient method we have to compute the values of the objective function and the values of the gradient of the objective function. According to the above discussion all of them can be expressed in terms of the values of probability distribution functions belonging to normal distributions, where the dimension of the distribution is N and $N - 1$ according that we compute the values or the gradient values of the objective function of problem (4.2).

Algorithm II. This algorithm is the sequential unconstrained minimization technique (SUMT) with a logarithmic penalty function. As we mentioned previously it is enough to formulate the method for problem (2.1).

We consider the following function

$$\begin{aligned} &- \log P(a_k \leq \zeta_k - Z_k \leq b_k, \quad k = 1, \ldots, N) \\ &- r \sum_{k=1}^{N} [\log (K - z_k) + \log z_k], \end{aligned} \tag{4.14}$$

where $r > 0$ is a constant. In view of Theorem 4 the first term in (4.14) is a convex function of the variables $z_1, \ldots, z_N$. The second term is obviously concave in the same variables. Thus the function (4.14) is a convex function. The SUMT procedure works so that in each iteration we minimize the function (4.14) under a suitably selected but fixed $r > 0$. When minimizing a function it is very favourable if the function is convex primarily because in this case any local minimum is also a global minimum.

Let us choose a sequence $r_1, r_2, \ldots$ of r values with the property that $r_k > 0$, $r_k > r_{k+1}$, $k = 1, 2, \ldots$ and $\lim_{k \to \infty} r_k = 0$.

Using any method (e.g. the gradient method) for the minimization of the function (4.14), denote by $z^{(k)}$ the minimizing vector in case of $r = r_k$. Then the sequence of values of the objective function of problem (2.1) belonging to the vectors $z^{(1)}, z^{(2)}, \ldots$ converges to the negative logarithm of the (constrained) optimum value of problem (2.1). This immediately follows from the theory presented in [4], our functions standing in problem (2.1) satisfy the conditions required for the convergence.

The unconstrained minimizations can be carried out by several methods (see e.g. [6], [7]). The values and if necessary the gradient values of the multivariate normal probability distribution function can be computed by the Monte-Carlo method with a satisfactory precision even in the case if $N = 20$ (see [3]).

In Section 1 we mentioned that it is possible to formulate and solve further variants of the model presented in this paper. Among these we mention the following

$$0 \leqq z_k \leqq K, \quad k = 1, \ldots, N,$$
$$P(a_k \leqq \zeta_k - Z_k \leqq b_k, \ k = 1, \ldots, N) \geqq p, \tag{4.15}$$
$$\min K.$$

Here p is a prescribed probability near unity in practice (e.g. $p = 0.9$). Problem (4.15) means relative to the Lake Balaton that we want to find the minimum capacity of the Sió channel subject to the condition that the water level remains between the prescribed levels with at least probability p. General, (4.15) type stochastic programming models are investigated in [11].

FORTRAN programs prepared for the computer CDC 3300 are available for the described algorithms. The numerical experiences show that the computing time is reasonable even in the case $N = 12$ which size is completely satisfactory in case of the water level problem since we program at most one year ahead and to choose 12 periods in one year seems to be a satisfactory fine subdivision of the time in a year.

5. Conclusions

We gave a dynamic model system for a special type of control of a discrete time parameter Gaussian stochastic process. The process otherwise is arbitrary, no independence or homogeneity assumption is needed, the parameters can be arbitrary. The purpose of the control is to keep a certain level within prescribed limits. The model system is a stochastic programming model system, the method of solution is the gradient method and the SUMT. The theory is applied for the regulation of the water level of the Lake Balaton.

References

1. The Balaton. Institute for Water Economies, Technical University of Budapest, 1970.
2. *Anderson, T. W.*, An Introduction to Multivariate Statistical Analysis. Wiley, New York, 1958.
3. *Deák, I.*, Computer Evaluation of a Stochastic Programming Model (in Hungarian), Communications of the Computing Centre of the Hungarian Academy of Sciences **9** (1972) 33—49.
4. *Fiacco, A. V., McCormick, G. P.*, Unconstrained Minimization Techniques. Wiley, New York, London, 1968.
5. *Hadley*, G., Nonlinear and Dynamic Programming. Addison-Wesley, Reading, Mass., 1964.
6. *Kowalik, J., Osborne, M. R.*, Methods for Unconstrained Optimization Problems. American Elsevier Publishing Company, Inc. New York, 1968.

7. *Powell*, M. J. D., Recent Advances in Unconstrained Optimization, Mathematical Programming **1** (1971) 26—57.
8. *Prékopa*, *A.*, On Probabilistic Constrained Programming, Proceedings of the Princeton Symposium on Mathematical Programming. Princeton University Press, Princeton, New Jersey, 1970, 113—138.
9. *Prékopa*, *A.*, Logarithmic Concave Measures with Application to Stochastic Programming, Acta Scientiarum Mathematicarum **32** (1971) 301—316.
10. *Prékopa*, A., On Logarithmic Concave Measures and Functions, Acta Scientiarum Mathematicarum **34** (1973) 335—343.
11. *Prékopa*, *A.*, Constributions to the Theory of Stochastic Programming, Mathematical Programming **4** (1973) 202—221.
12. *Prékopa*, *A.*, Eine Erweiterung der sogenannten Methode der zulässigen Richtungen der nichtlinearen Optimierung auf den Fall quasikonkaver Restriktionsfunktionen, Mathematische Operationsforschung und Statistik **5** (1974).

Применение стохастического программирования для оптимального управления уровнем запаса

А. ПРЕКОПА

(Будапешт)

Для управления процессами с дискретными параметрами, распределенными по Гауссу, применяется метод стохастического программирования. Проблемы построения системы решаются методами нелинейного программирования. Предлагаемый метод может быть использован для динамического регулирования уровня воды озера Балатон.

A. Prékopa
Department of Mathematics
Technical University
H-111, Budapest, XI. Goldmann Gy. tér 3.
Hungary

and

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

Problems of Control and Information Theory, Vol. 4 (3), pp. 205–218 (1975)

ON MINIMAX CONTROL AND ESTIMATION STRATEGIES UNDER INCOMPLETE INFORMATION

A. B. KURZHANSKII

(Sverdlovsk)

(Received October 11, 1974)

Minimax problems of synthesizing control systems with only part of the state variables being observed are posed under the assumption that the information on the uncertain disturbances in the systems is restricted to the description of the corresponding admissible domains. The solution is based on the theory for differential game problems of dynamics and also on investigations on minimax observation and filtering problems. Solutions for a number of open-loop control problems under imperfect data are given together with conditions when these are effective for constructing the closed-loop strategies.

Introduction

The theory of feedback control under uncertainty in the inputs considers problems when information on the output coordinates is either complete or not. If in the first case it is generally assumed that the output phase vector which describes the process dynamics is completely accessible due to available measurements, the second case is frequently related to problems where a certain statistical description for the inputs is at hand. The present paper deals with the situations when complete measurements for the output coordinates are unavailable and the accessible information is restricted to the values of a part of the outputs. It is assumed that both the control and the measurement systems are subjected to uncertain disturbances for which there is no statistical description whatever. The information on these values is restricted to the description of the corresponding admissible domains of variation. Consequently, the accessible current state (the position) for each of the systems under consideration is no more determined by a phase vector but is formed of a certain domain in the phase space depending on the observed signal and containing the unknown actual output (the phase vector of the system). The aim of the synthesizing process is to specify a control law formed on the basis of the accessible information (the position of the system) in order to ensure a least possible value for a given function of the process terminal coordinates (despite the counteraction of uncertain factors). The situation described is formalized within the framework of feedback control problems with minimax payoff. The given paper is thus devoted to specific informational game-theoretic

problems of dynamics which combine an optimization of the measurement-estimation process for the system coordinates with that of the process for selecting the control law itself. The solution is based on the constructions of the theory for game problems of dynamics [1—3] and also on the results of papers [4—6]. A number of open-loop control problems under imperfect data are considered, their solutions being further applied to the closed-loop problems under the approach of [1—3]. The discussion is restricted to linear systems which allow a description of the solutions in terms of constructive relations within the sphere of applicability of numerical methods of convex analysis. Other classes of control problems under incomplete information were discussed in investigations [7—11].

1. Accessible information
Problem formulation and mathematical formalization

An n-dimensional system is given

$$\dot{x} = A(t)x + B(t)u - C(t)v, \qquad t_0 \leq t \leq \vartheta \tag{1.1}$$

where u and v are the controls of dimension p and q respectively. At each moment of time t starting from t_0 accessible due to measurement is the observed m-vector signal

$$y(t) = G(t)x(t) + F(t)\xi(t) \tag{1.2}$$

where the r-vector value ξ is the disturbance in the observation channel.

It is necessary to emphasize that the realizations $v(t)$, $\xi(t)$ and the initial value $x^0 = x(t_0)$ are unknown in advance. The information on these values is restricted to the formulation of conditions

$$v(t) \in Q(t), \quad \xi(t) \in R(t), \quad x^0 \in X^0, \tag{1.3}$$

where the sets $Q(t)$, $R(t)$ (for all $t \geq t_0$) and the set X^0 are given in advance and assumed to be compact in the finite-dimensional spaces $R^{(q)}$, $R^{(r)}$, $R^{(n)}$. In contrary the control u is sought for, hence the realization $u(t)$ with values in convex compact set $P(t)$ is presumed to be known. The given matrix functions $A(t)$, $B(t)$, $C(t)$, $F(t)$, $G(t)$ are taken to be continuous.

Notations are adopted as follows: G_ε is the ε-neighbourhood for set G, $\varrho(l \mid G) = \sup l'g$, $g \in G$ is the support function for G, the prime standing for the transpose and the symbol $\| \cdot \|$ for the Euclidean norm; $g_t(\cdot \mid \tau)$ is a functional element generated by function $g(s)$, $\tau \leq s \leq t$, $g(t, \cdot \mid \tau) = g(s)$, $\tau \leq s < t$; $\zeta_t(\cdot \mid \tau) = \{v_t(\cdot \mid \tau), \ \xi_t(\cdot \mid \tau), x\}$; $\mathfrak{D}_t(\cdot \mid \tau, X) = \{\zeta_t(\cdot \mid \tau) : v(s) \in Q(s), \xi(s) \in R(s), \tau \leq s \leq t; x \in X\}$.

Square brackets used instead of round ones in symbols of type $\zeta_t(\cdot \mid \tau)$, $u(t)$, $\mathcal{D}_t(\cdot \mid \tau, X)$, $g_t(\cdot \mid \tau)$, $g(t, \cdot \mid \tau)$ and in the coming symbols always denote a fixed realization of the respective value.

The aim of the control u synthesized depending on the observed realization $y_t(\cdot)$ is to ensure a least possible value for a given functional of the terminal state variables despite the counteraction of the uncertain factors $v(t)$, $\xi(t)$. Consequently a minimax differential game-type problem is formed which combines the optimization of both the feedback measurement and estimation process and the procedure of specifying the control forces.

A precise formulation of the problem is as follows. Assume that at time $t_1 \geq t_0$ the condition $x(t_1) \in X$ was given and functions $y_t[\cdot \mid t_1]$, $u[t, \cdot \mid t_1]$ have been realized throughout the interval $[t_1, t]$.

Definition 1. The set $X(t, \cdot \mid t_1, X) = X(t, y_t[\cdot \mid t_1], u[t, \cdot \mid t_1] \mid t_1, X)$ is said to be compatible with signal $y_t[\cdot \mid t_1]$ with given $u[t, \cdot \mid t_1]$ if it consists of those and only those vectors $\{x\} \in R^{(n)}$ for each of which there exists an element $\zeta_t(\cdot \mid t_1) \in \mathcal{D}_t(\cdot \mid t_1, X)$ such that the corresponding solution $x(t)$, $y(t)$ of (1.1), (1.2) (generated in the interval $[t_1, t]$ by values $x(t_1) = x^*$, $v = v_t(\cdot \mid t_1)$ $\xi = \xi_t(\cdot \mid t_1)$ forming the given $\zeta_t(\cdot \mid t_1)$ and also by given realization $u = u[t, \cdot \mid t_1]$) would satisfy the conditions $x(t) = x$, $y(s) \equiv y[s]$, $\tau \leq s \leq t$.

Further take the following notations

$$X(t, \cdot \mid t_0, X^0) = X(t, \cdot); \; g_t(\cdot \mid t_0) = g_t(\cdot); \; g(t, \cdot \mid t_0) = g(t, \cdot).$$

Let us discuss Definition 1, taking for example $t_1 = t_0$. We note at once that it is not an arbitrary $\zeta_t(\cdot)$ that did generate the realization $y_t[\cdot]$. The knowledge of this realization allows a posterior procedure for determining all of the values $\zeta_t(\cdot)$ that could generate the given $y_t[\cdot]$. The vectors $\{x(t)\}$ corresponding to all of the mentioned values $\zeta_t(\cdot)$ and to the given $u[t, \cdot]$ with respect to equation (1.1) constitute precisely the set $X(t, \cdot)$. The value $\{t, X(t, \cdot)\}$ derived on base of the accessible information forms the current position (the generalized state) of the system which serves as necessary data in forming the control force. We shall thus seek for the control in the form of a functional $u = u(t, X(t, \cdot))$.

Definition 2. A multivalued functional $U = U(t, X)$ is said to be an admissible strategy if for each t it is defined on the set M of all convex compacts of $R^{(n)}$, with its values being convex, compact sets included in $P(t)$ and if it is upper semicontinuous in t, X with $\Delta t \geq 0$ (i.e. for any $\varepsilon > 0$ there exists a $\delta > 0$ such that with $0 \leq \Delta t \leq \delta$, $X^{(1)} \subseteq X_\delta$, $X \subseteq X_\delta^{(1)}$ we have

$$U(t + \Delta t, X^{(1)}) \subseteq U_\varepsilon(t, X)).$$

Substituting $U(t, X(t, \cdot))$ in (1.1) instead of u, we arrive at the equation

$$\dot{x} \in A(t)\, x + B(t)\, U(t, X(t, \cdot)) - C(t)\, v(t) \tag{1.4}$$

with a multivalued right-hand side. The problem will further be described in terms of solutions of (1.4).

Definition 3. A function $x[t]$ is said to be a solution of Eq. (1.4) relative to system (1.2) ($\zeta_\vartheta[\,\cdot\,]$ given) if

(A) there exists a function $u[t] \in P(t)$ such that $x[t]$ is a solution for (1.1) with $u = u[t]$ and with given $\zeta_\vartheta[\,\cdot\,]$.

(B) for almost all $t \in [t_0, \vartheta]$ the inclusion $u[t] \in U(t, X[t, \cdot\,])$ is valid provided $X[t, \cdot\,]$ is the set compatible in the sense of Definition 1 with signal $y_t[\,\cdot\,]$ due to Eq. (1.2) formed of $x[t]$ with $u = u[t]$ and with given $\zeta_\vartheta[\,\cdot\,]$.

Lemma 1.1. Assume that the functional $U = U(t, X)$ in (1.4) is an admissible strategy. Then there exists at least one solution of (1.4) relative to (1.2).

The proof of the lemma may be attained by the scheme of paper [12]. The set of solutions of (1.4) with $U = U(t, X)$, $\zeta_\vartheta[\,\cdot\,]$ given shall further be denoted as $\{x(\,\cdot\,) \mid U(t, \cdot\,), \zeta_\vartheta[\,\cdot\,]\}$. Therefore each strategy $U(t, X)$ with given $\zeta_\vartheta[\,\cdot\,]$ generates a set of solutions $\{x(t)\} = \{x(\,\cdot\,) \mid U(t, \cdot\,), \zeta_\vartheta[\,\cdot\,]\}$ each of which corresponds to a certain $u[t]$ satisfying conditions (A) and (B). In turn each of the solutions $x(t)$ obtained for $u = u[t]$, $\zeta = \zeta_\vartheta[\,\cdot\,]$ generates an observed signal $y_t[\,\cdot\,]$ and further on a set $X[\vartheta, \cdot\,]$ derived of $y_t[\,\cdot\,]$ and compatible with this function. As shown in the sequel set $X[\vartheta, \cdot\,]$ is always an element of M.

Define a functional $\Phi(X)$ on set M, taking

$$\Phi(X) = \max_x \varphi(x),\ x \in X, \tag{1.5}$$

where $\varphi(x)$ is a proper convex function in $R^{(n)}$ [13]. The problem will consist in specifying a strategy $U(t, X)$ in order to ensure the relation

$$\begin{aligned} \mathcal{F}^0(t_0, X^0) &= \mathcal{F}(t_0, X^0, U^0(t, \cdot\,), \zeta^0_\vartheta[\,\cdot\,]) = \\ &= \min_{U(t,\cdot)} \max_{\xi_\vartheta(\cdot)} \mathcal{F}(t_0, X^0, U(t, \cdot\,), \zeta_\vartheta(\,\cdot\,)) \qquad (1.6) \\ \mathcal{F}(t_0, X^0, U(t, \cdot\,), \zeta_\vartheta(\,\cdot\,)) &= \max_{\{x(\cdot) \mid U(t,\cdot),\, \xi_\vartheta(\cdot)\}} \Phi(X(\vartheta, \cdot\,)), \end{aligned}$$

over all admissible $U(t, \cdot\,)$, $\zeta_\vartheta(\,\cdot\,) \in \mathcal{D}_\vartheta(\,\cdot \mid t_0, X^0)$.

Along with strategies $U(t, \cdot\,)$ we shall also consider a class of piecewise open-loop single-valued Δ-strategies of the form $u = u(t, \tau_i, X(\tau_i, \cdot\,))$ which correspond to a given subdivision $\Delta = \{\tau_i\}$ of the time interval $[t_0, \vartheta]$. Selecting

Δ it may be achieved that the given class generates an approximation of the desired solution.

Under certain conditions the desired closed-loop strategies may be constructed by solutions of special open-loop control problems. A number of these conditions is further formulated starting from strongest assumptions and forming a sort of a scale.

2. Estimation of compatible domains. Minimax filtering. The extrapolation problem

A notable point in the solution is the description of sets $X(t, \cdot)$ forming the current position for the system.

Lemma 2.1. The sets $X(t, \cdot)$ are convex and compact in $R^{(n)}$. The formula is true

$$\varrho(l \mid X(t, \cdot)) = \inf_{l_1, l_2} \inf_{\lambda(t,\tau)} \{\varrho(l_1 S(t_0, t) \mid X^0) + \\ + \int_{t_0}^{t} [\varrho(-s^{(1)}(\tau \mid \lambda(t, \cdot), t_0) \mid C(\tau) Q(\tau) - B(\tau) u[\tau]) + \\ + \varrho(\lambda(t, \tau) F(\tau) \mid R(\tau) + y[\tau])] d\tau\}, \quad l_1 + l_2 = l, \tag{2.1}$$

under restriction $s^{(0)}(t \mid \lambda(t, \cdot), t_0) = l_2$. Here $s^{(1)}(\tau \mid \lambda(t, \cdot), t_0)$ is the solution of the following equation

$$\dot{s} = -sA(\tau) + \lambda(t, \tau) G(\tau), \tag{2.2}$$

with $s(t_0) = l_1 S(0, \vartheta)$; $s^{(0)}(\tau \mid \lambda(t, \cdot), t_0)$ is the solution of (2.2) with $s(t_0) = 0$; $\lambda(t, \tau)$ is a function square integrable in τ, $S(\tau, \vartheta)$ is the normed transition matrix (fundamental matrix) for the homogeneous system (2.2).

The scheme for deriving (2.1) is given in [5, 6].

Take

$$G(t_2, t_1, X, u_{t_2}[\cdot \mid t_1]) = \bigcup_{x, v} \{S(t_1, t_2) x + \int_{t_1}^{t_2} S(\tau, t_2) (B(\tau) u[\tau] - C(\tau) v[\tau]) d\tau\}$$

over all $x \in X$, $v(\tau) \in Q(\tau)$. The given set is the domain of attainability in the control v for system (1.1) from set X with u fixed. Realization $u[t_2 \cdot \mid t_1]$ defined in $[t_1, t_2)$, we have

$$G(t_2, t_1, X, u[t_2, \cdot \mid t_1]) = G(t_2, t_1, X, u_{t_2}[\cdot \mid t_1]).$$

The dynamics for sets $X(t, \cdot)$ are characterized by following properties [5, 6].

Lemma 2.2. Assume that control $u[t, \cdot\,]$ and signal $y_t[\,\cdot\,]$ generated by certain $\zeta_t[\,\cdot\,]$ are the realizations for the interval $[t_0, t]$. Then the inclusion is true $(t_0 \leq t_1 \leq t)$

$$X(t, \cdot\,) \subseteq G(t, t_1, X(t_1, \cdot\,), u_t[\,\cdot \mid t_1]) \cap X(t, \cdot \mid t_1, R^{(n)}) = X^*(t, \cdot \mid t_1). \quad (2.3)$$

For every $u[t, \cdot\,]$ l, there exists a signal $y_t[\,\cdot\,]$ which ensures a strict equality $\varrho(l \mid X(t, \cdot\,)) = \varrho(l \mid X^*(t, \cdot \mid t_1))$.

Realization $u[\vartheta, \cdot\,]$ fixed, the inclusion is true $(t \leq \vartheta)$

$$G(\vartheta, t, X(t, \cdot\,), u[\vartheta, \cdot \mid t]) \subseteq G(\vartheta, t_1, X(t_1, \cdot\,), u[\vartheta, \cdot \mid t_1]).$$

A multivalued function $G[t]$ is said to be discontinuous at $t = t_1$ if function $\varrho(l \mid G[t])$ is discontinuous at point $t = t_1$ for at least one l.

Lemma 2.3. Each of the realizations $X(t, \cdot\,) = X[t]$ (admissible relative to restrictions on systems (1.1), (1.2)) may have but a countable set of discontinuities.

The determination of the best in sense of minimax estimators for the output vector based on observation measurements of function $y_t[\,\cdot\,]$ is the object of minimax filtering theory (in contrary to stochastic filtering [7, 8, 14] where v, ξ are taken to be Gaussian white noise with given statistical parameters). The above-mentioned estimator $x^0(t, \cdot\,)$ may be determined e.g. as the Chebyshev center for the set $X(t, \cdot\,)$, i.e. the solution for the problem

$$\delta^0(t, \cdot\,) = \min_x \max_z ||x - z|| = \max_z ||x^0(t) - z||$$

$$x \in X(t, \cdot\,),\ z \in X(t, \cdot\,)$$

where $\delta^0(t, \cdot\,)$ is the guaranteed bias error of observation.

The construction of closed-loop controllers requires the solutions of extrapolation problems.

Assume that at time t the set $X(t, \cdot\,)$ was realized. Then for any possible future realization of $u_\vartheta[\,\cdot \mid t]$, $\zeta_\vartheta[\,\cdot \mid t] \in \mathfrak{D}_\vartheta(\,\cdot \mid t, X(t, \cdot\,))$ one may specify the corresponding signal $y_\vartheta[\,\cdot \mid t]$ and further the set $X(\vartheta, \cdot \mid t, X(t, \cdot\,)) = X(\vartheta, y_\vartheta[\,\cdot \mid t], u_\vartheta[\,\cdot \mid t] \mid t, X(t, \cdot\,))$, derived in result of the observation. Denote $\mathfrak{L}(t, l, Q) = \varrho(l'S(t, \vartheta) \mid Q)$, and $x(t; u_\vartheta(\,\cdot \mid t_1), v_\vartheta(\,\cdot \mid t_1), x, t_1)$ as the solution of (1.1) generated by $u = u_\vartheta(\,\cdot \mid t_1)$, $v = v_\vartheta(\,\cdot \mid t_1)$, $x(t_1) = x$.

Lemma 2.4. The formula is true

$$\varrho(l \mid X(\vartheta, \cdot \mid t, X(t, \cdot\,))) = l'x(\vartheta; u_\vartheta[\,\cdot \mid t_1], v_\vartheta[\,\cdot \mid t_1], x^*, t_1) +$$

$$+ K(l, X(t, \cdot\,), \zeta_\vartheta[\,\cdot \mid t]), \quad (2.4)$$

$$K(l, X(t, \cdot), \zeta_\vartheta[\cdot \mid t]) = \inf_{l_1, l_2} \inf_{\lambda(l,\tau)} \{\mathcal{K}(l_1, t, X(t, \cdot) - x^*]) +$$

$$+ \int_t^\vartheta [\varrho(-s^{(1)}(\tau \mid \lambda(\vartheta, \cdot), t) C(\tau) \mid Q(\tau) - v[\tau]) +$$

$$+ \varrho(-\lambda(\vartheta, \tau) H(\tau) \mid R(\tau) - \xi[\tau])] d\tau\}$$

with $s^{(1)}(t \mid \lambda(\vartheta, \cdot), t) = l_1 S(t, \vartheta)$, $s^{(0)}(\vartheta \mid \lambda(\vartheta, \cdot), t) = l$, $l_1 + l_2 = l$, where $\zeta_\vartheta[\cdot \mid t] = \{v_\vartheta[\cdot \mid t], \xi_\vartheta[\cdot \mid t], x^*\}$ is the value which generated the signal $y_\vartheta[\cdot \mid t]$ along with $u_\vartheta[\cdot \mid t]$.

Varying all possible values of $\zeta_\vartheta[\cdot \mid t] \in \mathcal{D}_\vartheta(\cdot \mid t, X(t, \cdot))$ in advance we achieve for each $u_\vartheta[\cdot \mid t]$ the set of all possible future realizations of sets $X(\vartheta, \cdot \mid t, X(t, \cdot))$.

3. Open-loop control under imperfect data

In this section the open-loop control problems proven to be effective for the construction of the closed-loop strategies are described.

Problem 1. Specify a control $u^0(t) = u^0(t; t_0, X^0)$, $t_0 \leq t \leq \vartheta$, so as to ensure the relation

$$\beta^0(t_0, X^0) = \max_{v, x^0} \varphi(x(\vartheta; u_\vartheta^0(\cdot), v_\vartheta(\cdot), x^0, t_0) =$$

$$= \max_{v, x^0} \min_u \varphi(x(\vartheta; u_\vartheta(\cdot), v_\vartheta(\cdot), x^0, t_0) \tag{3.1}$$

$$u(t) \in P(t), \quad v(t) \in Q(t), \quad x^0 \in X^0.$$

Assume $X(\vartheta, \cdot \mid t_0, X^0)$ to be the set described in Lemma 2.3 and related to the pair $u_\vartheta[\cdot]$, $\zeta_\vartheta[\cdot]$ that generated $y_\vartheta[\cdot]$. Take $\Phi(X)$ to be a functional of type (1.5) defined in M.

Problem 2. Specify a control $u^{00}(t) = u^{00}(t; t_0, X^0)$, $t_0 \leq t \leq \vartheta$ so as to ensure the relation

$$\beta^{00}(t_0, X^0) = \max_{\zeta_\vartheta(\cdot)} \Phi(X(\vartheta, u_\vartheta^{00}(\cdot), y_\vartheta[\cdot] \mid t_0, X^0)) =$$

$$= \max_{\zeta_\vartheta(\cdot)} \min_{u[t]} \Phi(X(\vartheta, u_\vartheta(\cdot), y_\vartheta[\cdot] \mid t_0, X^0) \tag{3.2}$$

$$\zeta_\vartheta(\cdot) \in \mathcal{D}_t(\cdot \mid t_0, X^0), \quad u(t) \in P(t).$$

Here all the open-loop control programs $u_\vartheta[\cdot]$, $\zeta_\vartheta[\cdot]$ are taken in advance to define the variety of sets $X(\vartheta, \cdot\,; t_0, X^0)$ when the minimax solution is further specified. Assume

$$X(t; u_\vartheta(\cdot), v_\vartheta(\cdot)) = \bigcup_{x^0} x(t; u_\vartheta(\cdot), v_\vartheta(\cdot), x^0, t_0), \qquad x^0 \in X^0.$$

Problem 3. Specify a control $u^*(t) = u^*(t; t_0, X^0)$, $t_0 \leq t \leq \vartheta$ so as to satisfy the relation

$$\beta^*(t_0, X^0) = \max_v \Phi\big(X(\vartheta; u^*_\vartheta(\cdot), v_\vartheta(\cdot)) = \\ = \max_v \min_u \Phi\big(X(\vartheta; u_\vartheta(\cdot), v_\vartheta(\cdot))\big) \tag{3.3}$$

$$v(t) \in Q(t), \quad u(t) \in P(t)\,.$$

Here the maxmin solution is to be obtained for the whole ensemble $X(t; u_\vartheta(\cdot), v_\vartheta(\cdot))$ of trajectories of (1.1) generated by X^0. Particularly if $\varphi(x) = r(x, M_0)$ is the Euclidean distance from convex set M_0, then $\beta^0(t_0, X^0)$ is the maxmin value for the neighbourhood of M_0 which includes all of the set $X(\vartheta; u_\vartheta(\cdot), v_\vartheta(\cdot))$.

Take $\psi(\tau, l) = \mathcal{L}(l, \tau, C(\tau)Q(\tau) - B(\tau)P(\tau))$

$$\Psi(t, l) = \int_t^\vartheta \psi(\tau, l)\, d\tau\,,$$

$\varphi^*(l) = \sup_x \{l'x - \varphi(x)\}$ ($\varphi^*(l)$ is the convex conjugate relative to $\varphi(x)$ [13]). Consider the problem

$$\gamma_0(t_0, X^0) = \sup_l \gamma_1(t_0, X^0, l), \tag{3.4}$$

$$\gamma_1(t_0, X^0, l) = \Psi(t_0, l) + \mathcal{L}(-l, t_0, X^0) - \varphi^*(-l)\,.$$

Assume conditions $\varphi^*(0) \neq \gamma^0(t_0, X^0)$ (of nondegeneracy) and

$$-\gamma_1(t_0, X^0, l) \to \infty, \qquad (\|l\| \to \infty) \tag{3.5}$$

to be fulfilled. The maximum for (3.4) is then attained at a compact set $\Gamma^0(t_0, X^0)$ of elements $l^0(t_0, X^0) \neq 0$. The given conditions are fulfilled if e.g. we have $\beta^0(t_0, X^0) \neq 0$ with $\varphi(x) = r(x, M_0)$. (Necessary and sufficient conditions for the attainability of a maximum in (3.4) may be formulated in terms of [15].)

Theorem 1. (i) The equality $\gamma^0(t_0, X^0) = \beta(t_0, X^0)$ is true.

(ii) The control $u^0(t) = u^0(t; t_0, X^0)$ of Problem 1 satisfies the necessary maximum condition

$$\mathcal{L}(l^0(t_0, X^0), t, B(t)u^0(t)) = \mathcal{L}(l^0(t_0, X^0), t, B(t)P(t))\,. \tag{3.6}$$

(iii) With $l^0(t_0, X^0) \neq 0$ and system (1.1), ($v(t) \equiv 0$) strongly controllable, [16, 17], condition (3.6) is also sufficient for $u^0(t)$ to solve Problem 1.

Denote with (conc f) (l) the concave hull (the upper envelope) for function $f(l)$ on $R^{(n)}$. Consider the problem

$$\gamma^{00}(t_0, X^0) = \sup_l \gamma_2(t_0, X^0, l) \tag{3.7}$$

$$\gamma_2(t_0, X^0, l) = \max_{\zeta_\vartheta[\,\cdot\,]} K_1(t_0, X^0, l, \zeta_\vartheta[\,\cdot\,]);\ \zeta_\vartheta[\,\cdot\,] \in \mathfrak{D}_\vartheta[\,\cdot \mid t_0, X^0],$$

$$K_1(t_0, X^0, l, \zeta_\vartheta[\,\cdot\,]) = \int_{t_0}^{\vartheta} \mathfrak{L}(l, \tau, -B(\tau)P(\tau) + C(\tau)v[\tau])\,d\tau -$$

$$- l'S(t_0, \vartheta)\,x^* + (\text{conc } h)(-l, X^0, \zeta_\vartheta[\,\cdot \mid t_0]);\ h(l, X^0, \zeta_\vartheta[\,\cdot \mid t_0]) =$$

$$= K(l, X^0, \zeta_\vartheta[\,\cdot \mid t_0]) - \varphi^*(l).$$

Let $\gamma_2(t_0, X^0, l)$ satisfy the condition (3.5). Then the maximum for (3.7) is attained at a compact set $\Gamma^{00}(t_0, X^0)$ of elements $l^{00}(t_0, X^0)$.

Theorem 2. (i) The condition $\gamma^{00}(t_0, X^0) = \beta^{00}(t_0, X^0) \geq \beta^0(t_0, X^0)$ is true.

(ii) The control $u^{00}(t) = u^{00}(t;\, t_0, X^0)$ of **Problem** 2 satisfies the necessary maximum condition

$$\mathfrak{L}(l^{00}(t_0, X^0), t, B(t)\,u^{00}(t)) = \mathfrak{L}(l^{00}(t_0, X^0), t, B(t)P(t)). \tag{3.8}$$

(iii) With $h(l, X^0, \zeta_\vartheta[\,\cdot \mid t_0])$ concave in l we have

$$\gamma_2(t_0, X^0, l) \equiv \gamma_1(t_0, X^0, l);\ \beta^{00}(t_0, X^0) = \beta^0(t_0, X^0)$$

and relations (3.7) and (3.8) coincide.

It is necessary to note that for each $\zeta_\vartheta[\,\cdot \mid t_0]$ the value $K(l, X^0, \zeta_\vartheta[\,\cdot\,|t_0]) = \varrho(l \mid G(X^0, \zeta_\vartheta[\,\cdot \mid t_0]))$ is the support function for a convex compact set $G(X^0, \zeta_\vartheta[\,\cdot \mid t_0])$ about the origin. The condition of Theorem 2 (iii) denotes in particular that with $\varphi(x) = r(l, M_0)$ the difference $\varrho(l \mid M_0) - \varrho(l \mid X(\vartheta, \cdot \mid t_0, X^0)$ is convex in l for all $u_\vartheta[\,\cdot\,], \zeta_\vartheta[\,\cdot \mid t] \in \mathfrak{D}_\vartheta(\,\cdot \mid t, X)$.

Consider the problem

$$\gamma^*(t_0, X^0) = \sup_l \gamma_3(t_0, X^0, l) \tag{3.9}$$

$$\gamma_3(t_0, X^0, l) = -\Psi(t_0, l) + (\text{conc } g)(l, t_0, X^0),$$

$$g(l, t_0, X^0) = \mathfrak{L}(-l, t_0, X^0) - \varphi^*(-l).$$

Assume $\gamma_3(t_0, X^0, l)$ satisfies condition of type (3.5). Then the maximum for (3.9) is attained at a compact set $\Gamma^*(t_0, X^0)$ of elements $l^*(t_0, X^0)$.

Theorem 3. (i) The condition $\gamma^*(t_0, X^0) = \beta^*(t_0, X^0) \geq \beta^{00}(t_0, X^0)$ is true.

(ii) The control $u^*(t) = u^*(t;\, t_0, X^0)$ of Problem 3 satisfies the necessary maximum condition

$$\mathfrak{L}(l^*(t_0, X^0), t, B(t)\, u^*(t)) = \mathfrak{L}(l^*(t_0, X^0), t, B(t)\, P(t)). \tag{3.10}$$

(iii) With $g(l, t_0, X^0)$ concave in l the condition of Theorem 2 (iii) is fulfilled, $\gamma_3(t_0, X^0, l) \equiv \gamma_2(t_0, X^0, l) = \gamma_1(t_0, X^0, l)$; $\beta^*(t_0, X^0) = \beta^{00}(t_0, X^0) = \beta^0(t_0, X^0)$, and relations (3.6), (3.8) and (3.10) coincide.

The concavity of $g(l, t_0, X^0)$ with $\varphi(x) = r(x, M_0)$ yield that the differences $\varrho(l \mid M_0) - \mathfrak{L}(l, t_0, X^0)$ are convex.

4. Estimates for feedback dynamics

Assume at time the system (1.1), (1.2) was at state $\{t, X(t, \cdot)\}$ and was transferred in time Δt from given state to position $\{t + \Delta t, X(t + \Delta t, \cdot)\}$. Solving Problem 1 one may determine the value $\gamma^0(t, X(t, \cdot))$. Our aim is to evaluate the increment

$$\Delta\gamma^0(t, \cdot) = \gamma^0(t + \Delta t, X(t + \Delta t, \cdot)) - \gamma^0(t, X(t, \cdot))$$

taking that in the interval $[t, t + \Delta t]$ there were realized the control $u[\tau]$ and the value $\zeta_{t+\Delta t}[\cdot \mid t] = \{v_{t+\Delta t}[\cdot \mid t],\ \xi_{t+\Delta t}[\cdot \mid t],\ x^*\}$ all of which had generated the measured signal $y_{t+\Delta t}[\cdot \mid t]$ of which in turn the set $X(t + \Delta t, \cdot)$ had been derived. Following [6] we arrive at the conclusion:

Lemma 4.1. Increment $\Delta\gamma(t, \cdot)$ satisfies the equality

$$\Delta\gamma^0(t, \cdot) = \max_l \Big\{ \int_t^{t+\Delta t} \big[\mathfrak{L}\big(l, \tau, B(\tau)\,(P(\tau) - u[\tau]) - C(\tau)\,(Q(\tau) - v[\tau])\big)\big]\, d\tau - \tag{4.1}$$

$$- \mathfrak{L}(-l, t, -x^* + X(t, \cdot)) + \varrho\big(-l \mid G(X(t, \cdot), \zeta_{t+\Delta t}[\cdot \mid t])\big) + 0(\Delta t)$$

where $l \in \Gamma^0(t, X(t, \cdot))$; $0(\Delta t)\,(\Delta t)^{-1} \to 0$ with $\Delta t \to 0$ and

$$\int_t^{t+\Delta t} \mathfrak{L}(l, \tau, C(\tau)\, v[\tau])\, d\tau - l' S(t, \vartheta)\, x^* + \tag{4.2}$$

$$+ \varrho\big(-l \mid G(X(t, \cdot), \zeta_{t+\Delta t}[\cdot \mid t]\big) \leq \mathfrak{L}(-l, t + \Delta t, G(t + \Delta t, t, X(t, \cdot), 0)).$$

Condition (4.2) follows from Lemmas 2.2 and 2.4. Inequality (4.1) generalizes some of the formulas of [18, 19] to specific functional relations of this paper.

Assumption A. Function $\varphi(x) = r(x, M_0)$. Function

$$\Psi(t, l) + \mathfrak{L}(-l, t, X(t, \cdot)) - \varphi^*(-l)$$

is concave in l.

With sets $P(t)$, $Q(t)$ being similar it suffices to verify the concavity property of Assumption A for $\{t_0, X^0\}$ only.

It follows from Lemma 2.3 that at points of discontinuity for realizations $X(t, \cdot) \equiv X[t]$ the functions $\gamma^0(t, X[t])$, $\gamma^{00}(t, X[t])$, $\gamma^*(t, X[t])$ are non-increasing.

Lemma 4.2. Assumption A being fulfilled, the vector $l^0(t, X(t, \cdot))$ is unique and with $u^0[t] \equiv u^0(t; t_0, X^0)$ satisfying condition (3.6) taken for given $l^0(t, X(t, \cdot))$ the inequality

$$\left.\frac{d\gamma^0(t, \cdot)}{dt}\right|_{u=u^0} = \lim_{\Delta t \to +0} \left.\frac{\Delta\gamma(t, \cdot)}{\Delta t}\right|_{u[t]=u^0[t]} \leq 0 \tag{4.3}$$

is true for any realization $\zeta_t[\,\cdot\,]$.

Similar estimates are further derived for $\gamma^{00}(t, X(t, \cdot))$, $\gamma^*(t, X(t, \cdot))$. Take $\mathfrak{D}(\vartheta, t, l) = \{\zeta^0_\vartheta(\,\cdot \mid t)\}$ to denote the set of elements maximizing $K_1(t, X(t, \cdot), l, \zeta_\vartheta[\,\cdot \mid t])$ and $\zeta^0_\vartheta[\,\cdot \mid t]$ to denote the element which coincides in $[t, t + \Delta t]$ with the fixed $\zeta_{t+\Delta t}[\,\cdot \mid t]$ that generated $X(t + \Delta t, \cdot)$ and coincides with the element $\zeta^0_\vartheta(\,\cdot \mid t)$ in $(t + \Delta t, \vartheta]$.

Lemma 4.3. The increment $\Delta\gamma^{00}(t, \cdot)$ satisfies the equality

$$\Delta\gamma^{00}(t, \cdot) = \max_l \Big\{ \int_t^{t+\Delta t} \mathfrak{L}\big(l, \tau, B(\tau)(P(\tau) - u[\tau])\big)\, d\tau + \tag{4.4}$$

$$+ \max_{\zeta_\vartheta(\cdot \mid t)} [K_1(t, X(t, \cdot), -l, \zeta^0_\vartheta[\,\cdot \mid t]) - K_1(t, X(t, \cdot), -l, \zeta^0_\vartheta(\,\cdot \mid t))] + 0_1(\Delta t)$$

$$\big(l \in \Gamma^{00}(t, X(t, \cdot)); \ \zeta^0_\vartheta(\,\cdot \mid t) \in \mathfrak{D}(\vartheta, t, l); \ 0_1(\Delta t)\,(\Delta t)^{-1} \to 0, \ \Delta t \to 0\big).$$

Assumption B. The function $\gamma_2(t, X(t, \cdot), l)$ is concave in l; $\varphi(x) = r(x, M_0)$.

Lemma 4.4. Assumption B being fulfilled the vector $l^{00}(t, X(t, \cdot))$ is unique and with $u^{00}[t] = u^{00}(t; t, X(t, \cdot))$ satisfying condition (3.8) taken for given $l^{00}(t, X(t, \cdot))$ the inequality

$$\left.\frac{d\gamma^{00}(t, \cdot)}{dt}\right|_{u=u^0} = \lim_{\Delta t \to +0} \left.\frac{\Delta\gamma^{00}(t, \cdot)}{\Delta t}\right|_{u[t]=u^{00}[t]} \leq 0 \tag{4.5}$$

is true for any realization $\zeta_t[\,\cdot\,]$.

Lemma 4.5. The increment $\Delta\gamma^*(t, \cdot)$ satisfies the equality

$$\Delta\gamma^*(t, \cdot) = \max_l \Big\{ \int_t^{t+\Delta t} [\mathcal{L}(l, \tau, B(\tau)(P(\tau) - u[\tau]) - C(\tau)(Q(\tau) - v[\tau]))]\, d\tau - \\ - \mathcal{L}(l, t, x^*) + [(\operatorname{conc} h)(-l, X(t, \cdot), \zeta_{t+\Delta t}[\cdot \mid t]) - \\ - (\operatorname{conc} g)(-l, t, X(t, \cdot))] + 0_2(\Delta t) \tag{4.6}$$

where $l \in \Gamma^*(t, \bullet)$; $0_2(\Delta t)(\Delta t)^{-1} \to 0$, $\Delta t \to 0$.

Assumption C. The function $\Psi(t, l) + (\operatorname{conc} g)(-l, t, X(t, \cdot))$ is concave in l; $\varphi(x) = r(x, M_0)$.

Lemma 4.6. Assumption C being fulfilled the vector $l^*(t, X(t, \cdot))$ is unique and with $u^*[t] = u^*(t; t, X(t, \cdot))$ satisfying condition (3.10) taken for given $l^*(t, X(t, \cdot))$ the inequality

$$\left.\frac{d\gamma^*(t, \cdot)}{dt}\right|_{u=u^*} = \lim_{\Delta t \to +0} \left.\frac{\Delta\gamma^*(t, \cdot)}{\Delta t}\right|_{u[t]=u^*[t]} \leq 0 \tag{4.7}$$

is true for any realization $\zeta_t[\cdot]$.

Inequalities (4.3), (4.5) and (4.7) are effective for points of continuity of realizations $X(t) \equiv X(t, \cdot)$. Note that these points constitute the whole interval $[t_0, \vartheta]$ with the exception of but a countable set.

The given estimates, achieved with the aid of results due to [3, 6] lead to a solution of the primary feedback control problem.

5. Solution of feedback control problem

Theorem 4. Assume that for any $\zeta_{t+\Delta t}[\cdot \mid t] \in \mathcal{D}_{t+\Delta t}(\cdot \mid t, X(t, \cdot))$ there exists a realization $u_{t+\Delta t}[\cdot \mid t]$ which ensures the inequality $\Delta\gamma^0(t, \cdot) \leq 0(\Delta t)$ related to (4.1). Then the given condition is necessary and sufficient for the existence of a subdivision Δ and a piecewise open-loop strategy $u^0(t, \cdot) = u^0(t; \tau_i, X(\tau, \cdot))$ which ensures for any $\varepsilon > 0$ the result

$$\mathcal{F}(t_0, X^0, u^0(t, \cdot), \zeta_\vartheta[\cdot]) \leq \gamma^0(t_0, X^0) + \varepsilon$$

whatever is the realization $\zeta_\vartheta[\cdot]$. The strategy $u^0(t; \tau_i, X(\tau_i, \cdot))$ is determined for each of the intervals $\Delta\tau_i = [\tau_i, \tau_{i+1})$ by condition $\Delta\gamma(\tau_i, \cdot) \leq 0(\Delta\tau_i)$ for position $\{\tau_i, X(\tau_i, \cdot)\}$ according to (4.1).

Analogously the existence of realizations $u_{t+\Delta t}[\,\cdot \mid t]$ which ensure the estimate $\Delta\gamma^{00}(t, \cdot) \leq 0_1(\Delta t)$, $(\Delta\gamma^*(t, \cdot) \leq 0_2(\Delta t))$ yields the existence of a subdivision Δ and a piecewise open-loop strategy $u^{00}(t, \tau_i, X(\tau_i, \cdot))$, $(u^*(t; \tau_i, X(\tau_i, \cdot)))$ which results in a value arbitrary close to $\gamma^{00}(t_0, X^0)$ $(\gamma^*(t_0, X^0))$. For the case of unique vectors $l(t_0, X(t, \cdot))$, $l^{00}(t, X(t, \cdot))$, $l^*(t, X(t, \cdot))$ denote the sets of all extremal elements due to (3.6), (3.8) and (3.10) as $U^0(t, X(t, \cdot)) = \{u^0(t; t, X(t, \cdot))\}$; $U^{00}(t, X(t, \cdot)) = \{u^{00}(t; t, X(t, \cdot))\}$; $U^*(t, X(t, \cdot)) = \{u^*(t; t, X(t, \cdot))\}$ respectively. Strategies $U^0(t, X(t, \cdot))$, $U^{00}(t, X(t, \cdot))$, $U^*(t, X(t, \cdot))$ (said to be "extremal") are admissible in the sense of Definition 1.

Theorem 5. (i) Under Assumption A we have

$$\mathfrak{F}(t_0, X^0, U^0(t, X(t, \cdot)), \zeta_\vartheta(\cdot)) \leq \gamma^0(t_0, X^0) = \mathfrak{F}^0(t_0, X^0).$$

If in addition the condition of Theorem 2 (ii) (of Theorem 3 (iii)) does hold then $\gamma^0(t_0, X^0) = \gamma^{00}(t_0, X^0)$ $(\gamma^0(t_0, X^0) = \gamma^*(t_0, X^0))$.

(ii) Under Assumption B we have

$$\mathfrak{F}(t_0, X^0, U^{00}(t, X(t, \cdot)), \zeta_\vartheta(\cdot)) \leq \gamma^{00}(t_0, X^0).$$

(iii) Under Assumption C we have

$$\mathfrak{F}(t_0, X^0, U^*(t, X(t, \cdot)), \zeta_\vartheta(\cdot)) \leq \gamma^*(t_0, X^0).$$

In general we have

$$\gamma^0(t, X(t, \cdot)) \leq \gamma^{00}(t, X(t, \cdot)) \leq \gamma^*(t, X(t, \cdot)).$$

Thus along with the worsening of convexity properties for certain functions characterizing the problem (from Assumption A to Assumptions B and C) the values γ^0, γ^{00}, γ^* for the payoff guaranteed by respective extremal strategies are generally increasing.

Acknowledgement. The author is grateful to academician N. N. Krasovskii for a valuable discussion of the paper.

References

1. *Krasovskii, N. N.*, Game Problems on the Encounter of Motions, Nauka, M., 1970.
2. *Krasovskii, N. N.*, Game Problems of Dynamics I, II. Izv. Akad. Nauk SSSR. Tehn. Kibernet. *5* 1969; *1* 1970.
3. *Krasovskii, N. N.*, The Differential Game of Approach-Evasion I, II. Izv. Akad. Nauk SSSR. Tehn. Kibernet. *2, 3* 1973.
4. *Kurzhanskii, A. B.*, Differential Games of Observation. Dokl. Akad. Nauk SSSR, Vol. 207, *3* 1972.
5. *Kurzhanskii, A. B.*, On the Theory of Feedback Observation. Izv. Akad. Nauk SSSR, *5* 1973.

6. *Kurzhanskii, A. B.*, Optimal Systems Combining Control and Observation. Prikl. Mat. Mekh. Vol. 38, *1* 1974.
7. *Bucy, R. S., Joseph, P. D.*, Filtering for Stochastic Processes with Application to Guidance. J. Wiley & Sons Inc. 1968.
8. *Bryson, A. E., Ho, Y. C.*, Applied Optimal Control. Blaisdell. Publ. Co., 1969.
9. *Krasovskii, N. N., Osipov, Yu. S.*, On Theory of Differential Games under Incomplete Information. Dokl. Akad. Nauk SSSR, Vol. 215, *4* 1974.
10. *Krotov, V. F.*, Optimization Methods for Control Processes with Minimax Criteria. I. Aut. i Telemekh. *12* 1973.
11. *Melikian, A. A., Chernousko, F. L.*, On Differential Games with Varying Informational Conditions. Dokl. Akad. Nauk SSSR, Vol. 203, *1* 1972.
12. *Castaing, Ch.*, Sur les équations differentielles multivoqués. C. R. Acad. Sci. Paris, T. 263, *2* 1966, Ser. A, p. 63.
13. *Rockafellar, R. T.*, Convex Analysis. Princeton Univ. Press, 1970.
14. *Kalman, R. E., Bucy, R. S.*, New Results in Linear Filtering and Prediction Theory. Trans. ASME, ser. D, Vol. 83, 1961.
15. *Rockafellar, R. T.*, State Constraints in Convex Control Problems of Bolza. SIAM J. Control, Vol. 10, *4* 1972.
16. *Krasovskii, N. N.*, Theory of Control of Motion. Nauka, 1968.
17. *Lee, E. B., Marcus, L.*, Foundations of Optimal Control Theory. J. Wiley & Sons Inc. 1967.
18. *Pshenitchniy, B. N.*, E-strategies in Differential Games. Topics in Differential Games. New York—London—Amsterdam, 1973.
19. *Demianov, V. F., Malozemov, V. V.*, Introduction to Minimax. Nauka, M., 1971.

О минимаксных стратегиях управления — оценивания при неполной информации

А. Б. КУРЖАНСКИЙ

(Свердловск)

Рассматривается задача минимаксного синтеза системы управления по результатам наблюдения части фазовых координат в предположении, что системы измерения и управления подвержены возмущениям неопределенной природы, для которых отсутствует какое-либо статистическое описание (информация об этих величинах исчерпывается заданием допустимых областей их изменения). Доступное текущее состояние (позиция) для каждой из рассматриваемых систем здесь определяется не фазовым вектором, но слагается из некоторой области фазового пространства, определяемого путем обработки наблюдаемого сигнала и содержащей неизвестный истинный выход (фазовый вектор) системы. Целью синтеза является нахождение закона управления, формулируемого по доступной информации (позиции системы) и обеспечивающего наименьшее по возможности значение заданной функции координат конечного состояния процесса (вопреки противодействию неопределенных факторов). Указанная ситуация формализуется в рамках специфических игровых информационных задач, в которых сочетается оптимизация процесса измерения — оценки координат системы и процесса выбора самого управляющего воздействия. Решение опирается на конструкции теории игровых задач динамики [1—3] и на результаты работ [4—6] по минимаксному наблюдению и фильтрации. Приводится решение ряда задач программного управления по неполным данным и условия их использования для построения искомого синтеза. Изложение ограничено линейными системами, для которых решение описывается в рамках конструктивных соотношений, охватываемых численными методами выпуклого анализа.

А. Б. Куржанский
Институт математики и механики УНЦ АН СССР
СССР Свердловск К-66,
ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 4 (3), pp. 219–229 (1975)

AN APPROACH TO THE MODELLING OF UNCERTAINTIES IN THE SIMULATION OF QUASIDETERMINISTIC DISCRETE EVENT SYSTEMS

A. JÁVOR

(Budapest)

(Received September 20, 1974)

Introduction

Modelling, which always means an abstraction, is generally divided — in the case of discrete event systems — into two categories: viz. deterministic and stochastic systems. These systems can be described as deterministic and probabilistic automata [1].

This categorisation is an abstraction in itself since practically every physical phenomenon is of stochastic nature; it is only the numerical values of the probabilities of the events involved that determine how we regard the system. Namely if these probabilities are very close to "1" or "0" we may treat the system as deterministic, otherwise as stochastic. The examination of real systems often shows, that the question of how they should be treated depends on the "resolution" or "level" of the examination, and with gradually more detailed examination, i.e. increased "resolution", the system can be interpreted alternately as stochastic or deterministic [2, 3].

In this paper we shall introduce a third — intermediate — category; the concept of *Quasideterministic Systems* (QDS).

This, of course, is an abstraction, too, by which we shall understand the following

Definition 1. *We shall designate a system as "quasideterministic" if it is such that in the majority of cases (or time) its behaviour can be regarded as deterministic and in the minority of cases (or time) as stochastic.*

These systems were generally intended to be deterministic by their designers, or hoped to be such by their investigators; at some level at least.

To illustrate the problem let us consider two examples.

1. A network of digital logic circuit elements with given binary input sequences can be treated basically as a deterministic system (at some level at least). Nevertheless having for example only a single RS flip-flop in our system driven by the — forbidden — $R = S = 1$ value, the resulting state

of the system becomes indeterminate. Considering the transient states of the system — always taking finite time — by increasing the "resolution" of the examination leads to indeterminacies as well [4].

2. The layered construction of an operating system of a computer is intended to remove the basically nondeterministic nature of the hardware and make "a more or less deterministic automaton" out of it [3].

The more layers smoothing the original behaviour, the less indeterminacy will remain. For instance a computer having two unsynchronized card readers, started simultaneously, as input devices for two programs and a single printer for their output may produce the results in unpredictable order, but the operator and/or operating system may provide for a deterministic order. So on this level we already have a deterministic system regarding the example above.

Such and similar examples may well be found in other fields of discrete event systems as well.

After having determined our concept of QDS's let us envisage our approach to its modelling. This can be formulated briefly. In this paper no attempt has been made to provide a theoretically complete descriptive model construction of QDS's but rather the development of a practical means for using it in computer simulation. The limits of the description should be well defined, and within these limits true and correct mapping of the real system should be supplied, providing as much relevant information at the cost of complexity (computing time, storage space, etc.) as is reasonable and economical.

Description of real discrete systems

A formal description of discrete systems may easily be given, using the common notations of automata theory [1]. Regarding the system as a deterministic finite automaton it can be described by the quintuple

$$\mathfrak{M} = \langle Q, X, Y, \Psi, \Phi \rangle, \tag{1}$$

where Q, X and Y represent the internal, input and output alphabet respectively, while Ψ is the next state function and Φ the output function performing the mappings.

$$q(t+1) = \Psi[q(t), x(t)] \tag{2}$$

and

$$y(t) = \Phi[q(t), x(t)] \tag{3}$$

where

$$q(t), q(t+1) \in Q;\ x(t) \in X;\ y(t) \in Y.$$

A generalization of the ordinary deterministic automata results in the probabilistic automata. Here an input letter $x \in X$ transfers the automaton from state $q_i \in Q$ into any state $q_j \in Q$ with constant, time independent probabilities $\Pi(x, q_i, q_j)$. As we do aim at the realistic description of systems in simulation, the description very often takes the form of a very important subclass; the automata with delay. Here the only difference to the formulae above is that the function Φ takes the form $\lambda(q)$, i.e.

$$y(t) = \lambda[q(t)] \tag{4}$$

and the present value of y does not depend on the present value of x. This can be interpreted as the mapping of reality considering the fact that changes in physical systems never occur instantaneously although in a number of cases this time delay can be ignored.

Let us now more closely consider our problem which consists of developing, operating, observing and evaluating a discrete system. The main points of our task related to the topic of this paper are the following:

§ 1. The simulation model in many cases consists of — a generally large number of — interconnected "element modules" with a priori well-known algorithms describing them [5—7].

This means that the behaviour of the systems can be determined theoretically in advance characterised either by distribution functions or by deterministic parameters; only the complexity of the systems causes them to defy analytic solution necessitating that their behaviour be determined by means of simulation [8, 9].

§ 2. The complexity of the system excludes an all-out investigation of its behaviour i.e. simulation for the entire space $Q \times X$ and only allows that for a subset $Q' \times X'$; $Q' \subset Q$, $X' \subset X$.

This is usually known as testing under "operating conditions".

Such a test has of course to be designed very carefully in order to promote the production of the utmost *relevant* information possible. This problem is beyond the scope of this paper but it should be mentioned that no complete solution of the problem for practical purposes is available [10, 11].

However, it should be mentioned that the rejection of an all-out investigation generally is not caused purely and simply by the difficulties of its realisation but that the amount of results given would be hardly manageable in several cases and enormous efforts would be needed for data compression in order to extract the relevant data from that which is redundant. What is really of interest is the behaviour of the system under "operating conditions" and when this is satisfactory the behaviour beyond it is irrelevant.

§ 3. The operation of the models in simulation is controlled by mechanisms of time advancement [12, 13] different from those used in the classical description of finite automata, since in the former case some time advancement algorithms are based on unequally quantized time steps, by skipping those time increments where no change in the existing $\langle q_i, y_i \rangle$ takes place.

Thus the following problem may well arise.

At t_1 the letter $x_1 \in X$ schedules the next change to t_2 from $q_1 \in Q$ to $q_j \in Q$ and analogously from $y_i \in Y$ to $y_j \in Y$. At time instant t_3 where $t_1 \leq t_3 \leq t_2$ (or in delayed automaton representation $t_1 \leq t_3 < t_2$) another letter $x_2 \in X$ changes the anticipated next states $\langle q_j, y_j \rangle$ to $\langle q'_j, y'_j \rangle$ and the anticipated time of the change to t'_2.

More precisely speaking; generally there are scheduled subsets $Q_2 \subset Q$ and $Y_2 \subset Y$ respectively to the set of future time instants T_2, i.e. for which the relation holds

$$(\forall\, t_{2i} \in T_2)\ (t_{2i} \geq t_1) \tag{5}$$

or

$$(\forall\, t_{2i} \in T_2)\ (t_{2i} > t_1) \tag{6}$$

for delayed automata.

The letter x_3 may change the scheduled subsets to $Q'_2 \subset Q$ and $Y'_2 \subset Y$ and their respective sets of times to T'_2 in the case

$$(\exists\, t_{2i} \in T_2)\ (t_3 \leq t_{2i}) \tag{7}$$

or, for delayed automata

$$(\exists\, t_{2i} \in T_2)\ (t_3 < t_{2i}). \tag{8}$$

This means that in the implementation of the model in simulation the information about the present and scheduled conditions, with their uncertainties, have to be stored and the algorithms performing transformations on them have to be chosen in such a way that if the "forecasts" are modified repeatedly the "last should be the best", i.e. the last transformation applied should supply the correct state or output of the system.

§ 4. Finally a problem which has grave consequences from the point of view of practical realisation in digital computer simulation arises due to the indeterminacies or uncertainties present in the system.

In average probabilistic systems (which may be represented by probabilistic automaton) this cannot really be treated as a difficulty but rather as the reason necessitating simulation. In such systems [7, 14—16] the results of the simulation are self-evidently in forms of distribution functions and the time of the simulation is determined by the level of confidence needed.

The problem arises in the simulation of QDS's where the anticipated result would be the more or less deterministic description of the operation of the model; however the indeterminacies make it difficult to achieve. Great efforts have been made to overcome this difficulty [4, 17—19], mainly in the field of binary logic system simulation using three and four valued simulation, but the problem of unmanageability was still there, since it has been shown [17] that handling the combination of indeterminacies is practically impossible in such a system.

It is obviously very difficult to determine the magnitude of the complexity of handling such a system in general, therefore just to illustrate the problem let us take a quick glance at a binary digital system with the following restrictions. We shall consider a k level digital network without feedback paths (outside the element blocks), where the number of elements, N is considerably higher than the inputs, and, further not coupled outputs of the whole network. Inside the network every $I/0$ point of the elements will be assumed to be connected somewhere. It follows than an element output will feed an average of $\frac{i}{w}$ inputs assuming i inputs and w outputs on average for an element.

Not knowing on which level an indeterminacy originates we shall calculate an average for all k possibilities and thus determine the number v of nodes in the network where it appears; to be written as

$$v = \frac{1}{k} \sum_{m=0}^{k-1} \sum_{l=0}^{m} \left(\frac{i}{w}\right)^l \tag{9}$$

from which the formula

$$v = \frac{\frac{i}{w}\left[\left(\frac{i}{w}\right)^k - 1\right] - k\left[\left(\frac{i}{w}\right) - 1\right]}{k\left[\left(\frac{i}{w}\right) - 1\right]^2} \tag{10}$$

can be derived.

Substituting the values $i = 4$, $w = 1$, $k = 5$ for an average network, we get the value $v = 90.6$ equivalent to 2^{90} possible next Q states. This indicates the magnitude of different continuations of operation caused by a single uncertainty due to its propagation which, if we wanted to track the behaviour of the system means as much branching as 2^{90} and consequently as many simulation runs from every "branching" point quite impossible to realize in modelling large systems.

Representation of QDS's based on the separate handling of indeterminacies of various kinds

Let us now, after determining the nature of the problem and the related environmental conditions in which it has to be solved, try to find a solution for handling uncertainties considering thereby the facts given in § 1—§ 4. As the requirements for getting as much information from the model as possible, and reducing the complexity of its implementation at the same time are contradictory the solution obviously has to be a compromise.

Thus here we should like to give a general philosophy and guidelines for the interpretation of QDS's thereby permitting a relatively wide range of possible alternatives to enable the matching of a realization to the boundary conditions given in a particular case.

The problem can be formulated as finding optimal realizations for the functions and parameters in Eq. (1) and we shall tackle it by a closer examination of some of the terms involved.

The most important of the terms for us is "uncertainty" or "indeterminacy" itself. The situation can be cleared considerably and significant simplifications can be achieved by dividing these situations — based on the nature of their origin — in two groups as follows

Definition 2. *Indeterminacies of first order* (IFO).

Uncertainties in the system due to transient periods between deterministic states $q_i, q_j \in Q$.

Definition 3. *Indeterminacies of second order* (ISO).

Uncertainties, meaning that the system is in a deterministic state (or in a condition to be described by IFO*) but the state (or the direction of the transient state described by the* IFO*) is not known and the state of the system is to be found in a subset of states:*

$$Q^* \subset Q.$$

Another term is the *relevancy* of the information yielded by the simulation. Considering the types of indeterminacies and bearing in mind the features of QDS's already mentioned the following can be stated.

(a) The occurrence of IFO's have to be described entirely according to the system modelled since this type of indeterminacy occurs frequently practically in every system and cannot be considered as irregular operation.

Their occurrence has to be tracked however because on the one hand they may cause ISO's, and on the other hand in some particular cases they should not be allowed during operation.

Their inclusion in the model however does not cause too much trouble since they are transient states by nature and their values and existence are clearly defined by the transient times and states in between which the change takes place. It should be noted that in several cases for "worst case design" the maximum possible changeover times should be taken into consideration.

(b) The representation of ISO's should however be restricted both in time and "space". By this we mean that the occurrence of an ISO at a time t should not influence the ISO signalised at $t' > t$.

This representation can be arranged in a way that if an ISO occurs the next state of the system is chosen arbitrarily as

$$q^*_{i0} \in Q^*$$

according to Definition 3.

Using the formalism given in Eqs (1), (2) and (3) the description of a QDS can be performed by replacing the values q, x, y by vectors

$$\bar{q} = \{q_s, q_{\mathrm{IFO}}, q_{\mathrm{ISO}}\} \tag{11}$$

$$\bar{x} = \{x_s, x_{\mathrm{IFO}}, x_{\mathrm{ISO}}\} \tag{12}$$

$$\bar{y} = \{y_s, y_{\mathrm{IFO}}, y_{\mathrm{ISO}}\}. \tag{13}$$

The indexes of the components designate thereby

s	value of the parameter (deterministic description),
IFO	IFO value of the parameter (i.e. transient value),
ISO	ISO state of the parameter (i.e. being actually in some subset of the whole set of possible values).

Equations (2) and (3) may be rewritten in the forms of

$$\bar{q}(t+1) = \Psi[\bar{q}_r(t), \bar{x}_r(t)] \tag{14}$$

and

$$\bar{y}(t) = \Phi[\bar{q}(t), \bar{x}_r(t)] \tag{15}$$

where, in the case of delayed automata, the term $\bar{x}(t)$ is obviously deleted from Eq. (15). The notations $\bar{q}_r(t)$ and $\bar{x}_r(t)$ mean reduced forms of the state and input letters where the last components for ISO representation are assumed to equal zero, i.e. are deleted.

In the description according to Eqs (14) and (15) the substitution of symbol $\bar{q}_r$ provide for the restriction of propagation in time, and $\bar{x}_r$ for the restriction in space. Regarding the latter it should be noted (see § 1) that the whole system may be represented as formed from a network of automata and the output of an elementary automata $\bar{y}_{ei}$ will provide the input $\bar{x}_{ej}$ of others.

Alternative applications of the description and some remarks

1. The parameters labelled by IFO and ISO in Eqs (11), (12) and (13) can be chosen differently corresponding to the demands of the simulation to be performed. A single value of an IFO parameter may correspond to the following meanings

$$a_i \to a_j.$$

$\alpha)$ $a_i, a_j \in A$ both are fixed

$\beta)$ $a_i \in A$ and a_j is fixed

$\gamma)$ $a_i \in A_i \subset A$ and $a_j \in A_j \subset A$

A_i, A_j being both fixed

$\delta)$ $a_i \in A$ and $a_j \in A_j \subset A$

A_j being fixed.

The values a and A standing for corresponding values of q, x, y and Q, X, Y respectively.

2. The subset of values (q, y) given in Definition 3 may be determined by

(a) The physical properties of the system modelled.

(b) Considerations regarding convenience and simplicity of handling. Thereby the sizes of distinct subsets may be increased at extremum to the whole set of possible values. (This remark naturally applies to the IFO representations as well.)

3. At the occurrence of an ISO due to the restrictions imposed on its further consequences the following measures may be taken to use the information acquired

— restarting the simulation from appropriate input vector $\bar{x}_l$ and appropriately chosen state vector $\bar{q}_l$ after examining — due to the QDS nature — the probably rare and unwanted situation,

— ignoring it after convincing proof that it is irrelevant to us.

This approach follows that of the generally accepted philosophy that simulation is an interactive man-machine method [19].

4. An advantage of our approach is that an analysis (by the machine during the run performing the algorithms, or by the investigator evaluating the results) is well descriptive using the three component vector description.

— The first component supplies the state assumed;

— the second indicates whether it is in a stable condition or has already begun to change its state and the direction of the change over. (The previous value of the first vector components should be changed only at the end of an ISO condition);

— the third component indicates major possible discrepancies in the deterministic operation.

5. Concerning the handling of ISO's it is obvious that some information on the cases in which they are liable to occur has to be determined prior to to the dynamic phase, i.e. run of the simulation. This can be arranged in two ways.

(a) The elements of which the whole model consists must themselves have — well-known behaviour and the algorithms describing them have to provide for the representation of ISO's — which they inherently cause.

(b) The other class of origins giving rise to ISO's is due to the outlay of the topology of the whole system consisting of element modules (i.e. feedback paths). The solution in such cases may be a prior analysis of the system thereby

α) determining those segments of the network the topology of which enables ISO's,

β) performing an all-out analysis on the behaviour of these — generally small — segments and including them in the whole system as subsystems with additional algorithms to be provided for handling the ISO's of topological origin.

Then additional algorithms may be attached to the elements in the subsystem or, alternatively, to the subsystem as a whole.

6. Our next remark concerns the term "state" in simulation model descriptions. It should be noted (referring to § 3) that what is understood here as "state" (i.e. has to be stored) consists of the present physical state, the scheduled future state, together with some associated time values (sometimes even the value of the length of time that the last state was valid has to be known, too).

7. Our last remark concerning § 3 is that in simulation models, although consecutive forecastings may overrule each other both in future $\bar{q}$ and $\bar{y}$ values and their corresponding schedule time values, the end result must be unambiguous independently of their order. This is very important since due to the sequential operation of digital computers even the simultaneous "model events" have to be performed sequentially, sometimes in a random order, from the point of view of the model operation. This poses a condition that the operators corresponding to Ψ and Φ have to be commutative, i.e. the end result has to be independent of the order of their multiple application with the same parameter sets.

Conclusions

Summarizing briefly our considerations and their results, it can be said that by analysing the nature of uncertainties or indeterminacies in discrete event systems a method for their representation has been outlined which can be used advantageously in the simulation of quasideterministic systems where not only the occurrences of "indeterminacies of second order" are rare but are generally to be treated in anexeptional way. The method proposed, together with the alternative variations, simplifies the process of the simulation without significantly reducing the amount of relevant information from the point of view of system analysis.

References

1. *Trakhtenbrot, B. A., Barzdin, Ya. M.*, Finite Automata. North Holland Publishing Company, Amsterdam—London, American Elsevier Publishing Company, Inc. New York, 1973.
2. *Heisenberg, W.*, Das Naturbild der Heutigen Physik. Rowohlt Verlag, Hamburg, 1955.
3. *Dijkstra, E. W.*, Hierarchical Ordering of Sequential Processes. Acta Informatica 1, 115—138 (1971). Springer Verlag.
4. *Jávor, A., Benkő, M. Mrs.*, On the Problems of Timing and Hazards in the Simulation of Digital Networks. Proc. of the "1974 European Conf. on Circuit Theory and Design" 23—26 July 1974. London (Conf. Publ. Number 116) pp. 257—262.
5. *Nutt, G. J.*, Evaluation Nets for Computer System Performance Analysis. Proc. of the Fall Joint Computer Conference, 1972. pp. 279—286.
6. General Purpose Simulation System/360 Introductory User's Manual. IBM Application Program 5th edition, 1969. Nov. GH 20-0304-04.
7. *Jávor, A., Csákány, A.*, Some Achievements in the Investigation of Data Transmission Networks Using a Flexible Simulation System. XVIII. Rassegna Internazionale Elettronica e Nucleare Congresso Scientifico Internazionale per l'Elettronica D-5, Roma, Marzo, 1971. pp. 331—337.
8. *Буленко Н. П.* Моделирование сложных систем. "Наука", Москва.
9. *Tocher, K. D.*, The Art of Simulation. Van Nostrand Co., Inc., Princeton, N. J. 1963. p. 160.
10. *Liskov, B. H.*, A Design Methodology for Reliable Software Systems. Proc. of the Fall Joint Computer Conf. 1972. pp. 191—199.

11. *Mize, J. H., Cox, J. G.*, Essentials of Simulation. Prentice-Hall, Inc. Englewood Cliffs, N. J. 1968.
12. *Breuer, M. A.*, Design Automation of Digital Systems. vol. 1. Prentice-Hall, Inc. Englewood Cliffs, N. J. 1972.
13. *Jávor, A.*, On the Simulation of Time in the Case of Quasisynchronous Discrete Event Distributions. Report of the Central Research Institute for Physics of the Hungarian Academy of Sciences. KFKI–74–9. Febr. 1974.
14. *Goshal, A.*, Some Aspects of Queueing and Storage Systems. Springer Verlag, Berlin—Heidelberg—New York, 1970.
15. *Gordon, G.*, System Simulation. Prentice-Hall Inc. Englewood Cliffs, New Jersey, 1969.
16. *Csákány, A., Jávor, A.*, Investigation of Throughput Efficiency of Multilevel Data Networks by Means of Simulation. XIX. Rassegna Internazionale Elettronica Telecinematografica, Roma, 1972. B–12, pp. 253—258.
17. *Breuer, M. A.*, A Note on Three Valued Logic Simulation. IEEE Trans. on Computers Vol. C–21. No. 4. Apr. 1972. pp. 399—402.
18. *Teets, J. J.*, The Role of Simulation in LSI Design. Spring Joint Computer Conference, 1972. pp. 1065—1070.
19. *Feuvrier, Ch. V.*, La simulation des systèmes. Dunod, Paris, 1971.

О приближенном моделировании квазидетерминированных дискретных систем

А. ЯВОР

(Будапешт)

В работе рассматривается задача представления недетерминированных состояний при моделировании квазидетерминированных систем. Эта проблема, возникающая во многих случаях на практике, обсуждается с учетом таких важных при моделировании вопросов как потребность в машинном времени и объем запоминающего устройства.

Предлагаемый метод основывается на разделении на категории недетерминированных состояний и на их отдельной обработке. Показано, что в процессе моделирования может быть достигнуто упрощение системы, причем количество информации, важной для анализа системы, практически не изменяется.

A. Jávor
Central Research Institute for Physics
Hungarian Academy of Sciences
H-1525 Budapest XII., Konkoly Thege M. út
Hungary

Problems of Control and Information Theory, Vol. 4 (3), pp. 231—239 (1975)

REMARKS ON THE ANALYSIS OF A STABILIZED ANALOG-DIGITAL CONVERTER

K. TARNAY

(Budapest)

(Received October 14, 1974)

Introduction

The results of the measuring of analog processes can be significantly affected by instability of the A/D converter, therefore we have provided an analysis of a stabilized A/D converter. Initially, the conversion factors and the algorithm of drift correction are determined then the relationship between the converter parameters and the stability criterion, respectively the spectrum distortion are represented using computer simulation. Finally a possible treatment of the stabilized A/D converter is presented employing concepts of information theory [1, 5].

Conversion factors of a drift-corrected A/D converter

Semiconductor detectors are very much used nowadays in gamma spectroscopy [2]. These high resolution detectors have made it necessary to improve the stability of the amplifier and A/D converter system. The measuring systems are usually stabilized by drift correction feedback loops. The amplitude-digital conversion [3] can be characterized by the algorithm

$$N = \text{entier } (K \cdot U - A) \tag{1}$$

where U is the input voltage,
K the slope conversion factor,
A the base level conversion factor.

The drift correction can be realized by the alteration of the conversion factors. Figure 1 illustrates a possible realization of the drift correction. The spectrum stabilization is ensured by a converter with a two-loop feedback control. One loop alters the zero point of the conversion characteristic, the second loop stabilizes the slope of the conversion characteristic. Each control loop has a specific reference signal. N_{10} channel number belongs to the input voltage (U_{10}) in the first loop, N_{20} channel number — to the input voltage

(U_{20}) of the second loop. Whereby if the conversion characteristic alters, the channel number of the reference signals also changes (N_{11} instead N_{10} and N_{22} instead N_{20}). This channel number difference controls the stabilization and alters the value of conversion factors until the channel number difference is less than 1.

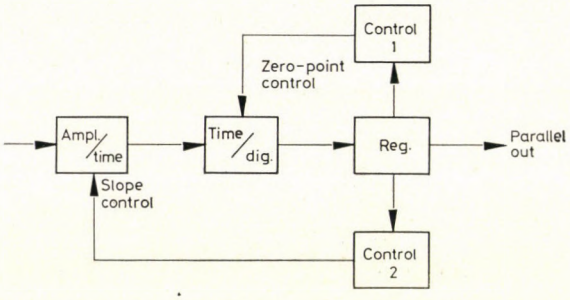


Fig. 1

The conversion factors of Wilkinson-type analog-digital converters are the following

$$K = \frac{C}{I^* T}, \quad A = \frac{D}{T}, \tag{2}$$

where C is the capacity of stretcher,

I the discharging current,

D the base subtraction,

T the repetition time of the clock generator.

The alteration of the slope conversion factor can be realized by an additional current (I_1), the alteration of the base level conversion factor by an additional time subtraction (D_1). The drift-corrected conversion factors are therefore given by

$$K = \frac{C}{I + I_1} * \frac{1}{T}, \quad A = \frac{D + D_1}{T}. \tag{3}$$

The algorithm of drift correction can be determined with the knowledge of the additional coefficients as the function of control steps $I_1(s), D_1(s)$

$$N_{11} = \left[C \frac{\Delta U_1}{I + I_1(s)} - D - D_1(s) \right] * \frac{1}{T}$$
$$N_{22} = \left[C \frac{\Delta U_2}{I + I_1(s)} - D - D_1(s) \right] * \frac{1}{T}, \tag{4}$$

where s is the number of control steps,
$I_1(s)$ the additional discharge current,
$D_1(s)$ the additional base subtraction.

The voltage differences are the following:

$$\Delta U_1 = U - U_{10}$$
$$\Delta U_2 = U - U_{20},$$

where U is the disturbing signal referring to the input.

Simulation of conversion and control

The exhaustive analysis of expected measuring results as a function of converter parameters and disturbing signals can be accomplished by computer simulation. The SINBAD program package (developed at the Central Research Institute for Physics, Budapest, Hungary) gives a good possibility of determining the questionable values [4]. The change in the reference signals is characterized in the program package by the functional relations which are the most frequently encountered in measuring practice (step, exponential, linear functions). In addition, we have introduced the concept of relative conversion factors meaning the difference between the conversion factor at the base state and the factor altered by the feedback control. Figure 2 shows the simulation of the converter control system if the disturbing signal is a step function which causes the input signal to change by 1% in both stabilizing loops. The channel number differences are seen in Fig. 2a as a function of stabilizing steps. Figure 2b illustrates the variation of the relative base level conversion factor and the variation of the relative slope conversion factor.

In Fig. 3 we have shown a case which is difficult to be formulated theoretically, but is nevertheless encountered in practice, that is, the arrival of a new disturbing signal during the stabilizing process. It is assumed that the primary disturbing signal step function is followed by a secondary, Dirac function signal which alters the base level. The broken line in the figure represents the stabilizing process if only primary disturbing signals are present. The program package is also able to furnish information on the appropriate choice of further converter parameters.

In the following the disturbing signals and the stabilization will be related to the spectrum distortion. The normal distributions forming the basic gamma spectrum were generated by Monte Carlo method. The statistic of radioactive decays were characterized by Poisson distribution, that of the time intervals between two successive particles by exponential distribution. These distributions were also generated by Monte Carlo method. The peak distortion

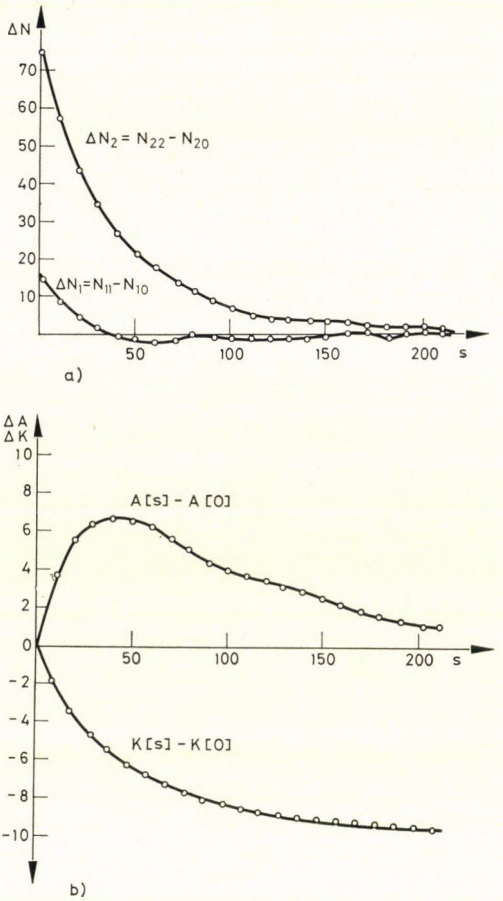

ΔN
70
60
50
40
30
20
10
ΔN2 = N22 - N20
ΔN1=N11-N10
50
100
150
200
s
a)
ΔA
ΔK
10
8
6
4
2
0
-2
-4
-6
-8
-10
A[s] - A[0]
K[s] - K[0]
50
100
150
200
s
b)

Fig. 2

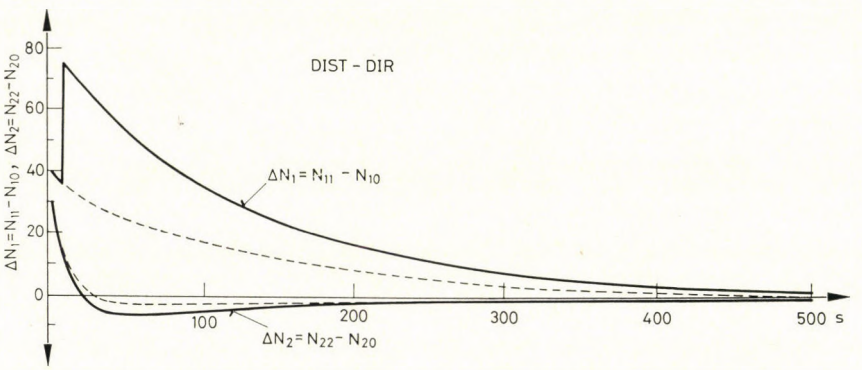

ΔN1=N11-N10, ΔN2=N22-N20
80
60
40
20
0
DIST - DIR
ΔN1 = N11 - N10
100
200
300
400
500 s
ΔN2=N22-N20

Fig. 3

associated with reference signal indicating a sudden change in the input voltage is shown in Fig. 4. The stabilization is performed in 100 steps, the total number of counts in the area below the curve is 100 000. The conversion factors were determined under the condition specified in the previous chapter, thus, $K = -8.909 \times 10^{-3}$ and $A = 0.5623$. The basic curve is to be seen in Fig. 4a,

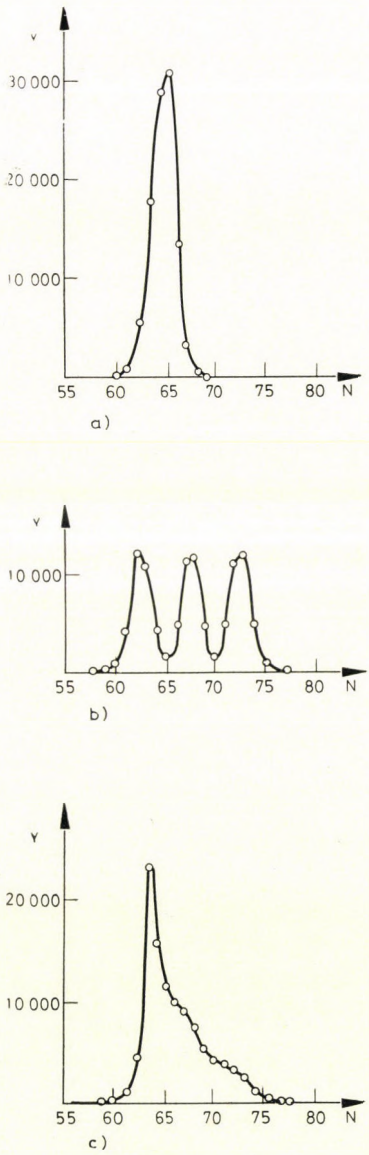


Fig. 4

the curve obtained without stabilization in Fig. 4b and that with stabilization in Fig. 4c. The curve for stabilized operation would give a better approximation to the normal distribution if more stabilizing steps had been assumed. However these diagrams illustrate more spectacularly the result of the simulation.

Figure 5 shows the spectrum distortion minimized by the drift correction analysed in Fig. 2. The increase in input voltage is $+1\%$. The value of the drift given in channel number and the variation of the relative conversion

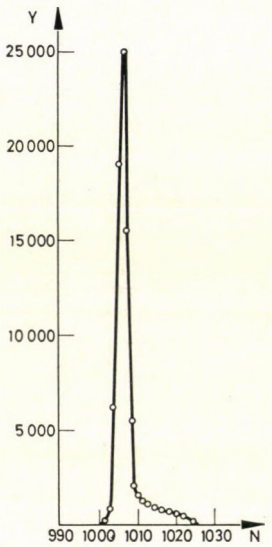


Fig. 5

factors due to the number of stabilizing steps are seen in Fig. 2. The spectrum distortion to be seen in Fig. 5 proves that this result of the simulation is already in good agreement with the measured values.

Entropy models of a stabilized A/D converter

Let us assume that the alterations caused by disturbing effects are limited and therefore the number of control steps is finite. We shall investigate the probability that the reference signal falls in the jth channel by the $n+1$-th control step assuming that it was in the ith channel by the nth control step. This conditional probability is

$$P_{ij}(n) = P\{\xi(n+1) = j \mid \xi(n) = i_n,\ \xi(n-1) = i_{n-1}, \ldots, \xi(1) = i_1\} \quad (5)$$

and may be substituted by the transitional probability of a Markov chain with a good approximation according to the simulation results. The transitional probability is then

$$p_{ij}(n) = P\{\xi(n+1) = j \mid \xi(n) = i_n\}, \tag{6}$$

where the $\xi(n)$ random variable is always given as a channel number.

The random variables can be expressed as a function of the converter parameters. The initial distribution in the case of ideal channel profile and zero differential nonlinearity is

$$P_i(0) = P\{\xi(0) = i\}, \tag{7}$$

and the random variable

$$\xi(0) = \text{entier } (K \cdot U - A) \tag{8}$$

according to (1).

In an ideal case K and A are constant, but in practice each has a Gaussian distribution. If the channel profile is not ideal, then u is substituted by the $w = f(u)$ random function characterizing the channel profile.

The $\xi(n)$ random variable can be expressed according to (4) by

$$\xi(n) = \text{entier } \left\{[\varkappa(n) * U_1 - \nu(n)] * \frac{1}{T}\right\}, \tag{9}$$

where U_1 is the input step function characterizing the disturbing effect. The random variables are the following

$$\varkappa(n) = K * \left\{\sum_{m=1}^{n} [\xi(m) - \xi(0)]\right\} \tag{10}$$

$$\nu(n) = D + \sum_{m=1}^{n} [\xi(m) - \xi(0)],$$

where $\nu(n)$ means the base level alteration, $k(n)$ the slope alteration. Relations (8), (9) and (10) are the first results, the course of possible following steps are only outlined.

The entropy models are determined by the matrix of the transitional probabilities. That is

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}. \tag{11}$$

If the control corrects only the slope alteration, **P** is a symmetrical matrix. The elements of the matrix can be determined from the transitional probabilities. The SINBAD program package gives the value of $(p_{i1}, p_{i2}, \ldots, p_{in})$ with a good approximation, that is, the probability of the transition from the ith channel into the 1st, 2nd, . . . , nth channel. The determination of these values is based on the relative frequency of the earlier introduced random variables by means of the generation of many random distributions.

Therefore the entropy according to the ith channel is

$$\mathbf{H}_i = -\sum_{k=1}^{n} p_{ik} \log p_{ik}\,, \tag{12}$$

the entropy characterizing the whole control and conversion is

$$\mathbf{H} = \sum_{i=1}^{n} P_i H_i\,. \tag{13}$$

Many control steps exist generally by A/D conversion so the stabilization is described by the rth entropy assuming that the system returns in r steps to the basic state

$$\mathbf{H}^{(r)} = \sum_{i=1}^{n} P_i H_i^{(r)}. \tag{14}$$

*

Acknowledgements. Thanks are due to Dr. Gy. Ember and to Dr. E. Tóth for their critical remarks. The author wishes to express her gratitude to M. Sándory for supporting this work and to her colleagues L. Varga, L. Szabó and A. Csákány for helpful discussions.

References

1. *Shannon, C. E.*, The mathematical theory of communication. University of Illinois Press, Urbana, 1949.
2. *Quittner, P.*, Gamma-ray spectroscopy. Akadémiai Kiadó, Budapest, 1972.
3. *Tarnay, K.*, Simulation of a servostabilized analog-to-digital converter. Proc. 1974 Summer Computer Simulation Conference (Houston, Texas) pp. 492–496.
4. *Tarnay, K.*, The SINBAD program package. Report KFKI–74–52 (1974).
5. *Csákány, A.*, private communication, 1974.

Замечания к анализу стабилизированных аналого-цифровых преобразователей

К. ТАРНАИ

(Будапешт)

Некоторые параметры А/Ц преобразователей гаммаспектрометрии с большим разрешением могут невыгодно влиять на результат измерения. Алгоритм А/Ц преобразования:

$$N = \text{entier}\ (K \cdot U - A),$$

где U — входное напряжение,
K — коэффициент преобразования крутизны,
A — коэффициент преобразования исходного уровня.

Приводится упрощенная блок-схема А/Ц преобразователя с двумя петлями и обратной связью с коррекцией дрейфа (рис. 1) и задаются коэффициенты преобразования А/Ц преобразователей типа Вилькинсона. Алгоритм коррекции дрейфа

$$N(s) = \left[C\,\frac{\Delta U}{I + I_1(s)} - D - D_1(s)\right] \cdot \frac{1}{T},$$

где s — число шагов регулирования,
$I_1(s)$ — дополнительный ток разряда,
$D_1(s)$ — дополнительное вычитание исходного уровня,
ΔU — разница напряжения на входе,
I — ток разряда,
D — вычитание исходного уровня,
T — время генератора.

В случае различных сигналов помех с помощью имитации на ЭВМ были получены кривые, характеризующие преобразование и регулирование.

На рис. 2 показана разница числа каналов в зависимости от шагов регулирования в том случае, если сигнал помехи является ступенчатой функцией.

На рис. 3 видно влияние второго сигнала помехи. На рис. 4 и 5 приведены искажения спектра.

В заключение описывается одна из возможностей объяснения действия стабилизированных А/Ц преобразователей на основе теории информации.

Переходная вероятность, характерная для регулирования,

$$p_{ij}(n) = P\{\xi(n+1) = j \mid \xi(n) = i_n\},$$

а случайная величина

$$\xi(n) = \text{entier}\left\{\left[\varkappa(n) \cdot U - \vartheta(n)\right] \cdot \frac{1}{T}\right\}.$$

С помощью переходных вероятностей может быть определена энтропия А/Ц преобразования на r-том шаге.

K. Tarnay
Central Research Institute for Physics
Hungarian Academy of Sciences
H-1525 Budapest XII., Konkoly Thege M. út
Hungary

Problems of Control and Information Theory, Vol. 4 (3), pp. 241—257 (1975)

ОПТИМАЛЬНОЕ НАБЛЮДЕНИЕ БОЗОННЫХ СИГНАЛОВ В КВАНТОВЫХ ГАУССОВСКИХ КАНАЛАХ

В. П. БЕЛАВКИН

(Москва)

(Поступила в редакцию 4 апреля 1974 г.)

Формулируется проблема оптимизации наблюдения в неидеальных каналах, подчиняющихся квантовой статистике. Доказывается, что в линейных гауссовских каналах со статистикой Бозе—Эйнштейна оптимальным наблюдением, минимизирующим бейесовский квадратичный риск оценивания гауссовского входного сигнала, является G-когерентное измерение, определяющее оптимальное линейное оценивание [1] в общем (негауссовском) бозонном случае.

Формулируется комплексное обобщение этой квадратично-гауссовской проблемы, допускающее индефинитные коммутационные соотношения $[y, y^+] = C$. Путем разложения $T = T_+ + T_-$, оптимально дефинитизирующего матрицу C, находится общий $A = A_+ + A_-$, $A_\pm = T_\pm \left(R \pm \frac{1}{2} C\right)^{-1}$ оптимального линейного преобразования Ay выходных наблюдаемых y, обобщающий классическое $TR^{-1}y$.

Введение

Статья посвящена решению задачи оптимизации наблюдения (т. е. изменения, оценивания и фильтрации) гауссовых сигналов в линейных каналах, описываемых на квантовом уровне. Можно считать, что рассматриваемые сигналы являются электромагнитными полями на апертуре приёмной антенны, хотя такая конкретизация вовсе не обязательна — изложение как и в [1] ведётся для произвольных бозонных сигналов, которые являются квантовыми аналогами выходов классических динамических систем. Результаты статьи имеют прямое отношение к оптимизации обработки оптических когерентных сигналов в лазерных системах связи, для которых соответствующие классические результаты, не учитывающие квантовую природу электромагнетизма, могут приводить к существенным ошибкам [2]. Подобные квантовые задачи возникают также при оптимизации наблюдения и управления микросистемами, а также макросистемами в условиях предельно высокой точности («микронаблюдение» и микроуправление»), и имеют принципиальное значение для общей теории систем.

В статье доказывается, что линейное оценивание и фильтрация [1], как и классическом случае, являются абсолютно оптимальными при квадратичной функции штрафов, гауссовом априорном распределении, и гауссово-аддитивном шуме. Предварительное решение этой задачи для вещественной ситуации

дано в [3], где оптимальность доказывается в предположении единственности решения соответствующего уравнения (единственность доказывается в [3] лишь в квазиклассическом приближении). В отличие от классического случая оптимальные оценки должны основывается, однако, не на результатах прямого наблюдения, а косвенного [3].

Излагаемый здесь метод оптимизации отличается от метода [3] большей универсальностью и дает исчерпываышее решение квадратично-гауссовской проблемы также и в смысле единственности. Для простейшего одномодового случая, допускающего нормальное комплексное представление, оптимальность линейного оценивания и соотвествующего косвенного иэмерения установлена также в [4,5]

1. Оптимальное наблюдение квантовых сигналов

Доступный наблюдению квантовый сигнал (например, квантовое поле на апертуре приемной антенны), как и в [1], будем обозначать семейством символов $y=\{y_\tau, \tau \in T\}$, образующих относительно алгебраических операций некоммутативную по умножению алгебру наблюдаемых $\mathfrak{I}$. Не ограничивая общности, как это принято в квантовой теории (см., например, [7]), алгебру $\mathfrak{I}$ будем отождествлять с алгеброй операторов, действующих в некотором гильбертовом пространстве $\mathfrak{H}$. Статистическое состояние на алгебре наблюдаемых $\mathfrak{I}$ является функцией передаваемой информации $\vartheta \in \Theta$, что находит отражение в параметрической зависимости от ϑ оператора плотности (о. п.) $\varrho(\vartheta)$. Семейство $\{\varrho(\vartheta), \vartheta \in \Theta\}$, описывающее при каждом ϑ некоторое состояние на операторной алгебре $\mathfrak{I}$, играет роль условных плотностей вероятностей $p(y/\vartheta)$. Уместно напомнить, что всякий положительный (т. е. неотрицательно определенный) оператор $\varrho \geq 0$, имеющий единичный след $\mathrm{Tr}\,\varrho = 1$, есть о. п. некоторого состояния, и определяет математические ожидания $\langle f \rangle$ наблюдаемых $f \in \mathfrak{I}$ по формуле: $\langle f \rangle = \mathrm{Tr}\, f\varrho$.

Произвольные (включая косвенные) наблюдения квантовых сигналов описываются операторнозначными мерами П на пространстве ожидаемых результатов измерений $\lambda \in \Lambda$. Операторные вероятности (о. в.) Π (см. [8]) являются квантовым обобщением условных распределений вероятностей $P(\cdot|y)$ на Λ, и во всех отношениях аналогичны им: они определяются как положительные счетно-аддитивные отображения $\Pi(\cdot)$ σ — алгебры подмножеств множества Λ в операторную алгебру $\mathfrak{I}$, удовлетворяющие условию нормировки:

$$\int_{\lambda} \Pi(d\lambda) = 1, \tag{1.1}$$

где 1 — единичный оператор алгебры $\mathfrak{I}$. Как и в классическом случае, вероятность $P(\Delta)$ события $\lambda \in \Delta$ есть математическое ожидание $\langle \Pi(\Delta) \rangle$. В случае,

когда о. в. удовлетворяет условию ортогональности $\Pi(\Delta)\Pi(\Delta') = 0$ для любых непересекающихся $\Delta, \Delta' \subset \Lambda$, она описывает прямое (не косвенное) наблюдение, т. е. идеальное измерение наблюдаемых $x = \int \lambda \Pi(d\lambda)$ алгебры $\mathfrak{I}$, имеющих общие ортопроекторы $\Pi(d\lambda)$. В противном случае наблюдение является косвенным даже если элементарные события $\lambda \in d\lambda$ по-прежнему описываются о. в. $\Pi(d\lambda) = \pi(\lambda)\, d\lambda$ пропорциональными некоторым проекторам $\pi(\lambda)$ на непересекающиеся (но неортогональные) подпространства $\mathfrak{H}_\lambda \subset \mathfrak{H}$. Косвенные наблюдения по отношению к квантовому сигналу y называются также *квазиизмерениями*.

В классическом случае принимаемый сигнал является в принципе полностью наблюдаемым, и любую содержащуюся в нём информацию можно получить в результате сколь угодно точной его регистрации. В квантовом случае такого универсального полного наблюдения не существует, и при решении задач синтеза в первую очередь возникает специфически квантовая задача *оптимизации наблюдения*. Очевидно, что для нахождения оптимальных статистических решений достаточно рассматривать лишь наблюдения, множества ожидаемых результатов которых совпадает с пространством $\mathfrak{U}$ допустимых оценок (каждая оценка есть результат наблюдения!). При этом о. в. являются одновременно и решающими операторами. Такой подход был с успехом применен в [4, 5, 9, 10] при решении некоторых бейесовских задач оптимизации обработки квантовых сигналов.

Бейесовские оптимальные наблюдения описываются о. в. Π^0, которые минимизируют средний риск

$$\mathfrak{R} = \iint C(\lambda, \vartheta) \cdot P(d\lambda/\vartheta)\, P(d\vartheta), \qquad P(d\lambda/\vartheta) = \operatorname{Tr} \Pi(d\lambda)\, \varrho(\vartheta)\,. \tag{1.2}$$

Здесь $C(\lambda, \vartheta)$ — функция штрафов, а $P(d\vartheta)$ — заданное априорное распределение вероятности информационных параметров (или гипотез) ϑ. Условиями минимальности риска (1.2), обобщающими условия, сформулированные в [2] для задач различения гипотез, есть условие эрмитовости оператора

$$\Lambda = \int \mathfrak{R}\,(\lambda)\, \Pi^0(d\lambda)\,, \tag{1.3}$$

где

$$\mathfrak{R}(\lambda) = \int C(\lambda, \vartheta)\, \varrho(\vartheta),\, P(d\vartheta),$$

которое вместе с условием положительности

$$\mathfrak{R}(\lambda) - \Lambda \geq 0 \tag{1.4}$$

является и достаточным условием оптимальности, если о. в. Π^0 удовлетворяет условию ортогональности (оптимизация в классе *косвенных* наблюдений в [2] не рассматривалась).

В общем случае решение Π^0 задачи (1.3), (1.4) в классе ортогональных о. в. может и не существовать. Для выделения оптимального Π^0 в классе

произвольных наблюдений условие (1.3) следует дополнить условием

$$(\mathcal{R}(\lambda) - \Lambda) \cdot \Pi(d\lambda = 0)\,, \tag{1.5}$$

выполняющимся автоматически лишь для ортогональных о. в. Необходимое условие оптимальности (1.5) было получено в [4, 5, 10], и в классическом случае вместе с (1.4) означает, что оптимальное решающее правило $\Pi(d\lambda/y)$ является нерандомизированным, и сосредоточено лишь в тех точках $\lambda = \gamma(y)$, в которых апостериорный риск $\mathcal{R}(\lambda)$ имеет при каждом y минимальное значение. В квантовом случае условия (1.4), (1.5) имеют аналогичный смысл, однако они не означают, что оптимальные о. в. при разных λ ортогональны и соответствующие им наблюдения являются некосвенными. Легко доказать [11], что условия (1.4)—(1.5) являются достаточными.

2. Вещественный квадратично-гауссовский случай

Пусть передаваемый сигнал ϑ (набор подлежащих оценке параметров $\vartheta = \{\vartheta_\alpha\}$) есть гауссовский вектор (столбец) со значениями в вещественном пространстве $\mathcal{U}$ размерности r, нулевым средним, и известной корреляционной матрицей $S = \| \langle \vartheta_\alpha \vartheta_\beta \rangle \|$. Пусть принимаемый вещественный бозонный (см. [1] сигнал y является вещественно-линейно модулированным, и при каждом ϑ гауссовским. Это означает, что $y = \{y_\tau\}$ есть суперпозиция $y = D\vartheta + \xi$, где $D = \|D_{\tau\alpha}\|$ — вещественная матрица, а $\xi = \{\xi_\tau\}$ — гауссов вещественный ($\xi_\tau^* = \xi_\tau$) бозонный шум канала с нулевым средним $\langle \xi \rangle = 0$, определяемый матрицами

$$N = \left\| \frac{1}{2} \langle \xi_\sigma \xi_\tau + \xi_\tau \xi_\sigma \rangle \right\|, \quad C = \| \xi_\sigma \xi_\tau - \xi_\tau \xi_\sigma \| \tag{2.1}$$

(кратко: $N = \langle \xi \xi^T \rangle_+$, $C = [\xi, \xi^T]$, ξ — столбец, ξ^T — строка).

Требуется найти оптимальное по критерию

$$c(\lambda, \vartheta) = \frac{1}{2} (\lambda - \vartheta)^T G (\lambda - \vartheta), \quad G = \|G_{\alpha\beta}\| > 0 \tag{2.2}$$

наблюдение, реализующее бейесовскую (раздел 1) оценку $\lambda \in \mathcal{U}$.

Предложение. 1. В сформулированных условиях оптимальное наблюдение сводится к введенному в [12] оптимальному когерентному измерению повторно коммутирующих наблюдаемых $u = \{\sum_\tau A_{\alpha\tau} y_\tau\} \equiv Ay$, где матрица A удовлетворяет уравнению [3, (3.10)], комплексный вариант которого представлен в (3.6). Это измерение в дальнейшем называем G — когерентным.

Для доказательства требуется следующее представление бозонного канала, определяемое разложением $\xi = a + a^*$, ортогональным относительно

некоторой симметрической положительно-определенной матрицы $H = \|H_{\alpha\tau}\|$: $a^T Ha = 0$, и удовлетворяющим коммутационным соотношениям $[a, a^T] = 0$, $[a, a^+] \geq 0$ ($a^+ = \{a_\tau^*\}$ — строка, сопряженная к столбцу $a = \{a_\tau\}$). Разложение $\xi = a + a^*$, называемое H-нормальным разложением по (ненормированным) операторам уничтожения a_τ и сопряженным операторам рождения a_τ^*, определяется проекционной матрицей $P: a = P\xi, a^* = P^*\xi$, удовлетворяющей условиям $P + P^* = I$, $P^+ HP^* = 0$, и дефинитизирующей (антисимметрическую, мнимую) матрицу $C: PC = CP^+ \geq 0$. Если матрица C невырождена (что в дальнейшем для простоты подразумевается), то матрица P однозначно определяется матрицей $H: P = (I + \operatorname{sign}(CH))/2$.

Определение. Представление операторов ϱ функционалами $p(v)$ на вещественном пространстве $\mathfrak{E}$ векторов $v = \{v_\tau \in T\}$ в виде нормально упорядоченных выражений $\varrho =: p(a + a^*)$: (т. е. таких, что все операторы $\{a_\tau^*\}$ действуют после операторов $\{a_\tau^*\}$ называется H-нормальным, если разложение $\xi = a + a^*$ H-нормально.

Если функционал $p_\xi(v)$ представляет о. п. ϱ_ξ аддитивного бозонного шума ξ канала относительно H-нормального разложения $\xi = a + a^*$, то функционал $p(v) = p_\xi(v - D\vartheta)$ представляет о. п. $\varrho(\vartheta)$ выхода $y = D\vartheta + \xi$ этого (линейного) канала относительно H-нормального разложения $y = b + b^*$ (преобразования сдвига $b = a + PD\vartheta$, упорядоченность не изменяют).

В H-нормальном представлении о. п. гауссовского бозонного шума описывается функционалом

$$p(v) = \det |CN_H^{-1}|^{1/2} \exp\left\{-\frac{1}{2} v^T N_H^{-1} v\right\}, \tag{2.3}$$

где матрица N_H есть H-антинормально упорядоченная корреляционная матрица шума $\xi = a + a^*$:

$$N_H = \langle a_\sigma a_\tau + a_\sigma a_\tau^* + a_\tau a_\sigma^* + a_\sigma^* a_\tau^* \rangle .$$

Она связана с матрицами (2.1) соотношением

$$N_H = N + \operatorname{Re} PCP^+ = N + \frac{1}{2}|CH| H^{-1}, \tag{2.4}$$

которое нетрудно получить, переходя к симметризационному (вигнеровскому) упорядочению с помощью коммутационных соотношений $[a, a^+] = PCP^+$, и, учитывая, что матрица $2 \operatorname{Re} PCP^+ H = \operatorname{sign}(CH)CH$ есть модуль $|CH|$ H-эрмитовой матрицы CH.

Подставляя в (2.3) вместо ν операторы $y-D\vartheta$, и интегрируя квадратично-гауссовское выражение

$$c(\lambda,\vartheta):p_\xi(y-D\vartheta):p(\vartheta)\quad\left(p=\det(2\pi S)^{-1/2}\exp\{-\vartheta^T S^{-1}\vartheta/2\}\right)$$

по ϑ, нетрудно найти оператор апостериорного риска в H-нормальном представлении:

$$\mathcal{R}(\lambda)=:\mathcal{R}\,p_y(y):+\frac{1}{2}:(\lambda-Ay)^T G\,(\lambda-Ay)\,p_y(y)\,. \tag{2.5}$$

Здесь $\mathcal{R}$ есть след $\mathrm{Sp}\,SG(I-AD)/2$ в $\mathcal{U}$, A — решение уравнения

$$A(DSD^T)+N+\frac{1}{2}A\,|CH|\,H^{-1}=SD^T\,, \tag{2.6}$$

а $p_y(.)$ — гауссова плотность

$$p_y(\nu)=\det|C(DSD^T+M)^{-1}|^{1/2}\exp\left\{-\frac{1}{2}\nu^T(DSD^T+M)^{-1}\nu\right\},$$

определяющая о. п. $\varrho=\int\varrho(\vartheta)\,P(d\vartheta)=:p_y(y)$: в виде нормально упорядоченного относительно операторов $b=Py$, $b^*=P^*y$ выражения.

Доказательство предложения 1. Рассмотрим вначале случай, когда размерности пространств $\mathcal{E}$ и $\mathcal{U}$ совпадают. Тогда, *если уравнение* (2.6) *при подстановке* $H=A^TGA$ *имеет невырожденное решение* A^0, *то оно определяет единственную о. в.* Π^0, *удовлетворяющую необходимым и достаточным условиям оптимальности* (1.3)—(1.5).

Действительно, полагая в (2.5) $\Lambda=:\mathcal{R}\,p(y):=\mathcal{R}\,\varrho$, $A=A^0$, $y=b+b^*$, расставляя, операторы $b=\{b_\tau\}$, $b^*=\{b_\tau^*\}$ в нормальном порядке, и учитывая, что $(b-\beta)^T A^{0T}GA^0(b-\beta)=0$ при $H=A^{0T}GA^0$ ($\beta=PA^{0-1}\lambda$), представим оператор $\mathcal{R}(\lambda)-\Lambda$ в следующем виде

$$\mathcal{R}(\lambda)-\Lambda=(b-\beta)^+A^{0T}GA^0\,\varrho\,(b-\beta)\,, \tag{2.7}$$

в котором его положительность очевидна. Оператор (2.7) при каждом β имеет единственный собственный вектор $\varphi(\beta)\in\mathfrak{H}$, отвечающий нулевому собственному значению — вектор когерентного состояния, являющийся и правым собственным вектором операторов уничтожения $\{b_\tau\}$ с собственными числами $\beta=\{\beta_\tau\}:b_\tau\varphi_{(\beta)}=\beta_\tau\varphi_{(\beta)}$. Как известно [13], система когерентных векторов $\{\varphi_\beta\}$ является неортогональной и переполненной и, следовательно, определяет единственное оптимальное квазиизмерение, описываемое семейством когерентных проекторов $\{\pi^0(\lambda)=\varphi_{(\beta)}\varphi_{(\beta)}^*\}$, $\lambda=\beta+\beta^*\in\mathcal{U}$. Любое такое

семейство можно представить в виде предела $\lim\limits_{\theta\to 0}\pi^\theta(\lambda)$ семейства гауссовских о. п.

$$\pi^\theta(\lambda)=\det\left|2\,\mathrm{sh}\,\frac{1}{2\theta}\,ACA^TG\right|^{\frac12}\mathrm{e}^{-\frac{1}{2\theta}(\lambda-Ay)^TG(\lambda-Ay)},\quad \lambda\in U. \qquad (2.8)$$

Каждое семейство (2.8) определяет некоторую *гауссовскую* о. в. $\Pi^\theta(d\lambda)=$ $=\pi^\theta(\lambda)\;d_\mu(\lambda)/\det|ACA^T|^{1/2}$, нормированную по мере $d\mu(\lambda)=\prod\limits_{\alpha=1}^{r}(d\lambda_\alpha/\sqrt{2\pi})$, и описывает при $\theta\to 0$ оптимальное по критерию $(\lambda-u)^TG(\lambda-u)$ *когерентное* измерение линейных функционалов $u=Ay$ [12]. Итак, оптимальные проекторы $\pi^0(\lambda)$ имеют форму (2.8), полностью определяются решением $A=A^0$ уравнения (2.6) при $H^{-1}=A^TGA$, и реализуют минимальный риск $R^0=$ $=\dfrac12\,\mathrm{Sp}\,SG(I-A^0D)$, который совпадает с найденным в [3]. В отличие от классического случая оптимальное наблюдение существенно зависит от вида квадратичного критерия (2.2), т. е. от матрицы G, почему мы и называем его *G-когерентным наблюдением (измерением) некоммутирующих наблюдаемых* $u^0=A^0y$.

В общем случае матрица A в (2.6) имеет область значений $A\mathfrak{E}$, принадлежащую $\mathfrak{U}:A\mathfrak{E}\subseteq\mathfrak{U}$, и ранг $r(A)\leq r\;(r=\dim\mathfrak{U})$. Если $r(A^TGA)<r(C)$, неопределённость обращения $(A^TGA)^{-1}$ при подстановке $H=A^TGA$ в (2.6) можно попытаться раскрыть, вводя функцию $\mathrm{sign}\,(x)=|x|/x$:

$$A\,|CH|\,H^{-1}=A\,\mathrm{sign}\,(CH)\,C\overset{H=A^TGA}{=\!=}\mathrm{sign}\,(ACA^TG)\,AC\,.$$

В результате уравнение (2.6) переходит в уравнение (3.6), выведенное в [3] для гауссовского вещественного случая минимизацией среднеквадратичного риска в классе операторов $u=\int\lambda\Pi(d\lambda)$ с использованием обобщённого неравенства Гейзенберга $\langle uu^T\rangle_+\geq\dfrac12\langle[u,u^T]\rangle$. Это уравнение позволяет выделить решение $A=A^0$, если оно удовлетворяет условию невырожденности матрицы A^0CA^{0T} в подпространстве $A^0\mathfrak{E}\subseteq U$, в противном случае матрица $\mathrm{sign}\,(A^0CA^{0T}G)$ на $A^0\mathfrak{E}$ не определена. Если такое A^0 существует, то оно определяет оптимальное наблюдение — G-когерентное измерение наблюдаемых $u^0=A^0y$. В самом деле, так как невырожденность матрицы A^0CA^{0T} в $A^0\mathfrak{E}$ эквивалентна невырожденности матрицы C_1, индуцируемой матрицей C в подпространстве $A^{0T}GA^0\mathfrak{E}$, в качестве определяющих H-нормальное представление пространства и матрицы H можно взять подпространство $\mathfrak{E}_1=A^{0T}GA^0\mathfrak{E}$ и $H=A^{0T}GA^0$. Представляя о. п. $\varrho(\vartheta)$ в виде нормально упорядоченного выражения от соответствующих операторов $b=Py$, $b^*=$

$= P^* y$, образующих "достаточную статистику" $y_1 = b + b^* = E_1 y : A^0 y = A^0 y_1$ (проектор $E_1 = P + P^*$ из $\mathfrak{H}$ в $\mathfrak{H}_1$, являясь единицей I в E_1, удовлетворяет условию $A^0 E_3 = A^0$), нетрудно найти, что оператор апостериорного риска и в этом случае будет иметь вид (2.5), где $A = A^0$. Апорядоченность по остальным, коммутирующим с $y_1 = E_1 y$ операторам $y_2 = y - C E_1 C_1^{-1} y_1$ и входящим только в $: p_y(y) : = \varrho$ для выявления неотрицательной определённости $\mathcal{R}(\lambda) - \Lambda \geq 0$ является несущественной. Действительно, полагая в (2.5) $A = A^0$, учитывая, что $A^0 y = A^0 y_1 = A^0(b + b^*)$, и расставляя операторы b, b^* в нормальном порядке, получим положительно определённый оператор

$$\mathcal{R}(\lambda) - \Lambda = (b - \beta)^+ A^{0T} G A^0_\varrho (b - \beta) + \frac{1}{2} \lambda_2^T G \lambda_2 \varrho , \tag{2.9}$$

где λ_2 — G-ортогональное дополнение вектора $\lambda_1 = A^0(\beta + \beta^*)$: $\lambda = \lambda_1 + \lambda_2$, $\lambda_2^T G \lambda_1 = 0$. В соответствии с этим оптимальное наблюдение описывается о. в., сосредоточенной на подпространстве $A^0 \mathfrak{H} \subseteq \mathcal{U} : \Pi^0(d\lambda) = 0$ при $\lambda_2 \neq 0$, на котором по прежнему описывается G-когерентными проекторами $\pi^0(\lambda_1)$, представимыми операторами (2.8) при $\lambda \in A^0 \mathfrak{H}$ и $\theta \to 0$.

И, наконец, если решения A^0 уравнения (3.6), удовлетворяющего условию невырожденности матрицы, индуцированной матрицей $A^0 C A^{0T}$ в $A^0 \mathfrak{H}$ не существует, следует дополнить это уравнение уравнением $A C A^T G \mathcal{U}_0 = 0$ на некотором подпространстве $\mathcal{U}_0 \subseteq A \mathfrak{H}$, дающим условие для нахождения неопределённого значения матрицы $B = \operatorname{sign}(A C A^T G)$ на $\mathcal{U}_0$. Такая система уравнений при подходящем $\mathcal{U}_0$ уже имеет решение. (В частности, всегда существует решение при $\mathcal{U}_0 = \mathcal{U}$, соответствующее полной коммутативности $[u, u^T] = = A C A^T = 0$). Среди всевозможных решений следует отобрать оптимальные A^0 в соответствии с неравенством $|A^0 C A^{0T} G| \geq B A^0 C A^{0T} G$, что соответствует доопределению значений сигнатуры в интервале $-I \leq B \leq I$ (знак равенства должен иметь место по крайней мере на G-ортогональном дополнении $\mathcal{U}_1 = = A \mathfrak{H} \ominus \mathcal{U}_0$ к $\mathcal{U}_0$, на котором сигнатура определена). Используя нормальное упорядочение по операторам $b = Py$, $b^* = Py$ ($P + P^* = E_1$ — единица $(A^{0T} G A^0)^{-1}$ — ортогонального дополнения $\mathfrak{H}_1 = A^{0T} G A^0 \ominus \mathfrak{H}_0$ к ядру индуцированной матрицы C в $A^{0T} G A^0 \mathfrak{H} = \mathfrak{H}_0 \oplus \mathfrak{H}_1$), и y_0-представление по коммутирующим между собой и с $y_1 = E_0 y$ операторам $y_0 = E_0 y$ (E_0 — проектор в $\mathfrak{H}_0$), нетрудно найти, что в операторе (2.9) будет присутствовать ещё и третий неотрицательный член $(\lambda_0 - A^0 y_0)^T G \varrho (\lambda_0 - A^0 y_0)/2$, соответствующий идеальному прямому измерению коммутирующих наблюдаемых $A^0 y_0$ с результатами $\lambda_0 \in \mathcal{U}_0$. Отметим, что решение A^0 уравнения (3.6), понимаемого в вырожденных случаях в указанном обобщённом смысле, может быть получено решением более общего матричного уравнения, полученного в (1) минимизацией риска в классе гауссовских о. в. (2.8) при конечном $\theta > 0$ в виде предела $A^0 = \lim\limits_{\theta \to 0} A^0$.

3. Комплексное обобщение и общее решение уравнения оптимального линейного оценивания

1. В разделе 2 при доказательстве оптимальности когерентного наблюдения была продемонстрирована полезность комплексных представлений, обусловленная наличием у квантовых сигналов естественной комплексной структуры. С точки зрения релятивистской теории квантованных полей вещественные сигналы являются частным (т. н. истинно нейтральным) случаем комплексных квантовых сигналов. В общем случае[1] бозонный сигнал описывается комплексным вектором $y = \{y_\tau\}$, компоненты y_τ которого повторно коммутируют с сопряженными y_τ^*:

$$[y_\sigma, y_\tau^*] = C_{\sigma\tau}, \quad [y_\sigma, y_\tau] =]y_\sigma^*, y_\tau^*] = 0, \quad [y, C_{\sigma\tau}] = 0. \tag{3.1}$$

Поскольку $(y_\sigma^* \cdot y_\tau)^* = y_\tau^* \cdot y_\tau$, матрица C комплексных чисел $C_{\sigma\tau}$ эрмитова $C^+ = C$, и, вообще говоря, индефинитна. В частности, если матрица C антисимметрична: $C^T = -C$, то y есть комплексификация $y = \frac{1}{\sqrt{2}}(y_R + iy_I)$ вещественного бозонного сигнала y_R той же размерности $[y_R, y_R^T] = C$ (y_I — копия y_R : $[y_I, y_I^T] = C$; $[y_I, y_R^T] = 0$).

Пусть передаваемый сигнал представляется случайным вектором $\vartheta = \{\vartheta_\alpha\}$ с комплексными значениями и гауссовским распределением[2]

$$P(d\vartheta) = p(\vartheta)\, d\mu(\vartheta),$$

где

$$p(\vartheta) = \det S^{-1} \cdot \mathrm{e}^{-\vartheta^+ S^{-1}\vartheta}, \quad d_\nu(\lambda) = \Pi \frac{1}{\pi} d \operatorname{Re} \vartheta_\alpha\, d\mu I_m \vartheta_\alpha. \tag{3.2}$$

Требуется найти о. в. $\Pi(d\lambda)$ на векторном пространстве $\mathfrak{U}$ комплексных значений оценок $\lambda = \{\lambda_\alpha\}$, описывающую оптимальное по квадратическому критерию

$$C(\lambda, \vartheta) = (\lambda - \vartheta)^+ G\,(\lambda - \vartheta), \qquad G > 0 \tag{3.3}$$

наблюдение суперпозиции $y = D\,\vartheta + \xi$ линейно-комплексно модулированного полезного сигнала ϑ с комплексным гауссовским бозонным шумом ξ. Последнее означает, что наблюдаемый сигнал y описывается комплексными

[1] Соотношения коммутации (3.1) можно представить и самосопряженными образующими, произведя овеществление: $y_R = \frac{1}{\sqrt{2}}(y + y^*)$; $y_I = \frac{1}{\sqrt{2}}(y - y^*)$, однако размерность овеществленного сигнала удвоится.

[2] Более общие гауссовские сигналы с плотностями, содержащими помимо членов $\vartheta_\alpha^* \cdot \vartheta_\beta$ члены $\vartheta_\alpha \cdot \vartheta_\beta, \vartheta_\alpha^* \cdot \vartheta_\beta^*$ в комплексном представлении рассматривать нецелесообразно. Их следует представлять вещественными векторами удвоенной размерности.

образующими $\{y_\tau\}$, удовлетворяющими соотношениям (3.1) и при каждом ϑ имеет гауссовский о. п.:

$$\varrho(\vartheta) = \det|1 - e^{-QC}|\, e^{-(y-D\vartheta)^+ Q(y-D\vartheta)},\ Q = \|Q_{\sigma\tau}\| > 0. \tag{3.4}$$

Предложение 2. *Решение сформулированной комплексной квадратично-гауссовской задачи оптимального наблюдения полностью определяется некоторой комплексной матрицей* $A = \|A_{\alpha\tau}\|$, *и описывается когерентной о. в., которую удобно представить в виде предела* $\Pi^0 = \lim \Pi^\theta$ *при* $\Theta \to 0$ *несингулярной гауссовский* о. в. $\Pi^\theta(d\lambda) = \pi^\theta(\lambda)\, d_\mu(\lambda)$:

$$\pi^\theta(\lambda) = \det|ACA^+\left(1 - e^{-\frac{1}{\theta} GACA^+}\right)^{-1}|^{-1}\, e^{-\frac{1}{\theta}(\lambda - Ay)^+ G(\lambda - Ay)}, \tag{3.5}$$

$d\mu(\lambda) = \prod\limits_\alpha \dfrac{1}{\pi}\, d\,\mathrm{Re}\ \lambda_\alpha\, d\,Im\ \lambda_\alpha$. *Матрица* A, *минимизирующая среднеквадратический риск линейного оценивания, удовлетворяет уравнению*

$$AR + \frac{1}{2}\,\mathrm{sign}\,(ACA^+ G)\, AC = T, \tag{3.6}$$

где R, T — *корреляционные матрицы*

$$R = \left\| \frac{1}{2} \langle y_\sigma y_\tau^* + y_\tau^* y_\sigma \rangle \right\|, \quad T = \|\langle \vartheta_\alpha y_\tau^* \rangle\|,$$

которые выражаются через исходные корреляционные матрицы сигнала S и шума $N = \dfrac{1}{2} C\,\mathrm{cth}\,QC$ обычным образом: $R = DSD^+ + N$, $T = SD^+$.

Доказательство проводится точно так же, как и в вещественном случае с использованием для о. п. (3.4) представления в виде подходящего нормально упорядоченного выражения, и повторять его мы не будем. Отметим лишь, что переход от упорядочения (3.4) к некоторому нормальному упорядочению в математическом отношении заключается в известном переходе [14] от индефинитной метрики $\{\varkappa', \varkappa\} = \varkappa^+ C\, \varkappa'$ к положительной $|\varkappa', \varkappa| = \varkappa_+^+ C\, \varkappa'_+ - \varkappa_-^+ C\, \varkappa'_-$ разложением $\mathcal{H} = \mathcal{H}_+ \oplus \mathcal{H}_-$ комплексного пространства $\mathcal{H}$ (области определения операторнозначного функционала $y(\varkappa) = \varkappa^+ y$) на ортогональные относительно некоторого скалярного произведения $(\varkappa', \varkappa) = \varkappa^+ H^{-1} \varkappa'$ области положительности $\varkappa_+^+ C\, \varkappa_+ > 0$, $\varkappa_+ \in \mathcal{H}_+$ и отрицательности $\varkappa_-^+ C\, \varkappa_- < 0$, $\varkappa_- \in \mathcal{H}_-$. Единственным отличием от вещественного случая является лишь то, что операторы уничтожения $b_+ \equiv y(\varkappa_+)$ и рождения $b_-^* \equiv y(\varkappa_-)$ в разложении $y(\varkappa) = b_+ + b_-^*$ коммутируют и относятся к различным модам (в вещественном случае $b_+ = b_-$ для вещественных H и $\varkappa = \varkappa_+ + \varkappa_-$).

2. Для выделения явного решения уравнения* (3.6), удовлетворяющего условию невырожденности матрицы ACA^+ в $A\mathcal{H}$, введём обозначения ;

$$K = \|\langle y_\tau^* y_\sigma \rangle\| = R - \frac{1}{2} C, \quad L = \|\langle y_\sigma y_\tau^* \rangle\| = R + \frac{1}{2} C, \tag{3.7}$$

и произведем факторизацию $T = T_+ + T_-$ матрицы $T = \|\langle \vartheta_\alpha y_\tau^* \rangle\|$, удовлетворяющую следующим условиям: $T_\pm^+ G T_+ = 0$,

$$T_+(K^{-1} - L^{-1}) T_\pm^+ \geq 0, \quad T_-(L^{-1} - K^{-1}) T_\pm^+ \geq 0, \quad T_+(K^{-1} - L^{-1}) T_\pm^+ = 0. \tag{3.8}$$

Такое G-ортогональное разложение всегда существует, причём матрицы $T_\pm$ однозначно определяются формулами

$$T_\pm = \frac{1}{2}\left(I \pm \operatorname{sign}\left(T(K^{-1} - L^{-1}) T^+ G\right)\right) T, \tag{3.9}$$

если матрица $T(K^{-1} - L^{-1}) T^+$ является невырожденной на подпространстве $T\mathcal{H} \subseteq \mathcal{U}$ значений матрицы T. В случае $T\mathcal{H} \subset \mathcal{U}$ матричное неравенство (3.11), обеспечивающее в соответствии с (3.13) эту невырожденность, следует понимать как строгую положительность лишь на подпространстве $T\mathcal{H}$.

Предложение 3. *Решение уравнения* (3.6) *может быть представлено в виде*

$$A^0 = T_+ L^{-1} + T_- K^{-1}, \tag{3.10}$$

если только матрицы K, L *и факторизация* (3.9) *матрицы* T *удовлетворяют неравенству*

$$T_+(L^{-1} - L^{-1} K L^{-1}) T_\pm^+ + T^- + T_-(K^{-1} - K^{-1} L K^{-1}) T_\pm^+ > 0. \tag{3.11}$$

Доказательство. Действительно, допустим что среди матриц A, для которых матрица ACA^T невырождена, хотя бы на подпространстве значений $A\mathcal{H}$, существует такая, которая обращает уравнение (3.6) в тождество. Тогда ее можно единственным образом так факторизовать $A = A_+ + A_-$, чтобы выполнялись условия

$$A_\pm^+ G A_+ = 0, \; A_+ C A_\pm^+ \geq 0, \; A_- C A_\pm^+ \leq 0, \; A_+ C A_\pm^+ = 0: \tag{3.12}$$

$$A_\pm = \frac{1}{2}\left(I \pm \operatorname{sign}\left(ACA^+ G\right)\right) A.$$

Производя в уравнении (3.6) замену $R = (K + L)/2$, $C = L - K$, получим уравнение относительно A_+, A_- : $A_+ L + A_- K = T$. Решение последнего

уравнения может быть представлено в виде $A_+ = T_+L^{-1}$, $A_- = T_-K^{-1}$ с помощью факторизации $T = T_+ + T_-$, удовлетворяющей условиям $T_-^{\pm} G T_+ = 0$,

$$T_+(L^{-1} - L^{-1}KL^{-1})T_+^{\pm} \geq 0,\ T_-(K^{-1} - K^{-1}LK^{-1})\ T_-^{\pm} \geq 0,$$
$$T_+(K^{-1} - L^{-1})\ T_-^{\pm} = 0, \tag{3.12a}$$

получаемым подстановкой $A_+ = T_+L^{-1}$, $A_- = T_-K^{-1}$ в (3.12). Условия (3.12a) гарантируют выполнимость определяющих матрицы (3.9) условий (3.8), так как имеют место следующие матричные неравенства

$$K^{-1} - L^{-1} > L^{-1} - L^{-1}KL^{-1},\quad L^{-1} - K^{-1} > K^{-1} - K^{-1}LK^{-1},$$

являющиеся следствиями тождеств

$$L^{-1} - L^{-1}KL^{-1} = K^{-1} - L^{-1} - (K^{-1} - L^{-1})K(K^{-1} - L^{-1}),$$
$$K^{-1} - K^{-1}LK^{-1} = L^{-1} - K^{-1} - (L^{-1} - K^{-1})L(L^{-} - K^{-1}). \tag{3.13}$$

Это доказывает предложение 3, т. н. неравенства (3. 12a) в $\mathcal{H}$ эквивалентны неравенству (3.11) в $T\mathcal{H}$.

Решение (3.10) с учетом (3.7), (3.9) представим в виде

$$A^0 = [TR^{-1} - \operatorname{sign}(TR^{-1}E(1 - E^2)^{-1}T^+ G)TR^{-1}E](1 - E^2)^{-1}, \tag{3.14}$$

где $E = \frac{1}{2}CR^{-1}$. Отметим, что найденное оптимальное линейное преобразование A^0 из $\mathcal{H}$ в $\mathcal{U}$ ранее было определено [1] лишь для простых случаев, допускающих нормальное комплексное представление: (при $C > 0$, очевидно $A^0 = TL^{-1}$, при $C < 0$ $A^0 = TK^{-1}$), а также в одномодовом случае

$$C = \hbar \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

$$A^0 = \frac{1}{1 - \varepsilon^2} T \cdot (R^{-1} - \varepsilon \cdot \delta \cdot (T^+ GT)^{-1}),\quad \text{где}\quad \varepsilon = \frac{1}{2}\sqrt{\det |CR^{-1}|},$$

$$\delta = \sqrt{\det (T^+ GTR^{-1})}.$$

3. Приведем другой прямой способ решения квадратично-гауссовской задачи комплексного измерения, основанный на использовании естественного упорядочения, при котором операторы y^* действуют слева от y. В указанном представлении о. п. (3.4) описывается некоторой гауссовской плотностью $p_\xi(\nu)$, нормированной на единицу $\int p_\xi(\nu)\, d\mu(\nu) = 1$: $\varrho(\vartheta) = \det |C|\ : p_\xi(y - D\vartheta)$: и оператор апостериорного риска, вычисляемый интегрированием $\int C(\lambda, \vartheta)\ : p_\xi(y - D\vartheta) : p(\vartheta)$ и квадратичной функции штрафов (3.3) с гауссовским априор-

ным распределением (3.2), описывается квадратично-гауссовским выражением

$$\mathcal{R}(\lambda) = \det|C|\,[\mathcal{R} : p_y(y) : + : (\lambda - TL^{-1}y)^+ G(\lambda - TL^{-1}y)\, p_y(y) :] =$$
$$= \mathcal{R}\,\varrho + (\lambda - TL^{-1}y)^+ G\varrho(\lambda - TL^{-1}y), \tag{3.15}$$

где $\mathcal{R} = \operatorname{Sp} G(S - TL^{-1}T^+)$, а $p_y(v) = \det L^{-1} \exp\{-v^+L^{-1}v\}$ — усреднённая с распределением (3.2) плотность $p_\xi(v - D\vartheta)$, определяющая гауссовский о. п. $\varrho = \int \varrho(\vartheta)\,P(d\vartheta)$:

$$\varrho = \det|CL^{-1}| : \mathrm{e}^{-y^+L^{-1}y} : = \det|CL^{-1}|\,\mathrm{e}^{-y^+Hy}. \tag{3.16}$$

$(H = -C^{-1}\ln(I - CL^{-1}))$. Если операторы $u = TL^{-1}y$ имеют неотрицательно определённую матрицу коммутаторов: $[u, u^+] = T(L^{-1} - L^{-1}KL^{-1})T^+ \geq 0$, то упорядоченность этих операторов в (3.15) является нормальной, и оптимальное наблюдение сводится к их G-когерентному измерению, описываемому в комплексной ситуации операторами (3.5) при $A = TL^{-1}$ и $\theta \to 0$.

В общем случае для перехода к нормальному упорядочению оператор (3.15) следует умножить слева и справа на $\varrho^{-\frac{1}{2}}$ и учесть формулу $\varrho^{\frac{1}{2}} y \varrho^{-\frac{1}{2}} = \sqrt{LK^{-1}}\,y$. Преобразованный оператор $\varrho^{-\frac{1}{2}}\mathcal{R}(\lambda)\,\varrho^{-\frac{1}{2}}$ представляется квадратичной формой относительно операторов $v = TM^{-1}y$, где $M^{-1} = L^{-1}\sqrt{LK^{-1}}$, нормально упорядочить которую не составляет труда:[3]

$$\varrho^{-\frac{1}{2}}\mathcal{R}(\lambda)\,\varrho^{-\frac{1}{2}} = \mathcal{R}^0 + (\lambda_+ - T^+M^{-1}y)^+ G(\lambda_+ - T_+M^{-1}y) +$$
$$+ (\lambda_- - T_-M^{-1}y)^T G^*(\lambda_- - T_-M^{-1}y)^*. \tag{3.17}$$

Здесь $T_\pm$ — матрицы (3.9), осуществляющие ортогональное $T_-^+GT_+ = 0$ разложение матрицы $[v, v^+] = T(K^{-1} - L^{-1})T^+$ на положительную и отрицательную части, $\lambda_\pm = \big(1 \pm \operatorname{sign}(T(K^{-1} - L^{-1})T^+G)\big)\lambda/2$, а

$$\mathcal{R}^0 = \operatorname{Sp} G(S - T_+L^{-1}T_+^+ - T_-K^{-1}T_-^+) \tag{3.18}$$

совпадает с риском $R^0 = \operatorname{Sp} G(S - A^0T)$, соответствующим решению (3.10). Произведя обратное преобразование, получим:

$$\mathcal{R}(\lambda) = R^0\,\varrho + (\lambda_+ - T_+L^{-1}y)^+ G\varrho(\lambda_+ - T_+L^{-1}y) +$$
$$+ (\lambda_- - T_-K^{-1}y)^T G^*\,\varrho(\lambda_- - T_-K^{-1}y)^*.$$

[3] Представление (3.17), вообще говоря, имеет место лишь при $\lambda \in T\mathcal{H}$, и единственно, если матрица $T(K^{-1} - L^{-1})T^+$ невырождена на подпространстве $T\mathcal{H} \subseteq \mathcal{U}$. В случае $T\mathcal{U} \subset \mathcal{H}$ всегда достаточно ограничиться рассмотрением оценок λ из $T\mathcal{H}$.

При этом нормальная упорядоченность сохранится лишь при условии, что матрицы T, K, L удовлетворяют неравенству (3.16), означающему неотрицательность матриц коммутаторов $[u_+, u_+^*]$, $[u_-^*, u_-^T]$ операторов $u_+ = T_+ L^{-1} y$, $u_- = T_- K^{-1} y$. Следовательно, решение задачи оптимального комплексного измерения при условии (3.16) действительно описывается когерентными проекторами $\pi^0(\lambda)$, удовлетворяющими уравнениям $u_+ \pi^0(\lambda) = \lambda_+ \pi^0(\lambda)$, $u_-^* \pi^0(\lambda) = \lambda_-^* \pi^0(\lambda)$, и представимыми в виде (3.5) при $A = T_+ L^{-1} + T_- K^{-1}$ и $\theta = 0$.

4. В заключение отметим, что если размерность пространства сигналов $\mathcal{H}$ не превышает размерности $\mathcal{U}$, решение комплексной квадратично-гауссовской задачи оптимального наблюдения можно представить с помощью проекторов $\varrho^0(\lambda)$, удовлетворяющих уравнениям $(V_+ - \lambda_+)\, \varrho^0(\lambda) = 0$, $(V_- - \lambda_-)^* \varrho^0(\lambda) = 0$, $V_\pm = T_\pm M^{-1} y$, по формуле

$$\Pi^0(d\lambda) = \varrho^{-\frac{1}{2}}\, \varrho^0(\lambda)\, \varrho^{-\frac{1}{2}} P(d\lambda), \tag{3.19}$$

являющейся квантовым аналогом формулы Бейеса. Здесь $P(d\lambda)$ — вероятностная мера, задающая «глауберовское» разложение

$$\varrho = \int \varrho^0(\lambda)\, P(d\lambda) \tag{3.20}$$

по когерентным проекторам $\varrho^0(\lambda)$. В таком виде решение задачи оптимального оценивания для вещественного квадратично-гауссовского случая и при равных размерностях пространств $\mathcal{U}$ и $\mathcal{H}$ было указано Б. А. Гришаниным [9].

Когерентные проекторы $\varrho^0(\lambda_+ + \lambda_-^*)$ являются собственными для оператора (3.17), отвечающими минимальному собственному значению $\mathfrak{R}^0$ (т. е. «вакуумными» для квадратичного оператора «энергии» $\eta_+^+ G \eta_+ + \eta_-^T G \eta_-^*$, где $\eta_\pm = T_\pm M^{-1} y - \lambda_\pm$), и описываются какой-либо корреляционной матрицей операторов $\eta = TM^{-1} y - \lambda$, например симметризационной $R_\eta = \langle \eta \eta^+ \rangle_+$

$$R_\eta = \frac{1}{2}\left(T_+ (K^{-1} - L^{-1})\, T_+^+ + T_- (K^{-1} - L^{-1})\, T^+\right) =$$

$$= \frac{1}{2}\, T\, |T(K^{-1} - L^{-1})\, T^+ GT|\, (T^+ GT)^{-1}\, T^+, \tag{3.21}$$

лежащей на границе обобщенного [3] неравенства Гейзенберга

$$R_\eta \geq \pm\, C_\eta / 2 \quad \left(C_\eta = [\eta, \eta^+] = T(K^{-1} - L^{-1})\, T^+\right).$$

Распределение $P(d\lambda)$ сосредоточено на подпространстве векторов $\lambda = \lambda_+ + \lambda_-^*$, совпадающем с подпространством значений матрицы T, на

котором описывается положительной гауссовской плотностью $p(\lambda)$, $\lambda \in T\mathcal{H}$. Условие положительности $p(\lambda) > 0$, сводящееся в гауссовском случае к условию положительной определенности на $T\mathcal{H}$ корреляционной матрицы $R_\lambda = \langle \lambda\lambda^+ \rangle$, определяет область оптимальности о. в. (3.19). Учитывая, что сумма $R_\eta + R_\lambda$ в соответствии с (3.20) равна определяющей гауссовский о. п. (3.16) корреляционной матрице $R_\nu = \langle \nu\nu^+ \rangle_+ = T(K^{-1} + L^{-1})T^+/2$ операторов $\nu = TM^{-1}y = \eta + \lambda$, запишем условие положительности матрицы $R_\lambda = R_\nu - R_\eta$ на $T\mathcal{H}$ в следующем виде: $R_\nu > R_\eta$. Последнее с учетом (3.21) можно записать в виде матричного неравенства

$$(K^{-1} + L^{-1})T^+GT^+ > |(K^{-1} - L^{-1})T^+GT|, \tag{3.22}$$

эквивалентность которого неравенству (3.11) при рассматриваемом соотношении размерностей $\dim \mathcal{H} > \dim \mathcal{U}$ нетрудно доказать использованием тождеств (3.13). В одномодовом случае неравенство (3.22) сводится к числовому неравенству

$$\varepsilon > t - \sqrt{t^2 - 1}, \quad \text{где} \quad t = \operatorname{Sp} T^+GTR^{-1}/2\sqrt{\det T^+GTR^{-1}},$$

полученному в [1]. В общем случае неравенство (3.11) в отличие от (3.22) дает необходимое и достаточное условие невырожденности оптимального решения (3.10) и минимальности соответствующего риска (3.18) также и при более естественном обратном соотношении размерностей $\dim \mathcal{H} > \dim \mathcal{U}$.

Заключение

Итак, оптимальные наблюдения в гауссовских линейных бозонных каналах, дающие минимальную среднеквадратическую ошибку оценивания гауссовских входных величин, являются квазиклассическими (т. е. когерентными). Однако в отличие от классического случая они являются косвенными, существенно зависят от метрической матрицы G, и в некоторых случаях для индефинитных коммутационных соотношений могут вырождаться. Примеры таких вырождений, а также другие физические примеры подробно рассмотрены в [1]. В настоящей статье доказана оптимальность G-когерентного измерения оптимального линейного преобразования A^0y, найдено условие его невырожденности и явный его вид (предложение 3) применительно к общему (многомерному, комплексному, индефинитному) случаю.

Таким образом, проблему статистического наблюдения в квантовых гауссовых каналах, частные аспекты которой были исследованы в [1—5, 9], можно считать решенной.

Литература

1. *Belavkin, V. P.*, Optimal linear random filtration of quantum boson signals. Problems of Control and Information Theory, 1974, 1.
2. *Helstrom, C. W., Liu, J. W., Gordon, J. P.*, Quantum-mechanical communication theory. Proc. of IEEE, **10** (1970) 186—207.
3. *Белавкин В. П., Гришанин Б. А.* Исследование задачи оптимального оценивания в квантовых каналах методом обобщенного неравенства Гейзенберга. Проблемы передачи информации, *9, 3* (1973) 47—52.
4. *Stratonovich, R. L.*, The quantum generalisation of optimal statistical estimation and hypothesis testing. Stochastics, **1** (1973) 83—126.
5. *Холево А. С.* Оптимальные квантовые измерения. III Международный симпозиум по теории информации. Таллин, 1973, 1.
6. *Белавкин В. П.*, Стратонович Р. Л. Об оптимизации обработки квантовых сигналов по информационному критерию. Радиотехника и электроника, *18, 9* (1973) 1839—1844.
7. *Фон Нейман Дж.* Математические основы квантовой механики. М., «Наука», 1967. *Von Neumann, J.*, Mathematical Foundation of Quantum Mechanics. Princeton Univ. Press, Princeton, NJ. 1955.
8. *Davis, E. B., Lewis, J. T.*, An operational approach to quantum probability. Comm. Math. Phys., **17** (1970) 239—260.
9. *Гришанин Б. А.* Некоторые методы решения квантовых задач обнаружения и измерения. Изв. АН СССР, Техническая кибернетика, *5* (1973) 127—137.
10. *Yuen, H., Lax, M.*, Multiple parameter quantum estimation and measurement of non-self-adjoint operators. IEEE International Symposium on Information Theory, Asilomar, CA, Jan. 31, 1972.
11. *Белавкин В. П., Ванцян А. Г.* О достаточных условиях оптимальности обработки квантовых сигналов. Радиотехника и электроника, *19, 7* (1974) 1391—1395.
12. *Белавкин В. П.* Линейное оценивание некоммутирующих квантовых переменных при косвенном их измерении. Радиотехника и электроника, *17, 12* (1972) 2534—2540.
13. *Клаудер Дж., Сударшан Э.* Основы квантовой оптики. «Мир», М., 1970.
14. *Надь К.* Пространства состояний с индефинитной метрикой. М., «Мир», 1969.

Optimal observation of boson signals in quantum Gaussian channels

V.P. BELAVKIN

(Moscow)

On the framework of classical theory of systems any system characteristics are on principle completely observable, and it can always be believed that the output signal is the result of state vector measurement of some accessible for observation subsystem called output system. It is well known that the classical theory gives only approximate, sometimes very rough description of so-called boson systems, and precise description of systems with any physical nature is possible only on the framework of quantum theory. From this point of view any universal observation like the state vector measurement of classical system does not exist, and it is necessary to choose among the quantum measurements in output system an optimal observation according to the aim, for which the information is used.

In this paper the general formulation of the optimization problem of static observation of quantum signals in output systems with the aim of parameters estimation of their statistical states is given. As an impotent simple model leading to the analytical solution it is considered Gaussian multiple boson channel, i.e. a linear stochastic system $y = Dx + \xi$ where D is an affine transformation of input vector x and ξ is the vector of boson Gaussian noise. It is considered both the real and complex boson channel (the latter with the indefinite commutation relations is the most general one). A notion of G-coherent measurement of second-commuting observables Ay is introduced, and optimality of such observation with the vector result $z = \lambda$ giving the estimate of Gaussian

vector parameter $x = \vartheta$ with the mean square criterion (3.2) is proved. The optimal affine transformation $A = A^0$ of output quantum signal y satisfy equation (3.6), which have been reduced for real situation in [3].

The method used for proof of this optimality is more universal and strict than the method of investigation [3] of this problem (formulated in [3] for real situation), and is based on the introduced notion of H-normal representation of boson signals.

The general solution (3.10), (3.9) of Eq. (3.6) is found when the inequality (3.11) is valid. For the simplest meal cases: single mode case and the case allowing a normal complex representation the solution of Eq. (3.6) (reduced by the optimization in the class of linear estimates) have been obtained earlier in [1], and optimal coherent (i.e. G-coherent) measurement by minimizing the mean square error in the class of all coherent measurement have been found in [6].

В. П. Белавкин
Московский Государственный Университет им. М. В. Ломоносова, физический факультет
СССР, Москва В—234,
Ленинские горы

Problems of Control and Information Theory, Vol. 4 (3), pp. 259—274 (1975)

SOME EXPERIENCES WITH THE PKF DISCRETE-TIME IDENTIFICATION METHOD

CS. BÁNYÁSZ, L. KEVICZKY

(Budapest)

(Received October 15, 1974)

Simultaneously with the widespread use of computer technique modern, effective, discrete-time identification methods have been developed. There are nowadays several methods of parametric identification of linear dynamic systems for model building. Most of these require an iterative technique for the minimization of the loss function. Perhaps only the priori knowledge fitting (PKF) or tally principle (TP) method gives the estimation of process parameters in explicit form for a wide class of the noise models. In our paper the parameter estimation of SISO systems by PKF method is considered, then the relationships are extended for MIMO systems, too. An algorithm is given by means of which the TP method is extended to estimate also the parameters of the noise model. Besides the known relationships, we show the possibility of masking in the parameter estimation and its importance. Several simulation examples are presented from which conclusions can be drawn regarding the accuracy and the necessary computation time of the TP method. Additional examples illustrate the help of masking in the right choice of the structure.

1. Introduction

Nowadays simultaneously with the widespread use of computer technique in most cases also the identification methods apply discrete-time models and the corresponding parameter estimation principles. The model structures forming the basis of modern methods and the maximum likelihood (ML) estimation technique were elaborated by Åström and Bohlin [1] and since then several methods have been developed for different noise models by various authors [2].

The so-called Åström model, by taking also the noise parameters into account, makes nonlinear parameter estimation possible likely to be solved by different iterative minimization methods. Though — as experience has shown — this model gives the most accurate results, in many cases — because of the big computation time — other methods are desirable.

The "priori knowledge fitting" (PKF) or "tally principle" (TP) method elaborated by Peterka and his coworkers, enables the process parameters to be estimated in explicit form but it does not give the parameters of the noise model [3]. As it is known, this method avails itself of the apriori knowledge of the assumption made for the independence of the input signal and output noise. In this paper — after the discussion of the single output — single output

PKF method — its generalization for the multiple input-multiple output case will be considered and in both cases shown how the parameters of the noise model can be estimated.

Then some experiences will be given in connection with the application of the PKF methods.

2. PKF method for single input-single output systems

Let a single input-single output (SISO) system be as in Fig. 1, i.e. its system equation be

$$y(t) = \frac{B(z^{-1})}{A(z^{-1})} u(t) + w(t) = v(t) + w(t) . \tag{1}$$

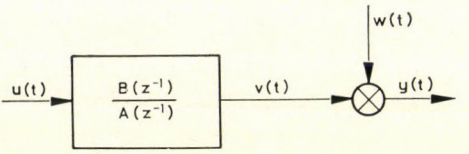


Fig. 1

Here $\{y(t), u(t), w(t); t = 1, 2, \ldots, N\}$ are the observed output signal, the applied input signal, the output disturbance sequence, respectively. z^{-1} is the backward shift operator and $A(z^{-1})$, $B(z^{-1})$ are the polynomials of z^{-1}:

$$\left.\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_n z^{-n} = 1 + \tilde{A}(z^{-1}) \\ B(z^{-1}) &= b_0 + b_1 z^{-1} + b_2 z^{-2} + \ldots + b_n z^{-n} \end{aligned}\right\} . \tag{2}$$

It is assumed that $u(t)$ is persistently exciting [2]. $w(t)$ has zero mean, it is an ergodic random signal independent of the input signal. Equation (1) can also be written in the form of the following difference equation:

$$y(t) = B(z^{-1})\, u(t) - \tilde{A}(z^{-1})\, y(t) + A(z^{-1})\, w(t) =$$

$$= \sum_{i=0}^{n} b_i\, u(t - i) - \sum_{i=1}^{n} a_i\, y(t - i) + \varepsilon(t) \tag{3}$$

where

$$\varepsilon(t) = \sum_{i=1}^{n} a_i\, w(t - i) + w(t) \tag{4}$$

is the so-called equation error. Since the output noise and the input signal are independent, so it can be easily seen [3] that the equation error is zero mean and also independent of the input signal, i.e.

$$E\{\varepsilon(t) u(t - i)\} = 0 \tag{5}$$

for all i (here $E\{\ldots\}$ means the expectation).

In the PKF method the knowledge of independence of the equation error $\varepsilon(t)$ and the input signal $u(t)$ is used, i.e. the assumption

$$E\{\varepsilon(t) \mid u(t), u(t-1), \ldots, u(t-m)\} =$$

$$= E\{\varepsilon(t)\} + \sum_{i=0}^{m} d_i u(t-i) = E\{\varepsilon(t)\} = 0 \tag{6}$$

must be held for every m. This can be possible only if all d_i are zeros (TP) or in this estimation problem, if they have minimum absolute values. (It has to be noted that the assumption $d_i = 0$ cannot be held exactly not only because of an estimation from the finite size of the samples but also because the system equation (6) contains the linear independence of a given order of the equation error and input signal and it does not include higher order interactions.) The basic idea of the PKF method is to determine the estimation of parameters a_i and b_i in such a way that the loss function

$$V_1 = \sum_{i=0}^{m} |\hat{d}_i|^2 = \hat{\mathbf{d}}^T \hat{\mathbf{d}} \tag{7}$$

obtained from the linear decomposition by (6) of the equation error should be minimum. Here

$$\hat{\mathbf{d}} = [\hat{d}_0, \hat{d}_1, \hat{d}_2, \ldots, \hat{d}_m]^T \tag{8}$$

where T means the transposition. Considering that $\hat{d}_i$ are estimated, we can write on the basis of (6) that

$$E\{\varepsilon(t) \mid u(t), u(t-1), \ldots, u(t-m)\} = \hat{\varepsilon}(t) + r(t) \tag{9}$$

where $r(t)$ is the value of the uncorrelated residual and $\hat{\varepsilon}(t)$ is a least-squares (LS) estimation. Using (3) and (9) we can rewrite the system equation for the residuals:

$$r(t) = \sum_{i=1}^{m} d_i u(t-i) - y(t) - \sum_{i=1}^{n} a_i y(t-i) + \sum_{i=0}^{n} b_i u(t-i)\,. \tag{10}$$

Let us introduce the following notations for the N measurements of the input and output mentioned with Eq. (1):

$$\mathbf{p} = [b_0, b_1, \ldots, b_n, a_1, \ldots, a_n]^T \tag{11}$$

$$\mathbf{r} = [r(1), r(2), \ldots, r(N)]^T \tag{12}$$

$$\mathbf{y} = [y(1), y(2), \ldots, y(N)]^T \tag{13}$$

$$\mathbf{Z} = \begin{bmatrix} \mathbf{q}^T(1) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{q}^T(N) \end{bmatrix} \tag{14}$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{f}^T(1) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{f}^T(N) \end{bmatrix}. \tag{15}$$

Here

$$\mathbf{f}(t) = [u(t), u(t-1), \ldots, u(t-n), -y(t-1), \ldots, -y(t-n)]^T \tag{16}$$

and

$$\mathbf{q}(t) = [u(t), u(t-1), \ldots, u(t-m)]^T. \tag{17}$$

By these notations the vector equation of (10) for N measurements is

$$\mathbf{r} = \mathbf{Fp} + \mathbf{Zd} - \mathbf{y}. \tag{18}$$

According to the principle of least-squares now the loss function

$$V_2 = \mathbf{r}^T\mathbf{r} \tag{19}$$

is minimized by $\mathbf{d}$.

We obtain from the minimum condition that

$$\hat{\mathbf{d}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{Fp}) \tag{20}$$

i.e. this is the function of even $\mathbf{p}$. Minimizing the loss function V_1 defined in (7) by $\mathbf{p}$, we obtain the estimation of process parameters:

$$\hat{\mathbf{p}} = [\mathbf{F}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{F}]^{-1}\mathbf{F}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}\,. \tag{21}$$

Peterka proved that this estimation is unbiased and consistent. Equation (21) can be considered either as an instrumental variable method having instrumental matrix $\mathbf{F}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ or as an LS estimation having weighting matrix $\mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$.

The necessary condition of the inversion of the matrix in brackets is $N \geq 2n$ and $m \geq 2n$.

The advantage of the PKF method can be seen from Eq. (21), i.e. it provides an explicit solution for the parameter estimation. $\mathbf{Z}^T\mathbf{Z}$ and similar forms can be computed easily. Similarly it is also very simple to perform the estimation under restriction $p_i \equiv 0$ for any given parameter vector. This can be realized by a so-called masking vector whose elements are ones if the parameter estimate is required, and are zeros, if the estimate is not required.

Since the estimation is of the form $\hat{\mathbf{p}} = \mathbf{H}^{-1}\mathbf{h}$, so it is sufficient to make the ith column and row in $\mathbf{H}$, and ith element in $\mathbf{h}$ equal with zero, and to give value 1 to the element H_{ii}. This kind of masking can also be done in Eq. (20) relating to the estimation of $\mathbf{d}$, but — of course — it is not always necessary to estimate $\hat{\mathbf{d}}$ which otherwise can be obtained by replacing (21) into (20).

The disadvantages of the TP estimation by (21) are that in general the variances of the coefficients are greater than at the ML method [5] for the same N samples, and the parameters of the noise model cannot be estimated with its help. But it can be seen easily that assuming a given structure for the output noise $w(t)$, the noise parameters can be determined from the residuals by iterative technique. Let us suppose that the noise model is of the form C/A according to the Åström model, i.e.

$$w(t) = \frac{C(z^{-1})}{A(z^{-1})}\, e(t)\,. \tag{22}$$

This means, that the equation error or now the residual in the TP method is

$$r(t) = C(z^{-1})\, e(t = \widetilde{C}(z^{-1})\, e(t) + e(t)\,, \tag{23}$$

where

$$C(z^{-1}) = 1 + c_1 z^{-1} + \ldots + c_n z^{-n} = 1 + \widetilde{C}(z^{-1}). \tag{24}$$

On the basis of (18) the estimated residuals $r(t)$ can be determined by the estimated vectors $\hat{\mathbf{p}}$ and $\hat{\mathbf{d}}$. Then the parameters c_i can be estimated by a simple repeated LS method according to (23):

$$\hat{\mathbf{c}} = [\mathbf{G}^T(\hat{\mathbf{e}}_{j-1})\mathbf{G}(\hat{\mathbf{e}}_{j-1})]^{-1}\mathbf{G}^T(\hat{\mathbf{e}}_{j-1})\hat{\mathbf{r}}, \tag{25}$$

where

$$\hat{\mathbf{e}}_j = \hat{\mathbf{r}} - \mathbf{G}^T(\hat{\mathbf{e}}_{j-1})\,\hat{\mathbf{c}}_j, \tag{26}$$

$$\hat{} = [\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n]^T \tag{27}$$

and

$$\mathbf{G}(\hat{\mathbf{e}}_j) = \begin{bmatrix} \mathbf{g}_j^T(1) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{g}_j^T(N) \end{bmatrix}. \tag{28}$$

Here

$$\mathbf{g}_j(t) = [\hat{e}_j(t-1), \ldots, \hat{e}_j(t-n)]^T. \tag{29}$$

The iterative procedure is continued until a given accuracy.

3. Extension of the PKF method for multiple input-multiple output systems

The method formulated for SISO systems can be easily extended to multiple input-multiple output (MIMO) systems and the equations are very similar. First Valis presented the multivariable version of this method [4], though he applied subsystems because his purpose was to develop an on-line algorithm.

The MIMO system description applied is a vector difference equation

$$\mathbf{y}(t) = \sum_{i=0}^{n} \mathbf{B}_i \, \mathbf{u}(t-i) - \sum_{i=1}^{n} \mathbf{A}_j \, \mathbf{y}(t-i) + \boldsymbol{\epsilon}(t) \tag{30}$$

which can be considered as a direct generalization of (3). Here

$$\boldsymbol{\epsilon}(t) = \mathbf{w}(t) + \sum_{i=1}^{n} \mathbf{A}_i \, \mathbf{w}(t-i) \tag{31}$$

and the assumptions for the output noise $\mathbf{w}(t)$ are the same as for the SISO systems. Introducing notations

$$\mathbf{P} = [\mathbf{B}_0, \mathbf{B}_1, \ldots, \mathbf{B}_n, \mathbf{A}_1, \ldots, \mathbf{A}_n] \tag{32}$$

$$\mathbf{E} = [\boldsymbol{\epsilon}(1), \ldots, \boldsymbol{\epsilon}(N)] \tag{33}$$

$$\mathbf{D} = [\mathbf{D}_0, \mathbf{D}_1, \ldots, \mathbf{D}_m] \tag{34}$$

$$\mathbf{Y} = [\mathbf{y}(1), \ldots, \mathbf{y}(N)] \tag{35}$$

$$\mathbf{F} = [\mathbf{f}(1), \ldots, \mathbf{f}(N)] \tag{36}$$

$$\mathbf{Z} = [\mathbf{q}(1), \ldots, \mathbf{q}(N)] \tag{37}$$

where

$$\mathbf{f}(t) = [\mathbf{u}^T(t), \ldots, \mathbf{u}^T(t-n), -\mathbf{y}^T(t-1), \ldots, -\mathbf{y}^T(t-n)]^T \tag{38}$$

and

$$\mathbf{q}(t) = [\mathbf{u}^T(t), \ldots, \mathbf{u}^T(t-s)]^T \tag{39}$$

then using the generalized form of (18)

$$\boldsymbol{\epsilon} = \mathbf{PF} + \mathbf{DZ} - \mathbf{Y} \tag{40}$$

$$V_1 = [\mathrm{vec}(\boldsymbol{\epsilon})]^T \, \mathrm{vec}(\boldsymbol{\epsilon}) \to \min_{\mathrm{vec}(\mathbf{D})} \tag{41}$$

is

$$\hat{\mathbf{D}} = (\mathbf{Y} - \mathbf{PF}) \, \mathbf{Z}^T (\mathbf{Z}\mathbf{Z}^T)^{-1}. \tag{42}$$

The parameter estimation $\hat{\mathbf{P}}$ can be determined from the condition

$$V_2 = [\mathrm{vec}(\hat{\mathbf{D}})]^T \mathrm{vec}(\hat{\mathbf{D}}) \to \min_{\mathrm{vec}(\mathbf{P})} \tag{43}$$

and we get

$$\mathbf{P} = \mathbf{YZ}(\mathbf{ZZ}^T)^{-1}(\mathbf{ZZ}^T)^{-1}\mathbf{ZF}^T[\mathbf{FZ}^T(\mathbf{ZZ}^T)^{-1}(\mathbf{ZZ}^T)^{-1}\mathbf{ZF}^T]^{-1}. \tag{44}$$

Naturally, similarly to SISO systems it is also now possible to introduce a matrix of the same size as $\mathbf{P}$ to mask the parameters whose estimations are not required. Moreover, the masking of the finding of the structure has a much greater importance for MIMO systems. It is simple to achieve by this technique the selection of the structure and by its repeated use the best model can easily be found. It is of special importance to mask some columns of the parameter matrices because this is a direct possibility to determine the unambiguous vector difference equation.

In the following part our experiences relating to the application of PKF method are presented by simulation examples.

4. Simulation results

Let us consider a simple second order system:

$$y(t) = \frac{b_0 + 1.0\,z^{-1} + 0.5\,z^{-2}}{1 - 1.5\,z^{-1} + 0.7\,z^{-2}}\,u(t) + \\ + \lambda\,\frac{1 - 1.0\,z^{-1} + 0.5\,z^{-2}}{1 - 1.5\,z^{-1} + 0.7\,z^{-2}}\,e(t)\,. \tag{45}$$

The minimum value of the loss function V_2 obtained by TP estimation from $N = 100$ samples in case of $b_0 = 1.0$ and $\lambda = 0.4$ is illustrated in a function of memory m of the instrumental model on Fig. 2. In Fig. 3 the dependence of the diagonal elements of the inverse matrix in (21) on m is shown (these are proportional to the variances of the parameter estimates). From these two figures we can establish that the reasonable value of m is $3n$ or $4n$. For higher values the computation time is increasing considerably, while the accuracy does not improve significantly. In Fig. 4 the computation time in function m parametrized by the value of N is shown. We can conclude for the noise sensitivity of the TP method from the results of Table 1. For the simulation investigations PRB input signal of ± 1 was applied, $e(t)$, however, was an uncorrelated noise series with normal distribution and unit variance. In Table 2 we show that even in the case of $\lambda = 10$ we get an appreciable result for $N = 400$ and $R = 6$.

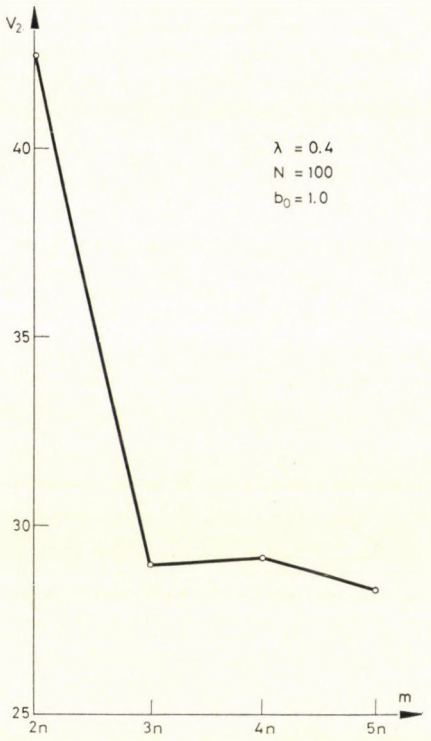


Fig. 2

Otherwise, by increasing N the accuracy of the extimates always improves. The changing of loss function V_1 in function of N is shown in Fig. 5.

In Table 3 it is shown how the right masking of only one parameter enables us to find the exact structure. In one case also the b_0 coefficient is estimated, in the other case its estimation is forbidden by masking.

Now let us consider a MIMO system of vector difference equation

$$\mathbf{A}(z^{-1})\,\mathbf{y}(t) = \mathbf{B}(z^{-1})\,\mathbf{u}(t) + \mathbf{C}(z^{-1})\,\mathbf{e}(t)\,, \tag{46}$$

where the covariance matrix of $\mathbf{e}(t)$ is Λ.

Let us choose Valis' example for simulation;

$$\mathbf{A}(z^{-1}) = \mathbf{I} + \mathbf{A}_1 z^{-1} + \mathbf{A}_2 z^{-2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} -0.829 & 0 \\ 0 & -0.606 \end{bmatrix} z^{-1} + \\ + \begin{bmatrix} 0.135 & 0 \\ 0 & 0 \end{bmatrix} z^{-2}.$$

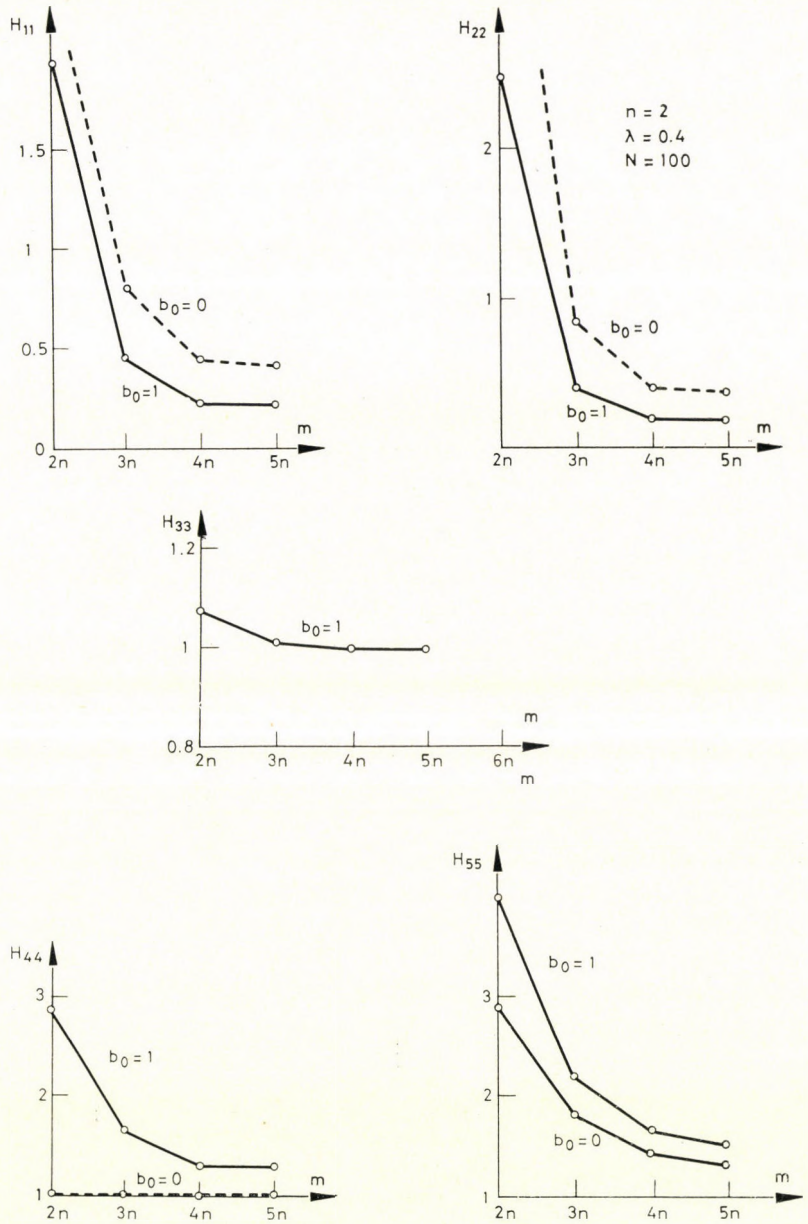

H11
1.5
1
0.5
0
b0 = 0
b0 = 1
m
2n
3n
4n
5n
H22
2
1
n = 2
λ = 0.4
N = 100
b0 = 0
b0 = 1
m
2n
3n
4n
5n
H33
1.2
1
0.8
b0 = 1
m
2n
3n
4n
5n
6n
m
H44
3
2
1
b0 = 1
b0 = 0
m
2n
3n
4n
5n
H55
3
2
1
b0 = 1
b0 = 0
m
2n
3n
4n
5n

Fig. 3

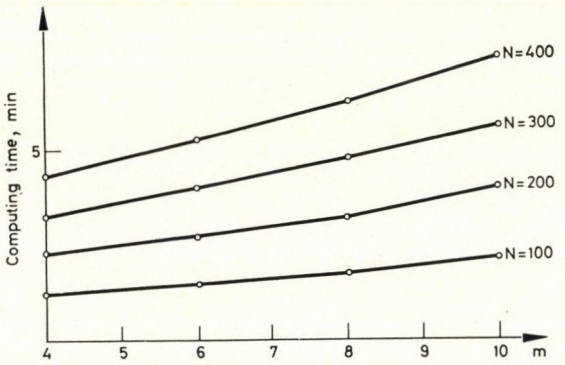


Fig. 4

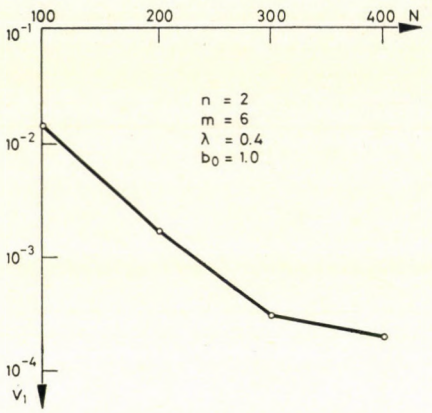


Fig. 5

$$\mathbf{B}(z^{-1}) = \mathbf{B}_0 + \mathbf{B}_1 z^{-1} + \mathbf{B}_2 z^{-2} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} +$$

$$+ \begin{bmatrix} 0.393 & 0.518 \\ 0.393 & 0.393 \end{bmatrix} z^{-1} + \begin{bmatrix} -0.088 & -0.314 \\ 0 & 0 \end{bmatrix} z^{-2}$$

and

$$\mathbf{C}(z^{-1}) = \mathbf{I} + \mathbf{C}_1 z^{-1} + \mathbf{C}_2^{-2} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} z^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} z^{-2},$$

where **I** denotes the unit matrix. This latter expression means a decoupled noise model at the output.

Table 1

$n = 2$ $N = 100$		$m = 3n = 6$		$m = 4n = 8$	
True parameters		Estimated parameters		Estimated parameters	
		$\lambda = 0.4$	$\lambda = 1.0$	$\lambda = 0.4$	$\lambda = 1.0$
a_1	−1.5	−1.5247	−1.5476	−1.5324	−1.5760
a_2	0.7	0.7174	0.7304	0.7251	0.7587
b_0	1.0	1.0003	1.0027	0.9971	0.9908
b_1	1.0	0.9241	0.8263	0.9125	0.7848
b_2	0.5	0.4182	0.3195	0.4067	0.2793
a_1	−1.5	−1.5348	−1.5594	−1.5448	−1.5997
a_2	0.7	0.7274	0.7406	0.7387	0.7867
b_0	0.0	0.0003	0.0039	−0.0030	−0.0100
b_1	1.0	0.9488	0.8752	0.9448	0.8605
b_2	0.5	0.4281	0.3502	0.4184	0.3136

Table 2

$n = 2$ $N = 400$ $m = 3n = 6$ $\lambda = 10.0$

True parameters		Estimated parameters	True parameters		Estimated parameters
a_1	−1.5	−0.9690	a_1	−1.5	−0.8831
a_2	0.7	0.2211	a_2	0.7	0.0906
b_0	1.0	1.0848	b_0	0.0	0.1038
b_1	1.0	1.1414	b_1	1.0	0.6298
b_2	0.5	1.5530	b_2	0.5	1.2798

Table 3

$n = 3$ $N = 200$ $m = 4n = 12$ $\lambda = 0.4$

True parameters		Estimated parameters	
		Without masking	With masking
a_1	−1.5	−1.2052	−1.3975
a_2	0.7	0.2634	0.5550
a_3	0.0	0.2011	0.0639
b_4	0.0	0.0456	0.0000
b_1	1.0	0.9656	0.9605
b_2	0.5	0.7556	0.5723
b_3	0.0	0.1892	0.0998

Figure 6 shows the computation time in function of m parametrizing by N.

In Fig. 7 the dependence of loss function V_2 on m is seen in noiseless case ($\Lambda = \mathbf{0}$) and $N = 100$.

Figure 7a illustrates the cases without masking and with the masking of submatrix $\mathbf{B}_0$. Figure 7b shows the case when each column containing only

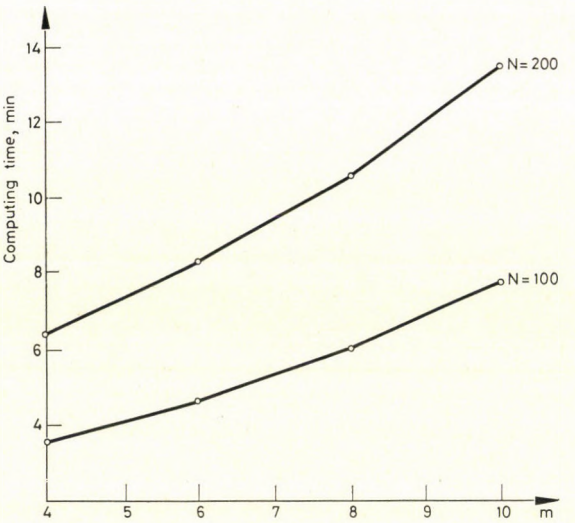


Fig. 6

zeros of parameter matrix $\mathbf{P}$ is masked. The improvement in the loss function is very striking. It should be noted that if every zero element of matrix $\mathbf{P}$ was masked, then we got worse results than if only the completely zero columns were masked. This symptom was also valid for simulation in case of $\Lambda \neq \mathbf{0}$.

From Table 4 this fact can be recognized via the estimated parameters. Table 5 shows the same thing for $\Lambda = 0.01\,\mathbf{I}$.

All of our simulation examples were run in ALGOL language by computer ODRA 1204.

5. Conclusions

The PKF method presented here for SISO and MIMO systems estimates the parameters of a discrete-time linear model of a process in explicit form. This solution — compared to the ML estimation — requires little computation

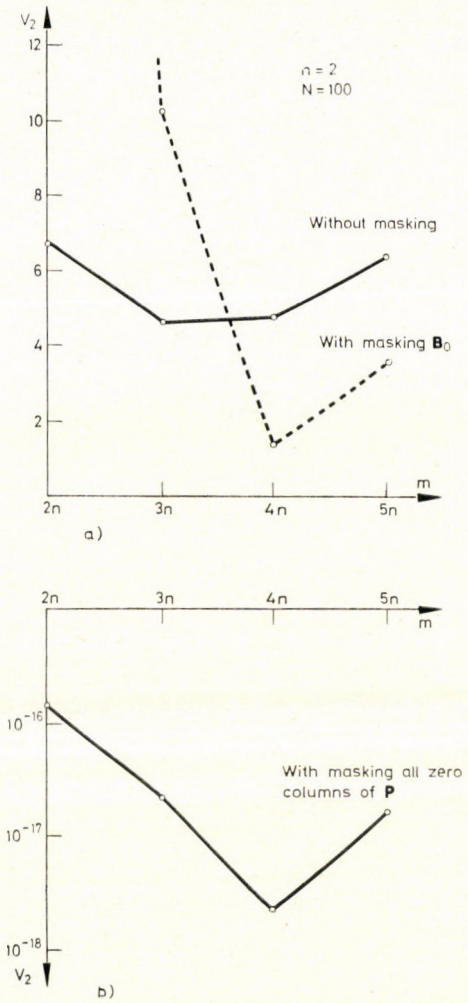


Fig. 7

time, its program can be easily prepared and enables the questions of the structure to be investigated efficiently.

Our simulation experiences are favourable, the accuracy of the obtained estimates is good enough for the purpose of primary investigations. We suggest to use these methods for the primary estimation of the structure and then, after having found the suitable structure, to continue the investigations by a more accurate estimation, e.g. ML method which, however, requires much more computation time.

Table 4

$n = 2$ $N = 100$ $m = 4n = 8$ $\Lambda = \mathbf{0}$

		$\mathbf{A}_1$		$\mathbf{A}_2$		$\mathbf{B}_0$		$\mathbf{B}_1$		$\mathbf{B}_2$	
True parameter matrices		−0.829	0	0.135	0	0	0	0.393	0.518	−0.088	−0.314
		0	−0.606	0	0	0	0	0.393	0.393	0	0
Estimated parameter matrices	without masking	−0.829	−0.718	0.135	0.093	$3.86 \cdot 10^{-12}$	$-2.95 \cdot 10^{-12}$	0.393	0.518	−0.484	0.518
		$1.86 \cdot 10^{-9}$	0.375	$-1.39 \cdot 10^{-9}$	−0.625	$1.81 \cdot 10^{-11}$	$9.09 \cdot 10^{-12}$	0.393	0.393	0.312	0.500
	with masking $\mathbf{B}_0$	−0.829	−0.406	0.135	0.234	0	0	0.393	0.518	−0.312	−0.492
		$2.32 \cdot 10^{-10}$	−0.125	$-4.65 \cdot 10^{-10}$	−0.250	0	0	0.393	0.393	0.312	0.375
	with masking all zero columns of $\mathbf{P}$	−0.829	$-1.16 \cdot 10^{10}$	0.135	0	0	0	0.393	0.518	−0.088	−0.314
		$-4.36 \cdot 10^{-11}$	−0.606	$-3.49 \cdot 10^{-10}$	0	0	0	0.393	0.393	$2.91 \cdot 10^{-10}$	$2.32 \cdot 10^{-10}$

Table 5

$n = 2$ $N = 200$ $m = 4n = 8$ $\Lambda = 0.01\ \mathbf{I}$

		$\mathbf{A}_1$		$\mathbf{A}_2$		$\mathbf{B}_0$		$\mathbf{B}_1$		$\mathbf{B}_2$	
True parameter matrices		−0.829	0	0.135	0	0	0	0.393	0.518	−0.088	−0.314
		0	−0.608	0	0	0	0	0.393	0.393	0	0
Estimated parameter matrices	without masking	−0.854	−0.351	0.147	0.195	−0.004	0.003	0.394	0.529	−0.221	−0.450
		−0.010	−0.724	0.004	0.063	−0.003	−0.001	0.380	0.391	−0.058	−0.047
	with masking $\mathbf{B}_0$	−0.855	−0.303	0.147	0.167	0	0	0.394	0.529	−0.203	−0.432
		−0.011	−0.725	0.004	0.065	0	0	0.380	0.391	−0.058	−0.048
	with masking all zero columns of $\mathbf{P}$	−0.841	−0.043	0.147	0	0	0	0.394	0.529	−0.099	−0.322
		−0.005	−0.624	0.004	0	0	0	0.380	0.391	−0.018	−0.006

References

1. *Åström, K. J., Bohlin, T.*, Numerical Identification of Linear Dynamic Systems from Normal Operating Records. IFAC Symp., Teddington, 1965.
2. *Åström, K. J., Eykhoff, P.*, System Identification — A Survey. IFAC Symp., Prague, 1970.
3. *Peterka, V., Šmuk, K.*, On-line Estimation of Dynamic Model Parameters from Input-Output Data. IFAC Congr., Warsaw, 1969.
4. *Valis*, J., On-line Identification of Multivariable Linear Systems of Unknown Structure from Input-Output Data. IFAC Symp., Prague, 1970.
5. *Gustavsson, I.*, Comparison of Different Methods for Identification of Industrial Processes. Report 7107(B), Lund Institute of Technology, Division of Automatic Control, 1971.
6. *Neudecker, H.*, Some Theorems on Matrix Differentiation with Special Reference to Kronecker Matrix Products. American Stat. Ass. Journal, 1969, September.
7. *Tracy, D. S., Dwyer, P. S.*, Multivariate Maxima and Minima with Matrix Derivatives. American Stat. Ass. Journal, 1969, December.

Применение метода дискретной идентификации, использующего априорную информацию о структуре объекта

Ч. БАНЯС, Л. КЕВИЦКИ

(Будапешт)

Широкое распространение вычислиетльной техники стимулировало развитие современных эффективных методов дискретной идентификации. В настоящее время известно несколько методов параметрической идентификации линейных дискретных систем. В большинстве из них применяются принципы приближения для минимизации функции потерь. Видимо, единственными методом, при использовании которого может быть получена оценка параметров процесса для широкого класса моделей с помехами, является метод дискретной идентификации, использующий априорную информацию о структуре объекта.

В настоящей статье указанный метод идентификации применяется для оценки параметров объектов с одним входом и одним выходом, а затем результаты обобщаются на объекты с несколькими входами и выходами. Разработаны алгоритмы, с помощью которых выше приведенный метод может быть распространен и для оценки параметров модели с помехами. Показана возможность и значение исключения некоторых членов матрицы параметров в процессе идентификации.

Представлено несколько примеров, показывающих точность метода и потребность машинного времени. На других примерах иллюстрируется методика исключения некоторых членов матрицы параметров и представлены преимущества такого подхода при выборе структуры объекта.

Cs. Bányász
Computer and Automation Institute
Hungarian Academy of Sciences
H-1502 Budapest XI., Kende u. 13—17.

L. Keviczky
Department of Automation
Technical University
H-1111 Budapest XI., Goldmann Gy. tér 3.
Hungary

BOOK REVIEW

DAN TEODORESCU: Sisteme automate-analiza și sinteza prin matrice specifice (Automatic Control Systems. The Specific Matrix Approach). Editura Facla, 1974, 511 pp. (In Roumanian)

Nowadays many new books dealing with modern control theory are published. The present work excels in the new concepts introduced for treating nonlinear systems in a computer-oriented manner.

The author introduces the term of the nonlinearity vector **z**, the components of which are the values of the nonlinear function at positive, equal and consecutive sampling intervals. He defines an equivalent gain vector **n**, which gives the amplitudes of the fundamental of the output as a function of the amplitude of the sinusoidal input. The connection between these two vectors may be given by the following transformation:

$$\mathbf{H}\,\mathbf{z} = \mathbf{n} \quad \text{or} \quad \mathbf{H}^{-1}\,\mathbf{n} = \mathbf{z}$$

where **H** is a numerically known matrix independent of the nonlinearity and based on some relationships of the Bessel functions.

This general expression, unusual for nonlinear system analysis, gives the possibility to develop a new method of analysis and synthesis. Tables given in the Appendix promote the calculation. E.g. a procedure for designing correction or compensation nonlinearities in a nonlinear system in order to reach a predetermined desired transient and/or steady-state behaviour is given. This matrix-plane approach leads to general computer-oriented methods. The author deals from this point of view with noisy nonlinear systems and quasioptimal systems as well.

To extend this specific matrix concept to linear systems, the author introduces the step-response vector **a** whose components are the points of the sampled step-response function. A mapping may be given between this vector and the real part of the frequency vector **u**, containing the sampled points of the real-frequency function.

$$\mathbf{A}\,\mathbf{u} = \mathbf{a} \quad \text{or} \quad \mathbf{A}^{-1}\,\mathbf{a} = \mathbf{u},$$

here **A** is the mapping matrix.

This connection between the time and the frequency domain can be utilized in solving optimization problems. Among others the author shows a solution to the following task: given a plant with its step-response vector; how to choose the structure of a feedback controller which ensures a time optimal response taking also some constraints into consideration. Similar problems are solved also for nonlinear systems with absolute stability (or hyperstability) requirements.

Matrices $\mathbf{A}$, $\mathbf{A}^{-1}$, $\mathbf{H}$, $\mathbf{H}^{-1}$ necessary for the design procedure are given in the Appendix.

Besides the new concepts and methods the book deals also with the most significant chapters of modern control theory, as the state-variable method, the analysis and synthesis of linear and nonlinear systems, optimal control systems, stochastic systems, time-variable and multivariable systems, etc.

The book is characterized by a clear style, and has also a pedagogical nature — it leads the reader from the theoretical concepts to the computational methods. The theory is illustrated by numerous examples.

The book is useful for engineers and mathematicians dealing with control theory. It would be worthwile to translate it to English.

R. Bars

P. L. Reichertz, G. Holthoff: Methoden der Informatik in der Medizin (Automatic Data Processing in Medicine). Springer Verlag, Berlin–Heidelberg–New York, 1975, 234 pp. (In German)

Automatic data processing is becoming more and more important in the practical and theoretical fields of medicine, where its applications appear to be more and more successful. Important and unsolved problems in medicine and in organization of hospitals have largely contributed to the fact that in all fields of medicine new projects are being carried out. In the same time the methods of data processing and mathematical analysis lead to further knowledge in medical research.

The Working Group "Medizinische Informatik" of the German Medical Society for Medical Documentation and Statistics has organized conferences on this subject. This book contains the papers which were presented at the conference in Hannover, 1974. The aim of the conference and the papers was not only to present the basic and solved problems but to allow the participants and the readers to get essential information on the operation of data processing systems in action as well as new projects being established in the future. The publisher intends to meet the continuously increasing interest and demand on this subject.

The main subjects covered in the book are:

- relationship between communication and information transmission
- principles of project-managements in medicine
- problems of complex information and control systems in hospital
- operations research problems
- hardware and software components of the on-line data acqusition systems in laboratories
- special measurement problems
- telemetric data transmission
- possibilities and use of statistical quality control in laboratory data processing, etc.

Each paper has its own Author's Index. The papers are arranged in accordance with the half-day sessions of the conference and each group of them is introduced by a short summary.

T. Gál

D. Seitzer: Arbeitsspeicher für Digitalrechner (Computer Storage Devices). Springer Verlag, Berlin–Heidelberg–New York, 1975, 168 pp. (In German)

The rapid progress in the design and application of digital computers and computer systems are partly due to the progress in memories. New kinds of memories are appearing and are being developed. The progress has been so rapid — especially in the field of the semiconductor memories since the invention of the integrated circuit technology — that it is difficult to keep in touch with the state of the art.

The purpose of this book is to give a college-level treatment of important principles of the main computer storages. (It contains the lectures delivered for students of electronic engineering in the 7th semester.) The reader is assumed to be aware of the essential knowledge of semiconductor physics and technology.

The first chapter describes the role of the main storage in a computer system, summarizes the important parameters and makes a comparison between them.

The second and third chapters are devoted to memories using magnetic materials. The important magnetic properties of the ferrite cores and the thin film materials are presented. In addition to the well-known $2D$, $3D$ and $2\frac{1}{2}D$ systems (D means here dimension), some special kinds of ferrite memories (the biax, the two-core-per-bit system, the flute memory and the laminated ferrite memory) are also treated. Among the thin film memories the planar thin film stores, the cylindrical thin film (plated-wire) stores and the twistor stores are described in detail.

Chapter 4 deals with the connected peripheral circuitry of the memories, e.g. the selection, decoding and drive circuits, sense amplifiers, etc.

The largest, and most important part of the book, Chapter 5 gives a very good summary of the semiconductor memories containing the most modern types, too. It is divided into four essential parts treating the random access memories (RAM's) the sequence access memories, the read only memories (ROM's) and the associative memories. Each of them is very comprehensive. The first part describes besides the well-known types the most modern types of the static and dynamic memories, among others the injection coupled memories, the static RAM's with normally off-type Schottky-barrier FET's, etc. It is followed by the discussion of the sequence access memories, the static, dynamic shift registers and the up-to-date charge coupled devices. The presentation of the read only memories includes the electrically programmable ROM's, and also the reprogrammable ROM's. The chapter is ended with the principles of the associative memories.

The last chapter is devoted to two kinds of bulk memories. The most promising types are the bubble memory and the optical memory.

At the end of each chapter is an Author's Index. The Subject Index completes the book.

Though the book was written first of all for students it may also be very valuable for engineers in practice, who as specialists in their fields are not always be able to follow the neighbouring fields at the required intensity. The comprehensive summary of semiconductor memories may be of interest even for specialists engaged in the field.

T. Gál

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY, VOL. 4 (3)
ENGLISH TRANSLATION OF THE COMMUNICATION IN RUSSIAN

OPTIMAL OBSERVATION OF BOSON SIGNALS IN QUANTUM GAUSSIAN CHANNELS

V. P. BELAVKIN

(Moscow)

The problem of observation optimization in non-ideal channels subject to quantum statistics is formulated. The G-coherent measurement defining the optimal linear estimation [1] in the general (non-Gaussian) boson case is proved to be the optimal observation minimizing the Bayes quadratic risk of estimating the Gaussian input signal in linear Gaussian channels with Bose—Einstein statistics.

A complex generalization of this quadratic-Gaussian problem is formulated allowing indefinite commutation relations $[y, y^+] = C$. Through decomposition $T = T_+ + T_-$ definitizing optimally matrix C, the general form $A = A_+ + A_-$, $A_\pm = T_\pm \left(R + \frac{1}{2} C\right)^{-1}$ of optimal linear transform Ay of output observables y generalizing the classical $TR^{-1} = y$, is found.

Introduction

In this paper the problem of observation optimization (i.e. measurement, estimation and filtration) is formulated and solved for Gaussian signals in linear quantum channels. One may believe that these signals are electromagnetic fields on the aperture of the receiver — but here such specialization is not necessary — any boson field may be considered [1]. The results can be applied for solving the problem of optimization of optical coherent signals processing in laser communication, whose corresponding classical results without taking into account the quantum nature of electromagnetism are not precise [2]. Such quantum problems arise also in optimization observation and control with microscopic systems, and also macroscopic system of extreme high precision. Solution of the problems has a principal significance for general system theory.

It is proved that linear estimation and filtration [1] like the classical case are absolute optimal for Gaussian *a priori* distribution, Gaussian additive noise and quadratic cost function. A previous solution of this problem is presented in [3], where the optimality is proved supposing the uniqueness of solution of the corresponding equation (uniqueness is proved in Ref. [3] only in quasi-classical approximation). Unlike the method of Ref. [3], that presented here is more universal and gives an exhaustive solution of the quad-

ratic-Gaussian problem also in the sense of uniqueness. For the simplest one-mode case allowing normal complex representation, optimality of linear estimation and corresponding indirect measurement were established also in [4,5].

1. Optimal observation of quantum signals

Like [1], an observable quantum signal (i.e. the quantum field at the aperture of a receiving antenna) will be denoted by a family of symbols $y = \{y_\tau, \tau \in T\}$ forming over algebraic operations a non-commutative, with respect to multiplication, algebra of observables $\mathcal{Y}$. As it is accepted in the quantum theory (see, e.g. [7]), without imposing restrictions on generality, algebra $\mathcal{Y}$ will be identified with an algebra of operators acting in Hilbert space $\mathcal{H}$. The statistical state over the algebra of observables $\mathcal{Y}$ is a function of transmitted information $\vartheta \in \Theta$, this fact being reflected by the parametric dependence of the density operator (d. o.) $\varrho(\vartheta)$ from ϑ. Family $\{\varrho(\vartheta, \vartheta \in \Theta\}$ describing for each ϑ a state over the operator algebra $\mathcal{Y}$ plays the role of conditional probability densities $p\{y \mid \vartheta\}$. It is in order to remind here that any positive (i.e. non-negatively defined) operator $\varrho \geq 0$ having a unitary trace $\mathrm{Tr}\,\varrho = 1$ is d.o. of some state and defines mathematical expectations $\langle f \rangle$ of observables $f \in \mathcal{Y}$ through formula $\langle f \rangle = \mathrm{Tr}\, f\varrho$.

Arbitrary (including indirect) observations of quantum signals are described by operator-valued measures Π over the space of expected measurement results $\lambda \in \Lambda$. Operator probabilities (o.p.) Π (see [8]) are the quantum generalization of conditional probability distributions $P(\,\cdot \mid y)$ to Λ and are in all respects their analog ϱ: they are defined as positive denumerably-additive mappings $\Pi(\,\cdot\,)$ of σ-algebra of subsets of set Λ into operator algebra $\mathcal{Y}$ satisfying the normalization condition

$$\int_{\Lambda} \Pi(d\lambda) = 1, \tag{1.1}$$

where 1 is a unitary operator of algebra $\mathcal{Y}$. Like the classical case, probability $P(\Delta)$ of event $\lambda \in \Delta$ is mathematical expectation $\langle \Pi(\Delta) \rangle$. In case of o.p. satisfying the orthogonality condition $\Pi(\Delta)\,\Pi(\Delta') = 0$ for all non-intersecting $\Delta, \Delta' \in \Lambda$, it describes direct (rather than indirect) observation, i.e. the ideal measurement of observables $\hat{\lambda} = \int \lambda \Pi(d\lambda)$ of algebra $\mathcal{Y}$ having common ortho-projectors $\Pi(d\lambda)$. Otherwise, observation is indirect even if elementary events $\lambda \in d\lambda$ are still described by o.p. $\Pi(d\lambda) = \Pi(\lambda)\,d\lambda$ proportional to some projectors $\pi(\lambda)$ into non-intersecting (but non-orthogonal) sub-spaces $\mathfrak{H}_\lambda \subset \mathfrak{H}$. Indirect, with respect to quantum signal y, measurements are also referred to as *quasi-measurements*.

In the classical case, the received signal is, in principle, completely observable, and any information contained in it may be obtained as the result of arbitrarily accurate registration of the signal. In the quantum case, there is no such a universal complete observation. Therefore, solution of synthesis problems gives rise to a specifically quantum problem of *observation optimization*. In order to find optimal statistical solutions it is evidently sufficient to consider only those observations whose sets of expected results coincide with space $\mathcal{Y}$ of allowable estimates (each estimate being observation result!). In doing so, o.p.'s are also decision operators. This approach has proved successful in solving some Bayes problems of quantum signal processing optimization [4, 5, 9, 10].

The Bayes optimal observations are described by o.p.'s minimizing mean risk

$$\mathcal{R} = \int\int C(\lambda, \vartheta) \cdot P(d\lambda \mid \vartheta)\, P(d\vartheta), \quad P(d\lambda \mid \vartheta) = \operatorname{Tr} \Pi(d\lambda)\, \varrho(\vartheta), \tag{1.2}$$

where $C(\lambda, \vartheta)$ is cost function, and $P(d\vartheta)$ is a given *a priori* distribution of probability of information parameters (or hypotheses) ϑ. The condition of risk (1.2) minimality, generalizing the conditions formulated in [2] for hypothesis discrimination problems, is that of operator

$$\Lambda = \int \mathcal{R}(\lambda)\, \Pi^{\circ}(d\lambda) \tag{1.3}$$

being Hermitian, where $\mathcal{R}(\lambda) = \int C(\lambda, \vartheta)\, \varrho(\vartheta)\, P(d\vartheta)$; together with the condition of positiveness

$$\mathcal{R}(\lambda) - \Lambda \geq 0 \tag{1.4}$$

it is also a sufficient condition of optimality if o.p. Π^0 satisfies the orthogonality condition (optimization in the class of *indirect* observations in [2] was not considered).

In the general case, there may be no solution Π^0 of problems (1.3), (1.4) in the class of orthogonal o.p.'s. To isolate the optimal Π^0 in the class of *arbitrary* observations, condition (1.3) should be supplemented with condition

$$(\mathcal{R}(\lambda) - \Lambda) \cdot \Pi(d\lambda) = 0 \tag{1.5}$$

which is satisfied automatically only for orthogonal o.p.'s. The necessary optimality condition (1.5) was obtained in [4, 5, 10]. Together with (1.4), it means, in the classical case, that the optimal decision rule $\Pi(d\lambda \mid y)$ is non-randomized and concentrated only in those points $\lambda = \gamma(y)$ where the *a posteriori* risk $\mathcal{R}(\lambda)$ has minimal value for each y. In the quantum case, conditions (1.4) and (1.5) have a similar meaning, however, they do not imply that under different λ optimal o.p.'s are orthogonal, and corresponding to them observations are non-indirect. Conditions (1.4) and (1.5) may be readily shown to be sufficient [11].

2. Real quadratic-Gaussian case

Let the transmitted signal ϑ (set of parameters $\vartheta = \{\vartheta_\alpha\}$ to be estimated) be a Gaussian vector (column) with values in real space $\mathcal{U}$ of dimensionality r, zero mean value, and known correlation matrix $S = \| \langle \vartheta_\alpha \vartheta_\beta \rangle \|$. Let received real boson (see [1]) signal y be real-linearly modulated and Gaussian at each ϑ. This means that $y = \{y_\tau\}$ is superposition $y = D\vartheta + \xi$ where $D = \| D_{\tau\alpha} \|$ is a real matrix, and $\zeta = \{\xi_\tau\}$ is Gaussian real ($\xi_\tau^* = \xi_\tau$) boson noise of a channel with zero mean value $\langle \xi \rangle = 0$ defined by matrices

$$N = \left\| \frac{2}{1} \langle \xi_\sigma \xi_\tau + \xi_\tau \xi_\sigma \rangle \right\|, \quad C = \| \xi_\sigma \xi_\tau - \xi_\tau \xi_\sigma \| \tag{2.1}$$

(briefly, $N = \langle \xi \xi^T \rangle_t$, $C = [\xi, \xi^T]$, ξ is a column, and ξ^T is a row).

An optimal, with respect to the criterion

$$C(\lambda, \vartheta) = \frac{1}{2} (\lambda - \vartheta)^T G (\lambda - \vartheta), \quad G = \| G_{\alpha,\beta} \| > 0, \tag{2.2}$$

observation is to be found realizing Bayes (Section 1) estimate $\lambda \in \mathcal{U}$.

Statement 1. Under the conditions formulated above, the optimal observation reduces to the optimal coherent measurement of repetitively commuting observables $u = \{ \sum_\tau A_{\alpha\tau} y_\tau \} = Ay$ as introduced in [12] where matrix A satisfies Eq. [3, (3.10)] whose complex form is shown in (3.6). Below we shall refer to this measurement as G-coherent.

The proof will require the following representation of the boson channel defined by decomposition $\xi = a + a^*$, orthogonal with respect to a symmetric positively defined matrix $H = \| H_{\alpha\tau} \| : a^T H a = 0$, and satisfying the commutation relations $[a, a^T] = 0$, $[a, a^+] \geq 0$ ($a^+ = \{a_\tau^*\}$ is a row associated with column $a = \{a_\tau\}$). Decomposition $\xi = a + a^*$ referred to as H-normal with respect to (non-normalized) annihilation operators a_τ and associated creation operators a_τ^*, is defined by a projection matrix $P : a = P\xi$, $a^* = P^*\xi$ satisfying conditions $P + P^* = I$, $P^+ H P^* = 0$ and definitizing (antisymmetric, imaginary) matrix $C : PC = CP^+ \geq 0$. If matrix C is non-singular (this is implied below for the sake of simplicity), matrix P is defined uniqely by matrix $H : P = \frac{1}{2}(I + \operatorname{sign} CH)$.

Definition. Representation of operators ϱ by functionals $p(\vartheta)$ over real space $\mathcal{E}$ of vectors $\vartheta = \{\vartheta_\tau, \tau \in T\}$ as normally ordered expressions $\varrho = : = : p(a + a^*):$ (i.e. such that all operators $\{a_\tau^*\}$ act after operators $\{a_\tau\}$) is called H-normal if decomposition $\xi = a + a^*$ is H-normal.

If functional $p_\xi(\vartheta)$ is d.o. ϱ_ξ of additive boson noise of a channel with respect to H-normal decomposition $\xi = a + a^*$, then functional $p(\vartheta) = p_\xi(\vartheta - D\vartheta)$ is d.o. $\varrho(\vartheta)$ of the output $y = D\vartheta + \xi$ of this (linear) channel with respect to H-normal decomposition $y = b + b^*$ (shift transforms $b = a + PD\vartheta$ do not change ordering).

In H-normal representation, d.o. of Gaussian boson noise is described by a functional

$$p_\xi(\vartheta) = \det |CN_H^{-1}|^{1/2} \exp\left\{-\frac{1}{2}\vartheta^T N_H^{-1}\vartheta\right\}, \tag{2.3}$$

where matrix N_H is H-antinormally ordered correlation matrix of noise $\xi = a + a^*$:

$$N_H = \langle a_\sigma a_\tau + a_\sigma a_\tau^* + a_\tau a_\sigma^* + a_\sigma^* a_\tau^* \rangle .$$

It is related to matrices (2.1) by relation

$$N_H = N + \mathrm{Re}\, PCP^+ = N + \frac{1}{2}|CH|\, H^{-1} \tag{2.4}$$

which may be readily obtained by passing to symmetrizing (Wigner) ordering through commutation relations $[a, a^+] = PCP^+$ and taking into account that matrix $2\mathrm{Re}\, PCP^+H = \mathrm{sign}(CH)CH$ is module $|CH|$ of H-Hermitian matrix CH.

By substituting in (2.3) operators $y = D\vartheta$ for ϑ and integrating quadratic-Gaussian expression

$$C(\lambda, \vartheta) \,:\, p_\xi(y - D\vartheta) \,:\, p(\vartheta) \quad (p(\vartheta) = \det(2\pi S)^{-1/2} \exp\{-\vartheta^T S^{-1}\vartheta/2\})$$

over ϑ, one easily finds the operator of *a posteriori* risk in H-normal representation:

$$\mathfrak{R}(\lambda) = :\mathfrak{R}p_y(y): + \frac{1}{2}:(\lambda - Ay)^T G(\lambda - Ay)\, p_y(y): \tag{2.5}$$

Here $\mathfrak{R}$ is trace of $S\mathrm{p}SG(2 - AD)/2$ in $\mathfrak{U}$, A is solution of equation

$$A(DSD^T + N) + \frac{1}{2}A|CH|\, H^{-1} = RD^T \tag{2.6}$$

and $p_y(\,\cdot\,)$ is Gaussian density

$$p_y(\gamma) = \det |C(DSD^T + M)^{-1}|^{1/2} \exp\left\{-\frac{1}{2}\vartheta^T (DSD^T + M)^{-1}\,\vartheta\right\}$$

defining d.o. $\varrho = \int \varrho(\vartheta)\, P(d\vartheta) = :p_y(y):$ as a normally ordered, with respect to operators $b = py$, $b^* = p^*y$, expression.

Proof of Statement 1. Consider first a case where dimensionalities of spaces $\mathfrak{E}$ and $\mathfrak{U}$ coincide. Then, *if Eq. (2.6) after substitution of* $H = A^TGA$ *has non-degenerate solution* A^0, *it defines a unique o.u.* Π^0 *satisfying necessary and sufficient optimality conditions* (1.3), (1.4) *and* (1.5).

Indeed, by taking in (2.5) $\Lambda = :\Re p(y): = \Re\varrho$, $A = A^0$, $y = b + b^*$, by placing operators $b = \{b_\tau\}$, $b^* = \{b_\tau^*\}$ in the normal order, and taking into account that $(b - \beta)^T A^{0T} G A^0 (b - \beta) = 0$ at $H = A^{0T} G A^0 (\beta = PA^{0-1}\lambda)$, one may represent operator $\Re(\lambda) - \Lambda$ in the following form

$$\Re(\lambda) - \Lambda = (b - \beta)^+ A^{0T} G A^0 \varrho (b - \beta) \tag{2.7}$$

where its positiveness is evident. Operator (2.7) for each β has a unique eigenvector $\varphi(\beta) \in \mathfrak{H}$ corresponding to the zero eigenvalue, a coherent state vector which is also the right eigenvector of annihilation operators $\{b_\tau\}$ with eigenvalues $\beta = \{\beta_\tau\} : b_\tau \varphi_{(\beta)} = \beta_\tau \varphi_\beta$. As shown in [13], system of coherent vectors $\{\varphi_\beta\}$ is nonorthogonal and overfilled and, consequently, defines a unique optimal quasi-measurement described by a family of coherent projectors $\{\pi^0(\lambda) = \varphi_{(\beta)}\varphi^*_{(\beta)}\}$, $\lambda = \beta + \beta^* \in \mathfrak{U}$. Any such a family may be represented as limit $\lim_{\theta \to 0} \pi^0(\lambda)$ of a family of Gaussian d.o.'s

$$\pi^\theta(\lambda) = \det \left| 2\mathrm{sh}\frac{1}{2\theta} ACA^TG \right|^{\frac{1}{2}} e^{-\frac{1}{2\theta}(-Ay)^T G(\lambda - Ay)} \tag{2.8}$$

$$\lambda \in \mathfrak{U}.$$

Each family (2.8) defines a Gaussian o.p. $\Pi^\theta(d\lambda) = \pi^\theta d\mu(\lambda)/\det |\, ACA^T \,|^{1/2}$ normalized with respect to measure $d\mu(\lambda) = \prod_{\alpha=1}^{r} (d\lambda_\alpha/\sqrt{2\pi})$ and describes for $\theta \to 0$ an optimal, with respect to criterion $(\lambda - u)^T G(\lambda - u)$ *coherent* measurement of linear functionals $u = Ay$ [12]. Thus, optimal projectors $\Pi^0(\lambda)$ have the form of (2.8), are completely defined by solution $A = A^0$ of Eq. (2.6) for $H^{-1} = A^TGA$ and realize minimal risk $\Re^0 = \frac{1}{2} SpSG(I - A^0D)$ coinciding with that found in [3]. As distinguished from the classical case, the optimal observation depends essentially on the form of quadratic criterion (2.2), i.e. on matrix G. That is why we call it G-coherent observation (measurement) of non-commutating observables $u^0 = A^0y$.

In the general case, matrix A in (2.6) has domain of values $A\mathfrak{E}$ belonging to $\mathfrak{U}: A\mathfrak{E} \subseteq \mathfrak{U}$ and rank $r(A) \leq r (r = \dim \mathfrak{U})$. If $r(A^TGA) < r(c)$, one may try to reveal the indeterminacy of conversion $(A^TGA)^{-1}$ arising at substitution of $H = A^TGA$ into (2.6) by introducing function $\mathrm{sign}(x) = |\, x \,|/x$:

$$A\,|CH|\,H^{-1} = A\,\mathrm{sign}(CH)\,C \overset{H=A^TGA}{=\!=} \mathrm{sign}(ACA^TG)\,AC.$$

As a result, Eq. (2.6) turns into Eq. (3.6) deduced in [3] for the Gaussian real case by minimization of the mean-square risk in the class of operators $u = \int \lambda \Pi(d\lambda)$ and by using the generalized Heisenberg inequality $\langle uu^T \rangle_+ \geq \geq \frac{1}{2}\langle [u, u^T] \rangle$. This equation enables one to isolate solution $A = A^0$ if it satisfies the condition of matrix $A^0 C A^{0T}$ non-singularity in sub-space $A^0\mathfrak{E} \subseteq \mathfrak{U}$, otherwise matrix $\mathrm{sign}(A^0 C A^{0T} G)$ is not determined over $A^0\mathfrak{E}$. If such A^0 does exist, it defines the optimal observation, G-coherent measurement of observables $u^0 = A^0 y$. Indeed, since the non-singularity of matrix $A^0 C A^{0T}$ in $A^0\mathfrak{E}$ is equivalent to the non-singularity of matrix C_1 induced by matrix C in sub-space $A^{0T} G A^0 \mathfrak{E}$; sub-space $\mathfrak{E}_1 = A^{0T} G A^0 \mathfrak{E}$ and $H = A^{0T} G A^0$ may be taken as defining H-normal representation of space and matrix H. By representing d.o. $\varrho(\vartheta)$ in the form of normally ordered expression of corresponding operators $b = Py$, $b^* = P^* y$ forming "sufficient statistics" $y_1 = b + b^* = E_1 y : A^0 y = A^0 y_1$ (being unit I in $\mathfrak{E}_1$, projector $E = P + P^*$ from $\mathfrak{E}$ to $\mathfrak{E}_1$ satisfies condition $A^0 E_1 = A^0$), operator of *a posteriori* risk may be readily shown to have in this case also the form of (2.5) where $A = A^0$. Orderliness with respect to the balance of operators $y_2 = y - CE$, $C_1^{-1} y_1$ commuting with $y_1 = E_1 y$ and included only into: $p_y / y := \varrho$ is unessential for detection of non-negative determinacy $\mathfrak{R}(\lambda) - \Lambda \geq 0$. Indeed, by assuming in (2.5) $A = A^0$, taking into account that $A^0 y = A^0 y_1 = A^0 (b + b^*)$ and arranging operators b in the normal order, obtain positively defined operator

$$\mathfrak{R}(\lambda) - \Lambda = (b - \beta)^+ A^{0T} G A^0 \varrho (b - \beta) + \frac{1}{2} \lambda_2^T G \lambda_2 \varrho \tag{2.9}$$

where λ_2 is G-orthogonal complement of vector $\lambda_1 = A^0(\beta + \beta^*) : \lambda = \lambda_1 + \lambda_2$, $\lambda_2^T G \lambda_1 = 0$. In accordance to this, the optimal observation is described by o.p. concentrated at $\lambda_2 \neq 0$ in sub-space $A^0\mathfrak{E} \subseteq \mathfrak{U} : \Pi^0(d\lambda) = 0$ where it is still described by G-coherent projectors $\pi^0(\lambda_1)$ representable by operators (2.8) at $\lambda \in A^0\mathfrak{E}$ and $\theta \to 0$.

Finally, if there is no solution A^0 of (3.6) satisfying the condition of non-singularity of the matrix induced by matrix $A^0 C A^{0T}$ in $A^0\mathfrak{E}$, this equation should be complemented by equation $ACA^T G \mathfrak{U}_0 = 0$ over some sub-space $\mathfrak{U}_0 < A\mathfrak{E}$ that gives a condition for determining the indefinite value of matrix $B = \mathrm{sign}(ACA^T G)$ over $\mathfrak{U}_0$. With appropriate $\mathfrak{U}_0$ this set of equations already has a solution. (In particular, at $\mathfrak{U}_0 = \mathfrak{U}$ there is always a solution corresponding to the complete commutativity $[u, u^T] = ACA^T = 0$). Among all possible solutions, optimal ones A^0 should be sorted out by means of inequality $| A^0 C A^{0T} G | \geq B A^0 C A^{0T} G$, this corresponds to determination of signature

values over interval $-I \leq B \leq I$ (equality sign should take place at least over G-orthogonal complement $\mathfrak{U}_1 = A\mathfrak{E} \ominus \mathfrak{U}_0$ to $\mathfrak{U}_0$ over which the signature is not determined). By using the normal ordering with respect to operators $b = Py$, $b^* = Py$ ($P + P^+ = E_1$ is unit of $(A^{0T}GA^0)^{-1}$ – orthogonal complement $\mathfrak{E}_1 = A^{0T}GA^0 \ominus \mathfrak{E}_0$ to the kernel of induced matrix C in $A^{0T}GA^0\mathfrak{E} = \mathfrak{E}_0 \oplus \mathfrak{E}_1$) and y_0 — representation with respect to commuting with each other and with $y_1 = E_1 y$ operators $y_0 = E_0 y$ (E_0 is projector into $\mathfrak{E}_0$), one easily finds that operator (2.9) will involve a third non-negative term $(\lambda_0 - A^0 y_0)^T G\varrho(\lambda_0 - A^0 y_0)/2$ corresponding to the ideal direct measurement of commuting observables $A^0 y_0$ with results $\lambda_0 \in \mathfrak{U}$. Note, that solution A^0 of (3.6) which is understood in degenerate cases in the mentioned above generalized sense, might be obtained through solution of a more general matrix equation obtained in [1] through minimization of risk in the class of Gaussian o.p.'s (2.8) for finite $\theta > 0$ as limit $A^0 = \lim_{\theta \to 0} A^0$.

3. Complex generalization and more general solution of the optimal linear estimation equation

1. In Section 2, while proving the optimality of coherent observation, usefulness of complex representation was demonstrated which stems form the natural complex structure of quantum signals. From the viewpoint of the relativity theory of quantified fields, real signals are a special (i.e. really neutral) case of complex quantum signals. In the general case,* the boson signal is described by complex vector $y\{y_\tau\}$ whose components y_τ are repeatedly commuting with conjugate y_τ^*:

$$[y_\sigma, y_\tau^*] = C_{\sigma\tau}, \; [y_\sigma, y_\tau] = [y_\sigma^*, y_\tau^*] = 0, \; [y, C_{\sigma\tau}] = 0. \tag{3.1}$$

Since $(y_\sigma^* \cdot y_\tau)^* = y_\tau^* \cdot y_\sigma$, matrix C of complex numbers $C_{\sigma\tau}$ is Hermitian, $C^+ = C$, and, generally speaking, indefinite. In particular, if matrix C is antisymmetric $C^T = C$, then y is a complexification $y = \frac{1}{\sqrt{2}}(y_R + iy_I)$ of real boson signal y_R of the same dimensionality $[y_R, y_R^T] = C$ (y_I is a copy of y_R: $[y_I, y_I] = C$; $[y_I, y_R^T] = 0$).

* Commutation relations (3.1) might be represented also by self-conjugate generators by making signals real, $y_R = \frac{1}{\sqrt{2}}(y + y^*)$; $y_I = \frac{1}{\sqrt{2i}}(y - y^*)$, however, the dimensionality of a signal made real would double.

Let transmitted signal be represented by a random vector $\vartheta = \{\vartheta_\alpha\}$ with complex values and Gaussian distribution**

$$P(\alpha\vartheta) = p(\vartheta)\, d\mu(\vartheta)$$

where

$$p(\vartheta) = \det S^{-1}\, e^{-\vartheta^T S^{-1}\vartheta},\ d_\mu(\vartheta) = \prod_\alpha \frac{1}{\pi} d\,\mathrm{Re}\,\vartheta_\alpha d\,\mathrm{Im}\,\vartheta_m. \tag{3.2}$$

An o.p. $\Pi(\lambda)$ is to be determined over vector space $\mathfrak{U}$ of complex values of estimates $\lambda = \{\lambda_\alpha\}$ that describes an optimal, with respect to quadratic criterion

$$C(\lambda, \vartheta) = (\lambda - \vartheta)^+ G(\lambda - \vartheta),\ G < 0, \tag{3.3}$$

observation of superposition $y = D\vartheta + \xi$ of linearly-complex modulated signal ϑ and complex Gaussian boson noise ξ. The latter means that the observed signal y is described by complex generators $\{y_\tau\}$ satisfying relations (3.1), and for each ϑ has Gaussian d.o.

$$\varrho(\vartheta) = \det |1 - e^{-QC}|\, e^{-(y-D\vartheta)^+ Q(y-D\vartheta)},\ Q = \|Q_{\sigma\tau}\| > 0. \tag{3.4}$$

Statement 2. Solution of the formulated complex quadratic-Gaussian optimal observation problem is completely determined by a complex matrix $A = \| A_{\alpha\tau}\|$ *and is described by coherent o.p. which may be conveniently represented as limit* $\Pi^0 = \lim \Pi^\theta$ *for* $\theta \to 0$ *of non-singular Gaussian o.p.* $\Pi^\theta(d\lambda) = \pi^\theta(\lambda)\, d\mu(\lambda)$:

$$\pi^\theta(\lambda) = \det |ACA^+(1 - e^{-\frac{1}{\theta} GACA^+})^{-1}|^{-1}\, e^{-\frac{1}{\theta}(\lambda - Ay)^+ G(\lambda - Ay)} \tag{3.5}$$

$$d\mu(\lambda) \prod_\alpha \frac{1}{\pi} d\,\mathrm{Re}\,\lambda_\alpha d\,\mathrm{Im}\,\lambda_\alpha.$$

Matrix A minimizing the mean-square risk of linear estimation satisfies equation

$$AR + \frac{1}{2}\,\mathrm{sign}\,(ACA^+ \cdot G)\, AC = T \tag{3.6}$$

where R, T *are correlation matrices*

$$R = \left\| \frac{1}{2} \langle y_\sigma, y_\tau^* + y_\tau^* y_\sigma \rangle \right\|, \qquad T = |\, \langle \vartheta_\alpha y_\tau^* \rangle\, |$$

** It is not reasonable to consider the complex representation of more general Gaussian signals with densities involving terms $\vartheta_\alpha \cdot \vartheta_\beta$, $\vartheta_\alpha^* \cdot \vartheta_\beta^*$ along with terms $\vartheta_\alpha^* \cdot \vartheta_\beta$. They should be represented by real vectors of doubled dimensionality.

expressed through the initial correlation matrices of signal S and noise $N = \frac{1}{2} C \operatorname{cth} QC$ in the conventional manner, $R = DSD^+ + N$, $T = SD^+$.

The proof follows the same reasoning as in the real case. For d.o. (3.4) a representation is used having the form of an appropriate normally ordered expression. We are not going to repeat it, mention only that the transition from ordering (3.4) to the normal ordering is, in mathematical terms, a known transition [14] from indefinite metric $\{\varkappa', \varkappa\} = \varkappa^+ C\varkappa'$ to positive one $|\varkappa', \varkappa| = \varkappa_+^+ C\varkappa_+' - \varkappa_-^+ C\varkappa_-'$ by means of decomposition $\mathcal{H} = \mathcal{H}_+ \oplus + \mathcal{H}_-$ of complex space $\mathcal{H}$ (domain of definition of operator-valued functional $y(\varkappa) = \varkappa^+ y$) into domains of positiveness $\varkappa_+^+ C\varkappa_+ < 0$, $\varkappa_+ \in \mathcal{H}_+$ and negativeness $\varkappa_-^+ C\varkappa_- < 0$, $\varkappa_- \in \mathcal{H}_-$ orthogonal with respect to a scalar product $(\varkappa', \varkappa) = \varkappa^+ H^{-1}\varkappa'$. The only difference from the real case is that operators of annihilation $b_+ \equiv y(\varkappa_+)$ and creation $b_-^* \equiv y(\varkappa^-)$ in decomposition $y(\varkappa) = b_+ + b_-$ commute and are related to different modes (in the real case $b_+ = b_-$ for real H and $\varkappa = \varkappa_+ + \varkappa_-$).

2. To isolate the explicit solution of Eq. (3.6) satisfying the condition of non-singularity of matrix ACA^+ in $A\mathcal{H}$ introduce the following notations:

$$K = \| \langle y_\tau^* y_\tau \rangle \| = R - \frac{1}{2} C, L = \| \langle y_\sigma y_\tau^* \rangle \| = R + \frac{1}{2} C \tag{3.7}$$

and carry out factorization $T = T_+ + T_-$ of matrix $T = \| \langle \vartheta_\alpha y_\tau^* \rangle \|$ satisfying the following conditions:

$$T_+(K^{-1} - L^{-1}) T_+^+ \geq 0, \; T_-(L^{-1} - K^{-1}) T_-^+ \geq$$
$$\geq 0, \; T_+(K^{-1} - L^{-1}) T_-^+ = 0. \tag{3.8}$$

There always exists such a G-orthogonal decomposition, matrices $T_\pm$ being defined uniquely through formulae

$$T_\pm = \frac{1}{2} \big(I \pm \operatorname{sign} (T (K^{-1} - L^{-1}) N^+ G)\big) T \tag{3.9}$$

if matrix $T(K^{-1} - L^{-1}) T^+$ is non-degenerate over subset $T\mathcal{H} < \mathcal{U}$ of matrix T values. In the case of $T\mathcal{H} < \mathcal{U}$, matrix inequality (3.11) providing this non-degeneracy in accordance with (3.13) should be regarded as a strict positiveness only over sub-space $T\mathcal{H}$.

Statement 3. Solution of Eq. (3.6) may be represented in the following form

$$A^0 = T_+ L^{-1} + T_- K^{-1} \tag{3.10}$$

only if matrices K, L and factorization (3.9) of matrix T satisfy inequality

$$T_+(L^{-1} - L^{-1}KL^{-1})T_+^+ + T_-(K^{-1} - K^{-1}LK^{-1})T_-^+ < 0. \qquad (3.11)$$

Proof. Indeed, assume that among matrices A for which matrix ACA^T is non-degenerate at least over sub-space of values of $A\mathcal{H}$, there is one such that it transforms Eq. (3.6) into identity. Then it may be factorized uniquely $A = A_+ + A_-$ so that conditions

$$A_\pm GA_+ = 0,\ A_+CA_+^+ \geqq 0,\ A_-CA_\pm \leq 0,\ A_+CA_\pm = 0: \qquad (3.12)$$

$$A_\pm = \frac{1}{2}\big(I \pm \operatorname{sign}(ACA^+G)\big)A$$

are satisfied. By making changes in Eq. (3.6) $-R = (K + L)/2$, $C = L - K$- obtain an equation with respect to A_+, $A_1 : A_+L + A_-K = T$. The solution of the latter equation may be represented as $A_+ = T_+L^{-1}$, $A_- = T_-K^{-1}$ through factorization $T = T_+ + T_-$ satisfying conditions $T_\pm GT_+ = 0$,

$$\begin{gathered} T_+(L^{-1} - L^{-1}KL^{-1})T_+^+ \geq 0,\ T_-(K^{-1} - K^{-1}LK^{-1})T_\pm \geq 0, \\ T_+(K^{-1} - L^{-1})T_\pm = 0 \end{gathered} \qquad (3.12a)$$

obtained by subsituting $A_+ = T_+L^{-1}$, $A_- = T_-K^{-1}$ into (3.12). Conditions (3.12a) assure that conditions (3.8) defining matrices (3.9) are satisfied because the following matrix inequalities hold:

$$K^{-1} - L^{-1} > L^{-1}KL^{-1},\ L^{-1} - K^{-1} > K^{-1} - K^{-1}LK^{-1}$$

which follow from identities

$$\begin{aligned} L^{-1} - L^{-1}KL^{-1} &= K^{-1} - L^{-1} - (K^{-1} - L^{-1})K(K^{-1} - L^{-1}), \\ K^{-1} - K^{-1}LK^{-1} &= L^{-1} - K^{-1}(L^{-1} - K^{-1})L(L^{-1} - K^{-1}). \end{aligned} \qquad (3.13)$$

This completes the proof of Statement 3 because inequalities (3.12a) in $\mathcal{H}$ are equivalent to inequality (3.11) in $T\mathcal{H}$.

Represent solution of (3.10) with allowance for (3.7), (3.9) as

$$A^0 = [TR^{-1} - \operatorname{sign}(TR^{-1}\mathcal{E}(1 - \mathcal{E}^2)^{-1}T^+G)TR^{-1}\mathcal{E}](1 - \mathcal{E}^2)^{-1} \qquad (3.14)$$

where $\mathcal{E} = \frac{1}{2}CR^{-1}$. Note that the established optimal linear transform A^0 from $\mathcal{H}$ into $\mathcal{U}$ was found previously [1] only for simple cases admitting normal

complex representation (evidently, for $C > 0$, $A^0 = TL^{-1}$ and for $C < 0$, $A^0 = TK^{-1}$) and also for the one-mode one

$$C = \hbar \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$$

$$A^\circ = \frac{1}{1 - \mathscr{E}^2} T \cdot \left(R^{-1} - \varepsilon \cdot \delta (T^+ G T)^{-1}\right)$$

where

$$\varepsilon = \frac{1}{2} \sqrt{\det |\, CR^{-1} |}\,, \qquad \delta = \sqrt{\det (T^+ G T R^{-1})}\,.$$

3. Demonstrate another, direct way to solve the quadratic-Gaussian complex measurement problem based on a natural ordering where operators y^* act to the left of y. In this representation, d.o. (3.4) is described by a Gaussian density $p_\xi(\vartheta)$ normalized to unit $\int p_\xi(\vartheta) d\mu(\vartheta) = 1 : \varrho(\vartheta) = \det |\, C \,| : p_\xi(y - D\vartheta) : ;$ and the *a posteriori* risk operator determined by integration $\int c(\lambda, \vartheta) : p_\xi(y - D\vartheta) : p(\vartheta) d\mu$ of the quadratic cost function (3.3) with Gaussian *a priori* distribution (3.2) is described by a quadratic-Gaussian expression

$$\begin{aligned} \mathscr{R}(\lambda) &= \det |\, C \,| \, [\mathscr{R} : p_y(y) : + : (\lambda - TL^{-1}y)^+ G (\lambda - TL^{-1}y) P_y(y) :] = \\ &= \mathscr{R}\varrho + (\lambda - TL^{-1}y)^+ G \varrho (\lambda - TL^{-1}y) \end{aligned} \tag{3.15}$$

where $\mathscr{R} = SpG(S - TL^{-1}T^+)$, and $p_y(\vartheta) = \det [\,^{-1} \exp\{-\vartheta^+ L^{-1} \vartheta\}$ is density $p_\xi(\vartheta - D\vartheta)$ defining Gaussian d.o. $\varrho = \int \varrho(\vartheta) P(d\vartheta)$ and averaged together with distribution (3.2):

$$\begin{gathered} \varrho = \det |\, CL^{-1} \,| : e^{-y + L^{-1} y} : = \det |\, CL^{-1} \,|\, e^{-y + Hy} \\ \left(H = -C^{-1} \ln (I - CL^{-1})\right). \end{gathered} \tag{3.16}$$

If operators $u = TL^{-1}y$ have non-negatively defined commutator matrix $[u, u^+] = T(L^{-1} - L^{-1}KL^{-1})T^+ \geq 0$ then orderliness of these operators in (3.15) is normal, and the optimal observation is reduced to their G-coherent measurement described, in the complex situation, by operators (3.5) at $A = TL^{-1}$ and $\theta \to 0$.

In the general case, to pass to the normal ordering one has to multiply operator (3.15) from the right and from the left by $\varrho^{-\frac{1}{2}}$ and take into account formula $\varrho^{\frac{1}{2}} y \varrho^{-\frac{1}{2}} = \sqrt{LK^{-1}}\, y$.

The transformed operator $\varrho^{-\frac{1}{2}} \mathscr{R}(\lambda) \varrho^{-\frac{1}{2}}$ is represented by a quadratic form with respect to operators $V = TM^{-1}y$ (where $M^{-1} = L^{-1}\sqrt{LK^{-1}}$) that

may be readily ordered*:

$$\varrho^{-\frac{1}{2}}\mathcal{R}(\lambda)\varrho^{-\frac{1}{2}} = \mathcal{R}_+^0 + (\lambda_+ - T_+M^{-1}y)^+G(\lambda_+ - T_+M^{-1}y) + \\ + (\lambda_- - T_-M^{-1}y)^T G^*(\lambda_- - T_-M^{-1}y)^*. \tag{3.17}$$

Here $T_\pm$ are matrices (3.9) performing orthogonal decomposition $T^+GT_+ = 0$ of matrix $[V, V^+] = T(K^{-1} - L^{-1})T^+$ into positive and negative parts; $\lambda_\pm = (1 \pm \operatorname{sign}(T(K^{-1} - L^{-1})T^+G))\,\lambda/2$; and

$$\mathcal{R}^0 = SpG(S - T_+L^{-1}T_+^+ - T_-K^{-1}T_\pm) \tag{3.18}$$

coincides with risk $\mathcal{R}^0 = SpG(S - A^0T)$ corresponding to solution (3.10). By inverse transform obtain

$$\mathcal{R}(\lambda) = R^0\varrho + (\lambda_0 - T_+L^{-1}y)^+ G\beta(\lambda_+ - T_+L^{-1}y) + \\ + (\lambda^- - T_-K^{-}y)^T G^*\varrho(\lambda_- - T_-K^{-1}y)^*.$$

In going so, the normal orderliness will be preserved only if matrices T, K, L satisfy inequality (3.16), this means that commutator matrices $[u_+, u_+^*]$, $[u^*, u^T]$ of operators $u_+ = T_+L^{-1}y$, $u_- = T_-K^{-1}y$ are non-negative. Consequently, solution of the optimal complex measurement problem, with regard to condition (3.16), indeed is described by coherent projectors $\Pi^0(\lambda)$ satisfying equations $u_+\pi^0(\lambda) = \lambda_+\pi^0(\lambda)$, $u^*\pi^0(\lambda) = \lambda_-^*\pi^0(\lambda)$ and representable as (3.5) at $A = T_+L^{-1} + T_-K^{-1}$ and $\theta \to 0$.

4. In conclusion note that if dimensionality of signal space $\mathcal{H}$ does not exceed that of $\mathcal{U}$, solution of the complex quadratic-Gaussian optimal observation problem is representable through projectors $\varrho^0(\lambda)$ satisfying equations

$$(V_+ - \lambda_+)\varrho^0(\lambda) = 0,\ (V_- - \lambda_-)^*\varrho^0(\lambda) = 0,\ V_\pm = T_\pm M^{-1}y,$$

by formula

$$\Pi^0(d\lambda) = \varrho^{-\frac{1}{2}}\varrho^0(\lambda)\,\varrho^{-\frac{1}{2}}P(d\lambda) \tag{3.19}$$

which is the quantum analog of the Bayes formula. Here $P(d\lambda)$ is a probabilistic measure giving "Glauber" decomposition

$$\varrho = \int \varrho^0(\lambda)\,P(d\lambda) \tag{3.20}$$

with respect to coherent projectors $\varrho^0(\lambda)$. It was B. L. Grishanin [9] who found solution of the optimal estimation problem in this form for the real quadratic-Gaussian case and equal dimensionalities of spaces $\mathcal{U}$ and $\mathcal{H}$.

* Generally speaking, representation (3.17) holds only for $\lambda \in T\mathcal{H}$ and only if matrix $T(K^{-1} - L^{-1})T^+$ is non-degenerate over sub-space $T\mathcal{H} \subseteq \mathcal{U}$. In the case of $T\mathcal{H} \subset \mathcal{U}$ it is always sufficient to restrict oneself to consideration of estimates λ from $T\mathcal{H}$.

Coherent projectors $\varrho^0(\lambda_+ + \lambda_-^*)$ are for operator (3.17) eigenprojectors corresponding to the minimal eigenvalue (i.e. "vacuum" for the quadratic "energy" operator $\eta_+ + G\eta_+ + \eta_\pm G\eta_-^*$ where $\eta_\pm = T_\pm M^{-1}y - \lambda_\pm$) and are described by a correlation operator matrix $\eta = TM^{-1}y - \lambda$, for instance, by a symmetrizing one $R_\eta = \langle \eta\eta^+ \rangle_+$

$$R_\eta = \frac{1}{2}\big(T_+(K^{-1} - L^{-1})T_+^+ + T_-(K^{-1} - L^{-1})T_-^+ = \\ = \frac{1}{2}T\,|\,T(K^{-1} - L^{-1})T^- GT\,|\,(T^+GT)^{-1} + T^+ \tag{3.21}$$

lying on the boundary of the generalized [3] Heisenberg inequality

$$R_\eta \geq \pm\, C_\eta/2\,(C_\eta = [\eta, \eta^+] = T(K^{-1} - L^{-1})T^+\big).$$

Distribution $P(d\lambda)$ is concentrated over sub-space of vectors $\lambda = \lambda_+ + \lambda_-^*$ coinciding with the sub-space of matrix T values where it is described by a positive Gaussian density $p(\lambda)$, $\lambda \in T\mathcal{H}$. The positiveness condition $p(\lambda) > 0$, which in the Gaussian case reduces to the condition of positive definiteness of correlation matrix $\mathcal{R}_\lambda = \langle \lambda\lambda^+ \rangle$ over $T\mathcal{H}$, defines the optimality domain of o.p. (3.19). Taking into account that, in accordance with (3.20), sum $R_\eta + R_\lambda$ is equal to correlation matrix $R_v = \langle VV^+ \rangle_+ = T(K^{-1} + L^{-1})T^+/2$ of operators $V = TM_y^{-1} = \eta + \lambda$ that defines Gaussian d.o. (3.16), write the condition of matrix $R_\lambda = R_v - R_\eta$ positiveness over $T\mathcal{H}$ as $R_v > R_\eta$. The latter may be written, with regard to (3.21), as a matrix inequality:

$$(K^{-1} + L^{-1})TGT^+ > |\,(K^{-1} - L^{-1})T^+GT\,| \tag{3.22}$$

whose equivalence to inequality (3.11) under taken relation between dimensionalities $\dim \mathcal{H} \leq \dim \mathcal{U}$ is readily shown through identities (3.13). In the one-mode case inequality (3.22) reduces to numerical inequality $\varepsilon > t - \sqrt{t^2 - 1}$ where $t = Sp\,T^+GT\,R^{-1}/2\,\sqrt{\det T^+GTR^{-1}}$ obtained in [1]. In the general case, inequality (3.11), unlike (3.22), gives a necessary and sufficient condition of the non-degeneracy of optimal solution (3.10) of the and minimality of corresponding risk (3.18) for a more natural, reverse relation between dimensionalities, $\dim \mathcal{H} > \dim \mathcal{U}$.

Conclusion

As might be expected, optimal observations in Gaussian linear boson channels giving minimal mean-square error for Gaussian input estimation are, thus, quasi-classical (i.e. coherent). However, unlike the classical case, they are indirect, depend essentially on metric matrix G and may degenerate in

some cases of indefinite commutation relations. Examples of such degenerations and some other physical examples are discussed in [1]. In the present paper optimality of G-coherent measurement of optimal linear transform A^0y is proved, condition of its non-degeneracy and its explicit form (Statement 3) are found as applied to the general (multidimensional, complex, indefinite) case. Thus the problem of statistical observation in quantum channels, whose special cases were discussed in [1–5, 9] may be regarded as exhausted.

Reference

1. *Belavkin, V. P.*, Optimal linear random filtration of quantum boson signals. Problems of Control and Information Theory, 1974, 1.
2. *Helstrom, C. W., Liu, J. W., Gordon, J. P.*, Quanti-mechanical communication theory. Proc. of IEEE, 1970, 10, pp. 186–207.
3. *Belavkin, V. P., Grishanin, B. A.*, Study of optimal estimation in quantum channels by the generalized Geisenberg inequality. Problemy Peredachi Informatsii, 1973, 9, 3, pp. 47–52 (in Russian).
4. *Stratonovich, R. L.*, The quantum generalization of optimal Statistical estimation and hypothesis testing. Stochastics, 1973, 1, pp. 83–126.
5. *Kholevo, A. S.*, Optimal quantum measurements. IIIrd International Symposium on Information Theory. Tallinn, June, 1973, 1 (in Russian).
6. *Belavkin, V. P., Stratonovich, R. L.*, On optimization of quantum signal processing with respect to the informational criterion, Radiotekhnika i Electronika, 1973, 18, 9, pp. 1839–1844. (in Russian).
7. *von Neumann, J.*, Mathematical foundations of quantum mechanics. Princeton Univ. Press, Princeton, NJ, 1955.
8. *Davis, E. B., Lewis, J. T.*, An operational approach to quantum probability. Comm. Math. Phys., 1970, 17, pp. 239–260.
9. *Grishanin, B. A.*, Some methods for solution of quantum detection and measurement problems. Izvestiya AN SSSR. Tekhnicheskaya Kibernetika, 1973, 5, pp. 127–137 (in Russian).
10. *Yuen, H., Lax, M.*, Multiple parameter quantum estimation and measurement of non-self-adjoint operators. IEEE International Symposium on Information Theory, Asilomar, CA, Jan. 31, 1972.
11. *Belavkin, V. P., Vantsyan, A. G.*, On sufficient conditions for optimal processing of quantum signals. Radiotekhnika i Electronika, 19, 1974, 7, pp. 1391–1395 (in Russian).
12. *Belavkin, V. P.*, Linear estimation of non-commuting quantum variables at their indirect measurement. Radiotekhnika i Elektronika, 1972, 17, 12, pp. 2534–2540 (in Russian).
13. *Klauder, J. R., Sudarshan, E. C. G.*, Fundamentals of quantum optics. W. A. Benjamin Inc., New York, Amsterdam, 1968.
14. *Nagy, K. L.*, State vector spaces with indefinite metric in quantum field theory. Akadémiai Kiadó, Budapest, 1966.

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 которы 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 гранок.

Статье должна предшествовать аннотация объемом до 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 высылаются гранки статьи, которые они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Prékopa, A.: Optimal control of a storage level using stochastic programming (*Прекопа А.* Применение стохастического программирования для оптимального управления уровнем запаса) 193

Kurzhanskii, A. B.: On minimax control and estimation strategies under incomplete information (*Куржанский А. Б.* О минимаксных стратегиях управления — оценивания при неполной информации) 205

Jávor, A.: An approach to the modelling of uncertainties in the simulation of quasideterministic discrete event systems (*Явор* А. О приближенном моделировании квазидетерминированных дискретных систем) 219

Tarnay, K.: Remarks on the analysis of a stabilized analog-digital converter (*Тарнаи К.* Замечания к анализу стабилизированных аналого-цифровых преобразователей) 231

Белавкин В. П. Оптимальное наблюдение бозонных сигналов в квантовых гауссовских каналах (*Belavkin, V. P.* Optimal observation of boson signals in quantum Gaussian channels) 241

Bányász, Cs., Keviczky, L.: Some experiences with the PKF discrete-time identification method (*Баняс Ч., Кевицки Л.* Применение метода дискретной идентификации, использующего априорную информацию о структуре объекта) 259

Book Review 275

316.920

VOL. 4 * NUMBER 4
ТОМ 4 * НОМЕР 4

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

1975

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 four issues published per year make up a volume of some 320 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

Президиумами Академии Наук СССР и Академии Наук ВНР было принято решение о совместном издании журнала

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал создан для обеспечения быстрейшей публикации материалов о новейших результатах научных исследований, разработок и кратких сообщений 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 издается четыре раза в год общим объёмом около 320 печатных страниц.

Журнал в первую очередь будет публиковать научные достижения 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. EMELIANOV (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

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 4 (4), pp. 275 – 287 (1975)

CONTROL OPTIMIZATION OF MULTIPLE OPERATION RESOURCES

A. A. VORONOV, E. P. PETRUSHININ

(Vladivostok–Moscow)

(Received March 10, 1974)

The problem of multiple operation resources control is considered. The corresponding mathematical model is studied. An example of optimization is given.

1. Statement of the problem

Resources necessary for jobs performing are divided into two groups: single-used (materials, energy, etc.) and repeatedly used (equipment, people, etc.). Unused for the time being, but ready for putting into operation resources of group 1 are considered to be stocks, controlling them requires a special study. Unused resources of group 2, but ready for operation are considered to be reserves, controlling them has some peculiarities [1]. This paper presents one of the optimal control problems.

There is given a set of resources

$$R = \{R_1, \ldots, R_n\} \tag{1.1}$$

taken for jobs performing at many objects

$$V = \{V_1, \ldots, V_m\}. \tag{1.2}$$

The amount of resources required for each object is defined by resource capacity (for example, capacity of labour, capacity of energy, etc.) $W_j (j = \overline{1, m})$. W_j is usually an additive resource expenditure function with different purposes and therefore each resource of the same purpose may be considered separately. Let R the resource having one purpose with its productivity K_i and quantity N_i for $R_i \in R$, $i = 1, n$. Let the resource capacity be expressed by the formula

$$W_j = \sum_{i=1}^{n} K_i N_i x_{ij}, \qquad j = \overline{1, m} \tag{1.3}$$

where x_{ij} is the time for using of the resource R_i at the object V_j. The values x_{ij}, $i = \overline{1, n}$; $j = \overline{1, m}$ are restricted by constraints

$$x_{ij} \geq 0 \tag{1.4a}$$

if using R for V_j is admissible, and by

$$x_j = 0 \tag{1.4b}$$

otherwise.

We start by looking for $\{x_{ij}\}$ satisfying constraints (1.3) and (1.4) and minimizing the total time T for performing operation. Resource transportation times to the objects at the first stage are not taken into account. The time T_i for using R_i at all the objects is equal to

$$T_i = \sum_{j=1}^{m} x_{ij}; \qquad i = \overline{1, n}. \tag{1.5}$$

If the resources are used independently then the goal function is equal to

$$T = \max_i \{T_i\} = \max_i \left\{ \sum_{j=1}^{m} x_{ij} \right\}. \tag{1.6}$$

Assume the resource R_s be used at all admissible for objects during the time $T_s = \max_i \{T_i\} = T$, and the resource R_l is used the time $T_e < T_s$, both kinds of resources being utilized at the object V_r. The time T_s can be decreased by the value Δx_s reducing time x_{sr} for resource using R_s at the object V_r by the value $\Delta x_{sr} = \Delta x_s$. But to keep resource balance (equality of resource capacity and resource expenditure) it is necessary simultaneously to increase the time x_{lr} for resource using R_l at the same object V_r by the value Δx_{lr} respectively, to satisfy the balance condition (1.3)

$$k_s N_s \Delta x_{sr} = k_l N_l \Delta x_{lr}; \tag{1.7}$$

Changing of Δx_{sr} may be done so that not to allow Δx_{sr} to break limits $0 \leq \Delta x_{sr} \leq x_{sr}$ on the one hand, and on the other may be increased until $T_s^{(1)} = T_s - \Delta x_{sr}$ becomes the largest of all T, as the further increasing does not allow T to decrease.

It is possible to decrease T by means of iterations being introduced by moving of value Δx_{ij}, which is multipled by the corresponding weight coefficients, on the solution table cells (Table 2) located at the vertices of a closed polygon formed by the "castle moves". The similar way it is done in the simplex method of linear programming. If these operations results in the solution

of the type, where all $T_i = \text{const} = T$, become identical then further T decreasing will not be possible. Hence the solution thus obtained will be optimal.

Really, let assume the solution be obtained, where $T_i = \text{const} = T$, $i = \overline{1, n}$. Then we try to decrease all T_i. First we shorten the time T_a for using some resource R_a at some object V_a by the value $\Delta x_{a\alpha}$. To keep the resource balance it is necessary to increase time T_b used at the same object of the other resource R_b by the value

$$\Delta x_{b\alpha} = \frac{k_a N_a}{k_b N_b} \Delta x_{a\alpha}$$

respect to (1.7), equal to vertical moving on the column α from the line a to the line b. As a result the time T_b has been increased to reduce it we must shorten the time $x_{b\beta}$ for using the same resource R_b but at the other object V_β by the value

$$\Delta x_{b\beta} \geq \Delta x_{b\alpha} = \frac{K_a N_a}{K_b N_b} \Delta x_{a\alpha}.$$

In the case of unequality the summing value T_b is being decreased, in the case of equality it is not changed (horisontal moving). To keep resource balance it is necessary to increase the time $x_{c\beta}$ for using the other resource R_c at the same object V_β by the value

$$\Delta x_{c\beta} \geq \frac{K_b N_b}{K_c N_c} \Delta x_{b\beta} \geq \frac{K_a N_a}{K_c N_c} \Delta x_{a\alpha}.$$

Further moving may be continued until all the resources are utilized or it may be stopped at any place. However in the long run we have to come back to the line of resource R_a at the object V_ξ having increased T_a by the value

$$\Delta x_{a\xi} = \frac{K_a N_a}{K_a N_a} \Delta x_{a\alpha} = \Delta x_{a\alpha}.$$

Mind that the unequality sign must not be used in the last expression, as increasing $\Delta x_{a\xi}$ will exceed its original decrease $\Delta x_{a\xi}$, so it should look like

$$\Delta x_{a\xi} = \Delta x_{a\alpha}.$$

To obtain this equality at the last iteration step we should substitute unequalities for equalities at all the previous steps as well. Thus, further decreasing T and simultaneous resource balance keeping is impossible. All T_i being equal one to another the solution is optimal.

If the solution is not optimal there must be some T_i less than T.

If in the optimal solution table there is even one closed rectangular polygon admitting the above movings, the optimal solution $\{x_{ij}\}$ will not be single one.

In optimal solution nm unknown quantities are connected by means of $n + m$ (Eqs. (1.3) and (1.5)). Since at $m > 2$, $n > 2$ the number of equations is less than the number of unknown quantities, the principal possibility of obtaining the sets of optimal solution is clear. To analyze the actual possibility of such solutions it is necessary to study the equation coefficients matrix (1.3) and (1.5) and the ability to satisfy the constraints (1.4). For the practical purposes it is usually more convenient to find in the solution table one of the closed rectangular contour in which the mentioned above movings are available without breaking the constraints (1.4). The vertices in these contours, where Δx_{kl} and x_{kl} are added up, alternate with the vertices, where Δx_{kl} is substracted from x_{kl}. It is necessary that the vertices having subtraction action to conform strictly to the positive values of x_{ij} and the value Δx_{kl} does not to exceed the smallest value x_{ij}.

2. The equivalent mathematical model and its properties

Theorem. To provide the existence of the value system $\{x_{ij}\}$, $i = \overline{1, n}$, $j = \overline{1, m}$, satisfying in the constraints

$$\sum_{i=1}^{u} K_i N_i x_{ij} = W_j, \qquad j = \overline{1, m} \tag{2.1}$$

$$x_{ij} \geq 0, \quad i = \overline{1, n}; \quad j = \overline{1, m} \tag{2.2}$$

and minimizing the value T (1.6) is necessary and sufficiently to have the existence of the function minimum.

$$\mathfrak{L} = \sum_{i=1}^{n} K_i N_i T_i^2 = \sum_{i=1}^{n} K_i N_i \left(\sum_{j=1}^{m} x_{ij} \right)^2 . \tag{2.3}$$

Proof. Let the values x_{ij}, $i = \overline{1, n}$; $j = \overline{1, m}$ be the problem solution of minimization of function $\mathfrak{L}$ and minimum value $\mathfrak{L}$ in this solution be equal to $\mathfrak{L}_0$.

The theorem proves this solution to be optimal relative to T as well. Let us assume the opposite: solution given here is not optimal relative to T and there exists another solution $\{x_{ij}^{(1)}\}$, where

$$\left.\begin{aligned} \mathfrak{L}^{(1)} &> \mathfrak{L}_0 \\ T_s^{(1)} &< T_s \end{aligned}\right\} \tag{2.4}$$

and

$$T_s = \max_i T_i = T.$$

If $\{x_{ij}\}$ is not optimal relative to T then as it is shown above, there exists if only one $T_l < T_s$ among T_i. To obtain the optimal solution we shall reduce T_s to the value $T_s^{(1)}$ and increase T_l to the same value, $T_s^{(1)}$, the constraints permitting or to the value $T_l^{(1)} < T_s^{(1)}$, if constraints do not permit. Thus, the system of unequalities has been obtained.

$$\left.\begin{aligned} T_s - T_s^{(1)} &= \Delta x_s > 0, \\ T_l - T_s &< 0, \end{aligned}\right\} \tag{2.5}$$

$$\left.\begin{aligned} T_l^{(1)} - T_l &= \Delta x_s > 0, \\ T_l^{(1)} - T_s^{(1)} &\leq 0. \end{aligned}\right\} \tag{2.6}$$

Let us make the difference from (2.3)

$$\Delta \mathfrak{L} = \mathfrak{L}^{(1)} - \mathfrak{L}_0 = k_l N_l (T_l^{(1)2} - T_l^2) + k_s N_s (T_s^{(1)2} - T_s^2) =$$

$$= k_l N_l \Delta x_l (T_l^{(1)} + T_l) - k_s N_s \Delta x_s (T_s^{(1)} + T_s);$$

or from (1.7)

$$\Delta \mathfrak{L} = k_s N_s \Delta x_s [T_l^{l1)} - T_s^{(1)} + T_l - T_s]. \tag{2.7}$$

As according to the conditions (2.5)—(2.6)

$$\Delta x_s > 0,\ T_l - T_s < 0,\ T_l^{(1)} - T_s^{(1)} \leq 0,$$

then

$$\Delta \mathfrak{L} < 0.$$

Thus, we have obtained the result: if $\{x_{ij}^{(1)}\}$ is optimal relative to T, then $\mathfrak{L}^{(1)} < \mathfrak{L}_0$, that contradicts the assumption (2.4). It means the supposition of solution to be optimal relative to T but not to $\mathfrak{L}$ is wrong. As $\Delta \mathfrak{L} = 0$, then the function value $\mathfrak{L}$ in the point optimal relative to T is less than any other function value in the point not optimal relative to T and in the optimal solution both T and L achieve the minimum simultaneously. Thus, the theorem has been proved.

According to the theorem the resource optimal control problem to meet criterion of the fastest jobs performing is equivalent to the resource optimal levelling problem by the quadratic goal function $\mathfrak{L}$. This means the problem to be solved by both methods: the linear and quadratic programming. The suggestion widely used that the linear methods are simpler than nonlinear ones in this case proved to be groundless.

Formulation of linear problem

Let us consider the linear function system

$$T_i(x) = \sum_{j=1}^{m} x_{ij}, \qquad i = \overline{1, n} \tag{2.8}$$

and the linear constraint system

$$W_j = \sum_{i=1}^{n} K_i N_i x_{ij}, j = 1, m. \qquad j = \overline{1, m}. \tag{2.9}$$

We have to find the solution $\{x^*_j\}$ satisfying these constraints, where

$$T(x^*) = \min_x \max_i |T_i(x)|. \tag{2.10}$$

This is the min max problem of Kantorovich's type which presents the piecewise linear programming problem. Taking the variable y

$$|T_i(x)| \leq y, \quad i = \overline{1, n} \tag{2.11}$$

we come to the linear programming problem which means minimizing the function

$$\mathfrak{L} = y \tag{2.12}$$

when

$$\left.\begin{aligned} \sum_{j=1}^{m} x_{ij} + y \geq 0, &\qquad i = \overline{1, n} \\ -\sum_{j=1}^{m} x_{ij} + y \geq 0, &\qquad i = \overline{1, n} \\ W_j = \sum_{i=1}^{n} K_i N_i x_{ij}, &\qquad j = \overline{1, m} \\ x_{ij} \geq 0, \qquad i = \overline{1, n}; &\qquad j = \overline{1, m} \end{aligned}\right\}. \tag{2.13}$$

Dimensionality of the problem $l = (m + 2n)mn$ is rather high and besides, as a set of calculations has shown the computing process often converges to the solution very slowly. It made us search the other ways. In [2] the method of quadratic programming, particularly the Beale method [3], is used for similar problems. It is connected with positive definiteness demand (or semidefiniteness) of the quadratic form $\mathfrak{L}$ in contrast to a lot of other variations of quadratic programming. The quadratic problem dimensionality $l = m^2 n$ is sufficiently lower than the equivalent linear problem dimensionality, the computing process converging hundred times as quick. Of course, to make the general conclusions concerning quantity estimation of efficiency methods, it is necessary to carry on the additional investigations.

For a particular case, when for resource using the constraints are imposed as follows

$$x_{ij} > 0, \quad i = \overline{1, n}; \quad j = \overline{1, m} \tag{2.14}$$

the optimal solutions is obtained at once without the programming procedure:

$$x_{ls}^{*} = \frac{W_s}{\sum_{i=1}^{n} K_i N_i}, \quad l = \overline{1, n}; \quad s = \overline{1, m}. \tag{2.15}$$

It comes from the Langrangian function

$$\mathfrak{L} = \sum_{i=1}^{n} K_i N_i \left(\sum_{j=1}^{m} x_{ij} \right)^2 - \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} K_i N_i x_{ij} - W_j \right). \tag{2.16}$$

The function extremum condition for $\mathfrak{L}$ is

$$2 \sum_{j=1}^{m} x_{ij} = \lambda_j, \quad i = \overline{1, n}; \quad j = \overline{1, m}. \tag{2.17}$$

Equation (2.17) is satisfied by a single set $\{x_{ij}^{*}\}$.

Particularly, if we take

$$x_{ls} = \frac{\lambda_s}{2m}; \quad l = \overline{1, n}; \quad s = \overline{1, m} \tag{2.18}$$

then by substituting (2.18) into (2.1) and by substituting its result again into (2.18) we obtain the expression (2.15).

Optimal levelling reduces the time wasted of a set of resources as much as possible. When levelling is ideal, i.e.

$$T_i = \text{const} = T.$$

The ideal time of all resources taken together is equal to zero.

Optimal solution $\{x_{ij}^{*}\}$ being not unique at the unique values $\{T_i^{*}\}$. It makes possible to select solutions optimal relative to additional criteria from a lot of solutions having equal constraints. In support of this the following example is given.

3. Example

The building company disposes three types of digging machines R_1, R_2, R_3 having productivity 100, 200 and 300 cubic metres, respectively. The quantities of machines N_1, N_2, N_3 are equal to 2, 1 and 1 respectively. It is necessary to dig out five foundation pits requiring the resources expenditure equal to 100, 120, 160, 200 and 300 machine time units, respectively, one

hour machine R_1 work being taken for a time unit. The constraints (2.1) having been simplified are reduced to

$$\left.\begin{aligned} x_{14} + x_{21} &= 50 \\ 2x_{14} + 3x_{34} &= 200 \\ 2x_{23} + 3x_{33} &= 160 \\ x_{32} &= 40 \\ 2x_{15} + 2x_{25} + 3x_{35} &= 300. \end{aligned}\right\} \tag{3.1}$$

The goal function becomes the following

$$\mathfrak{L} = 2(x_{11} + x_{14} + x_{15})^2 + 2(x_{21} + x_{23} + x_{25})^2 + \\ + 3(x_{32} + x_{33} + x_{34} + x_{35})^2. \tag{3.2}$$

The original data reduced to the basic machine R hours of action are put in Table 1. Blanks are taken for inadmissibility of utilizing the resources in respect with the technology process. Tables 2, 3 and 4 present the basis and goal

Table 1

R_i	N_i	K_i	W_j				
			100	120	160	200	300
R_1	2	1	x_{11}	—	—	x_{14}	x_{15}
R_2	1	2	x_{21}	—	x_{23}	—	x_{25}
R_3	1	3	—	x_{32}	x_{33}	x_{34}	x_{35}

Table 2

		x_{15}	x_{21}	x_{33}	x_{34}	x_{25}
x_{11}	50		−1			
x_{14}	100				− 3/2	
x_{23}	80			− 3/2		
x_{35}	100	− 2/3				− 2/3
$\mathfrak{L}$	116 600	+ 20	−140	+180	−30	−120
x_{15}	+ 20	+10/3	− 2	− 2	− 5	+ 4/3
x_{21}	−140	− 2	**+ 4**	− 3	+ 3	+ 2
x_{33}	+180	− 2	− 3	+ 15/2	+ 3	− 5
x_{34}	− 30	− 5	+ 3	+ 3	+15/2	− 2
x_{25}	−120	+ 4/3	+ 2	− 5	− 2	+ 10/3

+4 — basic element

Table 3

		x_{15}	U_1	x_{33}	x_{34}	x_{25}
x_{21}	35	+1/2	+1/4	+3/4	−3/4	−1/2
x_{11}	15	−1/2	−1/4	−3/4	+3/4	+1/2
x_{14}	100				−3/2	
x_{23}	80			−3/2		
x_{35}	100	−2/3				−2/3
£	111 700	−50		+75	+75	−50
x_{15}	−50	+7/3		−7/2	−7/2	+7/3
U_1			+1/4			
x_{33}	+75	−7/2		+21/4	+21/4	−7/2
x_{34}	+75	−7/2		+21/4	+21/4	−7/2
x_{25}	−50	+7/3		−7/2	−7/2	+7/3

Table 4

		U_2	U_1	x_{33}	x_{34}	x_{25}
x_{15}	21.4	+3/7		+3/2	+3/2	−1
x_{21}	45.7	−3/14	+1/4	+3/2		−1
x_{11}	4.3	−3/14	−1/4	−3/2		+1
x_{14}	100				−3/2	
x_{23}	80			−3/2		
x_{35}	85.7	−2/7		−1	−1	
£	110 628					
U_2		+3/7				
U_1			+1/4			
x_{33}						
x_{34}						
x_{25}						

function coefficients in optimization process according to Beale's algorithm [2], [3]. The vacant table cells are supposed to be zeros.

To get the optimal solution (Table 5) two iterations were required. The optimization has led to the ideal levelling where $T_1 = T_2 = T_3 = T = 125.7$ hours. According to Table 4 the goal function in optimal point does not depend

Table 5

	V_1	V_2	V_3	V_4	V_5	T_i
R_1	4.3	—	—	100	21.4	125.7
R_2	45.7	—	80	—	0	125.7
R_3	—	40	0	0	85.7	125.7

Table 6

R_1	4.3	—	—	0	121.4	125.7
R_2	45.7	—	80	—	—	125.7
R_3	—	40	0	$66^2/3$	$19^1/30$	125.7

Table 7

R_1	25.7	—	—	100	0	125.7
R_2	24.3	—	80	—	21.4	125.7
R_3	—	40	0	0	85.7	125.7

Table 8

R_1	50	—	—	0	75.7	125.7
R_2	0	—	80	—	45.7	125.7
R_3	—	40	0	$66^2/3$	$19^1/30$	125.7

Table 9

R_1	50	—	—	75.7	0	125.7
R_2	0	—	80	—	45.7	125.7
R_3	—	40	0	16.2	69.5	125.7

upon the variables x_{33}, x_{34}, x_{35} and changing them within the optimal basis we obtain the new optimal solutions.

Choice having many meanings makes it possible to use the additional optimization criteria. As far as every zero solution $x_{ij} = 0$ means no transportation of resource i to the object j, the maximum of optimal zero solutions is taken as the second stage criterion that corresponds to the minimum of resource transportations. Tables 6—9 give us all solutions having maximum

of zeros. There are five solutions and you can use one more criterion for the following variation choosing. Let us take the minimum time required for transportations to the objects as the third stage criterion. For this purpose we need the transportation graph. Let us assume the complete transportation graph is given (Fig. 1). In this graph every V_j has a node corresponding to index j and every link corresponds to the transportation time of resource t_{ij} from the object V_i to the object V_j. The problem is to find the transportation subgraph the using of which minimizes the function

$$t_i = \Sigma t_{k,l}\,, \qquad i = \overline{1, n}. \tag{3.3}$$

Taking into account the transportation time t_i and the using time for the resource R_i at the object V_j we find

$$T = \max_i \left\{ t_i + \sum_{j=1}^{m} x_{ij} \right\} = \max_i \{t_i\} + \max_i \left\{ \sum_{j=1}^{m} x_{ij} \right\}. \tag{3.4}$$

In the case the transportation is performed independently, i.e. t_i and x_{ij} do not influence each other, minimization can be done for components separately. Finding of $\min \max \Sigma x_{ij}$ is analyzed above.

Seeking for $\min \max t_i$ means finding the tree with minimum links length summed up inside the complete graph. For this solution it is possible to use Kruskal's algorithm [6] involving the consequent choice of the shortest links, which do not form the cycle with previously chosen links.

Taking into consideration the blanks (inadmissibility) of Table 5 and the existence of solution $x_{13} = 0$ in all other five tables we remove the nodes not admissable for each resource from the complete graph and obtain three initial data for graph optimization for resources R_1, R_2, R_3. Having remained the shortest links in it we shall have the trees shown in Fig. 2. To minimize the transportation time further we shall try to shorten the tree R_3 with the longest route by taking away the node 4, i.e. putting $x_{34} = 0$. Tables 5 and 7 satisfy this condition. Using any of them one will increase the work time by 10 hours because of transportations.

As and after this we have two more optimal variations. Let us use the criterion of the 4th stage which means reducing the member of the tone-

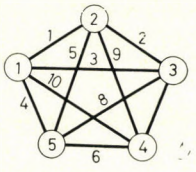


Fig. 1

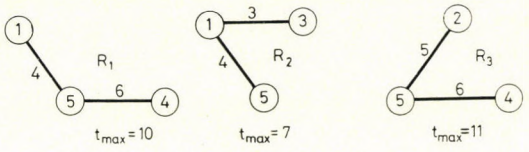


Fig. 2

kilometres. The weights and distances are not shown in our problem. Let us consider the weights of the load transported are proportional to productivity of the transported machines and the distances are proportional to the transportation times. Hence, according to Table 5 it is necessary to carry out the transportation of the resource $R_1 = 2$ for 10 distance units and the resource $R_2 = 2$ for 3 units, total $2 \times 10 + 2 \times 3 = 26$ and according to Table 3 the summed transportations will result in $2 \times 10 + 2 \times 7 = 34$. Thus, to meet criteria of all 4 stages the optimal solution is that given in Table 5. The corresponding transportation graphs are shown in Fig. 3, and the transportation schedules are shown in Fig. 4.

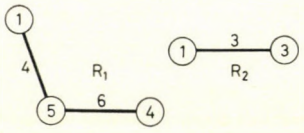


Fig. 3

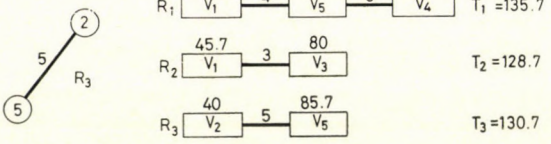


Fig. 4

References

1. *Воронов А. А., Петрушинин Е. П.* Задачи управления комплексом ресурсов на множестве объектов. Докл. 1 Всесоюзной конференции по исследованию операций. Минск, 1972.
2. *Beale, E. M.*, On Quadratic Programming. Naval Res. Logistic Quarterly, 6, 1959.
3. *Воронов А. А., Петрушинин Е. П.* Решение задачи оптимального распределения ресурсов методом квадратичного программирования. Автоматика и телемеханика, 11, 1966.
4. *Петрушинин Е. П.* Метод Лагранжа и задачи оптимального распределения ресурсов в системах сетевого планирования. Автоматика и телемеханика, 12, 1966.
5. *Петрушинин Е. П., Солодовник Р. П., Суни М. А.* Стандартная программа для решения задач квадратичного программирования на ЭВМ Минск-22. Карельское книжн. изд., Петрозаводск, 1968.
6. *Kruskal, L. B.*, On the Shortest Spanning Subtree of a Graph. Proc. Appl. Math. Soc., 7, 1956.

Оптимизация управления ресурсами многократного действия

А. А. ВОРОНОВ, Е. П. ПЕТРУШИНИН

(Владивосток — Москва)

Рассматривается задача управления комплексом ресурсов многократного действия на заданном множестве объектов, относящаяся к проблеме управления резервами. В качестве критерия оптимизации используется минимум времени выполнения работ на множестве объектов.

Соотношения (1.1), (1.2) определяют комплекс ресурсов R, участвующий в выполнении работ на множестве объектов V. Ограничения (1.3) задают трудоемкость объекта как линейную функцию времени использования ресурсов различного типа (при известном количестве и производительности ресурсов). Ограничения (1.4) являются технодогическими,

поскольку они характеризуют допустимость использования ресурса вида R_i на объекте V_j. Математически задача управления ресурсами состоит в нахождении минимума соотношения (1.6) при ограничениях (1.3)—(1.4).

В статье показано, что сформулированная задача может быть сведена к задаче кусочно-линейного программирования (2.12)—(2.13) размерности $l_2 = m^2 n + 2mn^2$ (m — число объектов, n — число видов ресурсов), однако более эффективный путь состоит в построении эквивалентной оптимизационной модели (2.1)—(2.3) с квадратичным критерием оптимизации. Доказана следующая теорема. Для существования системы величин $\{x_{ij}\}$, удовлетворяющих ограничениям (2.1), (2.2) и минимизирующих (1.6), необходимо и достаточно существование минимума квадратичной целевой функции (2.3). Оптимизационная модель с квадратичным критерием оптимизации имеет размерность $l_2 = m^2 n$ и, кроме того, скорость сходимости в квадратичном случае в сотни раз выше. Существующее предубеждение, что линейные методы проще нелинейных, в данном случае оказывается необоснованным.

В статье показано, что в частном случае, когда технологические ограничения заданы в виде (2.14), искомый оптимум может быть определен непосредственно по соотношению (2.15).

В заключение приведен пример оптимизации трех видов ресурсов на множестве из пяти объектов. Соответствующая оптимизационная модель записана в соотношениях (3.1), (3.2). Таблица 1 показывает технологические условия задачи, где прочерки соответствуют запретам использования ресурсов. Таблица 2 содержит запись исходной задачи в начальной точке. Таблицы 3 и 4 показывают процесс оптимизации, причем в таблице 4 записана оптимальная точка. Для завершения оптимизационного процесса потребовалось 2 итерации.

Таблица 5 показывает решение, полученное в процессе оптимизации. В таблицах 6—9 приведены новые оптимальные решения, полученные из исходного применением метода последовательной оптимизации (минимизация числа транспортных переброс ресурсов на множестве объектов).

Рисунки иллюстрируют оптимизацию по дополнительным критериям минимума времени транспортных перевозок на объекты и минимума числа тонно-километров. По критериям четырехступенчатой последовательной оптимизации оптимальным является решение, даваемое таблицей 5.

Е. П. Петрушинин
Институт электронных управляющих машин
СССР, Москва В-333
ул. Вавилова, д. 24

Problems of Control and Information Theory, Vol. 4 (4), pp. 289—297 (1975)

ON THE RATE DISTORTION FUNCTION OF STATIONARY SOURCES

K. MARTON

(Budapest)

(Received December 6, 1974)

The rate distortion function of (discrete or continuous time) stationary sources is expressed as the infimum of information rates over a set of pairs of random processes. For discrete time sources sufficient conditions are given under which infimum can be replaced by minimum.

1. Main results

Let us consider a (discrete or continuous time) stationary source with a single letter fidelity criterion [1, Section 7]. We denote by T the set of admissible values of time: T is the set of all integers (discrete time source) or the set of all real numbers (continuous time source). The input and reproducing alphabets of the source are complete separable metric spaces Z resp. U*, the distribution of the source is a stationary probability measure p on $X \triangleq \prod_{s \in T} Z$, and the single letter fidelity criterion is defined by a nonnegative measurable function ϱ on $Z \times U$. This source will be denoted by $[T, Z, U, \varrho, p]$.

Let $\xi = \{\xi(s): s \in T\}$ be a random process with values in Z and distributed according to p. If time is continuous then ξ is supposed to be a measurable process. Moreover, we assume that for some $s \in T$, $s \geq 0$, we have**

$$\lim_{t \to \infty} \frac{1}{t} I(\xi_{-\infty}^{0}, \xi_{s}^{t}) = 0 . \tag{1.1}$$

(Here we used the following notation. Let V be an arbitrary set, and*** $-\infty \leq s < t \leq \infty$. If $v_s = \{v(\tau): s < \tau \leq t\} \in \prod_{s<\tau\leq t} V$, and $s', t' \in T$, $s \leq s' < t' \leq t$, then we write $v_{s'}^{t'}$ for $\{v(\tau): s' < \tau \leq t\}$. If $t = \infty$ or $t' = \infty$ then the symbol $\leq$ in the definition of v_s^t resp. $v_{s'}^{t'}$ has to be replaced by $<$.)

Set

$$X_0^t = \prod_{0<s\leq t} Z, \qquad Y_0^t = \prod_{0<s\leq t} U \qquad (t > 0) .$$

* Complete separable metric spaces are also considered as measurable spaces, the σ-algebras of which consist of the Borel sets of the underlying space.

** I denotes mutual information.

*** s, t and τ always denote numbers in T; in this paragraph $s = -\infty$ and $t = \infty$ are allowed.

Define

$$\varrho_0^t(x_0^t, y_0^t) = \varrho(x_0^t, y_0^t) = \begin{cases} \dfrac{1}{t}\sum_{s=1}^{t} \varrho\big(x(s), y(s)\big), \\ \dfrac{1}{t}\int_0^t \varrho\big(x(s), y(s)\big), \end{cases}$$

corresponding to whether time is discrete or continuous. Here $x_0^t = \{x(s) : 0 < s \leq t\} \in X_0^t$, $y_0 = \{y(s) : 0 < s \leq t\} \in Y_0^t$, and x_0^t and y_0^t are assumed to be measurable functions of s in the case of continuous time. Let p_0^t $(t > 0)$ be the distribution induced by p on X_0^t.

The rate distortion function of the source $[T, Z, U, \varrho, p]$ is defined as

$$\mathfrak{R}(d) = \lim_{t\to\infty} \frac{1}{t} R^t(d), \quad d \geq 0,$$

where*

$$R^t(d) = \inf\{I(\xi_0^t, \eta_0^t) : \ P_{\xi_0^t} = p_0^t, \quad E\varrho(\xi_0^t, \eta_0^t) \leq d\}. \tag{1.2}$$

The inf in (1.2) is taken over all those random pairs (ξ_0^t, η_0^t) that take values in $X_0^t \times Y_0^t$, and for which ξ_0^t is distributed according to p_0^t and $E\varrho(\xi_0^t, \eta_0^t) \leq d$. If time is continuous then $\eta_0^t = \{\eta(s): 0 < s \leq t\}$ is supposed to be a measurable random process.

Let $\eta = \{\eta(s): s \in T\}$ be some random process (measurable if time is continuous) with values in U and such that the process (ξ, η) be stationary. (1.1) implies that the information rate

$$\mathfrak{I}(\xi, \eta) \triangleq \lim_{t\to\infty} \frac{1}{t} I(\xi_0^t, \eta_0^t)$$

exists, and**

$$\mathfrak{I}(\xi, \eta) = \frac{1}{t} EI(\xi_0^t, \eta \mid \xi_{-\infty}^0) = EI(\xi_0^1, \eta \mid \xi_{-\infty}^0)$$

for all $t > 0$. (Cf. [2, Theorem 7.4.2].)

For $d \geq 0$ denote by W_d the set of all stationary processes $(\xi, \eta) = \{\big(\xi(s), \eta(s)\big) : s \in T\}$ with values in $Z \times U$ and such that (i) ξ is distributed according to p; (ii) $E\varrho\big(\xi(0), \eta(0)\big) \leq d$.

Theorem 1. Let $[T, Z, U, \varrho, p]$ be a (discrete or continuous time) stationary source such that, if $\xi = \{\xi(s) : s \in T\}$ is a Z-valued random process distributed according to p, then (1.1) holds for some $s \geq 0$. Then for all $d \geq 0$

$$\mathfrak{R}(d) = \inf\{\mathfrak{I}(\xi, \eta): (\xi, \eta) \in W_d\}. \tag{1.3}$$

* E denotes mathematical expectation.
** EI denotes average conditional information.

Now we are going to formulate conditions under which inf in (1.3) can be replaced by min. For the rest of this section assume that time is discrete, i.e. T is the set of integers. Moreover, suppose that ϱ is continuous and satisfies the following

Compactness condition. For any compact set $C \subset Z$ and for any $M \geq 0$, the set

$$K(C, M) \triangleq \{(z, u) \in Z \times U\colon\ z \in C,\ \varrho(z, u) \leq M\}$$

is compact.

Theorem 2. Let $[T, Z, U, \varrho, p]$ a discrete time stationary source such that (i) if ξ is a Z-valued random process then (1.1) holds for some $s \geq 0$; (ii) ϱ is continuous and satisfies the compactness condition formulated above. Then for all $d \geq 0$

$$\mathcal{R}(d) = \min \{\mathcal{I}(\xi, \eta)\colon\ (\xi, \eta) \in W_d\}.$$

2. Proof of Theorem 1

We need the following

Lemma. Let ζ_1, ζ_2, ϑ_1 and ϑ_2 be random variables taking values in arbitrary measurable spaces. Suppose that ϑ_1, ζ_1, ζ_2 and ϑ_2 form a Markov chain (in this order). Then

$$I((\zeta_1, \zeta_2), (\vartheta_1, \vartheta_2)) \leq I(\zeta_1, \vartheta_1) + I(\zeta_2, \vartheta_2).$$

This lemma is proved (although not explicitely stated) in [3]. (Cf. Lemma 2.1 in [3].)

Proof of Theorem 1. The inequality

$$\mathcal{R}(d) \leq \inf \{\mathcal{I}(\xi, \eta)\colon\ (\xi, \eta) \in W_d\}$$

is obvious. In order to prove the reversed inequality, suppose that $\mathcal{R}(d) < \infty$, and let $\xi = \{\xi(s)\colon\ s \in T\}$ be a Z-valued random process distributed according to p. Choose an $\varepsilon > 0$, and let t be so large that the inequality

$$\frac{1}{t} R^t(d) < \mathcal{R}(d) + \varepsilon$$

holds. Let $\eta_0^t = \{\eta(s)\colon\ 0 < s \leq t\}$ be a U-valued random sequence or function, according to whether time is discrete or continuous; in the latter case let η_0^t be measurable. Suppose that

$$E\varrho(\xi_0^t, \eta_0^t) \leq d,$$

and

$$\frac{1}{t} I(\xi_0^t, \eta_0^t) < \mathcal{R}(d) + \varepsilon . \tag{2.1}$$

Define a U-valued random process $\eta = \{\eta(s)\colon s \in T\}$ such that: (i) for all $i = 0, \pm 1, \ldots$, the distribution of $(\xi_{(i-1)t}^{it}, \eta_{(i-1)t}^{it})$ is the same as that of (ξ_0^t, η_0^t); (ii) for fixed $\xi_{(i-1)t}^{it}$, $i = 0, \pm 1, \ldots$, $\eta_{(i-1)t}^{it}$ is conditionally independent of $(\xi, \eta_{-\infty}^{(i-1)t}, \eta_{it}^{\infty})$. (Observe that η is not stationary.)

Let ν be a random variable independent of (ξ, η) and uniformly distributed over the set $\{0, 1, \ldots, t-1\}$ if time is discrete, and over the interval $[0, t-1)$ if time is continuous. Define

$$\xi'(s) = \xi(s + \nu), \quad \eta'(s) = \eta'(s + \nu), \quad s \in T, \tag{2.2}$$

and set $\xi' = \{\xi'(s)\colon s \in T\}$, $\eta' = \{\eta'(s)\colon s \in T\}$.

We claim that $(\xi', \eta') \in W_d$. In order to give an idea how to carry out the proof ,we just show that the process (ξ', η') is stationary. We restrict ourselves to the continuous time case. Denote by ζ and ζ' the processes (ξ, η) and (ξ', η') respectively. Let n be a natural number, $t_1, t_2, \ldots, t_n \in T$, let A be a Borel set in $Z^n \times U^n$ and $s > 0$. Then

$$\Pr\{(\zeta'(t_1 + s), \ldots, \zeta'(t_n + s)) \in A\} =$$

$$= \frac{1}{t} \int_0^t \Pr\{(\zeta(t_1 + s + \tau), \ldots, \zeta(t_n + s + \tau)) \in A\}\, d\tau = \tag{2.3}$$

$$= \frac{1}{t} \int_s^{t+s} \Pr\{(\zeta(t_1 + \tau), \ldots, \zeta(t_n + \tau)) \in A\}\, d\tau =$$

$$= \frac{1}{t} \int_s^t + \frac{1}{t} \int_t^{t+s} = \frac{1}{t} \int_s^t + \frac{1}{t} \int_0^s = \tag{2.4}$$

$$= \frac{1}{t} \int_0^t \Pr\{(\zeta(t_1 + \tau), \ldots, \zeta(t_n + \tau)) \in A\}\, d\tau = \tag{2.5}$$

$$= \Pr\{(\zeta'(t_1), \ldots, \zeta'(t_n)) \in A\} .$$

((2.3) and (2.5) follow from (2.2) and from the independence of ν and (ξ, η), while (2.4) follows from the fact that the distribution of $(\zeta(t_1 + \tau), \ldots, \zeta(t_n + \tau))$ is the same as that of $(\zeta(t_1 + \tau - t), \ldots, \zeta(t_n + \tau - t))$. (2.5)

proves that $\zeta' = (\xi', \eta')$ is stationary. It can be proved in a similar way that ξ' is distributed according to p, and $E\varrho(\xi'(0), \eta'(0)) \leq d$. I.e. $(\xi', \eta') \in W_d$.

Now we are going to overbound $\Im(\xi', \eta')$. Observe first that ξ' is independent of ν. Indeed, let n be some natural number, $t_1, t_2, \ldots, t_n \in T$, $s \in T \cap$ $\cap [0, t)$ and let A be a Borel set in Z^n. Using (2.2), the independence of ξ and ν and the stationarity of ξ, we see that

$$\Pr\{(\xi'(t_1), \ldots, \xi'(t_n)) \in A \mid \nu = s\} = \Pr\{(\xi(t_1 + s), \ldots, \xi(t_n + s)) \in A\} =$$
$$= \Pr\{(\xi(t_1), \ldots, \xi(t_n)) \in A\}, \tag{2.6}$$

i.e. the conditional probability (2.6) does not depend on s.

Now we have for every natural number N

$$I(\xi_0'^{Nt}, \eta_0'^{Nt}) \leq I\big(\xi_0'^{Nt}, (\eta_0'^{Nt}, \nu)\big) = I(\xi_0'^{Nt}, \nu) + EI(\xi_0'^{Nt}, \eta_0'^{Nt} \mid \nu) \leq$$
$$\leq EI(\xi_0'^{Nt}, \eta_0'^{Nt} \mid \nu)\,. \tag{2.7}$$

Using the independence of (ξ, η) and ν, (2.2) implies that

$$EI(\xi_0'^{Nt}, \eta_0'^{Nt} \mid \nu) = \begin{cases} \dfrac{1}{t} \displaystyle\sum_{s=0}^{t-1} I(\xi_s^{Nt+s}, \eta_s^{Nt+s})\,, \\ \dfrac{1}{t} \displaystyle\int_0^t I(\xi_s^{Nt+s}, \eta_s^{Nt+s})\, ds\,, \end{cases}$$

corresponding to whether time is discrete or continuous, therefore in both cases

$$EI(\xi_0'^{Nt}, \eta_0'^{Nt} \mid \nu) \leq I(\xi_0^{(N+1)t}, \eta_0^{(N+1)t})\,. \tag{2.8}$$

From the definition of (ξ, η) and from the lemma stated above, we see that

$$I(\xi_0^{(N+1)t}, \eta_0^{(N+1)t}) \leq (N + 1)\, I(\xi_0^t, \eta_0^t)\,. \tag{2.9}$$

Comparing inequalities (2.7)–(2.9) yields

$$\Im(\xi'\, \eta') = \lim_{N \to \infty} \frac{1}{Nt} I(\xi_0'^{Nt}, \eta_0'^{Nt}) \leq$$
$$\leq \lim_{N \to \infty} \frac{1}{Nt} EI(\xi_0'^{Nt}, \eta_0'^{Nt} \mid \nu) \leq \frac{1}{t} I(\xi_0^t, \eta_0^t)\,. \tag{2.10}$$

(1.3) follows from (2.10) and (2.1).

3. Proof of Theorem 3

Suppose again that $\mathfrak{R}(d) < \infty$. We have to construct a pair of random processes $(\xi, \eta) \in W_d$ such that

$$\mathfrak{R}(d) = \mathfrak{I}(\xi, \eta). \tag{3.1}$$

Let us show first that all the functions ϱ_0^n, $n = 1, 2, \ldots$, satisfy the compactness condition imposed on ϱ. Let $C^n \subset X_0^n$ be compact and $M \geq 0$. It is easy to see that there exist compact sets $C_i \subset Z$, $i = 1, 2, \ldots, n$, such that $C^n \subset \coprod_{i=1}^{n} C_i$. Moreover, if for $x_0^n \in X_0^n$ and $y_0^n \in Y_0^n$ we have $\varrho_0^n(x_0^n, y_0^n) \leq M$, then $\varrho(x(i), y(i)) \leq nM$ for $i = 1, 2, \ldots, n$. Thus

$$K(C^n, M) \triangleq \{(x_0^n, y_0^n) \in X_0^n \times Y_0^n : x_0^n \in C^n, \varrho_0^n(x_0^n, y_0^n) \leq M\} \subset \coprod_{i=1}^{n} K(C_i, nM). \tag{3.2}$$

Since $K(C^n, M)$ is clearly closed, (3.2) implies the compactness of $K(C^n, M)$.

We need the following notation. If q is a probability measure on* $\coprod_{i=-\infty}^{\infty} V$, where V is a complete separable metric space, then we denote by q_0^n or by q the measure induced by q on $\coprod_{i=1}^{n} V$ $(n = 1, 2, \ldots)$.

Now let $\{(\xi_k, \eta_k): k = 1, 2, \ldots\} \subset W_d$ be a sequence of random processes such that

$$\mathfrak{R}(d) = \lim_{k \to \infty} \mathfrak{I}(\xi_k, \eta_k).$$

Denote by q_k the distribution of (ξ_k, η_k).

Let us show that the measures $(q_k)_0^1$, $k = 1, 2, \ldots$, are uniformly tight on $Z \times U$, i.e. for every $\varepsilon > 0$ there exists a compact set $K \subset Z \times U$ such that $q_k(K) > 1 - \varepsilon$ for all k. Choose a compact set $C \subset Z$ such that

$$p(C) > 1 - \frac{\varepsilon}{2},$$

and set $M = \dfrac{2d}{\varepsilon}$. We supposed that $K(C, M)$ is compact. Let us show that

$$q_k(K(C, M)) \geq 1 - \varepsilon, \quad k = 1, 2, \ldots.$$

Indeed,

$$q_k(Z \times U \setminus K(C, M)) = \Pr\{\xi_k(1) \notin C \quad \text{or} \quad \varrho(\xi_k(1), \eta_k(1)) > M\} \leq$$

$$\leq p(Z \setminus C) + \Pr\{\varrho(\xi_k(1), \eta_k(1)) > M\} \leq \frac{\varepsilon}{2} + \frac{d}{M} = \varepsilon.$$

* i always denotes integer numbers.

(The last inequality follows from the fact that $E\varrho(\xi_k(1), \eta_k(1)) \leq d$, and from Markov's inequality.)

It can be shown in the same way that for all n the measures $\{(q_k)_0^n : k = 1, 2, \ldots\}$ are uniformly tight on $X_0^n \times Y_0^n$.

Now [4, Theorem 6.7] implies that $\{q_k\}$ contains a subsequence $\{q_{k_j}\}$ with the following property: For all n the sequence $\{(q_{k_j})_0^n : j = 1, 2, \ldots\}$ weakly converges to some measure q_0^n on $X_0^n \times Y_0^n$; in symbols: $(q_{k_j})_0^n \Rightarrow q_0^n$. We shall suppose, to avoid double indexing, that

$$(q_k)_0^n \Rightarrow q_0^n, \tag{3.3}$$

as $k \to \infty$, for all fixed n. It is clear that the measures $\{q_0^n : n = 1, 2, \ldots\}$ satisfy Kolmogorov's consistency conditions, so they can be extended to a stationary probability measure q on $X \times Y$, where $Y = \prod_{i=-\infty}^{\infty} U$.

Let $(\xi, \eta) = \{(\xi(i), \eta(i)) : -\infty < i < \infty\}$ be a random process with values in $Z \times U$ and distributed according to q. We show that $(\xi, \eta) \in W_d$. It is clear that ξ is distributed according to p. Let us show that

$$E\varrho(\xi(1), \eta(1)) \leq d. \tag{3.4}$$

Set

$$\varrho_m(z, u) = \min \{m, \varrho(z, u)\}, \quad m = 1, 2, \ldots, z \in Z, u \in U.$$

The sequence $\varrho_m(z, u)$, $m = 1, 2, \ldots$, clearly increases to $\varrho(z, u)$ for all $z \in Z$, $u \in U$. Choose $\alpha < E\varrho(\xi(1), \eta(1))$. For m large enough we have

$$\alpha < E\varrho_m(\xi(1), \eta(1)),$$

and since ϱ_m is a bounded continuous function, it follows from (3.3) that for large enough k

$$\alpha < E\varrho_m(\xi_k(1), \eta_k(1)) \leq E\varrho(\xi_k(1), \eta_k(1)) \leq d.$$

(3.4) is proved.

Now we show that

$$\mathfrak{I}(\xi, \eta) \leq \mathfrak{R}(d). \tag{3.5}$$

Choose an $\varepsilon > 0$ and an s for which (1.1) holds. Let t be so large that

$$\frac{1}{t} I(\xi_s^t, \xi_{-\infty}^0) \leq \varepsilon,$$

and

$$\mathfrak{I}(\xi, \eta) \leq \frac{1}{t-s} I(\xi_0^{t-s}, \eta_0^{t-s}) + \varepsilon. \tag{3.6}$$

Write $n = t - s$. It is proved in [5, Lemma 2.2] that mutual information is lower semicontinuous with respect to weak convergence of measures. Therefore (3.3) implies that for large enough k

$$\frac{1}{n} I(\xi_0^n, \eta_0^n) \leq \frac{1}{n} I\big((\xi_k)_0^n, (\eta_k)_0^n\big) + \varepsilon . \tag{3.6}$$

On the other hand, we have

$$\begin{aligned} I((\xi_k)_0^n, (\eta_k)_0^n) &\leq I\big((\xi_k)_s^t, (\eta_k, (\xi_k)_{-\infty}^0)\big) = I((\xi_k)_s^t, (\xi_k)_{-\infty}^0) + EI((\xi_k)_s^t, \eta_k \mid (\xi_k)_{-\infty}^0) = \\ &= I(\xi_s^t, \xi_{-\infty}^0) + EI((\xi_k)_s^t, \eta_k \mid (\xi_k)_{-\infty}^0) \leq t\varepsilon + EI((\xi_k)_0^t, \eta_k \mid (\xi_k)_{-\infty}^0) = \\ &= t\varepsilon + t\,\vartheta(\xi_k, \eta_k) . \end{aligned} \tag{3.8}$$

Comparing (3.6)-(3.8), we see that for large enough t and k

$$\vartheta(\xi, \eta) \leq \frac{t}{t - s} [\vartheta(\xi_k, \eta_k) + 3\varepsilon] .$$

(3.5) is proved, and the proof of the theorem is completed.

Acknowledgement. The author is indebted to Dr. I. Csiszár for calling her attention to the problem and for his valuable remarks.

References

1. *Berger, T.*, Rate Distortion Theory: A Mathematical Basis for Date Compression. Prentice–Hall, Inc. Englewood Cliffs, New Jersey, 1971.
2. *Pinsker, M. S.*, Information and Information Stability of Random Variables and Processes. Holden–Day, Inc. San Francisco etc. 1964.
3. *Pinsker, M. S.*, Sources of messages, Problemy Peredachi Informatsii. **14**, 5–20. (In Russian).
4. *Parthasarathy, K. R.*, Probability Measures on Metric Spaces. Academic Press, New York–London, 1967.
5. *Csiszár, I.*, An extremum problem of information theory, Studia Math. Sci. Hung., (to be published).

Об ε-энтропии стационарных источников

К. МАРТОН

(Будапешт)

Рассматривается стационарный источник дискретного (или непрерывного) времени. Пусть T — совокупность допустимых значений времени. Предполагается,что входной и воспроизводящий алфавиты источника Z и U — полные сепарабельные метрические пространства, причем распределение на входе задаётся стационарной вероятностью p на $\prod_{s \in T} Z$, а критерий точности воспроизведения — измеримой неотрицательной функцией ϱ на $Z \times U$.

Пусть $\xi = \{\xi(s): s \in T\}$ — случайный процесс со значениями в Z и с распределением p. Предположим, что для некоторого $s \in T$, $s \geq 0$

$$\lim_{t \to \infty} \frac{1}{t} I(\xi_{-\infty}^{0}, \xi_{s}^{t}) = 0.$$

Здесь

$$\xi_s^t = \{\xi(\tau): s < \tau \leq t\}, \quad \xi_{-\infty}^{0} = \{\xi(\tau): -\infty < \tau < 0\},$$

а через $I(\cdot)$ обозначается взаимная информация.

Пусть $d \geq 0$. Обозначим через W_d совокупность стационарных процессов $(\xi, \eta) = \{(\xi(s), \eta(s)): s \in T\}$ со значениями в $Z \times U$ и таких, что ξ распределено по закону p и имеет место $E\varrho(\xi(0), \eta(0)) \leq d$. Обозначим через $\mathcal{R}(\cdot)$ ε-энтропийную функцию рассматриваемого источника. Тогда справедлива теорема.

Теорема 1. Для любого $d \geq 0$

$$\mathcal{R}(d) = \inf \{\mathfrak{I}(\xi, \eta): \quad (\xi, \eta) \in W_d\},$$

где $\mathfrak{I}(\xi, \eta)$ — информационная скорость процессов ξ и η.

В случае, когда время дискретно, а функция ϱ непрерывна и удовлетворяет следующему условию компактности.

Теорема 2. При условиях, сформулированных выше, для любого $d \geq 0$ имеет место

$$\mathcal{R}(d) = \min \{\mathfrak{I}(\xi, \eta): (\xi, \eta) \in W_d\}.$$

K. Marton
Mathematical Research Institute
Hungarian Academy of Sciences
H–1053 Budapest, V. Reáltanoda u. 13—15
Hungary

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 4 (4), pp. 299–310 (1975)

О СТОИМОСТИ ПЕРЕДАЧИ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Г. Л. ХОДАК

(Омск)

(Поступила в редакцию 12 мая 1974 г.)

Рассматривается задача передачи сообщений по каналу без шума. Источник сообщений есть произвольный дискретный стохастический процесс, а кодирование есть любое измеримое отображение множества входных последовательностей в выходные. Стоимость передачи определяется как среднее число выходных букв, необходимых для декодирования одной буквы сообщения. Такое определение отличается от традиционного. Доказывается, что при конечноавтоматном кодировании новая стоимость равна традиционной, если кодирование дешифруемо, и бесконечна в противном случае. Обобщается основная теорема для каналов без шума, по-видимому, на максимально широкий класс источников и кодирований.

1. Введение

К. Шенноном [1] была сформулирована основная теорема для каналов без шума, утверждающая, что стоимость передачи (обратная величина скорость) не может превосходить энтропии источника, делённой на логарифм числа букв в выходном алфавите, но может быть сделана как угодно близкой к этой величине. Отрицательная часть утверждения теоремы доказана: в [1] для кодирования любого источника автоматом, замещающим букву буквой; для блочного префиксного кодирования бернуллиевского источника в [2]; для произвольного блочного кодирования бернуллиевского источника в [3]; эти результаты обобщаются на стационарный источник [4]; для конечноавтоматного кодирования марковского источника в [5]; для стационарного источника в [6]; для весьма широкого класса источников и отображений входных слов в выходные в [7]. Во всех этих работах кодирование считается в том или ином смысле дешифруемым. Положительная часть теоремы доказана в [1] для стационарного источника. Кодирования со стоимостью равной нижней границе построены: в [8] для стационарного источника, в [7] для более широкого класса источников.

В упомянутых работах под стоимостью передачи понимается среднее число букв выходного алфавита, выработанных кодером в ответ на одну входную. Это показатель работы кодера. Возможна такая ситуация, что стоимость конечна (скорость не нулевая), но ни по какому коду нельзя восстановить сообщение. Так как стоимость передачи характеризует кодирование

не полностью, то от него требуется та или иная дешифруемость, т. е. возможность восстановления сообщения по его коду. Однако дешифруемость не гарантирует, что декодирование может вестись с той же скоростью, что и порождение кода.

В настоящей работе стоимостью передачи названо среднее число букв выходного алфавита, необходимых для декодирования одной буквы входного. Из определения видно, что стоимость будет служить показателем работы всей системы связи, а не только кодера. Так как стоимость определена через декодирование, никаких условий дешифруемости на кодирование не накладывается.

Под источником сообщений ниже понимается любой дискретный стохастический процесс, выборочными функциями которого (сообщениями) являются последовательности букв конечного или счётного входного алфавита, под кодированием — измеримое отображение входных последовательностей в выходные. Определение, наиболее близкое к этому, сформулировано в [6, 8]. Там кодированием считается либо стационарное, либо монотонное отображение последовательностей. В общем случае, как известно, даже блочное кодирование не является кодированием в этом смысле. В работах [1—5, 7] кодированием называется то или иное отображение множества слов во входном алфавите в множество слов в выходном. При условии дешифруемости отображение слов индуцирует единственное измеримое отображение последовательностей [7], и поэтому его можно считать частным случаем кодирования, определенного в настоящей работе. В то же время не всякое отображение последовательностей может быть индуцировано отображением слов. Для каналов без шума интересен случай только конечного выходного алфавита. В противном случае бесконечна пропускная способность. Из конечности выходного алфавита следует, что входной алфавит не может быть более, чем счётным. Без условия измеримости отображения последовательностей (кодирования) нельзя говорить о вероятностях слов на выходе кодера. Поэтому класс рассматриваемых источников и кодирований является максимально широким.

Доказывается (п. 6), что обычно традиционная и введённая стоимости совпадают, в частности: при дешифруемом конечноавтоматном кодировании марковских источников (при небольших ограничениях (теорема 1)); побуквенном дешифруемом кодировании бернуллиевских источников; при «декодируемом» [7] кодировании.

Показывается, что из конечности стоимости следует дешифруемость в обычных смыслах (п. 7). При конечной стоимости кодирование взаимно однозначно почти всюду (теорема 2), т. е. дешифруемо в смысле [6]. При конечноавтоматном кодировании марковского источника (с небольшими ограничения-

ми) стоимость конечна тогда и только тогда, когда кодирование дешифруемо (теорема 3).

Обобщается (п. 8) основная теорема для каналов без шума на описанный выше класс источников и кодирований. Для получения из доказанной теоремы результатов других авторов нужно установить равенство стоимостей. Как следствие получаются результаты [1—4], [7] для «декодируемого» кодирования и [5] при небольших ограничениях. Положительная часть теоремы доказывается аналогично [7].

2. Обозначения

Пусть $\mathfrak{A}$ — некоторый алфавит; $\mathfrak{A}^*$ — множество всех слов в алфавите $\mathfrak{A}$; Λ — пустое слово; $|A|$ — число букв в слове A; $\mathfrak{A}^\infty$ — множество всех последовательностей вида $\{a_i\}_{i=1}^{\infty}(a_i \in \mathfrak{A} : i = 1, 2, \ldots)$. Если $A_1 \in \mathfrak{A}^*$ и $A_2 \in \mathfrak{A}^* \cup \mathfrak{A}^\infty$, то $A_1 \cdot A_2$ означает результат приписывания A_2 справа к A_1. Если $A \in \mathfrak{A}^*$, $V \subset \mathfrak{A}^*$ и n — натуральное число, то $A^n = A \cdot A \ldots A$ (n раз), $A^0 = \Lambda$, $V^n = \{A_1 \ldots A_n : A_1, \ldots, A_n \in V\}$,

$$V^0 = \{\Lambda\},\ A^\infty = A \cdot A \ldots,\ V^\infty = \{A_1 A_2 \ldots : A_1, A_2, \ldots \in V\}.$$

Запись $A_1 < A_2$, где $A_1 \in \mathfrak{A}^*$ и $A_2 \in \mathfrak{A}^* \cup \mathfrak{A}^\infty$ означает что A_1 есть префикс (начало) A_2. Множество $V \subset \mathfrak{A}^*$ называется префиксным, если ни для каких двух различных слов из V одно не может быть началом другого. Если $A \in \mathfrak{A}^*$, $V \subset \mathfrak{A}^* \cup \mathfrak{A}^\infty$, то $A \prec V$ означает, что слово A есть префикс всех элементов множества V. $\| V \|$ — мощность множества V, $\varphi^{-1}(V)$ — полный прообраз V при отображении φ.

Всюду далее $\mathfrak{A}$ — конечный или счетный входной алфавит, $\| \mathfrak{A} \| \geq 2$; $\mathfrak{B}$ — конечный выходной алфавит мощности m, $2 \leq m < \infty$; $\sigma(\mathfrak{A})$ — некоторая σ — алгебра подмножеств $\mathfrak{A}^\infty$, содержащая все конечномерные цилиндры т. е. $A\,\mathfrak{A}^\infty \in \sigma(\mathfrak{A})$ для любого слова $A \in \mathfrak{A}^*$; $\sigma_0(\mathfrak{B})$ — σ — алгебра подмножеств $\mathfrak{B}^\infty$, порожденная всеми конечномерными цилиндрами; p — вероятностная мера на $\sigma(\mathfrak{A})$.

3. Источник сообщений и его энтропия

Источником по определению будем называть вероятностное пространство $\{\mathfrak{A}^\infty, \sigma(\mathfrak{A}), p\}$. Сообщением назовем последовательность в алфавите $\mathfrak{A}$, т. е. элементарное событие вероятностного пространства $\{\mathfrak{A}^\infty, \sigma(\mathfrak{A}), p\}$. Меру множества вида $A\,\mathfrak{A}^\infty$, где $A \in \mathfrak{A}^*$, обозначим через $p(A)$ и назовем вероятностью слова A. Назовем источник положительным, если положительны вероятности всех слов.

Обычным образом определим энтропию префиксов длины n ($n = 1, 2, \ldots$) $H(n) = -\sum_{A \in \mathfrak{A}^n} p(A) \ln p(A)$.* Функция $H(n)$ — монотонно неубывающая. Верхней и нижней энтропиями источника назовем: $\overline{H} = \overline{\lim_{n\to\infty}} \frac{1}{n} H(n)$ и $H = \underline{\lim_{n\to\infty}} \frac{1}{n} H(n)$, которые всегда существуют. Если $\overline{H} = H$, то говорят [5, 9], что источник имеет энтропию $H = \overline{H} = H$.

4. Кодирование

Кодированием назовем отображение $\varphi: \mathfrak{A}^\infty \to \mathfrak{B}^\infty$, измеримое относительно $\sigma(\mathfrak{A})$ и $\sigma_0(\mathfrak{B})$, то есть $\forall V \in \sigma(\mathfrak{B})\left(\varphi^{-1}(V) \in \sigma(\mathfrak{A})\right)$. Кодирования, совпадающие п. в. (почти всюду) назовем эквивалентными. Пусть $\sigma(\mathfrak{B}) = \{V \subset \mathfrak{B}^\infty : \varphi^{-1}(V) \in \sigma(\mathfrak{A})\}$. Как известно [9, стр. 54], $\sigma(\mathfrak{B})$ есть σ-алгебра. По определению кодирования $\sigma_0(\mathfrak{B}) \subset \sigma(\mathfrak{B})$. Кодирование переносит меру p с $\sigma(\mathfrak{A})$ на $\sigma(\mathfrak{B})$ [9, стр. 54]. По определению мера множества $V \in \sigma(\mathfrak{B})$ равна мере его полного прообраза, т. е. $p(\varphi^{-1}(V))$.

5. Стоимость передачи

Естественно считать, что префикс A некоторого сообщения A' декодируется по префиксу B последовательности $\varphi(A')$, если п. в. сообщения, коды которых начинаются в B, начинаются с A, то есть $A \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty)$ п. в. Такое определение декодируемости соответствует случаю, когда буквы кода сообщения поступают в декодирующее устройство равномерно, без каких-либо пробелов между словами. Значит для кодирований вида $\psi: \mathfrak{A}^* \to \mathfrak{B}^*$ при получении слова B не известно, является ли оно кодом некоторого слова A $(B = \psi(A))$ или префиксом некоторого другого слова A', для которого $A \not\prec A'$ и $B \prec \psi(A')$.

Для $n = 1, 2, \ldots$ и любой последовательности $B' \in \varphi(\mathfrak{A})^\infty$ положим $t(n, B')$ равным минимальной длине префикса последовательности B', по которому можно декодировать некоторое слово длины n. Если такого префикса нет, то $t(n, B') = +\infty$. Формально: положим

$$t(n, B') = \begin{cases} \min\limits_{B \in V(n, B')} |B|, & \text{если } V(n, B') = \varnothing; \\ +\infty, & \text{если } V(n, B') = \varnothing, \end{cases}$$

* При $p = 0$ произведение $p \cdot \ln p$ считается равным нулю. Сумма всегда определена, так как число слагаемых не более, чем счетно, и все они одного знака. Величина $H(n)$ может равняться ∞.

где

$$V(n, B') = \{B \in \mathfrak{B}^*: B \prec B' \,\&\, \exists\, A \in \mathfrak{A}^*(|A| = n \,\&\, A \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty)\ n.\ b.)\}.$$

Легко проверяется, что при любом натуральном n функция $t(n)$, определена п. в. на $\mathfrak{B}^\infty$ и измерима относительно $\sigma(\mathfrak{G})$. Если кодирования φ_1 и φ_{21} эквивалентны, то соответствующие им функции t_{φ_1} и t_{φ_2} равны.

Обозначим через $C(n)$ среднее число букв выходного алфавита, необходимых для декодирования первых n букв сообщения: $C(n) = \int\limits_{\mathfrak{B}^\infty} t(n,)\ dp =$ $= \int\limits_{\mathfrak{A}^\infty} t(n, \varphi(\ ))\, dp^*$ $(n = 1, 2, \ldots)$.* Функция $C(n)$—монотонно неубывающая. Верхней и нижней стоимостями передачи назовем:

$$\overline{C} = \overline{\lim_{n\to\infty}} \frac{1}{n} C(n) \text{ и } \underline{C} = \underline{\lim_{n\to\infty}} \frac{1}{n} C(n).$$

Эти пределы всегда существуют. Если кодирования φ_1 и φ_2 эквивалентны, то соответствующие им $\overline{C}_{\varphi_1}$, $\overline{C}_{\varphi_2}$ и $\underline{C}_{\varphi_1}$, $\underline{C}_{\varphi_2}$ попарно равны. Если $\overline{C} = \underline{C}$, то будем говорить, что стоимость передачи есть $C = \overline{C} = \underline{C}$.

При рассмотрении кодирований вида $\psi : \mathfrak{A}^* \to \mathfrak{B}^*$ стоимость передачи обычно называют $L = \lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{A \in \mathfrak{A}} p(A) \mid \psi(A) \mid$ (предполагается существование предела) [5, 7]. В простых случаях за определение принимают явное выражение предела [1—4].

Проиллюстрируем определение стоимости. Пусть $\mathfrak{A} = \{\alpha, \beta, \gamma\}$, $\mathfrak{B} = (0, 1\}$, $\sigma(\mathfrak{A}) = \sigma_0(\mathfrak{A})$, источник бернуллиевский $p(\alpha) = p(\beta) = p(\gamma) = \frac{1}{3}$, кодирование побуквенное $\varphi(\alpha) = 0$, $\varphi(\beta) = 01$, $\varphi(\lambda) = 11$, $\varphi(a_1 a_2 \ldots) = \varphi(a_1)$ $\varphi(a_2) \ldots$. Пусть $B' = 0^{k_1} \cdot 1^{k_2} \cdot 0^{k_3} \ldots$ — некоторая последовательность из $\varphi(\mathfrak{A}^\infty)$. Если $k_1 = 0, k_2 \neq 0$, то $V(1, \beta') = \{1, \ldots, 1^{k_2}, 1^{k_2}0, \ldots\}$ и $t(1, \beta') = 1$, так как $\gamma \prec \varphi^{-1}(1 \cdot \mathfrak{B}^\infty)$ и I — минимальный профикс B', по которому декодируется некоторая буква. Если $k_1 \geq 2$, то $\big(V(1, \beta^1) = \{00, \ldots, 0^{k_1}, 0^{k_1}1, \ldots\}, t(1, B' =$ $= 2$, т. к. $\alpha \prec \varphi^{-1}(00 \cdot \mathfrak{B}^\infty)$ и $\varphi^{-1}(0 \cdot \mathfrak{B}^\infty) = \alpha \cdot \mathfrak{A}^\infty \cup \beta \cdot \mathfrak{A}^\infty$ (по профиксу 0 нельзя декодировать ни одной буквы). Если $k_1 = 1$, k_2 — четное, то $V(1, B') =$ $= \{0 \cdot 1^{k_2}0, 0\, 1^{k_2}00, \ldots\}$ и $t(1, B') = k_2 + 2$, так как $\alpha \prec \varphi^{-1}(0 \cdot 1^{k_2} \cdot 0 \cdot \mathfrak{B}^\infty)$ и $\varphi^{-1}(0 \cdot 1^{k_2} \cdot \mathfrak{B}^\infty) = \alpha\gamma^{k_2/2} \cdot \mathfrak{A}^\infty \cap \beta \cdot \gamma^{k_2/2} \cdot \mathfrak{A}^\infty$. Если $k_1 = 1$ и k_2 — неточное, то аналогично $V(1, B') = \{0, 1^{k_2}0, \ldots\}$ и $t(1, B') = k_2 + 2$. Если $B' = 0 \cdot 1^\infty$, то $V(1, B') = \varnothing$ и $t(1, B') = \infty$. Имеем $p(1 \cdot B^\infty) = p\big(\varphi^{-1}(1 \cdot B^\infty)\big) = p(\gamma \cdot$ $\cdot \mathfrak{A}^\infty) = \frac{1}{3}$, $p(00 \cdot B^\infty) = p(\alpha^2 \cdot \mathfrak{A}^\infty \cup \alpha\beta\, \mathfrak{A}^\infty) = \frac{2}{y}$, при $k = 1, 2, \ldots$ $p(0 \cdot$

* Интеграл определен даже если $t(n) = \infty$ на множестве положительной меры, так как $t \geq 0$. В этом случае $C(n) = +\infty$.

$\cdot 1^{2k} \cdot 0 \cdot \mathfrak{B}^\infty) = p(\alpha\gamma^k \cdot \alpha \cdot \mathfrak{A}^\infty \cup \alpha \cdot \gamma^k \cdot \beta \cdot \mathfrak{A}^\infty) = \frac{2}{3^{k+2}}$, $p(0 \cdot 1^{2k-1} \cdot 0 \cdot \mathfrak{B}^\infty) =$
$= p(\beta \cdot \gamma^{k-1} \cdot \alpha \cdot \mathfrak{A}^\infty \cup \beta \cdot \gamma^{k-1} \cdot \beta \cdot \mathfrak{A}^\infty) = \frac{2}{3^{k+1}}$. Поэтому $C(1) = 1 \cdot \frac{1}{3} +$
$+ 2 \cdot \frac{2}{9} \sum_{k=1}^{\infty} (2k+1) \frac{2}{3^{k+1}} + \infty \cdot 0 = \frac{16}{9}$. Для декодирования первой буквы сообщения требуется в среднем $\frac{16}{9}$ выходных букв. Входное же устройство выработает в ответ на первую входную $L = \frac{1}{3} + 2 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3} = \frac{5}{9}$, что меньше предыдущей величины. Расхождение объясняется тем, что, зная только код буквы α, нельзя сказать, с какой буквы начинается сообщение. Легко проверить, что при любых: натуральном n, слове $A \in \mathfrak{A}^\infty$ и последовательности $B' = \varphi(A) \cdot B'' \in (A \cdot \mathfrak{A}^\infty)$ $t(n+1, B') = |\varphi(A)| + t(2, B'')$.

Поэтому $C(n+1) = nL + C(1)$.

6. Связь между различными определениями стоимости

Лемма 1. Если источник положительный и определена стоимость L, то $C \geq L$.

Утверждение леммы следует из того, что никакое слово $a_1 \ldots a_n \in \mathfrak{A}^*$, не может быть декодировано по $\psi(a_1 \ldots a_{n-1})$.

Просто доказывается, что если кодирование «декодируемо» в смысле [7], то $\overline{C} \leq L$. Если к тому же источник положительный, то в силу леммы 1 $C = L$.

В работе [5] доказано, что для дешифруемого конечноавтоматного кодирования марковских источников существует стоимость L. Определения, относящиеся к автоматному кодированию, стандартны и приведены в работе ватора [10]; здесь они не повторяются.

Теорема 1. Если источник марковский произвольного порядка связности и положительный, автоматное кодирование дешифруемо и автомат сильно связный, то стоимость C существует и равна стоимости L.

Доказательство. Пусть выполнены все условия теоремы, $\varphi : \mathfrak{A}^* \cup \mathfrak{A}^\infty \to$
$\to \mathfrak{B}^* \cup \mathfrak{B}^\infty$ — автоматное кодирование и U — универсальное декодирующее слово, существование которого доказано в лемме 5 [10]. Положим

$$P = \min_{A \in \mathfrak{A}^*} \frac{p(A \cdot U)}{p(A)} = \min_{A \in \mathfrak{A} \cup \{\lambda\}} \frac{p(A \cdot U)}{p(A)} > 0. \quad (6.1)$$

* При рассмотрении конечноавтоматного кодирования алфавит $\mathfrak{A}$ считается конечным. Будем считать, что $U \neq \lambda$, в противном случае в качестве U можно взять любое непустое слово.

Фиксируем произвольно натуральное число n. Каждую последовательность $a_1 \cdot a_2 \ldots \in \mathfrak{A}^\infty$ представим в виде $a_1 \ldots a_n \cdot A_1 \cdot A_2 \ldots$, где $|A_1| = |A_2| = \ldots = |U|$. Если при некотором i (а такое i найдется с вероятностью единица) $A_i = U$, то $a_1 \ldots a_n$ (и даже $a_1 \ldots a_n A_1 \ldots A_{i-1}$) декодируется по $\varphi(a_1 \ldots a_n \cdot A_1 \ldots A_i)$ и поэтому

$$t(n, \varphi(A)) \leq |\varphi(a, \ldots a_n \cdot A_1 \ldots A_i)| \leq |\varphi(a_1 \ldots a_n)| + M \cdot i|U|, \quad (6.2)$$

где M — длина максимального слова, сопоставляемого одной букве.

Интегрируем почленно неравенство (6.2)

$$C(n) \leq \sum_{A \in \mathfrak{A}^n} p(A)\,|\varphi(A)| + M \cdot |U| \cdot \sum_{i=0}^{\infty} (i+1)(1-P)^i =$$

$$= \sum_{A \in \mathfrak{A}^n} p(A)\,|\varphi(A)| + M \cdot |U| \cdot P^{-2}. \quad (6.3)$$

Делим обе части неравенства (6.3) на n и переходим к пределу при $n \to \infty$, получаем $\overline{C} \leq L$.

Таким образом, из леммы 1 следует утверждение теоремы.

Следствие 1. Если источник бернуллиевский (не обязательно положительный), входной алфавит конечен, а кодирование побуквенное и дешифруемое в смысле [3], то стоимость C существует и равна стоимости L.

Для доказательства достаточно рассмотреть кодирование на подалфавите из букв с ненулевыми вероятностями.

7. Конечность стоимости и дешифруемость

Теорема 2. Если нижняя стоимость конечна, то сужение кодирования на некоторое множество полной меры взаимно однозначно.

Эскиз доказательства. Из определения стоимости и условия теоремы следует, что можно выбрать множество $V \subset \mathfrak{A}^\infty$ так, что для всех $A = a_1 a_2 \ldots \in V$, всех $n = 1, 2, \ldots$ и всех $B \in V(n, \varphi(A))$ выполняются соотношения $t(n, \varphi(A)) < \infty$; $a_1 \ldots a_n \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty) \cap V$ (*bcygy*); $p(V) = 1$. Отображение $\varphi: V \to \varphi(V)$ будет взаимно однозначным.

Кодирование $\varphi: \mathfrak{A}^\infty \to \mathfrak{B}^\infty$ было названо П. Биллингслеем [6] дешифруемым, если оно взаимно однозначно п. в.

Следствие 2. Если нижняя стоимость конечна, то существует эквивалентное кодирование, взаимно однозначное п. в. (дешифруемое в смысле [6]).

Доказательство. По предыдущей теореме отображение $\varphi: V \to \varphi(V)$ взаимно однозначно. Если $V = \mathfrak{A}^\infty$, то следствие выполняется. Пусть $V \neq \mathfrak{A}^\infty$.

Изменим кодирование на $\mathfrak{A}^\infty \setminus V$. Всегда найдется последовательность $B \in \mathfrak{B}^\infty$ вероятности ноль. Положим $\varphi(A) = B$ для всех $A \in \mathfrak{A}^\infty \setminus V$. Тогда любая последовательность из $V \setminus \{\varphi^{-1}(B)\}$ однозначно определяется по своему коду и $p(V) \setminus \{\varphi^{-1}(B)\}) = 1$. Следствие доказано.

Теорема 3. Если источник марковский произвольного конечного порядка связности и положительный, кодирование осуществляется сильно связным автоматом, то оно дешифруемо тогда и только тогда, когда стоимость C конечна (определения приведены в [10]).

Доказательство. Пусть выполнены все условия теоремы. По теореме 1 для дешифруемого кодирования $C = L < \infty$. Покажем, что для не дешифруемого кодирования $C = \infty$. Пусть $t(A)$ — заключительное состояние автомата при подаче на вход слова $A \in \mathfrak{A}^*$. Если кодирование не дешифруемо, то найдутся слова $A_1, A_2 \in \mathfrak{A}^*$, для которых

$$A_1 \neq A_2, \quad \varphi(A_1) = \varphi(A_2), \quad t(A_1) = t(A_2). \tag{7.1}$$

Из (18) следует, что при любой последовательности $A \in \mathfrak{A}^\infty$

$$\varphi(A_1 \cdot A) = \varphi(A_2 \cdot A). \tag{7.2}$$

Из (19) и положительности вероятностей $pA_{(1)}$ и $p(A_2)$ следует, что не существует взаимно однозначного сужения кодирования на множество полной меры. Поэтому в силу теоремы $2\, C = C = \infty$. Теорема доказана.

Следствие 3. Если источник бернуллиевский, входной алфавит конечен и кодирование побуквенное, то оно дешифруемо в смысле [3] тогда и только тогда, когда стоимость C конечна.

Для доказательства достаточно рассмотреть сужение кодирования на подалфавит, состоящий из букв с ненулевыми вероятностями.

8. Основная теорема для каналов без шума

Теорема 3 A. При любых источниках и кодировании $\overline{C} \geq H \cdot \ln^{-1} m$, $C \geq H \ln^{-1} m$.

Доказательство. Фиксируем произвольно источник и кодирование. Будем считать, что $C < \infty$ (8.1). Иначе утверждение теоремы тривиально. Положим

$$V_0(n) = \bigcup_{B \in \varphi(\mathfrak{A}^\infty)} V(n, B) \tag{8.2}$$

$$V(n) = \{B \in V_0(n) \colon \forall B_1 \in V_0(n)\, (B = B_1 \vee B_1 \not\prec B)\}, \quad (n = 1, 2, \ldots), \tag{8.3}$$

где $V(n, B)$ — множество, определенное в пункте 5. Множество $V(n)$ состоит из всех префиксов кодов сообщений, по которым можно впервые декодировать некоторое слово длины n; этом ножество префиксное. Если последовательность $B' \in \varphi(\mathfrak{A}^\infty)$ принадлежит $B \cdot \mathfrak{B}^\infty$ при некотором $B \in V(n)$, то минимальная длина префиксов последовательности B, по которому можно декодировать некоторое слово длины n,

$$t(n, B') = |B| . \tag{8.4}$$

Если B' не принадлежит

$$\bigcup_{B \in V(n)} B \cdot \mathfrak{B}^\infty, \text{ то } t(n, B') = \infty . \tag{8.5}$$

Из (8.4), (8.5) следует, что*

$$C(n) = \sum_{B \in V(n)} p(B)\,|B| + \infty \cdot p\Big(\mathfrak{B}^\infty \setminus \bigcup_{B \in V(n)} B \cdot \mathfrak{B}^\infty\Big) \qquad (n = 1, 2, \ldots). \tag{8.6}$$

Фиксируем произвольно натуральное число n. Для $A \in \mathfrak{A}^n$ положим

$$V(A) = \{B \in V(n):\ A \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty)\, n \cdot b\} . \tag{8.7}$$

Из (8.2), (8.3), (8.7) получаем

$$\bigcup_{A \in \mathfrak{A}^n} V(A) = V(n) . \tag{8.8}$$

Из (8.7) следует, что $A \cdot \mathfrak{A}^\infty \supset \bigcup_{B \in V(A)} \varphi^{-1}(B \cdot \mathfrak{B}^\infty)$ и поэтому

$$p(A) \geq \sum_{B \in V(A)} p(B) \qquad (A \in \mathfrak{A}^n) . \tag{8.9}$$

Из (8.1), (8.6) имеем $p\Big(\mathfrak{B}^\infty \setminus \bigcup_{B \in V(n)} B \cdot \mathfrak{B}^\infty\Big) = 0$, поэтому

$$\sum_{B \in V(n)} p(B) = 1 . \tag{8.10}$$

Из (8.6) следует, что если $p(B) > 0$ $(B \in V(n))$, то $B \in V(A)$ только для одного $A \in \mathfrak{A}^n$. Поэтому из (8.9), (8.10) и того, что $\sum_{A \in \mathfrak{A}^n} p(A) = 1$, следует, что

$$p(A) = \sum_{B \in V(A)} p(B) \qquad (A \in \mathfrak{A}^n) . \tag{8.11}$$

* По определению $\infty \cdot x = \begin{cases} \infty, & \text{если } x > 0; \\ 0, & \text{если } x = 0. \end{cases}$

По свойству энтропии [4 стр. 62] имеем

$$-\sum_{B\in V(n)} p(B)\ln p(B) \geq -\sum_{A\in\mathfrak{A}^n} p\big(V(A)\cdot\mathfrak{B}^\infty\big)\ln p\big(V(A)\cdot\mathfrak{B}^\infty\big) =$$

$$-\sum_{A\in\mathfrak{A}^n} p(A)\ln p(A) = H(n) \qquad (n<1,\ 2,\ \ldots). \tag{8.12}$$

Из (8.12) и префиксности множества $V(n)$ следует, что

$$C(n) = \sum_{B\in V(n)} p(B)\,|B| \geq \ln^{-1} m\cdot\sum_{B\in V(n)} p(B)\ln p(B) \geq \ln^{-1} m\cdot H(n). \tag{8.13}$$

Делим (8.13) на n и переходим к пределу при $n\to\infty$. Получаем утверждение теоремы.

Проиллюстрируем доказательство на том же примере, что и в п. 5. Из сказанного там следует, что $V_0(1)=\{1,\ 00,\ 010,\ 0110,\ldots\}$. При $n=1$ $V(\alpha)=\{00,\ 0110,\ 011110,\ \ldots\}$,

$$V(B)=\{010,\ 01110,\ldots\},\ V(\gamma)=\{1\}.$$

Имеем $p(\alpha)=p(00)+p(0110)+\ldots=\sum_{k=0}^{\infty}\frac{2}{3^{k+2}}=\frac{1}{3}$, $p(\beta)=p(010)+\ldots=$
$=\sum_{k=1}^{\infty}\frac{2}{3^{k+1}}=\frac{1}{3}$,

$$p(\gamma)=p(1)=\frac{1}{3}.$$

При $n>1$ $V_0(n)=\varphi(\mathfrak{A}^{n-1})\cdot V_0(1)$ и при любом слове $A\in A_1\cdot a\in\mathfrak{A}^n$ $V(A)=\varphi(A_1)\cdot V(a)$.

Теорема 3 B. Для любого источника существует кодирование, при котором $\overline{C}=\overline{H}\cdot\ln^{-1} m$, $\underline{C}=\underline{H}\ln^{-1} m$.

Дадим эскиз доказательства. Каждое сообщение $A\in\mathfrak{A}^\infty$ представляется в виде $A_1(A)\cdot A_2(A)\ldots$, где $|A_n(A)|=n^2$ $(n=1,\ 2,\ \ldots)$. Для всех сообщений с общим началом $A_1\ldots A_{n-1}$ таким, что $p(A_1\ldots A_{n-1})>0$, слова $A_n(A)$ кодируются префиксным кодом аналогичным коду К. Шеннона [1], построенным по условному распределению вероятностей

$$p_{A_1\ldots A_{n-1}}(A)=\frac{p(A_1\ldots A_{n-1}\cdot A)}{p(A_1\ldots A_{n-1})}.$$

Для этого кодирования $C\big(\sum_{i=1}^{n} i^2\big)\leq\ln^{-1} m\cdot H\big(\sum_{i=1}^{n} i^2\big)+n$ при любом $n=1,2,\ldots$. Переходя к пределу с учетом монотонности C( ) и H( ), получаем утверждение теоремы.

Заключение

Приведенные выше результаты показывают, что данное в работе определение стоимости не только интуитивно естественно, но также согласуется с традиционным определением и позволяет сравнительно просто обобщить предшествующие результаты, в частности, основную теорему для каналов без шума.

Литература

1. *Шеннон К.* Математическая теория связи, Работы по теории информации и кибернетике, ИЛ, М., 1963, 243—332.
2. *Файнстейн А.* Основы теории информации, ИЛ., М., 1960.
3. *Мак-Миллан Б.* Два неравенства, обусловленные однозначностью расшифрования, Кибернетический сборник (1961), 3, 88—92.
4. *Фано Р.* Передача информации. Статистическая теория связи., «Мир», 1965.
5. *Сидельников В. М.* О статистических свойствах преобразований, осуществляемых конечными автоматами, Кибернетика (1965), 6, 1—14.
6. *Биллингслей П.* Эргодическая теория и информация, «Мир», 1969.
7. *Чисар И.* О каналах без шума, Проблемы передачи информации (1970), 4, 6, 3—15.
8. *Billingsley, P.*, On the coding theorem for the noiseless channel, Ann. Math. Statistics (1961), 32, 2, 594–601.
9. *Невё Ж.* Математические основы теории вероятностей, «Мир», М., 1969.
10. *Ходак Г. Л.* О средней задержке декодирования дешифруемых кодов, ДАН СССР (1972), 207, 6, 1284—1287.

On the cost sequence transmission

G. L. KHODAK

(Omsk)

The transmission of message over noiseless channel is considered. The source of messages is any discrete stochastic process. The coding is any measurable mapping of a set of input sequences to a set of output ones. In contrast to other authors the cost of transmission is defined as average number of output symbols which are necessary for decoding of one symbol of a message. This definition more corresponds to an intuition and permits to generalize the fundamental theorem for the noiseless channels comparatively simple and the class of sources and coding is apparently maximum wide. It is proved that for finite automata coding the new cost is equal to the traditional one if a coding is decipherable one is equal to infinity otherwise.

Definition. Let $\mathfrak{A}$ be finite or numerable input alphabet and B finite output one. $\mathfrak{A}^*$ and $\mathfrak{A}^\infty$ are mean sets of all words and all sequences of letter from $\mathfrak{A}$, $|A|$ is a number of letters in $A \in \mathfrak{A}^*$ its length, for $A_1 \in \mathfrak{A}^*$, $A_2 \in \mathfrak{A}^* \cup \mathfrak{A}^\infty$ $A_1 < A_2$ means that A_1 is a prefix of A_2, for $V \subset \mathfrak{A}^* \cup \mathfrak{A}^\infty$. $A_1 < V$ means that $\forall A \in V (A_1 < A)$ $||V||$ is power of a set V, $A_1 V = \{A_1 \cdot A : A \in V\} \cdot \sigma(\mathfrak{A})$ means any σ-algebra of subsets of $\mathfrak{A}^\infty$ containing $A \cdot \mathfrak{A}^\infty$ for all $A \in \mathfrak{A}^*$, $\sigma_0(\mathfrak{B})$ means the σ-algebra of subsets of $\mathfrak{B}^\infty$ produced by $\{B \cdot \mathfrak{B}^\infty : B \in \mathfrak{B}^*\}$, p is a stochastic measure on $\sigma(\mathfrak{A})$, $p(A) = p(A \cdot \mathfrak{A}^\infty)$ $(A \in \mathfrak{A}^*)$.

The source of message is the probability space $\{\mathfrak{A}^\infty \sigma(\mathfrak{A}), p\}$. Upper entropy of the source is $\overline{H} = \overline{\lim_{n \to \infty}} - \frac{1}{n} \sum_{|A|=n} p(A) \ln p(A)$ and lower one is $H = \lim_{n \to \infty} - \frac{1}{n} \sum_{|A|=} \cdot$ $\cdot p(A) \ln p(A)$.

The coding is any measurable mapping $\varphi : \mathfrak{A}^\infty \to \mathfrak{B}^\infty$ i.e. $\forall\, V \in \sigma_0(\mathfrak{B})\,(\varphi^{-1}(V) \in \in \sigma(\mathfrak{A}))$. The coding transfers the measure p from $\sigma(\mathfrak{A})$ to

$$\sigma(\mathfrak{B}) = \{V \subset \mathfrak{B}^\infty : \varphi^{-1}(V) \in \sigma(\mathfrak{A})\} \supset \sigma_0(\mathfrak{B})\,.$$

A prefix A of some message A' may be decoded with a prefix B of sequence $\varphi(A')$ if $A \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty)$ almost everywhere (a.e.). For all $n = 1, 2, \ldots$ and all $B' \in \varphi(\mathfrak{A}^\infty)$ $t(n, B')$ means the minimum length of prefixes of B' with which some words of length n may be decoded,

$$t(n, B') = \begin{cases} \min\limits_{B \in V(n, B')} |\, B \,| & \text{if } V(n, B') \neq \varnothing \\ +\infty & \text{if } V(n, B') = \varnothing \end{cases}$$

where

$$V(n, B') = \{B \in \mathfrak{B}^* : B \prec B' \,\&\, \exists\, A \in \mathfrak{A}^*(|\, A \,| = n \,\&\, A \prec \varphi^{-1})\, B \cdot \mathfrak{B}^\infty)\, a \cdot e)\}\,.$$

The function $t(n)$, is measurable

The upper cost of the coding is $c = \lim\limits_{n \to \infty} \dfrac{1}{n} \int\limits_{B^\infty} t(n,)\, dp$ and the lower one is $c = \lim\limits_{n \to \infty} \dfrac{1}{n} \int\limits_{\mathfrak{B}^\infty} t(n,)\, dp$. The traditional definition of the cost is $L = \lim\limits_{n \to \infty} \dfrac{1}{n} \sum\limits_{|A|=n} \cdot$ $\cdot\, p(A)\,|\,\varphi(A)\,|$ for the coding φ: $\mathfrak{A}^* \to \mathfrak{B}^*$.

It is proved that in classical cases $\bar{c} = c = L$ if a coding is decipherable.

The fundamental theorem A. For all sources and coding $\bar{c} \geq \overline{H} \ln^{-1} ||\, \mathfrak{B} \,||$, $c \geq H \cdot \ln^{-1} ||\mathfrak{B}\,||$. B. For any source there exists the coding such that $c = H \ln^{-1} ||\, \mathfrak{B} \,||$, $c = H \cdot \ln^{-1} ||\, \mathfrak{B} \,||$.

Г. Л. Ходак
Политехнический институт
СССР, Омск, просп. Мира, д. 11.

Problems of Control and Information Theory, Vol. 4 (4), pp. 311—316 (1975)

THE EXTENSION OF THE EXTRAPOLATION TECHNIQUE IN THE SUMT METHOD

L. GERENCSÉR

(Budapest)

(Received October 6, 1974)

The extension of the extrapolation technique will be discussed. Instead of a single curve $x^*(r)$ we consider a family of curves, which is defined by the equation $\nabla f(x) + r\nabla I(x) + r\Theta = 0$. Here $f(x)$ is the objective function, $I(x)$ is a logarithmic barrier function, r, Θ are parameters. These parameters are considered as dependent variables of the variable x, and a search process is given.

1. Introduction

We shall consider the problem of nonlinear programming in the form

$$\min f(x) \tag{1.1}$$

$$g_i(x) \geqq 0 \quad i = 1, \ldots, m. \tag{1.2}$$

Here the functions $f(x)$, $-g_i(x)$ are twice continuously differentiable convex functions of the n-dimensional vector variable x. These functions are defined on an open convex set containing the feasible region Ω, defined by condition (1.2). We suppose that Ω is bounded and the set of interior points of Ω is not empty and it coincides with the set Ω_0

$$\Omega_0 = \{x\colon\ g_i(x) > 0\}.$$

We assume that the problem has a single solution x^*. Since we are going to use extrapolation, some regularity conditions should be satisfied.

The set of active indices in x^* will be denoted by $I(x^*)$ i.e.

$$I(x^*) = \{i\colon\ g_i(x^*) = 0\}.$$

The first condition which must be fulfilled is

Condition a). The vectors $\nabla g_i(x^*)$, $i \in I(x^*)$ are linearly independent.

The multipliers in the Kuhn–Tucker condition will be denoted by λ_i^* $(i = 1, \ldots, m)$. Our second requirement is expressed in

Condition b). For every $i = 1, \ldots, m$ the numbers λ_i^* and $g_i(x^*)$ are not both zero. (Strict complementary slackness.)

The third condition refers to the Lagrangian function of the problem which we denote by

$$L(x, \lambda) = f(x) - \sum_{i=1}^{m} \lambda_i g_i(x).$$

Here λ is a vector with components $\lambda_1, \ldots, \lambda_m$.

Condition c). The second-order sufficient conditions are satisfied, that is for any vector $v \neq 0$, which satisfies

$$\nabla g_i'(x^*)v = 0 \quad i \in I(x^*),$$

the inequality

$$v' \nabla^2 L(x^*, \lambda^*)v > 0$$

holds. (Second-order sufficient condition holds.)

We assume that Conditions a), b) and c) are satisfied. We are going to solve the problem with the SUMT method. We choose the version where a logarithmic barrier function is used, which is defined by

$$I(x) = -\sum_{i=1}^{m} \ln g_i(x).$$

Define the function

$$P(x, r) = f(x) + rI(x).$$

The following assertions are contained in the book of Fiacco and McCormick [1]. For sufficiently small r the function $P(x, r)$ is strictly convex and attains its minimum at a unique point, say $x^*(r)$. The points $x^*(r)$ converge to x^* when $r \to 0$, and the trajectory $x^*(r)$ is continuously differentiable for $r \geqq 0$. An extrapolation can be used for the determination of x^*, if only a few values of $x^*(r)$ has been computed. However the computation of $x^*(r)$ is quite difficult, when r is small.

2. Extension of the extrapolation technique

A direct way of the determination of x^* would be the integration of a differential equation which is satisfied by $x^*(r)$. Such a differential equation can be built by differentiating in r the equality

$$\nabla P(x, r) = \nabla f(x) + r \nabla I(x) = 0. \tag{2.1}$$

This equality holds along $x^*(r)$. We get

$$\left(\nabla^2 f(x) + r \nabla^2 I(x)\right) \frac{dx^*(r)}{dr} = -\nabla I(x) = r^{-1} \nabla f(x)$$

where for x the point $x^*(r)$ should be substituted.

The matrix $\nabla^2 I(x)$ can be written in details as

$$\nabla^2 I(x) = -\sum_{i=1}^{m} \frac{\nabla^2 g_i(x)}{g_i(x)} + \sum_{i=1}^{m} \frac{\nabla g_i(x)\, \nabla' g_i(x)}{g_i^2(x)}.$$

Introduce the notations

$$\lambda_i(x) = \frac{r}{g_i(x)}.$$

Then we can make the following simplifications in (2.2)

$$\nabla^2 f(x) + r\nabla^2 I(x) = \nabla^2 f(x) - \sum_{i=1}^{m} \lambda_i(x)\, \nabla^2 g_i(x) + r^{-1} \sum_{i=1}^{m} \lambda_i^2(x)\, \nabla g_i(x)\, \nabla' g_i(x).$$

Introduce the notation

$$\Gamma(x) = \sum_{i=1}^{m} \lambda_i^2(x)\, \nabla g_i(x)\, \nabla' g_i(x).$$

Then we get using the Lagrangian function:

$$\nabla^2 f(x) + r\nabla^2 I(x) = \nabla^2 L(x, \lambda(x)) + r^{-1}\Gamma(x)$$

where $\lambda(x)$ is a vector with components $\lambda_1(x), \ldots, \lambda_m(x)$.

Finally introduce the matrix

$$D(x) = r\nabla^2 L(x, \lambda) + \Gamma(x) \tag{2.3}$$

where we substitute $\lambda = \lambda(x)$.

Our hypotheses implies that $D(x)$ has an inverse (see [1]) and we find that

$$\frac{dx^*}{dr} = -D^{-1}(x)\, \nabla f(x). \tag{2.4}$$

The numerical integration of this equation is not possible, since the right-hand side is defined only along the curve $x^*(r)$. Therefore we shall replace the curve $x^*(r)$ by a family of curves $x^*(r, \Theta)$. This family is defined by the following equation:

$$\nabla f(x) + r\nabla I(x) + r\Theta = 0. \tag{2.5}$$

Here Θ is a vector with components $\Theta_1, \ldots, \Theta_m$. The vector Θ is hold constant while r is positive. The solution to this system of nonlinear equations will be denoted by $x^*(r, \Theta)$. There exists a single solution, since the solution of Eq. (2.5) is equivalent to the problem of minimizing in x the function

$$P(x, r, \Theta) = f(x) + rI(x) + r\Theta' x.$$

If r is sufficiently small then this function is the sum of a strictly convex and a linear function of the variable x. Thus $P(x, r, \Theta)$ is strictly convex in x, and attains its minimum in a single point $x^*(r, \Theta)$.

Simple modifications of the usual argument show that

$$x^*(r, \Theta) \to x^* \quad \text{when} \quad r \to 0.$$

Moreover the curve $x^*(r, \Theta)$ is continuously differentiable for $r \geqq 0$ for every fixed 0. The tangent vector to the curve $x^*(r, \Theta)$ is expressed in formula (2.4), where the matrix $D(x)$ should be computed in the point $x^*(r, \Theta)$.

Now we can use the above idea in the following way. The curves $x^*(r, \Theta)$ fill in some part of the feasible region. Thus for an arbitrary x we can find some curves passing through this point. For this reason we have only to solve Eq. (2.5), in which the unknown variables are now $r, \Theta_1, \ldots, \Theta_n$. Since we have n equations and $n + 1$ unknowns we introduce an additional equation. A simple way for this is to use the equation

$$\nabla' f(x)\Theta = 0.$$

We then get the solution

$$r = r(x) = -\frac{\nabla' f(x)\, \nabla f(x)}{\nabla' I(x)\, \nabla f(x)}.$$

The value of Θ is not needed in the computation. If the value of $r(x)$ is positive, we go on making an extrapolation.

This means the we use a first-order Taylor-formula in the point $x = x^*(r, \Theta)$ for the computation of x^*:

$$x^* = x^*(r, \Theta) - r\frac{dx^*}{dr}(r, \Theta) + \Theta(r^2).$$

This formula is valid if the curve is uniformly differentiable for $r \geqq 0$. Introduce the notation

$$h(x) = -\frac{dx^*}{dr}(r, \Theta).$$

This direction is computed by formula (2.4), i.e.

$$h(x) = -D^{-1}(x)\, \nabla f(x).$$

We compute $D(x)$ according to (2.3), where

$$r = r(x) \qquad \text{and} \qquad \lambda_i(x) = \frac{r_i(x)}{g_i(x)}.$$

The search proceeds in the direction of $h(x)$. We apply the formula

$$x_{n+1} = x_n + r(x_n)\, h(x_n).$$

If x_{n+1} does not belong to the feasible region, then we replace r by $\frac{r}{2}$.

3. Computational aspects

The main difficulty is the inversion of the matrix $D(x)$. Some suggestions are given by Fletcher and McCann [2]. One reason for this is the fact that the matrix $D(x)$ converges to the matrix

$$\Gamma^* = \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) \nabla' g_i(x^*)$$

when x converges to x^* along some trajectory $x^*(r, \Theta)$. The matrix Γ^* is singular, if the number of active constraints is less than n.

Better results can be eventually obtained, when we have n active constraints, that is the optimum is reached in a vertex. Namely in this case the matrix Γ^* is nonsingular. Suppose that the active indices are $1, \ldots, n$. Then x^* can also be determined from the equation

$$\begin{aligned} g_1(x^*) &= 0 \\ &\vdots \\ g_n(x^*) &= 0. \end{aligned}$$

There are several methods for solving a system of nonlinear equation, however, we prefer to use extrapolation.

We tested the algorithm for the following numerical example:

$$\min [(x-1)^2 + (y-1)^2 + (z-1)^2]$$

subject to

$$1 - x^2 - 100y^2 - 100z^2 \geqq 0.$$

The initial point was

$$x = 0. \quad y = 0. \quad z = 0.08.$$

After five iterations we obtained the points

$$x = 0.724 \quad y = 0.037 \quad z = 0.057. \tag{3.1}$$

The value of the objective function in this point is 1.890, while the solution is 1.875. The problem is rather ill conditioned, therefore the solution (3.1) is far from being accurate, while the objective function has an agreable value.

Acknowledgement. I express my thanks to Professor András Prékopa for his encouragement to write this paper.

References

1. *Fiacco, A. V., McCormick, G. P.*, Nonlinear programming: Sequential unconstrained minimization techniques, Wiley, 1968.
2. *Fletcher, R., McCann, A. P.*, Acceleration techniques for nonlinear programming. In "Optimization" (ed. by R. Fletcher), Academic Press, 1969.

Развитие экстраполяционных алгоритмов в методе штрафных функций

Л. ГЕРЕНЧЕР

(Будапешт)

Рассматривается развитие экстраполяционных алгоритмов в методе штрафных функций. В частности, предлагается вместо одной кривой $x^*(r)$ ввести семейство кривых, определяемых посредством уравнения $\nabla f(x) + \nabla(x) + r\Theta = 0$, где $f(x)$ — целевая функция; $I(x)$ — логарифмическая штрафная функция; r, Θ — параметры. Параметры могут зависеть от x. Формулируется алгоритм поиска.

L. Gerencsér
Computer and Automation Research Institute
Hungarian Academy of Sciences
H–1502 Budapest. XI. Kende u. 13–17
Hungary

Problems of Control and Information Theory, Vol. 4 (4), pp. 317—326 (1975)

A SEMIINVARIANT METHOD OF CLOSING THE EQUATIONS FOR MOMENTS IN STATISTICAL ANALYSIS OF NONLINEAR SYSTEMS

M. L. DASHEVSKY

(Moscow)

(Received November 26, 1974)

Introduction

References [1, 2, 3] recommend the use of differential equations for semiinvariants in statistical analysis of nonlinear dynamic systems. Since these equations generally form an infinite set (the right-hand member of equations for semiinvariants of order k depend on semiinvariants of an order higher than k) to close these equations it is suggested that senior semiinvariants be neglected by assuming that the semiinvariants whose order exceeds a certain number r are zero. It will be shown that at $r = 2$ the equations for semiinvariants coincide with those for the mathematical expectation vector and a covariance matrix of system state coordinates obtained by statistical linearization [4]. Increasing r, or with the number of the semiinvariants taken in account, the analysis accuracy increases as seen in numerous examples. In a number of cases results are significantly more accurate than those obtained in statistical linearization. Although applicability of this approach (the semiinvariant method) to analysis of nonlinear systems. seems to be confined to systems where probability distributions are near-Gaussian*, the method is found to be useful in a number of practical problems. However, equations for semiinvariants are often cumbersome and hard to formalize. Their derivation in a multi-dimensional case is a tiresome and labor-consuming process. This is one of the disadvantages of the method. Equations for moments are much easier to derive. In the case where the plant to be analyzed is described by stochastic differential equations with polynomial coefficients the equations for moments are generally linear. For these equations simple formulae are known [5]. Equations for central moments are also easily formalized. Equations for moments also form an infinite system but unlike semiinvariants mere neglect of higher-order moments for closing the system results in

* It is well known that in normally distributed random variable the semiinvariant of order higher than 2 are equal to zero. This suggests that in near-Gaussian distributions the significance of senior semiinvariants decreases.

large analysis errors. Therefore this paper proposes a combined method whereby nonlinear dynamic plants are statistically analyzed by a "semiinvariantly closed" system of differential equations for moments. To do that the notion of "r-truncated" moments is introduced, or of an expression of senior moments (of order $k > r$) through lawer-order moments (of order $p \leq r$) at which the semiinvariants of order higher than r are assumed to be zero. The "semiinvariant" closing of moment equations consists in isolating a finite system of equations for moments from one to r inclusive with replacement of moments higher than r which are incorporated into the right-hand parts of equations for moments up to order r by "r-truncated" moments. Equations obtained in this way are equivalent, as follows from their construction, to equations for semiinvariants up to order r which are obtained from the initial infinite system of semiinvariants by eliminating the semiinvariants of order higher than r. Therefore as far as accuracy is concerned the method proposed is equivalent to the method of semiinvariants. It is attractive because of good formalizability of equations, ease of computation and the widespread knowledge of the notion of moments among researchers.

1. Statement of the problem

Let a dynamic plant be described by the differential equation

$$\frac{dx(t)}{dt} = a\big(t, x(t)\big) + \sigma(t)\,\xi(t) \tag{1.1}$$

where $x(t) = ||x_i(t)||_{i=\overline{1,n}}$ — column vector of plant phase coordinates; $a(t, x(t)) = ||a_i(t, x(t))||_{i=\overline{1,n}}$ column vector-valued function of equation coefficients; $\sigma(t) = ||\sigma_{ij}(t)||_{i,j=\overline{1,n}}$ matrix; $\xi(t) = ||\xi_i(t)||_{i=\overline{1,n}}$ column vector of white Gaussian independent noises:

$$M[\xi_i(t)] = 0, i = \overline{1, n}; M[\xi_i(t)\,\xi_j(t)] = \begin{cases} \delta(t-\tau), & i = j \\ 0, & i \neq j \end{cases}.$$

Assume that the coefficients $a_i(t, x)$ are polynomials:

$$a_i(t, x) = \sum_{p_1+\ldots+p_n=0}^{N_i} a_{ip_1\ldots p_n}(t)\, x_1^{p_1} \ldots x_n^{p_n} \tag{1.2}$$

$$i = \overline{1, n}; \quad p_i = 0, 1, 2, \ldots; \quad N_i = 0, 1, 2, \ldots$$

and such that a solution of Eq. (1.1) exists at $t > t_0$. Assume that the moments

$$m_{k_1 \ldots k_n}(t) = M[x_1^{k_1}(t) \ldots x_n^{k_n}(t)], \quad k_1 \ldots k_n = 0, 1, 2, \ldots \tag{1.3}$$

of the process $x(t)$ at time $t = t_0$* exists and are known.

It is required to find mathematical expectations

$$m_i(t) = M[x_i(t)] = m_{\underbrace{00\ldots1}_{i}0\ldots0}(t)$$

and the second-order matrix of moments,

$$\Gamma(t) = || \Gamma_{ij}(t) ||_{i,j=\overline{1,n}}$$

$$\Gamma_{ij}(t) = M[x_i(t)x_j(t)] = m_{\underbrace{0\ldots1}_{i}0\ldots1\overbrace{0\ldots0}^{i}}(t) \tag{1.4}$$

or the covariance matrix

$$\mu(t) = || \mu_{ij}(t) ||_{i,j=\overline{1,n}}; \quad \mu_{ij}(t) = M[(x_i(t) - m_i(t))(x_j(t) - m_j(t))]$$

of the process $x(t)$ at times $t > t_0$.

These parameters can be computed at a certain accuracy through differential equations for moments of the process $x(t)$.

2. Equations for moments

Equations for moments (1.3) of the process $x(t)$ satisfying (1.1) are long since known [5]. They can be arived at in different ways. They are often derived through a differential equation for the characteristic function

$$f(t, z_1 \ldots z_n) = M\left[e^{i\sum_{k=1}^{n} x_k(t) z_k}\right] \tag{2.1}$$

$$\frac{\partial}{\partial t} f(t, z_1, \ldots, z_n) = iM\left[e^{i\sum_{k=1}^{n} x_k(t) z_k} \sum_{k=1}^{n} z_k a_k(t, x(t))\right] -$$

$$- \frac{1}{2} \sum_{l,m=1}^{n} b_{lm}(t) z_l z_m f(t, z_1, \ldots, z_n), \tag{2.2}$$

where $b_{lm}(t)$ — elements of the matrix

$$B(t) = \sigma(t)\sigma^*(t) \tag{2.3}$$

* — matrix transposition symbol.

* The number $k = k_1 + \ldots + k_n$ is referred to as order of the moment $m_{k_1 \ldots k_n}$.

In doing so the formula

$$m_{k_1 \ldots k_n}(t) = (-i)^k \frac{\partial^k f(t, z_1, \ldots, z_n)}{\partial z_1^{k_1} \ldots, \partial z_n^{k_n}} \Bigg|_{z_1 = \ldots = z_n = 0;\ k = k_1 + \ldots + k_n} \tag{2.4}$$

is used. In our case equations for moments obtained by Eqs. (2.2) and (2.4) are of the form

$$\begin{aligned}
\frac{dm_i}{dt} &= \sum_{p_1 + \ldots + p_n = 0}^{N_i} a_{ip_1 \ldots p_n}(t)\, m_{p_1 \ldots p_n} \\
\frac{d\Gamma_{ij}}{dt} &= \sum_{p_1 + \ldots + p_n = 0}^{N_i} a_{p_1 \ldots p_n}(t)\, m_{p_1 \ldots p_{i-1},\, p_i + 1,\, p_{i+1}, \ldots,\, p_n} + \\
&+ \sum_{p_1 + \ldots = p_n = 0}^{N_j} a_{jp_1 \ldots p_n}(t)\, m_{p_1 \ldots p_{j-1},\, p_j + 1,\, p_{j+1} \ldots p_n} + b_{ij}(t)
\end{aligned} \tag{2.5}$$

$$\begin{aligned}
\frac{dm_{k_1 \ldots k_n}}{dt} &= \sum_{i=1}^{n} k_i \sum_{p_1 + \ldots + p_n = 0}^{N_i} a_{ip_1 \ldots p_n}(t) \times \\
&\times m_{p_1 + k_1, \ldots,\, p_{i-1} + k_{i-1},\, p_i + k_i - 1,\, p_{i+1} + k_{i+1}, \ldots,\, p_n + k_n} + \\
&+ \frac{1}{2} \sum_{i=1}^{n} k_i (k_i - 1)\, b_{ii}(t)\, m_{k_1 \ldots k_{i-1},\, k_i - 2,\, k_{i+1}, \ldots,\, k_n} + \\
&+ \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} k_i k_j b_{ij}(t)\, m_{k_1 \ldots k_{i-1},\, k_i - 1,\, k_{i+1}, \ldots,\, k_{j-1},\, k_j - 1,\, k_{j+1} \ldots k_n} \\
&k_1 + \ldots + k_n = k > 2.
\end{aligned}$$

As follows from Eq. (2.5), equations for moments are linear and this is their advantage over equations for semiinvariants. Besides, they are well-formalized since Eq. (2.5) is a relatively simple expression for equations of moments of an arbitrary order k.

Equations (2.5), however, generally form an infinite set of interrelated equations and their numerical integration is thus out of the question. Simple neglect of higher order moments for closing the system of equations for lower-order moments leads to large computation errors $m_i(t)$, $\Gamma_{ij}(t)$, $t > t$. This is the disadvantage of equations for moments: although semiinvariants also form an infinite set of equations, neglect of higher order semiinvariants is often less detrimental.

Equations for central moments

$$\mu_{k_1 \ldots, k_n}(t) = M[(x_1(t) - m_1(t))^{k_1} \ldots (x_n(t) - m_n(t))^{k_n}] \tag{2.6}$$

of arbitrary order k are also not hard to obtain. Unlike equations for moments, these equations are nonlinear but, as (2.5), they are well-formalized and this is what makes them more attractive than semiinvariant equations. Equations for central moments also form an infinite set.

3. Truncating the moments

The infinity of the system of differential equations for moments hampers their use in solving problems of statistical analysis for nonlinear dynamic plants. If there were precise formulae interrelating higher and lower order moments, the problem of closing the moment equations would be a simple one. Expressing the moments whose order is greater than a certain number r through moments whose order $P \leq r$, using Eq. (2.5) we would have an accurate finite closed system of differential equations for moments of order up to r. There are, however, no such formulae*, therefore higher-order moments will be expressed just approximately through lower-order moments. This can be probably done in many ways [6]. This paper proposes one approach reducible to the so-called "truncation" of moments by using the relation of moments and semiinvariants

$$\lambda_{q_1 \ldots q_n}(t) = (-i)^q \left. \frac{\partial^q \ln f(t, z_1, \ldots, z_n)}{\partial z^{q_1} \ldots \partial z_n^{q_n}} \right|_{z_1 = \ldots = z_n = 0}, \tag{3.1}$$

where $q = q_1 + \ldots + q_n$ — order of the semiinvariant. In the one-dimensional case such a formula is convenient [7]:

$$m_k = \lambda_k + \sum_{q=1}^{k-1} c_{k-1}^{q-1} \lambda_q m_{k-q}. \tag{3.2}$$

Using the evident equality

$$\frac{\partial^{k_1 - 1 + k_2}}{\partial z_1^{k_1 - 1} \partial z_2^{k_2}} \left(\frac{\partial f(z_1, z_2)}{\partial z_1} \right) = \frac{\partial^{k_1 - 1 + k_2}}{\partial z_1^{k_1 - 1} \partial z_2^{k_1}} \left(\frac{\partial \psi(z_1, z_2)}{\partial z_1} f(z_1, z_2) \right)$$

where $\psi(z_1 z_2) = \ln f(z_1, z_2)$ and in the light of (2.4) and (3.1) at $n = 2$ it is easy to obtain an analogous formula for a two-dimensional random vector:

$$m_{k_1 k_2} = \lambda_{k_1 k_2} + \sum_{q_1 = 1}^{k_1 - 1} c_{k_1 - 1}^{q_1 - 1} \lambda_{q_1 k_2} m_{k_1 - q_1, 0} + \\ + \sum_{q_2 = 0}^{k_2 - 1} c_{k_2}^{q_2} \sum_{q_1 = 1}^{k_1} c_{k_1 - 1}^{q_1 - 1} \lambda_{q_1 q_2} m_{k_1 - q_1, k_2 - q_2}. \tag{3.3}$$

* An exception is the case where the form of the distribution law for the vector $x(t)$ is known. Thus in statistical linearization techniques the distribution is assumed to be Gaussian.

For an n-dimensional vector $x = (x_1, \ldots x_n)$ we can find that

$$m_{k_1 k_2 \ldots k_n} = \sum_{q_n=0}^{k_n} c_{k_n}^{q_n} \sum_{q_{n-1}=0}^{k_{n-1}} c_{k_{n-1}}^{q_{n-1}} \cdots \sum_{q_2=0}^{k} c_{k_2}^{q_2} \times$$

$$\times \sum_{q_1=1}^{k_1} c_{k_1-1}^{q_1-1} \lambda_{q_1 q_2 \ldots q_n} m_{k_1-q_1,\, k_2-q_2,\, \ldots,\, k_n-q_n}. \tag{3.4}$$

The formulae relating the central moments (2.6) and the semiinvariants (3.1) easily follow from (3.2)–(3.4). Thus from (3.2) follows

$$\mu_k = \lambda_k + \sum_{q=2}^{k-2} c_{k-1}^{q-1} \mu_{k-q}. \tag{3.5}$$

From (3.3) follows

$$\mu_{k_1 0} = \lambda_{k_1 0} + \sum_{q_1=2}^{k_1-2} c_{k_1-1}^{q_1-1} \lambda_{q_1 0} \mu_{k_1-q_1,\, 0};\ \mu_{0 k_2} = \lambda_{0 k_2} + \sum_{q_2=0}^{k_2-2} c_{k_2-1}^{q_1-1} \lambda_{0 k_2} \mu_{0, k_2-q_2}$$

$$\mu_{k_1 k_2} = \lambda_{k_1 k_2} + (k_1 - 1)\, k_2\, \lambda_{k_1-1,\, k_2-1}\, \mu_{11} +$$

$$+ \sum_{q_1=1}^{k_1-2} c_{k_1-1}^{q_1-1} (k_2 \lambda_{q_1,\, k_2-1} \mu_{k_1-q_1,\, 1} + \lambda_{q_1 k_2} \mu_{k_1-q_1, 0}) + \tag{3.6}$$

$$+ \sum_{q_2=0}^{k_2-2} c_{k_2}^{q_2} \sum_{q_1=1}^{k_1} c_{k_1-1}^{q_1-1} \lambda_{q_1 q_2} \mu_{k_1-q_1,\, k_2-q_1};\ k_1, k_2 > 0;\ \sum_{q_1=1}^{k_1-2} = 0,\ k_1 = 1,2;\ \sum_{q_2=2}^{k_2-2} = 0,\ k_2 = 1.$$

Formulae (3.2)–(3.6) give an accurate expression relating higher-order moments with semiinvariants of the same order, lower-order semiinvariants and lower-order moments. For our purposes of closing the moment equations higher-order moments should be expressed only through lower-order moments. For this purpose introduce the notion of "r-truncated" moments.

Let $f(z_1, \ldots, z_n)$ be the characteristic function of the random vector $x = (x_1, \ldots, x_n)$. It is associated with the function

$$\psi(z_1, \ldots, z_n) = \ln f(z_1, \ldots, z_n) = \sum_{q_1+\ldots+q_n=1}^{\infty} i^{q_1+\ldots+q_n} \frac{z_1^{q_1} \ldots z_n^{q_n}}{q_1! \ldots q_n!} \lambda_{q_1 \ldots q_n} \tag{3.7}$$

where $\lambda_{q_1 \ldots q_n}$ — semiinvariants. Assuming in Eq. (3.7) that $\lambda_{q_1 \ldots q_n} = 0$ at $r < q_1 + \ldots + q_n = q \leq R$ leads to the function*

$$\hat{\psi}_{r,R}(z_1, \ldots, z_n) = \sum_{q=1}^{r} i^q \frac{z^q}{q!} + \sum_{q=R+1}^{\infty} i^q \frac{z^q}{q!} \lambda_q = \tag{3.8}$$

$$= \psi(z_1, \ldots, z_n) - \sum_{q=r+1}^{R} i^q \frac{z^q}{q!} \lambda_q.$$

* Here the notation is introduced: $z_1^{b_1} \ldots z_n^{q_n} = z^q$; $q_1! \ldots q_n! = q$; $\lambda_{q_1 \ldots q_n} = \lambda_q$.

Introduce a "truncated" characteristic function

$$f_{r,R}(z_1, \ldots, z_n) = e^{\psi_{r,R}(z_1, \cdots, z_n)} = f(z_1, \ldots, z_n) \exp\left\{ - \sum_{q=r+1}^{R} i^q \frac{z^q}{q!} \lambda_q \right\}. \quad (3.9)$$

By "r-truncated" moment of order $k = k_1 + \ldots + k_n$ we will understand the parameter

$$m^{(r,R)}_{k_1 \ldots k_n} (-i)^k \frac{\partial^k}{\partial z_1^{k_1} \ldots \partial z_n^{k_n}} f_{r,R}(z_1, \ldots, z_n) \Big|_{z_1 = \ldots = z_n = 0}. \quad (3.10)$$

Here $r \geq 2$ and $R \geq k > r$ are some natural numbers.

"Truncated" central moments $\hat{\mu}^{(r,R)}_{k_1 \ldots k_n}$ are obtained from (3.10) at $m_i = M[x_i] = 0$, $i = \overline{1, n}$.

From (3.8)–(3.10) it follows that "truncated" moments coincide with the actual ones only when semiinvariants of order q (where $r < q \leq R$) in actual distribution are indeed zero. This is true, for instance, for normal distribution whose semiinvariants are zero at $q > 2$. In a general case "r-truncated" moments differ from actual ones and may be associated with no distribution whatsoever.

Consequently, the moments of original distribution are "truncated" by assuming that some of its semiinvariants are zero:

$$\lambda_{q_1 \ldots q_n} = 0, \quad r < q = q_1 + \ldots + q_n \leq R.$$

The distribution itself with this "semiinvariant truncation" is, naturally, distorted.

Semiinvariantly truncated moments of order k can be expressed in terms of lower order ($p < k$) moments, in particular, by using Eqs. (3.2)–(3.6). Assuming, for instance, in Eq. (3.2) that $\lambda_q = 0, k \geq q > r$ gives

$$\hat{m}^{(r,k)}_k = \sum_{q=1}^{k-r-1} c^{q-1}_{k-1} \lambda_q \hat{m}^{(r,k)}_{k-q} + \sum_{q=k-r}^{r} c^{q-1}_{k-1} \lambda_q m_{k-q}. \quad (3.11)$$

The desired expressions for "truncated" moments may be obtained from Eq. (3.11) by replacing there the semiinvariants λ_q, $1 \leq q \leq r$ by their expressions in terms of moments m_p, $1 \leq p \leq r$. The formula relating semiinvariants and moments, convenient for this purpose, are given in [7]. Write down, for instance, the resultant moments of order 5 and 6 for a one-dimensional random variable x assuming that the semiinvariants $\lambda_5 = \lambda_6 = 0$

$$\hat{m}^{(4,5)}_5 = 5m_4 m_1 + 10 m_3 m_2 - 20 m_3 m_1^2 - 30 m_2^2 m_1 + 60 m_2 m_1^3 - 24 m_1^5 \quad (3.12)$$

$$\hat{m}^{(4,6)}_6 = 15 m_4 m_2 + 10 m_3^2 - 60 m_3 m_2 m_1 - 30 m_2^3 + 90 m_2^2 m_1^2 - 24 m_1^6.$$

Assuming that in Eq. (3.6) the semiinvariants $\lambda_{q_1,q_2,q_1+q_2} = q < 4$ are zeros, write down expressions for "truncated" central moments of order 5 for a two-dimensional random vector

$$\hat{\mu}_{50}^{(4,5)} = 10\mu_{30}\mu_{20} \qquad \hat{\mu}_{41}^{(4,5)} = 4\mu_{30}\mu_{11} + 6\mu_{21}\mu_{20}$$
$$\hat{\mu}_{32}^{(4,5)} = \mu_{30}\mu_{02} + 6\mu_{21}\mu_{11} + 3\mu_{20}\mu_{12} \tag{3.13}$$
$$\hat{\mu}_{14}^{(4,5)} = 4\mu_{03}\mu_{11} + 6\mu_{12}\mu_{02}; \quad \hat{\mu}_{05}^{(4,5)} = 10\mu_{03}\mu_{02}$$

The formula for "truncated" moments of an n-dimensional random vector can be obtained from Eq. (3.4) at $\lambda_{q_1 \ldots q_n} = 0$, $q_1 + \ldots + q_n = q > r$ where r — order of truncation.

Note that at $r = 2$, or when semiinvariants of order higher than 2 are assumed zero, the formulae for "truncated" moments coincide with the formulae interrelating higher and lower order moments for normal distribution of random variable.

4. Closing the equations for moments

Let us confine ourselves in Eq. (2.5) to a finite system of equations for moments from order one to order r inclusive where r — certain arbitrary number. The resultant system will be referred to as "r-truncated". The right-hand parts of equations in an "r-truncated" system may incorporate moments of order higher than r and this prevents closing the system. Let the maximal order of moments in the right-hand parts of equations of the "r-truncated" system be $R > r$. Then all moments $m_{k_1 \ldots k_n}$ whose order is $k = k_1 + \ldots + k_n > r$ are replaced by "r-truncated" moments $m_{k_1 \ldots k_n}^{(r,R)}$. What we have now is a finite closed system of differential equations for moments of order up to r inclusive. Integrating it under initial conditions specified by the moments at $t = t_0$ one can compute with a certain degree of accuracy the mathematical expectations $m_i(t)$, $i = \overline{1,n}$ and second-order moments, at $t > t_0$. Note that at $r = 2$ the resultant equations coincide with the equations which are obtained by the method of statistical linearization. At $r > 2$ one should expect more accurate results of analysis than those obtained in statistical linearization.

Let us give an elementary example illustrating the closing of equations for moments. Assume that the plant is described by the differential equation

$$\frac{dx(t)}{dt} = -ax^3(t) + b\xi(t) \tag{4.1}$$

* Semiinvariants of order 4 should be expressed through moments [7].

where $\xi(t)$ — white noise. Equations for moments of the process $x(t)$ according to Eq. (2.5) are of the form*:

$$\frac{dm_1}{dt} = -am_3 \qquad \frac{dm_2}{dt} = -2am_4 + b^2$$

$$\frac{dm_k}{dt} = -kam_{k+2} + \frac{k(k-1)}{2} b^2 m_{k-2}, \quad k = 3, 4, \ldots \tag{4.2}$$

Let us confine to the four first ($r = 4$) equations in (4.2)

$$\frac{dm_1}{dt} = -am_3 \qquad \frac{dm_2}{dt} = -2am_4 + b^2$$

$$\frac{dm_3}{dt} = -3am_5 + 3b^2 m_1; \qquad \frac{dm_4}{dt} = -4am_6 + 6b^2 m_2 . \tag{4.3}$$

Here the moments m_5, m_6 incorporated into equations for m_3 and m_4 hinder closing the system. Replace them by "4-truncated" moments $\hat{m}_5^{(4,6)}$, $\hat{m}_6^{(4,6)}$ and write down the system of equations thus obtained**

$$\frac{d\tilde{m}_1}{dt} = -a\tilde{m}_3 \qquad \frac{d\tilde{m}_2}{dt} = -2a\tilde{m}_4 + b^2 \tag{4.4}$$

$$\frac{d\tilde{m}_3}{dt} = -3a\hat{m}_5^{(4,6)} + 3b^2\tilde{m}_1; \qquad \frac{d\tilde{m}_4}{dt} = -4am_6^{(4,6)} + 6b^2\tilde{m}_2$$

The system (4.4) is closed since $\hat{m}_5^{(4,6)}$ $\hat{m}_6^{(4,6)}$ depend only on the moments $\tilde{m}_1, \ldots, \tilde{m}_4$ (see (Eq. (3.12)).

Conclusions

The method of statistical analysis for nonlinear dynamic systems of this paper is reduced to integrating a closed, in a special way, system of differential equations for moments. The equations are closed by "truncating" the moments using the hypothesis that higher-order semiinvariants are not very important in computing the lower-order moments of near-Gaussian random variables. The accuracy of the analysis results obtained by this method is equivalent to that of the method of semiinvariants***, but the method is simpler and the equations are easier to formalize.

* For simplicity let us write m_k instead of $m_k(t)$.

** The symbol $\sim$ signifies the difference of the approximate system (4.4) from the actual moments.

*** Therefore we do not give numerical examples illustrating its effectiveness. Such examples may be found in Refs [1, 2].

References

1. *Дашевский М. Л., Липцер Р. Ш.* Приближенный анализ нелинейных нестационарных динамических систем. Автоматика и телемеханика, 8, 1967.
2. *Дашевский М. Л.* Приближенный анализ точности нестационарныхнел инейных систем методом семиинвариантов. Автоматика и телемеханика. II, 1967.
3. *Дашевский М. Л.* Уравнения семиинвариантов нелинейной динамической системы. Автоматика и телемеханика 10, 1968.
4. *Синицин И. Н.* Методы статистической линеаризации. (Обзор). Автоматика и телемеханика 5, 1974.
5. *Пугачев В. С.* Применение теории марковских процессов к анализу точности автоматических систем. Энергетика и автоматика 3, 1961.
6. *Дёмух В. И.* Приближенный метод анализа точности нелинейных систем. Автоматика и телемеханика 6, 1965.
7. *Кендалл М., Стюарт А.* Теория распределений. М. «Наука» 1966.

Семиинвариантный метод замыкания уравнений для моментов в задачах статистического анализа нелинейных систем

М. Л. ДАШЕВСКИЙ

(Москва)

Рассмотрен метод замыкания дифференциальных уравнений для моментов в задачах статистического анализа нелинейных систем, в основе которого лежит предположение о возможности пренебрежения старшими семиинвариантами при вычислении младших моментов. Метод развит для систем, которые описываются нелинейными стохастическими дифференциальными уравнениями (1.1) с полиномиальными коэффициентами (1.4). Для начальных моментов процесса (1.1) выписана в общем виде бесконечная система дифференциальных уравнений (2.5). Из этой бесконечной системы уравнений выделена конечнаян система уравнений для моментов от 1-го до r-го порядка включительно, где r — некоторые целое число. Эта система уравнений названа «укороченной». В правые части уравнений укороченной системы могут входить моменты, порядок которых k превосходит r, что мешает замкнуть систему. Эти «лишние» старшие моменты предлагается заменять специально введенными «r-урезанными» моментами (3.10), которые соответствуют «урезанной» характеристической функции (3.9). «Урезанные» моменты выражаются только через младшие моменты (порядка $k \leq r$); именно это позволяет замкнуть укороченную систему уравнений. Процесс «урезания» моментов проводится «приравниванием» к нулю семиинвариантов порядка $p > r$ в формулах связи старших моментов с младшими семиинвариантами и младшими моментами (12) – (16). Интегрирование укороченной системы уравнений позволяет с той или иной степенью точности вычислить вектор математического ожидания и ковариационную матрицу процесса в момент времени $t > t_1$, если моменты при $t = t_1$ известны. Отмечается, что при $r = 2$ укороченная система уравнений совпадает с уравнениями, которые получаются методом статистической линеаризации коэффициентов (1.4). При $r > 2$ следует ожидать большей, чем при статистической линеаризации, точности результатов статистического анализа. Предлагаемый метод эквивалентен по точности методу семиинвариантов [1, 2, 3], но обладает преимуществом: уравнения для моментов легче формализуются, чем уравнения для семиинвариантов. Приведен пример, иллюстрирующий идею замыкания уравнений для моментов.

М. Л. Дашевский
Институт проблем управления
СССР, Москва В—485,
Профсоюзная ул., 81.

Problems of Control and Information Theory, Vol. 4 (4), pp. 327–337 (1975)

THE EFFICIENCY OF COMPLEX HIERARCHICAL MULTILEVEL SYSTEMS WITH SPECIAL REGARD TO RELIABILITY

J. BLIZNAKOV

(Sofia)

(Received September 20, 1974)

Introduction and task setting

In solving control problems of organizations and technical systems the question, how to improve the structure of the control, arises unavoidably. The task becomes more complex by the application of electronic computers in the productive-economical and the non-productive fields of human activities.

Such tasks are solved usually by the efforts of complete organizational collectives of complex hierarchical multilevel structures. Therefore the questions of the evaluation of the efficiency of the organizational structures play a special role. The need of modelling the work of the complex hierarchical multilevel systems (CHMS) and evaluating their efficiency are of a particular importance, when new, great systems are established, or existing ones developed, for selecting the most efficient solution of the system structure.

Certain experiences, with the corresponding theoretical developments in connection with individual aspects of the theory of modelling hierarchical systems, exist already [1]–[5], but basically the problem was solved for bi-level hierarchical systems [5] in two working conditions [4].

The aim of this paper is therefore to determine the efficiency of complex hierarchical multilevel systems with their reliability taken into consideration.

CHMS is a system consisting of hierarchically organized controlled and controlling elements (organs, subsystems), which may be in several (multiple) working conditions, and possessing not less than two control levels. Here the expression "hierarchical" signifies the presence of controlled and controlling subsystems; "multilevel" means not less than two control levels and "complex" denotes the existence of several different working states (the normal working state and a number of not fully working states, amlng them also the non-working state) of the individual elements of the system.

The efficiency of the CHMS is characterized by the average output effect of the system in unit time. In this sense the reliability of the CHMS elements is the characteristic degree of probability (readiness) of their unfailing

work during the working period of the system and the reliability of the CHMS is the functioú of the reliability of its organs.

For solving the present task the basic types of the CHMS must be delimited. Accordingly:

a) the highest level (rank) control may be, depending on the number of its organs (coordinators, elements),
   — finite — with one coordinator,
   — unfinite — with several coordinators;

b) according to the number of the higher level (rank) control organs as related to the given lower level (rank) organ it may be of
   — independent control, where each organ of the CHMS is controlled by only one higher level organ,
   — non-independent control, where the individual organs of the CHMS are controlled by several higher level organs;

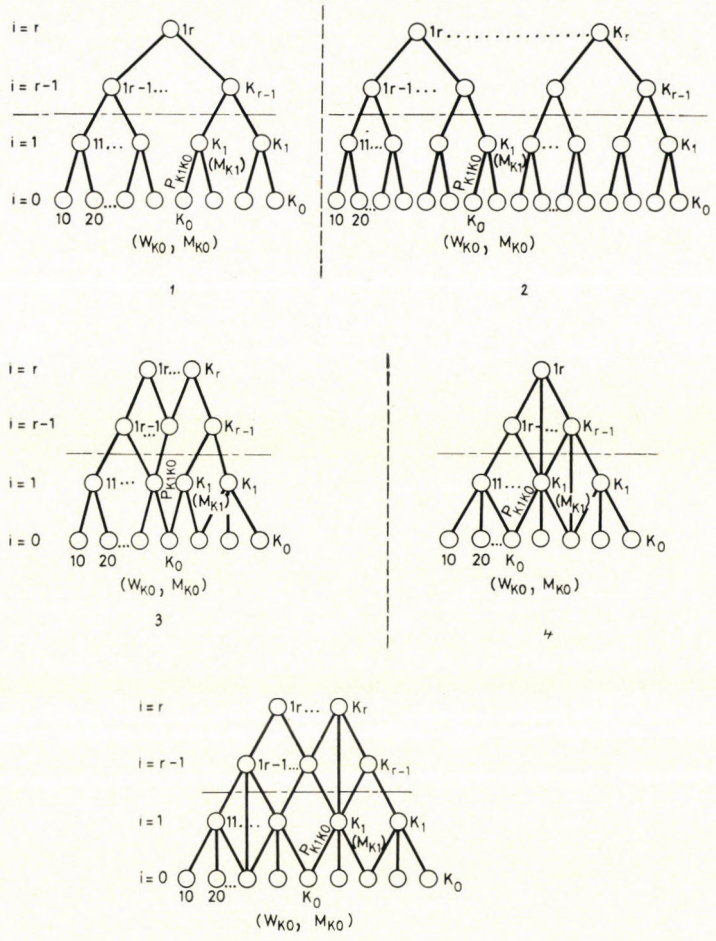
c) depending on the control channels:
   — direct control, with the presence of controlling communication channels only between the adjacent levels (ranks) of the CHMS,
   — indirect control, with communication channels of control also between the non-adjacent levels (ranks) of the CHMS.

On the basis of the above classification we shall determine the efficiency of some of the basic forms of CHMS in the order of increasing complicacy, as follows:

1. Finite CHMS with independent direct control (Fig. 1).
2. Non-finite CHMS with independent direct control (Fig. 2).
3. Non-finite CHMS with non-independent direct control (Fig. 3).
4. Finite CHMS with non-independent, indirect control (Fig. 4).
5. Non-finite CHMS with non-independent control (Fig. 5).

In practice the functioning organizational systems show all the types of the CHMS listed above, therefore the efficiency of all these forms of the CHMS must be determined, while on the basis of these 5 types the efficiency of other CHMS structures may also be determined in the majority of the cases, i.e. in this way the question of the efficiency of the CHMS with variable structures is solved within the classified range.

In all the cases of the efficiency determination of the CHMS identical aims are assumed, to whose realization the activities of the hierarchically organized control organs is oriented. Accordingly the aims of all the control organs of the CHMS must be selected correctly in order to reach the global aim of the CHMS and to obtain in this way the corresponding output effect.



Figs. 1–5

The formalization of the task and its theoretical contents

The CHMS is regarded as a system consisting of a finite number of organs, which may be in a finite number of states.

We introduce the following designations:

1. Rank (level) of the subsystem — i $(0 \leq i \leq r)$ in the case of the multilevel system $(r \geq 2)$.

2. Number of organ (elements) of the ith rank — k_i $(1 \leq k_i \leq K_i)$ $(k_i \in K_i)$.

3. Number of the organ of $(i - l)$th rank controlling the kth organ of ith rank — k_{i-l} $(1 \leq k_{i-l} \leq K_{i-l})$ $(k_{i-l} \in K_{i-l})$.

4. Number of the state of the kth organ of ith rank — s_{ki} $(0 \leq s_{ki} S_{ki})$ $(s_{ki} \in S_{ki})$; $s_{ki} = 0$ means the normal working state of the kth organ of ith rank.

5. The period, during which the kth organ of ith rank is in the s_{ki} state, $-t_{ski}$ $(0 \leq t_{ski} \leq T)$. T is the total working time of the CHMS.

6. The normal efficiency (capacity) of the work (normal amount of work done by the system in unit time), as a result of the functioning of the kth organ of ith rank in the state $s_{ki} = 0$, under the condition that all the other higher level organs are in the normal working state — W_{0ki};

7. The theoretical efficiency of the CHMS — E_T is determined, as the quantity of the output effect of the CHMS, when all its organs are in the normal working state, i.e. $E_T = W_{0k0}$.

8 The factual working efficiency (quantity of work by the system in unit time), as a result of the work of the kth organ of ith rank in the state $s_{ki} \neq 0$ — $K_{W_{ski}}$.

9. The relative efficiency coefficient of the work of the kth organ of ith rank in the state $s_{ki} \neq 0$, as related to the state $s_{ki} = 0$ — $K_{W_{ski}}$.

By definition:

$$K_{W_{ski}} = \frac{W_{ski}}{W_{0ki}}. \tag{1}$$

If there is a linear dependence between the duration (τ_{ski}) of finishing the unit work by the kth organ of ith rank in the state s_{ki} and its efficiency and between the work of the whole system, then the calculation may be reduced and (1) assumes the form of:

$$K_{W_{ski}} = \frac{\tau_{0ki}}{\tau_{ski}}, \tag{2}$$

where τ_{0ki} — duration of finishing the unit work by the kth organ of ith rank in the normal working state $s_{ki} = 0$. According to (1) and (2) $K_{W0ki} = 1$.

10. The probability density of the kth organ being in the s_{ki} state — $P_{ski}(t)$.

11. The probability of the kth organ (of ith rank) being in the s_{ki} state — $P_{ski} = \int_0^{\pi} P_{ski}(t)\, dt$. The probability density of the CHMS being in the μ state is — $p_{\mu}(t)$.

12. The stationary probability of the CHMS being in the μ state — P_μ.

13. The reliability (readiness) of the kth organ of ith rank as to the normal work efficiency (degree of readiness to the normal functioning):

$$M_{ki} = \sum_{s_{ki}=0}^{s_{ki}} P_{ski} \cdot K_{wski} \qquad (0 \leq M_{ki} \leq 1) . \tag{3}$$

14. The expected working efficiency of the kth organ of ith rank

$$W_{ki} = W_{0ki} \cdot M_{ki} . \tag{4}$$

15. The expected working efficiency of the kth organ of ith rank as related to that of the kth organ of $(i - l)$ rank:

$$W_{ki,ki-l} W_{0ki} \cdot P_{ki,ki-l} \cdot M_{ki} . \tag{4'}$$

We assume that the rest of the organs are in the normal working state. $P_{ki,ki-l}$ is determined as shown in No. 17.

16. The period of the kth organ of ith rank being in the state of controlling the kth organ of $(i - l)$th rank — $t_{ki,ki-l}$.

17. The probability that the kth organ of ith rank controls the k_{i-l}-th organ of $(i - l)$th rank — $P_{ki,ki-l}$.

During the continuous control period of the individual organs of $(i - l)$th rank:

$$P_{ki,ki-l} = \lim_{T \to \infty} \frac{\sum_{t=0}^{T} \sum_{l} t_{ki,ki-l}}{\sum_{t=0}^{T} \sum_{ki=1}^{ki} \sum_{l} t_{ki,ki-l}} = \lim_{T \to \infty} \frac{\sum_{t} \sum_{l} t_{ki,ki-l}}{T} ,$$

$$(l = 1, 2, \ldots, i - 2, i - 1) .$$

By definition $\sum_{ki-l} P_{ki,ki-l} = 1$ during the continuous control (in time) of the individual organs of $(i - l)$th rank. In the double communication channel mode of operation the reliability of the operation is assumed to be higher.

18. Output effect of the CHMS during the period t of the μth state — $E_\mu(t)$.

19. Average output effect of the CHMS during the period of a single μ state:

$$\bar{E}_\mu = \int_0^T E_\mu(t) \cdot p_\mu(t)\, dt . \tag{5}$$

20. The mathematically expected average length of one incidence of the CHMS in the μ state:

$$\bar{t}_\mu = \int_0^T t \cdot p_\mu(t)\, dt\,. \tag{6}$$

21. The mathematically expected average output effect of the CHMS during the period T of its functioning:

$$M[E(T)] = \sum \frac{p_\mu \cdot T}{\bar{t}_\mu} \bar{E}_\mu \quad \left(0 \leq \mu \leq \prod_{i=1}^{r} (s_{ki} + 1)\right). \tag{7}$$

22. The average efficiency of the CHMS with the reliability of its organs, from the 0-th to the ith rank including, taken into account — $\bar{E}_{0-i}$ $(0 \leq i \leq r)$. We assume that the organs of the other $(r - i)$th ranks of the CHMS are in the normal working state.

In the general case the mathematically expected random values of discrete distribution with the final set of realizations is the sum of the differential quotients of all the realizations according to their probabilities [6]. In the concrete case (7) this is changed into the sum of the differential quotients of the average output effect of the system during the period of one μ state over the average number of the incidences of the CHMS into the μ state during the time T.

The efficiency of the CHMS is directly influenced by the s_{ki} states of each kth organ of ith rank, which may be represented as the totality of the individual states, or

$$s_{ki} = s_{ki}^0 \cup s_{ki}^1 \cup \ldots \cup s_{ki}^{s_{ki}}. \tag{8}$$

During the functioning of the CHMS the states of its organs change. Consequently the state of the whole system at a given time moment t may be written by the vector

$$\bar{S}(t) = [s_{10}(t), s_{20}(t), \ldots, s_{11}(t), s_{21}(t), \ldots, s_{ki}(t), \ldots, s_{kr}(t)]\,. \tag{9}$$

Each state of the CHMS corresponds to a determined effect during the period of its being in that state.

So the average efficiency of the CHMS, i.e. the average output effect in unit time during the period T of the functioning of the CHMS is:

$$\bar{E} = \lim_{T \to \infty} \frac{M[E(T)]}{T} = \lim_{T \to \infty} \sum_{\mu} \frac{P_\mu \cdot T}{\bar{t}_\mu} \cdot \bar{E}_\mu\,. \tag{10}$$

where $\bar{E}$ = the average efficiency of the CHMS.

In the CHMS the output effect depends on the totality of the states s_{k0} of the executive (output) organs. So, in the case, where $E_\mu(s_{k0,t})$ is linear, we obtain [4]:

$$E_\mu(s_{k0}, t) = c_\mu t \qquad (0 \leq s_{k0} \leq S_{k0}), \tag{11}$$

where c_μ is the constant coefficient of the μ-th state of the system.

Consequently:

$$\bar{E} = \sum_\mu c_\mu \cdot P_\mu . \tag{11'}$$

As it is shown in [4], for systems, whose output organs can be only in two states, (11) turns to

$$\bar{E} = \bar{c} \sum_\mu y_\mu \cdot P_\mu , \tag{12}$$

where $\bar{c}$ is the average output effect of the system in unit time with one normally functioning organ; and y_μ is the number of the normally functioning output organs.

It is now evident that (12) represents only one special case of the present task. Therefore let us consider all the basic forms of the CHMS according to the given classification for obtaining algorithms suitable to evaluate the efficiencies of the same.

The efficiencies of the individual forms of the CHMS are studied with the help of the given graphs (Figs. 1–5) in the "order of the decreasing efficiencies".

The basic principle of the suggested approximation is the determination of the expected efficiency of the CHMS based on its theoretical efficiency, with the effect reducing the reliability of the hierarchically organized elements and the probability distribution of the controlling effects on the lower level elements of the system taken into account.

For determining the efficiency let us pass each graph of the controlling communication channels between the organs of the CHMS. This may be done in two ways:

1. *Horizontally*, i.e. going through all the possible communication channels in the horizontal section from each higher level organ to each lower level organ (starting with the lowest rank and ending up with the highest rank controlling subsystem) and summing up the expected output effects in unit time of each element of all ranks of the controlling subsystems.

2. *Vertically*, i.e. going through all possible controlling communication channels in the vertical section from each organ (element) to be controlled to each other organ (starting from the lowest rank and ending with the highest

rank organ to be controlled), and summing up the expected output effects in unit time of each organ.

It must be noted that if the controlling communication channel (serving for the transfer of the control effect) between two hierarchical organs of the ith and the $(i-l)$th ranks respectively coincides, then $P_{ki,ki-l}=0$.

In the final resort the method of going through the graphs does not change the obtained results.

1. Finite CHMS with independent direct control (Fig. 1). By the inductive method we obtain:

$$\left.\begin{aligned}
\bar{E}_{0-0} &= \sum_{k0} W_{k0}\cdot M_{k0}\\
\bar{E}_{0-1} &= \sum_{k0}\sum_{k1} W_{k0}\cdot M_{k0}\cdot P_{k1\,k0}\cdot M_{k1}\\
\bar{E}_{0-2} &= \sum_{k0}\sum_{k1}\sum_{k2} W_{k0}\cdot M_{k0}\cdot P_{k1\,k0}\cdot M_{k1}\cdot P_{k2\,k1}\cdot M_{k2}\\
&\text{---------------------------------}\\
\bar{E}_{0-i} &= \sum_{k0}\sum_{k1}\cdots\sum_{ki} W_{k0}\cdot M_{k0}\prod_{i=1}^{i} P_{ki\,ki-1}\cdot M_{ki}\\
&\text{---------------------------------}\\
\bar{E}_{0-r} &= \sum_{k0}\sum_{k1}\cdots\sum_{kr-1} W_{k0}\cdot M_{k0}\prod_{i=1}^{r} P_{ki\,ki-1}\cdot M_{ki}
\end{aligned}\right| \qquad (13)$$

$$\left(P_{ki\,ki-1}\,\middle|\begin{matrix}=0\\=1\end{matrix}\right).$$

2. Non-finite CHMS with independent direct control (Fig. 2). Analogously we determine:

$$\left.\begin{aligned}
\bar{E}_{0-i} &= \sum_{k0}\sum_{k1}\cdots\sum_{ki} W_{k0}\cdot M_{k0}\prod_{i=1}^{i} P_{ki\,ki-1}\cdot M_{ki}\\
&\text{---------------------------------}\\
\bar{E}_{0-r} &= \sum_{k0}\sum_{k1}\cdots\sum_{kr} W_{k0}\cdot M_{k0}\prod_{i=1}^{r} P_{ki\,ki-1}\,M_{ki}
\end{aligned}\right| \qquad (14)$$

$$\left(P_{ki\,ki-1}\,\middle|\begin{matrix}=0\\=1\end{matrix}\right).$$

3. Non-finite CHMS with non-independent direct control (Fig. 3).

$$\left.\begin{aligned}
\bar{E}_{0-i} &= \sum_{k0}\sum_{k1}\cdots\sum_{ki} W_{k0}\,M_{k0}\prod_{i=1}^{i} P_{ki\,ki-1}\,M_{ki}\\
&\text{---------------------------------}\\
\bar{E}_{0-r} &= \sum_{k0}\sum_{k1}\cdots\sum_{kr} W_{k0}\,M_{k0}\prod_{i=1}^{r} P_{ki\,ki-1}\,M_{ki}
\end{aligned}\right| \qquad (15)$$

$$(0\leq P_{ki\,ki-1}\leq 1).$$

4. Finite CHMS with non-independent indirect control (Fig. 4).

$$\left.\begin{aligned} \bar{E}_{0-1} &= \sum_{l} \sum_{k0} \sum_{k1} \cdots \sum_{ki} W_{k0} M_{k0} \prod_{i=1}^{i} P_{ki\,ki-l} M_{ki} \\ \bar{E}_{0-r} &= \sum_{l} \sum_{k0} \sum_{k1} \cdots \sum_{ki} W_{k0} M_{k0} \prod_{i=1}^{r} P_{ki\,ki-l} M_{ki} \end{aligned}\right| \qquad (16)$$

$$(l = 1, 2, \ldots, i-2, i-1) \quad (0 \leq P_{ki\,ki-l} \leq 1).$$

5. Non-finite CHMS with non-independent indirect control (Fig.5).

$$\left.\begin{aligned} \bar{E}_{0-1} &= \sum_{l} \sum_{k0}^{i} \sum_{k1} \cdots \sum_{ki} W_{k0} M_{k0} \prod_{i=1}^{i} P_{ki\,ki-l} M_{ki} \\ \bar{E}_{0-r} &= \sum_{l} \sum_{k0} \sum_{k1} \cdots \sum_{kr} W_{k0} M_{k0} \prod_{i=1}^{r} P_{ki\,ki-l} M_{ki} \end{aligned}\right| \qquad (17)$$

$$(l = 1, 2, \ldots, i-2, i-1) \quad (0 \leq P_{ki\,ki-l} \leq 1).$$

Discussion of the results and basic conclusions

By comparing the obtained algorithms (13)–(17) for the evaluation of the efficiencies of the CHMS of various forms, it may be established that all of them represent detailed cases of (17). The reason is that the generalized type of CHMS is proved to be the CHMS with non-independent indirect control. In this way (17) does not change, when in the course of the functoning of the CHMS their hierarchical structures change in the classified senses.

Consequently the obtained algorithm (17) for evaluating the efficiency of the non-finite CHMS with non-independent indirect control may be utilized also for evaluating the efficiency of CHMS with variable structures in the described senses.

So if a CHMS turns finite, then in (17) $\sum_{kr} = 1$ is obtained; if it turns into a CHMS of independent control, then $- P^{k}_{iki-l} \Big| \begin{matrix} = 0 \\ = 1 \end{matrix}$; and if it turns into a CHMS with direct control, then $- P_{ki\,ki-l} = P_{ki\,ki-l}$.

In determining the efficiencies of the CHMS the corresponding algorithms do not introduce any limitation concerning the application of the CHMS.

Therefore it is advised to utilize the deduced quantitative relationships for evaluating the efficiencies of the CHMS for the purpose of selecting the rational hierarchical structures of great organizations and technical systems

both in the productive and the non-productive fields of national economy, including the greater part of the intended automated control systems of technological processes, the automated control systems of the enterprises, the information systems, etc.

On the basis of the established efficiency criterion the structure of the system controls must be improved by selecting the CHMS variation, ensuring the maximum efficiency, i.e. where

$$\bar{E} \to \max. \tag{18}$$

Other approximations are also possible for the formulation of the efficiency of the CHMS, with due consideration to the expenses of their creation. But in any case the selected variation must always be the most efficient one.

References

1. *Гвишиани Д. М. (Gvishiani, D. M.)*, Организация и управление (Organization and control) Изд-во «Наука», М., 1972.
2. *Янг С. (Joung, S.)*, Системное управление организацией (Management: a system analysis) Изд-во «Сов. радио», М., 1972.
3. *Фаткин Ю. М. (Fatkin, Yu. M.)*, Оптимальное управление в иерархических структурах (Optimum control of hierarchial structures). ДАН СССР, М., 202, *1*, 1972.
4. *Вертлиб В. А., Вишневский В. М. (Vertlib, V. A., Vishnevskii, V. M.)* Эффективность иерархических систем с учетом надежности (Efficiency of hierarchial systems with consideration to reliability). Автоматика и телемеханика, М., *11*, 1970.
5. *Mesarovic, M. D., Macko, D., Takahara, Y.*, Theory of Hierarchical System. Academic Press, New York and London, 1970.
6. *Пугачев В. С. (Pugachev, V. S.)*, Стохастические системы (Stochastic systems). ИПУ, М., вып. 9, 1973.

Эффективность сложных иерархических многоуровневых систем с учетом надежности

Й. БЛИЗНАКОВ

(София)

При решении проблем управления большими организационными и техническими системами постоянно стоят задачи совершенствования структуры управления.

Под сложной иерархической многоуровневой системой (СИМС) понимается система, состоящая из иерархически организованных управляемых и управляющих элементов (органов, подсистем), которые могут находиться в нескольких (множестве) рабочих состоянях и имеющая не меньше двух уровней управления.

Эффективность СИМС характеризуется средним выходным эффектом системы в единицу времени. Надежность органов СИМС рассматривается как характеристика сте-

пени вероятности их безотказной работы на протяжении срока действия системы. Надежность СИМС является функцией надежности ее органов.

Классифицированы основные виды СИМС (рис. 1—5). Приводится оценка эффективности функционирования СИМС [см. (10)]. Получены алгоритмы для оценки эффективности всех основных видов СИМС с учетом надежности.

Дается приложение теории к рациональному выбору иерархических структур организационных и технических систем в народном хозяйстве.

J. Bliznakov
Committee for Scientific Technical Progress and High Education
Sofia
Bulgaria

Problems of Control and Information Theory, Vol. 4 (4), pp. 339—352 (1975)

STRUCTURAL CHARACTER RECOGNITION BY FORMING PROJECTIONS

P. BREUER, M. VAJTA, Jr.

(Budapest)

(Received December 6, 1974)

In this paper a procedure for recognizing handwritten numerals is suggested The procedure starts with placing the numerals onto a rasterfield. On this basis as usual, a matrix with zero and one entry is associated with the character. Data reduction (feaure extraction) is made in two steps: first, integer valued "projections" are defined, second, identical runs of appropriate size in these projections are used to form a feature of three components. The small size of the feature space permits us to make decision using dictionary. The features are very easy to calculate, which is the main advantage of the method. A formal description of the feature extraction is given by using a goal-oriented, context-free grammar. An account is given on some experiments using this method.

1. Introduction

From the side of the structure of the patterns one can sometimes achieve a better classification than using statistical methods. The linguistic methods — as they are called — are especially useful, when some parts of the patterns have specific form, or, when there is an important interaction among some parts.

There are many problems when the structural examination leads to better solution than the simple receptor/categorizer model, e.g. complicated, two dimensional pictures where the statistical pattern recognitions are either too difficult or impossible; line patterns, biomedical pictures, etc.

The main advantage of the linguistic methods is that they allow us to discribe the pattern in terms of its structure and, therefore, to make better recognition. That is why it is worthwile attacking the recognition problems by linguistic methods, whenever it is possible. Fortunately, this is the case, when investigating the problem of the recognition of handprinted numerals, which is our main concern in this paper.

It seemed necessary to recall basic ideas of the linguistic methods, which we have done in Section 2. We have tried to formulate our problem in Section 3, as briefly as we could. In Section 4 a new linguistic method for recognizing handprinted numerals is suggested and discussed. Section 5 contains some experimental results. Finally, we should like to mention that a machine based on the ideas of this paper is under construction.

2. Preliminaries

Primitives are the smallest parts into which an object is subdivided for the purpose of the description.

Description is an appropriate representation of an object.

Structure is a specific arrangement interrelation of the primitives of an object. Structure is considered as a property assigned by man or machine and based on physical or virtual separation of parts.

Structural description is the representation of the structure of an object which defines the organization on the primitives chosen to represent object [1].

Picture processing is the analysis or synthesis of pictures, in other words two dimensional objects, by computer with or without human help. It includes computer graphics and digital pattern recognition.

In the linguistic pattern recognition one can separate the following steps:

1. *Preprocessing*, including coding of pattern data. This is a sequence of operations which transform the original picture (α) into more new pictures ($\alpha_1, \ldots, \alpha_k$) for further processing.

2. *Primitive detection.* This often consists of two steps: (a) transformation of the preprocessed picture into function of one variable; (b) further transformation into series of real numbers which are strongly connected with the primitive set.

3. Building up a goal-oriented *grammar*.

4. *Parsing*, or in other words, the syntactic analysis. This already includes the *Decision*-making.

For details on linguistic pattern recognition we refer to [1–15].

3. Problem statement

Our problem is recognition of handprinted characters. There is no need to explain how useful an appropriate solution could be in many practical situations e.g. post-code and cheque decoding. There are many methods having quite a high succes in Optical Character Recognition [16–24] which refer to machine printed alfanumeric characters, but not in handprinted [25–28] or even more handwritten — cursive (italics) — ones [29, 30]. Some interesting approach based on phenomenological attributes can be found in papers of Blesser et al. [31, 32].

Different realizations of the same character can be quite dissimilar. Despite of this, people recognize them. One may expect that using appro-

priate methods, a computer (or machine) can do this job, too. Our aim is to construct such a method using a classification of the characters, which is based on suitably chosen primitives. Unfortunately, whatever classification is made, the classes will overlap due to the randomness involved (see Fig. 1). Our aim is to choose the primitives (and, thus, the classes) in a way that such overlapping should occur as rarely as possible, which, in turn allows us to keep the number of the defective decisions very low.

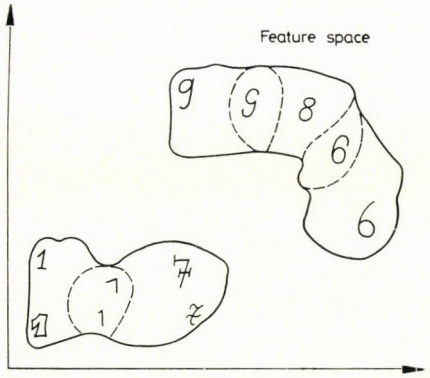


Fig. 1

4. A linguistic method for recognition handprinted numerals

Preprocessing, feature extracting. Indicated in the title of the paper, our preprocessing method has two main principles: forming mean and projection.

The procedure serves for recognition of numerals 0 to 9.

It was, however, tested for only six numerals because of technical reasons. (Some digits may look quite alike in different writing, therefore we should have need "finer" rasterfield, see below.)

The data-numerals are written onto a rasterfield or, better to say, squared wet (see Fig. 2) which is coded into binary form in a well-understandable way. (In Fig. 2 zeros are missing for better visual enjoyment.)

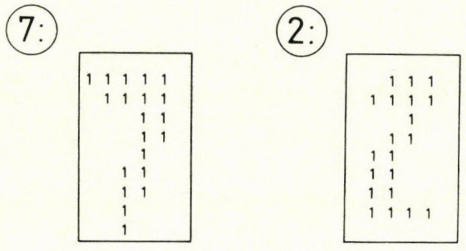


Fig. 2

An $n \times m$ matrix **A** with entries a_{ij} corresponds to the $n \times m$ rasterfield in the ordinary way, i.e.

$$a_{ij} = \begin{cases} 1 \text{ if the } (i,j)\text{-box of the raster contains of a segment of the character} \\ 0 \text{ otherwise.} \end{cases}$$

The mean of 1's in the ith row is defined as:

$$S_i = \frac{\sum_{j=1}^{m} j \cdot a_{ij}}{\sum_{j=1}^{m} a_{ij}}, \qquad \forall i, \quad j = 1, \ldots, n$$

where m is the number of columns.

If S_i is not an integer

$$S_i' = \text{entier } (S_i + 0.5) \tag{1}$$

The mean of 1's in the jth column is:

$$S_j = \frac{\sum_{i=1}^{n} i \cdot a_{ij}}{\sum_{i=1}^{n} a_{ij}}, \qquad \forall j, \; j = 1, \ldots, m$$

where n is the number of rows.

The definition of S_j' is now obvious (1).

One can expect that means provide a "good" skeleton of the figures. Actually this is the case e.g. with the numeral 3 (or see Figs 2 and 3a).

At the same time these means do not permit us to distinguish well e.g. between the numerals 8 and 0. Therefore, it seems necessary to divide the original raster into smaller units, and restrict the calculation of the mean to these small units. It seems reasonable to make this division "vertically" and "horizontally", that is, smaller units can be handle as rasterfields of the same type as the original was. In what follows, we will describe it through an example. In this example we suppose that n is divisible by 3 and m is divisible by 2 (which is an unimportant restriction). The matrix is divided vertically into two domains of size $n \times m/2$ each and horizontally into three ones of size $(n/3) \times m$. The mean is calculated for incomplete rows and columns Figs 3b, 3c). It is heuristically clear that better performance can be expected using the means of this incomplete matrices than those of the entire matrix For example in this rough subdivision one can distinguish between the digits

8 and 0 fairly well. The determination of these mean values terminates the processing.

The following step is the *primitive detection*. In this step we form vertical and horizontal projections of the cut-up-characters. The formal description of the projections goes as follows:

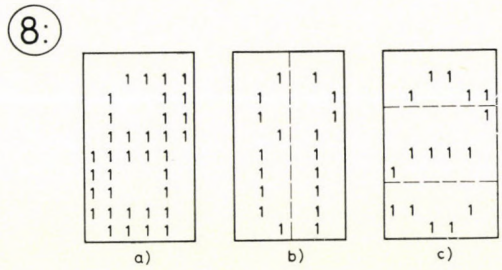


Fig. 3

Suppose that in general the raster (or, what is the same, the matrix **A**) is divided horizontally into r parts and vertically into s parts, such that the columns with subscripts

$$j \in J_t = \{j_t + 1, j_t + 2, \ldots, j_{t+1}\} \text{ form a unit, } t = 1, 2, \ldots, r$$

where

$$j_1 = 0 < j_2 < \ldots < j_{t+1} = m$$

and, similarly, the rows with subscripts

$$i \in I_t = \{i_t + 1, i_t + 2, \ldots, i_{t+1}\} \text{ form a unit, } t = 1, 2, \ldots, s$$

where

$$i_1 = 0 < i_2 < \ldots < i_{s+1} = n.$$

The preprocessing will provide us the mean values

$$S'_{i,t} \triangleq \text{entier}\left(\frac{\sum_{j \in J_t} j \cdot a_{ij}}{\sum_{j \in J_t} a_{ij}} + 0.5\right), \quad t = 1, 2, \ldots, r \qquad (1^*)$$

and

$$S''_{j,t} \triangleq \text{entier}\left(\frac{\sum_{i \in I_t} i \cdot a_{ij}}{\sum_{i \in I_t} a_{ij}} + 0.5\right), \quad t = 1, 2, \ldots, s, \qquad (2)$$

(In the case if any of these expressions is not meaningful, define it as zero.)

The mean values (1*) and (2) are used to define the vertical and horizontal projections, respectively:

Let

$$a^*(i,j) = \begin{cases} 1 & \text{if } j = S'_{i,t_0} \text{ for some } 1 \leq t_0 \leq r \\ 0 & \text{otherwise} \end{cases}$$

and

$$a^{**}(i,j) = \begin{cases} 1 & \text{if } i = S''_{j,t_0} \text{ for some } 1 \leq t_0 \leq s \\ 0 & \text{otherwise.} \end{cases}$$

The vertical and horizontal projections are defined

by $$v(i) = \sum_{j=1}^{m} a^*(i,j), \qquad 1 \leq i \leq n$$

and $$h(j) = \sum_{i=1}^{n} a^{**}(i,j), \quad 1 \leq j \leq m$$

respectively. Here, obviously $v(i)$ and $h(j)$ ranges over the integers $0, 1, \ldots, r$ and $0, 1, \ldots, s$, respectively. In our example above, $r = 2$ and $s = 3$. For the character of Fig. 3, we have

$$\underline{v} = [0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3]$$

$$\underline{h} = [3, 3, 3, 3, 3, 2].$$

The derivation of the horizontal and vertical projections above serves illustrative background rather than a computational rule. For this purpose the following definitions are used:

Definition 1. Let $v(i)$ denote the number of the classes J_t, $1 \leq t \leq r$, $i = \overline{1, n}$, for which there is a $j \in J_t$ such that $a_{ij} = 1$.

Let $h(j)$ denote the number of the classes I_t, $1 \leq t \leq s$ for which there is an $i \in I_t$ such that $a_{ij} = 1$, $j = \overline{1, m}$.

The vectors $\underline{v} = \{v(i), \ldots, v(n)\}$ and $\underline{h} = \{h(j), \ldots, h(m)\}$ are called vertical and horizontal projections, respectively.

It is clear that this definition is equivalent to the former one. The vectors $\underline{v} = \{v(i), \ldots, v(n)\}$ and $\underline{h} = \{h(j), \ldots, h(m)\}$ are used to form the primitives in our linguistic method. There are three primitives only, which yield a great reduction of the set of possible vectors. (We note that it is possible to define more primitives in the same fashion if necessary.)

Definition 2. Let $k^*(r, n)$ be a given number, and define the vertical upper primitive, VU as the greatest integer which is assumed at least k^* times consecutively by the coordinates $v(1)$, $v(2), \ldots, v\left(\frac{n}{2} - 1\right)$. If such an integer does not exist, let $VU = 0$. Similarly let the vertical lower primitive

VL be defined as the greatest integer which is assumed at least k^* times consecutively by the coordinates $v\left(\frac{n}{2}-1\right), \ldots, v(n)$, and let $VL = 0$ if such an integer does not exist.

Obviously, both VU and VL depend on k^*, which, in turn, depends on r and n. Furthermore $0 \leq VU,\ VL \leq r$.

Definition 3. Let $k_*(s, m)$ be a given number, and define the horizontal primitive H as the greatest integer which is assumed at least k_* times consecutively by the coordinates $h(1), \ldots, h(m)$, and let $H = 0$ if such an integer does not exist.

Clearly, H depends on k_*, which in turn depends on m, s. Furthermore $0 \leq H \leq s$.

It has to be emphasized, that the values m, n, r, s, k^*, k_* should be determined experimentally in such a way, that the classification would have as small error probability as possible.

Linguistic analysis. Now we introduce a context-free grammar to analyse the pattern syntactically (parsing). This goal-oriented grammar uses projection "thicknesses" as primitives.

With the usual terminology the grammar is defined by the fourtuple:

$$G = (V_N, V_T, S, P)$$

where G is the grammar,

V_N is a finite set of nonterminals (variables),

V_T is a finite set of terminals (here primitives) $(V_N \cap V_T = \Phi)$,

$S \in V_N$ is the start or "sentence" symbol,

P a finite set of productions or rewriting rules having the form of $\alpha \rightarrow \beta$ where α and β are strings over $V_N \cup V_T$ and α contains at least one nonterminal symbol.

In our illustrative example, where $r = 2$, $s = 3$, we have

$$V_N = \{S, H, VU, VL, 0X, 1X, 2X, 3X, 0Y, 1Y, 2Y\}$$

$$V_T = \{0, 1, 2, 3\}.$$

The language $\mathfrak{L}(G)$ generated by grammar G is a set of strings

$$\mathfrak{L}(G) = \left\{ Q \,|\, Q \in V_T, \text{ and } Q \text{ starts with either } 0, 1, 2 \text{ or } 3 \text{ and ends with either } 0 \text{ or } 1 \text{ or } 2 \right\}$$

P:

$$\begin{array}{llll} S \to H & H \to 0X & VU \to 0Y & VL \to 0 \\ X \to VU & H \to 1X & VU \to 1Y & VL \to 1 \\ Y \to VL & H \to 2X & VU \to 2Y & VL \to 2. \\ & H \to 3X & & \end{array}$$

For every pattern this grammar yields a terminal sentence of three numbers. These are used to make the decision. It is obvious that the decision rule must be established experimentally. Our experiments concerning the particular case $r = 2$, $s = 3$ resulted in the following parsing $\mathfrak{L}(G)$:

$$\begin{array}{ll} \mathfrak{L}(G): & 111 \Rightarrow 7 \\ & 112 \Rightarrow 7 \\ & 113 \Rightarrow 2 \\ & 122 \Rightarrow 6 \\ & 123 \Rightarrow 6 \\ & 212 \Rightarrow 9 \\ & 213 \Rightarrow 9 \\ & 222 \Rightarrow 0 \\ & 223 \Rightarrow 8. \end{array}$$

All patterns, having terminal sentence not listed here, are rejected.

Actually words of this language refer clearly to the form of patterns (thickness of projection show whole, single line etc.). Three numbers in a proper context form the terminal numeral. In this hierarchy of composition lies the power of linguistic method. The parsing leads to classifying the pattern, which terminates this syntax-controlled recognition process.

5. Experimental results

The linguistic method discussed above will be used in a complete numeral recognition system. A flying-spot scanning device is under building. Its output is connected through an A/D converter to a computer, where software program will make the recognition based on our method. The output of the A/D converter is the digitized picture in binary form.

In order to investigate the goodness of the method, we have made some experiments with the particular case of rasterfield 10×6, $r = 2$, $s = 3$ (see Fig. 4). The "roughness" of the rasterfield is due to the manual evaluation of the results. This is the reason of another restriction too, namely, that the procedure was tested on the digits 2, 6, 7, 8, 9, 0 only. It turned out that, when forming primitives it is the best looking for three identical consecutive elements horizontally and for four ones vertically.

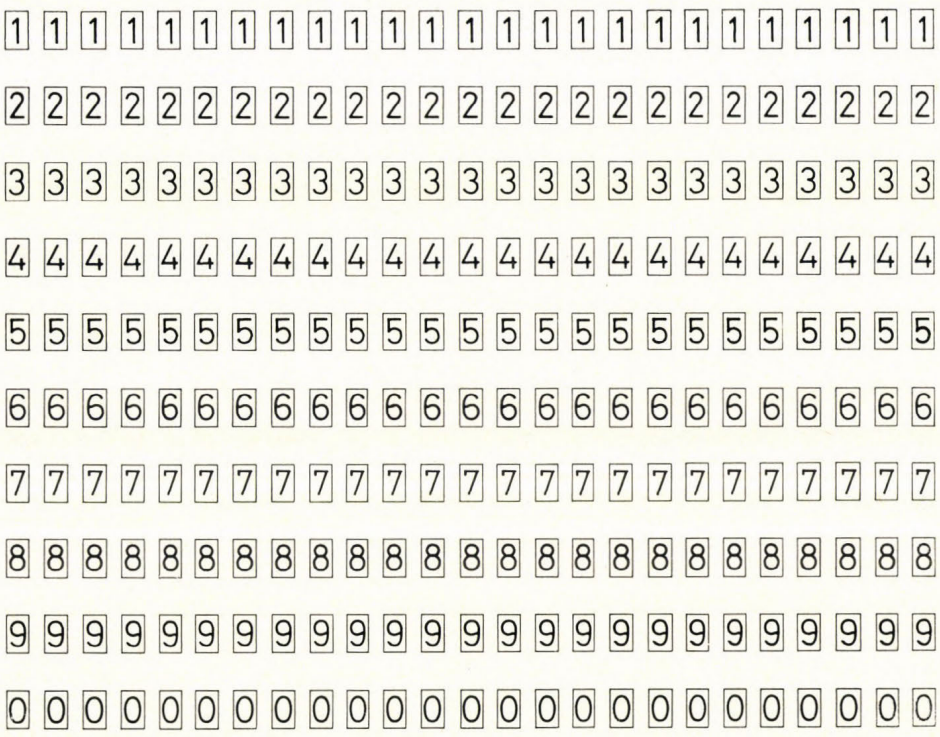


Fig. 4

Table 1 summarizes the results of an experiment with three persons, where each of them had written 25 numerals of all the six type.

There is no need to explain the dependence of the procedure on so-called "beauty" of writing, especially in the case of such a "rough" raster. The IIIrd person has written very beautiful numerals by all means for this program. The greatest problem was to recognize numeral 2. A high percentage of them was classified as 9 or as 7. The number of rejections is neglectable except for the numeral zero. The reason of the more rejections in this case was that

Table 1

	correct recogn.		reject		error	
	samples	%	samples	%	samples	%
Ist Person						
2 (25 samples)	13	52	0	0	12	48
6 (25 samples)	21	84	0	0	4	16
7 (25 samples)	24	96	1	4	0	0
8 (25 samples)	14	56	0	0	11	44
9 (25 samples)	21	84	0	0	4	16
0 (25 samples)	17	68	6	24	2	8
IInd Person						
2 (25 samples)	15	60	0	0	10	40
6 (25 samples)	20	80	0	0	5	20
7 (25 samples)	16	64	0	0	9	36
8 (25 samples)	24	96	0	0	1	4
9 (25 samples)	23	92	0	0	2	8
0 (25 samples)	18	76	0	0	7	24
IIIrd Person						
2 (25 samples)	16	64	0	0	9	36
6 (25 samples)	25	100	0	0	0	0
7 (25 samples)	25	100	0	0	0	0
8 (25 samples)	24	96	1	4	0	0
9 (25 samples)	25	100	0	0	0	0
0 (25 samples)	24	96	1	4	0	0
Sum total						
2 (75 samples)	44	58.3	0	0	31	41.7
6 (75 samples)	66	88	0	0	9	12
7 (75 samples)	64	85.4	1	1.33	10	13.33
8 (75 samples)	63	84	0	0	12	16
9 (75 samples)	62	82.7	0	0	13	17.3
0 (75 samples)	59	78.7	7	9.3	9	12

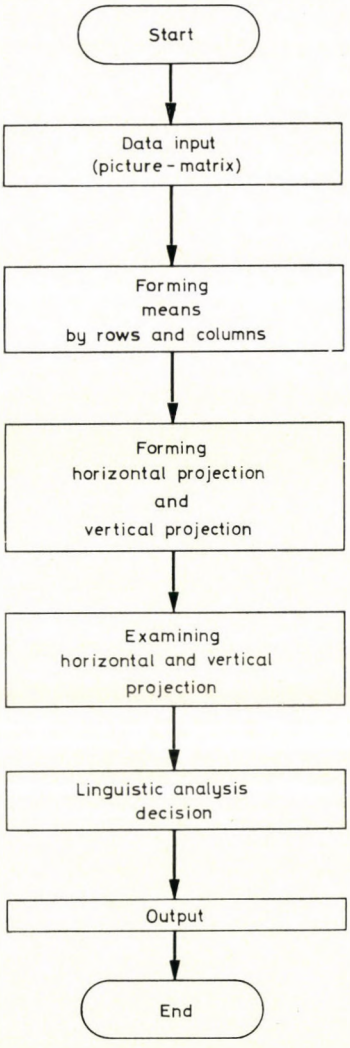


Fig. 5

person I has written too small characters. Figure 5 shows the flowchart of the program written in ALGOL for RASDAN–3 computer, and Table 1 presents the results.

6. Concluding remarks

On the basis of our experiments this method can be considered as a promising one. Its great advantage lies in the very simple computational work involved.

It is the task of further experiments to determine the most appropriate raster size, subunits (e.g. the numbers r and s), the length of the identical consecutive runs required in the projection vectors. Clearly, it is possible to construct a learning process to solve this task.

Our method actually does not make too much use of the skeleton, or the grammar. We used skeleton (means) description just to show the facts which motivated our investigations. On the other hand, the linguistic description may be useful in making the decision if there are more than 3 primitives. Also, it indicates the possibility of using stochastic grammars for noisy patterns.

Finally, it should be mentioned that our method is a goal-oriented syntax-controlled linguistic recognition system. Looking it from the viewpoint of formal language theory and considering the parsing one can see that our grammar is context-free.

*

Acknowledgement. We are very grateful to dr. T. Nemetz for the valuable discussion and very important remarks on the topics of this paper.

References

1. *Kaneff, S.*, Pattern cognition and organization of information, Frontier of Pattern Recognition (ed. by Watanabe), New York, Academic Press, 1972.
2. *Miller, W. F., Shaw, A. C.*, Linguistic methods in picture processing — A survey, Proc. Fall. Comput. Conf., 1968.
3. *Fu, K. S., Swain, P. H.*, On syntactic pattern recognition, Software Engineering (ed. by Tou), vol. II, New York, Acad. Press, 1971.
4. *Zavalishin, N. V., Muchnik, I. B.*, Linguistic (structural) approach to the pattern recognition problem (Review), Automation and Remote Control, No. 8, pp. 1263–1291, Aug. 1969.
5. *Grenander, U.*, A unified approach to pattern analysis, Advances in Computers, vol. 10, Acad. Press, 1970.
6. *Bierman, A. W., Feldman, J. A.*, A survey of results in grammatical inference, Frontier of Pattern Recognition (ed. by Watanabe), New York, Acad. Press, 1972.
7. *Narasimhan, R.*, Labeling schematic and syntactic description of pictures, Information and Control, vol. 7, pp. 151–179, 1964.
8. *Fu, K. S.*, On syntactic pattern recognition and stochastic languages, Frontier of Pattern Recognition (ed. by Watanabe), New York, Acad. Press, 1972.
9. *Eden, M.*, Handwriting and pattern recognition, IRE Trans. Inf. Theory, Vol. 8, pp. 160–166, 1963.
10. *Narasimhan, R.*, A linguistic approach to pattern recognition, Digital Computer Lab. Univ. of Illinois, Urbana, Ill., Rep. Nr. 121.
11. *Kirsch, R. A.*, Computer interpretation of English text and picture pattern, IEEE Trans. Electr. Comput., Vol. EC–13, pp. 363–376, 1964.
12. *Ledley, R. S. et al.*, FIDAC: Film input to digital automatical computer and associated syntax-directed pattern recognition programming system, Optical and Electrooptical Information Processing Systems (ed. by Tippet), Cambridge, Mass., M.I.T. Press, 1965.

13. *Feder, J.*, Language automata and classes of chain-encoded pattern, Techn. Rep. Nr. 400–165, New York Univ, Lab. for Electroscience Research, 1967.
14. *Pavlidis, T.*, Analysis of set patterns, Pattern Recognition Vol. 1, pp. 165–178, 1968.
15. *Kanal, L., Chandrasakaran, B.*, On linguistic, statistical and mixed models for pattern recognition, Frontier of Pattern Recognition (ed. by Watanabe), New York, Acad. Press, 1972.
16. *Chattan, J. B., Teacher, C. F.*, Character recognition techniques for address reading, Optical Character Recognition (ed. by Fischer et al.), New York, Spartan, 1968.
17. *Sherman, H.*, A Quasi-topological method for the recognition of line patterns, Int. Conf. on Inf. Process, Paris, 1959 (UNESCO).
18. *Sprick, W., Ganzhorn, K.*, An analogous method for pattern recognition by following the boundary, Int. Conf. on Inf. Process, Paris, 1959 (UNESCO).
19. *Ugher, S. H.*, Pattern detection and recognition, Proc. IRE, Vol. 47, pp. 1737–1752, 1959.
20. *Kasmierczak, H.*, The potential field as an aid to character recognition, Int. Conf. on Inf. Process, Paris, 1959 (UNESCO).
21. *Harmon, L. D.*, Automatic recognition of print and script, Proc. IEEE, Vol. 60, pp. 1165–1176, 1970.
22. *Harmon, L. D.*, Line-drawing pattern recognizer, Electronics, Vol. 33, pp. 39–43, 1960.
23. *Gábor, D.*, Character recognition by holography, Nature, Vol. 208, pp. 422–423, 1965.
24. *Booth, W. T. et al.*, Design consideration for stylized *Font* character readers, Optical Character Recognition (ed. by Fischer et al.), New York, Spartan, 1968.
25. *Bakis, R. et al.*, An experimental study of machine recognition of handprinted numerals, IEEE Trans. Syst. Sci. Cyb., Vol. SSC–4, pp. 119–132, 1968.
26. *Munson, J.*, Experiments in the recognition of handprinted text: Part I.: Character recognition, Proc. Fall Joint Comput. Conf., 1968, AFIPS Conf.: Vol. 33, pt 2, Washington, D. C., Spartan, 1968.
27. *Spinrad, R. J.*, Machine recognition of handprinting, Information and Control, Vol. 8, pp. 128–142, 1965.
28. *Parks, J. R. et al.*, An articulate recognition procedure applied to handprinted numerals, 2nd Int. Joint Conf. on Pattern Recognition, Copenhagen, 1974.
29. *Frischkopf, L. S., Harmon, L. D.*, Machine reading of cursive script, 4th London Symp. on Inf. Theory, London, Butterworth, 1961.
30. *Mermelstein, P., Eden, M.*, Experiments on computer recognition of connected handwritten words, Inf. and. Control Vol. 7, pp. 255–270, 1964.
31. *Blesser, B. et al.*, A theoretical approach for character recognition based on phenomenological attributes, Proc. of 1st Int. Joint Conf. on Pattern Recognition, Washington, D. C., 1973.
32. *Shillman, R. J. et al.*, Experimental methodologies for character recognition based on phenomenological attributes, Proc. of 2nd Int. Joint Conf. on Pattern Recognition, Copenhagen, 1974.

Стуктурное распознавание букв путем проектирования

П. БРЕЙЕР. М. ВАЙТА

(Будапешт)

В настоящее время имеется много методов решения проблемы распознавания букв. Большое развитие получил в последнее время лингвистический подход к распознаванию образов.

В статье рассматриваются основные проблемы лингвистического подхода к задаче распознавания и подчеркивается значение связи между методом и структурой образов. Рассматриваются отдельные этапы процесса распознавания.

Специфичным является то, что производится чтение букв, или точнее цифр, написанных вручную. Неизвестные цифры записываются на поле размером $n \times m$; после их преобразования образуется точка равновесия рисунка и определяются ее проекции на оси.

Рассматриваются теоретические основы данного метода и приводятся некоторые результаты, полученные на ЦВМ, а также делаются замечания о перспективах данного подхода.

P. Breuer
Research Institute for Telecommunication
H–1026, Budapest, II., Gábor Áron u. 65.

and

M. Vajta, Jr.
Department of Automation
Technical University
H–1111, Budapest, XI., Goldmann Gy. tér 3.
Hungary

Problems of Control and Information Theory, Vol. 4 (4), pp. 353–367 (1975)

О НЕКОТОРЫХ ВОПРОСАХ ОПТИМАЛЬНОГО БЫСТРОДЕЙСТВИЯ В СТОХАСТИЧЕСКИХ СИСТЕМАХ

В. Б. КОЛМАНОВСКИЙ

(Москва)

(Поступила в редакцию 31 мая 1974 г.)

Рассмотрена задача об оптимальном управлении по быстродействию в стохастических системах. Установлены условия существования оптимального и допустимого управления. Предложен способ приближенного решения задачи. Рассмотрены конкретные примеры.

1. Пусть управляемая система имеет вид

$$\dot{x}(t) = f(x(t)) + Bu(x(t)) + \sigma(x(t))\,\dot{\xi}(t), \qquad t \geq 0 \tag{1.1}$$
$$x(0) = x.$$

Здесь B — заданная постоянная матрица, $x(t)$ — вектор фазовых координат из евклидова пространства E_n, функция f и матрица σ удовлетворяют условиям, обеспечивающим существование решения уравнения (1.1) при $u = 0$, т. е. [1] при некоторой постоянной $c \geq 0$ и $x_1, x_2 \in E_n$

$$| f(x_1) - f(x_2) | + | \sigma(x_1) - \sigma(x_2) | \leq c\, | x_1 - x_2 |.$$

Наконец, $\xi(t)$ — винеровский процесс, вектор $x \in E_n$ задан, а $u(x) \in E_m$ — управление, причем при всех x

$$u(x) \subset U, \tag{1.2}$$

где U — заданное выпуклое ограниченное множество.

Пусть также задано некоторое множество Q (ограниченное или нет) в фазовом пространстве E_n. Граница Q далее обозначается через Q_1, а через $\tau_x(u)$ обозначается момент первого достижения множества Q процессом (1.1) с начальным условием $x(0) = x$, где произвольная начальная точка $x \in E_n$ при управлении u.

Если некоторые реализации процесса (1.1) не достигают Q ни за какое конечное время, то для этих реализаций случайная величина $\tau_x(u)$ полагается равной бесконечности.

Задача. Выбрать управление u, удовлетворяющее ограничению (1.2), минимизирующее функционал

$$M\,\tau_x(u), \tag{1.3}$$

где M — знак математического ожидания.

В теории оптимального управления детерминированными системами, т. е. системами (1.1) при $\sigma = 0$, задача быстродействия занимает особое место [2]. Для этой задачи получены довольно полные и законченные результаты. В частности, установлены теоремы существования и единственности оптимального по быстродействию управления, установлен вид оптимального управления, предложены алгоритмы приближенного построения управления, рассмотрены разнообразные примеры.

В то же время в поставленной выше задаче о минимизации среднего времени достижения управляемым стохастическим процессом заданного множества известные законченные результаты (см., например, [3]) сводятся к тому, что при некоторых условиях гладкости минимального значения $V(x)$ функционала (1.3) этот функционал удовлетворяет уравнению в частных производных Беллмана:

$$\begin{gathered}\min_{u\in U}\left(L_u V(x)+1\right)=0, \quad x\in Q_2\\ V(x)=0, \qquad\qquad x\in Q_1\\ L_u V=(f+BU)'\frac{\partial V}{\partial x}+\frac{1}{2}\,Tr\,\sigma\sigma'\frac{\partial^2 V}{\partial x^2}, \qquad V(x)\min_{u\in U} M\,\tau_x(u).\end{gathered} \tag{1.4}$$

Здесь Q_2 — дополнение $Q\cup Q_1$ до E_n; штрих — знак транспонирования; вектор $\partial V\,|\,\partial x$ имеет координаты $\partial V\,|\,\partial x_i$, матрица $\partial^2 V\,|\,\partial x^2$ имеет компоненты $\partial^2 V\,|\,\partial x_i$; $i, j = 1, 2, \ldots, n$; символ Tr — означает след матрицы.

Необходимо, однако, подчеркнуть, что задача (1.4) имеет не единственное решение. Поэтому, даже если $V(x)$ — достаточно гладкая функция, заранее не ясно, с каким именно решением уравнения Беллмана (1.4) она совпадает.

Разъясним сказанное на примере.

Пример. Пусть система (1.1) — скалярная и имеет вид

$$\dot{x}(t)=u(t)+\sqrt{2}\,\dot{\xi}(t),\ |\,u(t)\le 1,\ x(0)=x. \tag{1.5}$$

Множество Q есть интервал $(-\infty, \alpha]$ при некоторой постоянной α. При $x > \alpha$ уравнение Беллмана (1.4) имеет вид

$$\begin{gathered}\frac{d_2 V(x)}{dx^2}-\left|\frac{dV(x)}{dx}\right|+1=0, \quad x>\alpha\\ V(\alpha)=0,\end{gathered}$$

решением которого служит любая функция вида $x-\alpha+c\left(\exp(x-\alpha)-1\right)$, где c — произвольная неотрицательная постоянная. Заметим, что для этого примера оптимальное управление есть $u=-1$. Действительно, если $x(t, u)$ — решение задачи (1.5) при управлении u, то $x(t, -1) \leq x(t, u)$ с вероятностью 1 при любом управлении u, $|u| \leq 1$. При этом, как показано далее, оптимальное время перехода при управлении $u = 1$ равно $x - \alpha$ для любого $x \geq \alpha$.

Другое обобщение некоторых проблем оптимального быстродействия на линейные стохастические системы предложено в [4], где рассмотрена задача о минимизации суммарной вероятности пребывания управляемого случайного процесса вне заданной окрестности нуля за бесконечное время. Для решения этой задачи в [4] используется метод функций Ляпунова. Кроме того, в [4] обсуждаются способы приближенного решения задачи.

В разделе 2 настоящей работы установлены некоторые условия существования допустимого управления, формулируемые в терминах существования функций Ляпунова от траекторий системы, а в 3 рассмотрены условия существования оптимального управления, установлено, с каким решением уравнения Беллмана (1.4) совпадает минимальное значение функционала (1.3) и приведены некоторые примеры. Напомним [5], что процесс (1.1) при некотором управлении u называется возвратным, если $\tau_x(u) < \infty$ с вероятностью 1 и называется положительным, если $M\,\tau_x(u) < \infty$ для любых конечных x.

2. В этом параграфе рассмотрены условия существования допустимого управления.

Назовем управление u допустимым, если при этом управлении $M\,\tau_x(u) < \infty$ для любых конечных x и существует сильное решение уравнения (1.1), т. е. [1] существует случайный процесс $x(t)$, измеримый относительно $\xi(s)$, $0 \leq s \leq t$, удовлетворяющий с вероятностью 1 интегральному тождеству

$$x(t) - x = \int_0^t \left[f(x(s))\, Bu(x(s))\right] ds + \int_0^t \sigma(x(s)\, d\xi(s)\,.$$

Из [5] стр. 130 вытекает

Лемма 2.1. Пусть существуют в Q_2 функция $u(x)$ и дважды непрерывно дифференцируемая функция $V(x) \geq 0$ такие, что при управлении $u(x)$ существует решение системы (1.1) и

$$L_u\, V(x) \leq -c, \qquad x \in Q_2\,, \tag{2.1}$$

где c — некоторая положительная постоянная. Тогда $u(x)$ — допустимое управление, причем $M\,\tau_x(u) \leq c^{-1}\, V(x)$.

Пример 2.1. Пусть точка движется по прямой под действием управления u, $|u| \leq 1$ и случайной возмущающей силы $\sigma\dot{\xi}$, где σ — некоторая заданная постоянная. Уравнения движения имеют вид

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u + \sigma\,\dot{\xi}. \tag{2.2}$$

Требуется осуществить быстрейшее попадание на ось x_2, т. е. в данном случае Q есть прямая $x_1 = 0$. Покажем, что при любом σ допустимое управление есть

$$\begin{aligned} u &= 1, \quad x_1 < 0, \\ u &= -1, \quad x_1 > 0. \end{aligned} \tag{2.3}$$

Для этого зафиксируем некоторую постоянную $k_1 > \sigma^2/2$, а затем подберем k_2 так, чтобы при всех x_1, x_2 функция V, равная

$$V = k_2 + \frac{1}{2}x_2^2 + k_2 + x_1, \qquad x_1 > 0,$$

$$V = -k_1 x_2 + \frac{1}{2}x_2^2 + x_2 - x_1, \qquad x_1 < 0,$$

была неотрицательна. Вычисляя, получим с учетом (2.3), что всюду вне прямой $x_1 = 0$

$$L_u V = -k_1 + \frac{\sigma^2}{2}.$$

Значит, на основании леммы 2.1 при управлении (2.3) процесс (2.2) положителен. Далее можно показать, что управление (2.3) оптимальное. Действительно, при $x_1(0) > 0$ и любом управлении $u_1(x) \geq -1$ неравенство

$$M\,\tau_{u_1}(x) < M\,\tau_{-1}(x)$$

невозможно, т. к. с вероятностью 1

$$x_1(t, -1) \leq x_1(t, u_1),$$

где через $x_1(t, u)$ обозначено решение $x_1(t)$ системы (2.2) при управлении u. Подобным же образом убеждаемся в оптимальности управления $u = 1$ в области $x_1 < 0$.

Замечание. При исследовании вопроса о существовании допустимого управления в некоторых конкретных ситуациях требования леммы 2.1 могут быть несколько ослаблены. Точнее говоря, пусть в Q_2 найдутся функция $u(x)$ и непрерывная функция $V(x) \geq 0$ такие, что при управлении $u(x)$

существует решение системы (1.1) и для любых точек $t_1 \leq t_2$ на траекториях $x(t, u)$ системы (1.1)

$$M\, V(x(t_2, u)) - M\, V((x(t_1, u)) \leq -c(t_2 - t_1),$$

где c — некоторая положительная постоянная. Тогда $u(x)$ — допустимое управление. В частности, от функции V достаточно требовать, чтобы

$$\overline{\lim_{\Delta \to +0}} \frac{1}{\Delta} [MV(x(\Delta, u)) - V(x_0)] \leq -1. \tag{2.4}$$

При этом, если V удовлетворяет требованиям леммы 2.1, то предел в (2.4) есть $L_u\, V(x_0)$. Справедливость этих утверждений вытекает из результатов ([5] стр. 130).

3. В этом разделе рассматривается вопрос о существовании решения уравнения Беллмана и способах его определения. Всюду далее считается, что Q — ограниченная область в фазовом пространстве E_n с гладкой границей Q_1 типа Ляпунова ([7], стр. 346.) Обозначим через R_n последовательность таких шаров радиуса N в пространстве E_n, что $Q \subset R_N$. Границу R_n обозначим через r_N. Определим последовательность скалярных функций $V_N(x)$ из краевой задачи

$$\begin{aligned} &\min_{u \in U} (L_u\, V_N(x) + 1) = 0, \qquad x \in Q_2 \cap R_N, \\ &V_N(x) = 0, \quad x \in Q_1, \quad x \in r_N. \end{aligned} \tag{3.1}$$

Здесь выражение $L_u\, V_N(x)$ определяется формулой (1.4).

Приведем в виде леммы некоторые свойства функций $V_N(x)$, вытекающие из результатов [8—10].

Лемма 3.1. Предположим, что коэффициенты уравнения (3.1) удовлетворяют требованиям раздела 1, матрица $\sigma(x)$ невырождена при любом $x \in E_n$ и ограниченная область Q имеет границу Q_1 типа Ляпунова. Тогда существует единственное решение $V_N(x)$ задачи (3.1), причем $V_N(x) \geq 0$ и для любых $N_2 \geq N_1$

$$V_{N_2}(x) \geq V_{N_1}(x), \quad x \in Q_2 \cap R_{N_1}.$$

Доказательство. Существование единственного решения краевой задачи (3.1) вытекает из [8] (теорема 4.10.2), либо из [9] (теорема 3). Далее, в силу [10] (стр. 603) существует измеримая функция $u_{N_1}(x) \in U$, доставляющая минимум в уравнении (3.1). Значит, функция $V_N(x)$ удовлетворяет линейному уравнению

$$L_{u_N} V_N(x) + 1 = 0, \quad x \in Q_2 \cap R_N.$$

Отсюда, с учетом нулевых граничных условий (3.1) для $V_N(x)$ и [8] (стр. 160) вытекает, что $V_N(x) \geq 0$. Из неотрицательности всех $V_N(x)$ и теоремы 40.1 ([11]) следует справедливость последнего утверждения леммы 3.1.

Теорема 3.1. Пусть справедливы все требования леммы 3.1 и существует допустимое управление $\omega(x)$. Тогда

1) существует минимальное положительное решение внешней краевой задачи (3.1), обозначаемое через $V_0(x)$;

2) на любом ограниченном множестве изменения x последовательность $V_N(x)$ равномерно сходится к $V_0(x)$ при $N \to \infty$, монотонно не убывая;

3) ни при каком допустимом управлении нельзя получить значение функционала (1.3) меньше $V_0(x)$;

4) если при управлении $u_0(x)$, определяемом из условия

$$\min_{u \in U} L_u V_0(x) = L_{u_0} V_0(x),$$

существует решение системы (1.1), то $u_0(x)$ — оптимальное управление.

Доказательство. Положим

$$W(x) = M \tau_x(\omega).$$

Покажем вначале, что при всех N

$$V_N(x) \leq W(x). \tag{3.2}$$

Обозначим через $\tau_x^N(u)$ момент первого достижения процессом (1.1) границ множества $Q_2 \cap R_N$ при управлении u и $x \in Q_2 \cap R_N$, а через $x(t, u)$ — решение системы (1.1) при управлении u.

Поскольку, очевидно,

$$M \tau_x^N(\omega) \leq W(x), \tag{3.3}$$

то для доказательства (3.2) достаточно установить, что

$$V_N(x) \leq M \tau_x^N(\omega). \tag{3.4}$$

С этой целью применим формулу Дынкина ([12] стр. 191) к функции $V_N(x)$ и марковскому моменту $\tau_x^N(\omega)$, что возможно на основании определения $V_N(x)$, леммы 3.1 и конечности математического ожидания $\tau_x^N(\omega)$ в силу (3.3). Имеем с учетом (3.1) при $x \in Q_2 \cap R_N$

$$V_N\big(x(\tau_x^N(\omega), \omega)\big) - V_N(x) = M \int_0^{\tau_x^N(\omega)} L_\omega V_N\big(x(t, \omega)\big)\, dt \geq$$

$$\geq M \int_0^{\tau_x^N(\omega)} \min_{u \in U} L_u V_N\big(x(t, \omega)\big)\, dt = -M \tau_x^N(\omega).$$

Тем самым неравенство (3.4), а вместе с ним и (3.2) установлены.

Зафиксируем произвольную ограниченную область $Q_3 \subset E_n$ с границей того же типа, что и Q_1, причем $Q \subset Q_3$ и установим сходимость $V_N(x)$, $x \in Q_2 \cap Q_3$ при $N \to \infty$. Для этого заметим, что из условий теоремы и неравенств (3.2) с учетом априорной оценки [8] стр. 316 при любом N

$$\max_x \left\| \frac{\partial V_N(x)}{\partial x^2} \right\| \leq C_1, \quad x \in Q_3 \cap Q_2. \tag{3.5}$$

Здесь и далее C_i — некоторые неотрицательные постоянные.

Из (3.5), (3.2), невырожденности матрицы $\sigma(x)$ и справедливости для неё условия Липшица, используя оценку вторых производных решений уравнений (3.1) ([8] теорема 4.6.3), заключаем, что

$$\max_x \left\| \frac{\partial^2 V_N(x)}{\partial x^2} \right\| \leq C_2, \quad x \in Q_3 \cap Q_2. \tag{3.6}$$

Оценки (3.2), (3.5), (3.6) означают, что на любом ограниченном множестве $Q_3 \cap Q_2$ последовательность $V_N(x)$ вместе со своими двумя производными равномерно ограничена. Значит, ([13] стр. 240) последовательность $(V_N(x), \partial V_N(x) | \partial x)$ компактна в пространстве непрерывных функций для $x \in Q_3 \cap Q_2$. Будем считать, как и обычно, что и сама последовательность $(V_N(x), \partial V_N(x | \partial x))$ равномерно сходится. Обозначим предел $V_N(x)$ при $N \to \infty$ через $V_0(x)$. Тогда на основании [11] (стр. 176 теорема 42,9) и произвольности области Q_3 предел $V_0(x)$ есть решение задачи (1.4).

Установим теперь, что $V_0(x)$ есть минимальное положительное решение задачи (1.4). Предположим противное, т. е. предположим, что найдется другое положительное решение $V(x)$ задачи (1.4), причем

$$0 \leq V(x) \leq V_0(x). \tag{3.7}$$

Функция $V(x)$, будучи решением задачи (1.4), удовлетворяет уравнению (3.1) при любом N внутри области $x \in Q_2 \cap R_N$ и, кроме того,

$$V(x) = V_N(x) = 0, \quad x \in Q_1, \tag{3.8}$$

$$V(x) \geq V_N(x) = 0, \quad x \in r_N.$$

Из (3.7), (3.8) и принципа максимума для квазилинейных эллиптических уравнений (3.1) ([11] стр. 162, теорема 40,1) заключаем, что и при всех $x \in Q_2 \cap R^N$

$$V_N(x) \leq V(x). \tag{3.9}$$

Из (3.7), (3.9) и установленной выше сходимости $V_N(x)$ к $V_0(x)$ вытекает равенство $V(x) = V_0(x)$. Тем самым утверждения 1), 2) теоремы доказаны.

Для доказательства п. 3) теоремы предположим, что при некотором допустимом управлении получено значение $W_0(x)$ функционала (1.3) меньшее, чем $V_0(x)$. Тогда, аналогично оценке (3.2), устанавливается, что при любом N

$$V_N(x) \leq W_0(x).$$

Отсюда и из сходимости $V_N(x)$ к $V_0(x)$ убеждаемся в противном, что $W_0(x) \geq$ $\geq V_0(x)$. Наконец, для доказательства п. 4) теоремы положим в условиях леммы 2.1 функцию $V(x) = V_0(x)$ и $u(x) = u_0(x)$. Поскольку V_0, u_0 удовлетворяют уравнению Беллмана (1.4), то для них справедливо неравенство (2.1). На основании леммы 2.1

$$M\,\tau_x(u_0) \leq V_0(x).$$

С другой стороны, в силу установленного утверждения п. 3) теоремы

$$M\,\tau_x(u_0) \geq V_0(x).$$

Значит, утверждение п. 4) теоремы справедливо.

Теорема 3.1 доказана.

Замечание. Пусть в (1.1) $B = 0$ (т. е. система (1.1) неуправляема) и процесс (1.1) при $B = 0$ положителен. Тогда утверждение теоремы 3.1 о том, что время достижения есть минимальное положительное решение уравнения (1.4) при $B = 0$ следует из результатов [5, 6].

Доказанная теорема допускает ряд обобщений. Сформулируем некоторые из них в виде следствий.

Следствие 1. Пусть выполнены все условия теоремы, но функция $f(x)$ в (1.1) растет не быстрее некоторой степени $||x||$ при $||x|| \to \infty$ и удовлетворяет условию Липшица в каждом ограниченном множестве, т. е. при $||x_1|| \leq N$, $||x_2|| \leq N$, $x_1, x_2 \in E_n$

$$||f(x_1) - f(x_2)|| \leq c(N)\,||x_1 - x_2||, \tag{3.10}$$

где N — произвольное число, а $c(N)$ — некоторая заданная постоянная, зависящая от N.

Тогда все утверждения теоремы остаются в силе. Справедливость следствия 1 устанавливается повторением доказательства теоремы.

Следствие 2. Из доказательства теоремы видно, что её требование о существовании допустимого управления было использовано лишь при обосновании оценки (3.2). Значит, утверждения теоремы справедливы, если выполнены все её условия, либо условия следствия 1, за исключением требо-

вания о существовании допустимого управления, которое заменено любым из следующих предположений:

а) существует неотрицательное решение W_0 задачи (1.4);

б) последовательность $V_n(x)$ равномерно ограничена на любом ограниченном множестве изменения x;

в) последовательность $V_N(x)$ сходится поточечно к решению задачи (1.4).

Для доказательства достаточно заметить, что при любом из условий а) — в) справедлива оценка типа (3.2). Действительно, при условии а) в силу [11] (стр. 162) для любого N последовательность $V_N(x) \leq W_0(x)$, а из условия б) следует условие а). Установленные факты дают способ определения оптимального управления в задаче стохастического быстродействия. Для этого надлежит либо определить минимальное положительное решение задачи (1.4), либо построить последовательность $V_N(x)$ из (3.1) и перейти к пределу при $N \to \infty$. В любом из этих случаев, если минимум в (1.4) достигается на допустимом управлении, то оно и будет оптимальным.

Замечание. Пусть $V(x)$ — положительное решение уравнения Беллмана (1.4) и справедливы требования раздела 1. Тогда для любого $1 > \varepsilon > 0$ найдется такое допустимое управление $u_0(x)$, что

$$M\tau_x(u_0) \leq \frac{1}{1-\varepsilon} V(x) . \tag{3.11}$$

Для обоснования (3.11) заметим, что из доказательства лемм 2.1, 2.2 статьи [14] вытекает оценка

$$u_0' \frac{\partial V}{\partial x} \leq \min_{u \in U} u \frac{\partial V}{\partial x} + \varepsilon , \tag{3.12}$$

причем $u_0(x) \in U$ и удовлетворяет локальному условию Липшица (3.10). Значит, ввиду ограниченности U, при управлении $u_0(x)$ существует [15] единственное решение системы (1.1). Из (3.12) и уравнения Беллмана для V заключаем, что

$$L_{u_0} V \leq -1 + \varepsilon .$$

Следовательно, на основании леммы 2.1 управление u_0 допустимо и справедливо неравенство (3.11). Из (3.11) вытекает.

Теорема 3.2. Пусть выполнены предположения леммы 3.1. Тогда

1) если существует минимальное положительное решение V_0 задачи (1.4), то для любой фиксированной начальной точки x_0 и любого $\varepsilon > 0$ найдет-

ся такое допустимое управление u, что

$$M\,\tau_{x_0}(u) \leq V_0(x_0) + \varepsilon; \tag{3.13}$$

2) если задача быстродействия разрешима и минимальное время $V_0(x)$ — дважды непрерывно дифференцируемо, то V_0 — минимальное положительное решение уравнения Беллмана (1.4). Действительно, оценка (3.13) следует из (3.11). Для доказательства второго утверждения заметим, что минимальное время $V_0(x) \geq 0$ есть решение задачи (1.4). Поэтому в силу следствия 2 п. а) существует минимальное положительное решение W задачи (1.4). Если бы в некоторой точке x_0 было $W(x_0) < V_0(x_0)$, то ввиду п. 1) теоремы 3.2 нашлось такое допустимое управление u, при котором $M\,\tau_{x_0}(u) < V_0(x_0)$, что противоречит минимальности времени $V_0(x_0)$. Отметим, что теорема 3.1 доставляет также способ приближенного или численного определения времени быстродействия при помощи построения соответствующей последовательности $V_N(x)$.

Пример 3.1. Пусть система (1.1) — скалярная и имеет вид

$$\dot{x}(t) = f(x(t)) + u(t) + \sigma\dot{\xi}(t). \tag{3.14}$$

Здесь $\sigma \neq 0$ — накоторая постоянная, управление $u(t)$ подчинено требованию $|u(t)| \leq 1$, наконец, множество Q, как и в примере 1, есть интервал $(-\infty, \alpha]$.

Потребуем, чтобы

$$\int_x^\infty \frac{1}{b(s)}\,ds < \infty\,. \tag{3.15}$$

Здесь

$$b(t) = \exp\left\{-\int_0^t \frac{2}{\sigma^2}\,(f(s) - 1)\,ds\right\} \tag{3.16}$$

Требование (3.15) выполнено н. п., если $\sup\limits_{x \geq \alpha} f(x) < 1$.

Покажем, что минимальное время перехода в этом примере есть

$$W(x) = \frac{2}{\sigma^2}\int_\alpha^x b(t)\,dt \int_t^\infty \frac{ds}{b(s)}, \qquad x \geq \alpha\,. \tag{3.17}$$

Из (3.14)—(3.17) вытекает, что $L_{-1}\,W = -1$. Значит, управление $u = -1$ допустимо, а потому и оптимально, ибо на основании теоремы сравнения [16] (стр. 120) $x(t, -1) \leq x(t, u)$ с вероятностью 1 при любом допустимом u.

Отсюда и из теоремы 3.1 следует, что оптимальное время перехода $V(x)$, соответствующее управлению -1, определяется из соотношения

$$V(x) = \lim_{N \to \infty} V_N(x), \tag{3.18}$$

где функции $V_N(x)$ есть решения краевой задачи

$$\frac{\sigma^2}{2} \frac{d^2 V_N(x)}{dx^2} + (f(x) - 1) \frac{dV_N(x)}{dx} + 1 = 0, \qquad x > \alpha$$

$$V_N(x) = 0, \quad x = \alpha, \quad x = N.$$

Вычисляя, получим

$$V_N(x) = \frac{2}{\sigma^2} \left[\int_\alpha^N b(t)\, dt \right]^{-1} \int_\alpha^x b(t_1)\, dt_1 \int_x^N b(t_2)\, dt_2 \int_{t_1}^{t_2} \frac{dt_3}{b(t_3)} \tag{3.19}$$

Но процесс (3.14) возвратен при управлении $u = -1$. Поэтому ([5] стр. 138)

$$\lim \int_\alpha^N b(t)\, dt = \infty .$$

Отсюда, переходя к пределу при $N \to \infty$ в (3.18) с учетом (3.15), (3.17) и правила Лопиталя, заключаем, что оптимальное время перехода $V(x) = W(x)$.

Отметим, что в этом примере при $u = 0$ процесс (3.14) может быть даже невозвратен, н. п. в силу [5], если $f(x) = \frac{1}{2}$ при $x \geq \alpha$.

Пример 3.2. Рассмотрим управляемое движение твердого тела под воздействием случайных возмущений. Пусть x_i, $i = 1, 2, 3$ — компоненты кинетического момента относительно жестко связанной с телом системы координат, оси которой совпадают с главными центральными осями инерции тела. Уравнения движения Эйлера имеют вид

$$\begin{aligned} \dot{x}_1 &= x_2 x_3 (a_2 - a_3)\ (a_3 a_2)^{-1} + u_1 + \sigma \xi_1 \\ \dot{x}_2 &= x_1 x_3 (-a_1 + a_3)\ (a_1 a_3)^{-1} + u_2 + \sigma \xi_2 \\ \dot{x}_3 &= x_1 x_2 (a_1 - a_2)\ (a_1 a_2)^{-1} + u_3 + \sigma \xi_3 . \end{aligned} \tag{3.20}$$

Здесь a_i — главные центральные моменты инерции, σ — постоянная, управление $u = (u_1, u_2, u_3)$ удовлетворяет при постоянной $b > 0$ ограничению

$$|u| = (u_1^2 + u_2^2 + u_3^2)^{1/2} \leq b .$$

Отметим, что система (3.20) с ограничением (3.21) допускает, например, следующую интерпретацию [17]. Имеется в распоряжении одна пара управляемых газовых струй, создающих вращающие моменты, причем, газовые сопла можно установить под любым углом относительно твердого тела и воздействие управления на систему одинаково в каждом направлении.

Для системы (3.20) поставим задачу о наискорейшем в смысле функционала (1.3) приведении её в сферу $Q = |x| \leq r$, где r — заданное положительное число.

Решение этой задачи получим с помощью следствия 2 к теореме в соответствии с которым необходимо построить вначале последовательность решений V_N краевой задачи (3.1), затем определить предел V_0 последовательности V_N при $N \to -$ и, наконец, по V_0 построить оптимальное управление. Конкретизируем соотношения (3.1) применительно к рассматриваемому случаю. Имеем при $r < |x| < N$

$$\frac{\sigma^2}{2}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial x_3^2}\right)V_N(x)+\min_{u,\,|u|\leq b}\left(u_1\frac{\partial V_N(x)}{\partial x_1}+u_2\frac{\partial V_N(x)}{\partial x_2}+u_3\frac{\partial V_N(x)}{\partial x_3}\right)+$$

$$+x_2x_3\frac{a_2-a_3}{a_3a_1}\frac{\partial V_N(x)}{\partial x_1}+x_1x_3\frac{a_3-a_1}{a_1a_3}\frac{\partial V_N(x)}{\partial x_2}+$$

$$+x_1x_2\frac{a_1-a_2}{a_1a_2}\frac{\partial V_N(x)}{\partial x_3}=-1,\quad V_N(x)=0,\quad |x|=r,\quad |x|=N. \tag{3.22}$$

Определим число $r_1 = r_1(N)$ как корень уравнения

$$\int_r^{r_1} z(t)\,dt\int_t^{r_1}\frac{du}{z(u)}=\int_{r_1}^{N} z_1(t)\,dt\int_{r_1}^{t}\frac{du}{z_1(u)}. \tag{3.23}$$

Здесь введены обозначения

$$z(t)=\left(\frac{r}{t}\right)^2 e^{\frac{2b}{\sigma^2}(t-r)}$$

$$z_1(t)=\left(\frac{r}{t}\right)^2\exp\left[\frac{2b}{\sigma^2}(r-t)\right]. \tag{3.24}$$

Левая часть (3.23) при $r_1 = r$ равна нулю и с учетом (3.24) монотонно возрастает. Правая же часть (3.23) равна нулю при $r_1 = N$ и монотонно убывает при $r \leq r_1 \leq N$. Значит, уравнение (3.23) при любом фиксированном N имеет

и притом единственное решение $r_1(N)$, причем

$$r < r_1(N) < N. \tag{3.25}$$

Подстановкой в (3.22) убеждаемся, что решение (3.22) задается формулами

$$V_N(x) = \frac{2}{\sigma^2} \int_r^{|x|} z(t)\,dt \int_r^{r_1} \frac{dt_1}{z(t_1)} \qquad \text{при} \qquad r \le |x| \le r_1,$$

$$V_N(x) = \frac{2}{\sigma^2} \int_{|x|}^{N} z_1(t)\,dt \int_{r_1}^{t} \frac{dt_1}{z_1(t_1)} \qquad \text{при} \qquad r_1 \le |x| \le N.$$

Покажем теперь, что

$$\lim_{N\to\infty} r_1(N) = \infty. \tag{3.26}$$

Вычислим производную $dr_1(N)\,|\,dN$, существующую в силу [18] (стр. 452), для чего продифференцируем по N обе части равенства (3.23). Имеем

$$\frac{dr_1(N)}{dN} = \left[z_1(N) \int_{r_1}^{N} \frac{dt}{z_1(t)}\right] \cdot \left[\frac{1}{z_1(r_1)} \int_{r_1}^{N} z_1(t)\,dt + \frac{1}{z(r_1)} \int_r^{r_1} z(t)\,dt\right]^{-1}. \tag{3.27}$$

Отсюда и из (3.25) вытекает, что последовательность $r_1(N)$ монотонно возрастает, т. е. существует её предел при $N \to \infty$. Предположим, что равенство (3.26) неверно. Это означает, что существует такое число r_0, $r < r_0 < \infty$, что $r_1(N) \le r_0$ при всех N. Покажем, однако, что в этом случае найдется такая постоянная $c_1 > 0$, что для всех $N > r_0$ справедлива оценка

$$\frac{dr_1(N)}{dN} \ge c_1, \tag{3.28}$$

которая, очевидно, противоречит неравенству $r_1(N) \le r_0$.

В силу (3.24) первый сомножитель в правой части (3.27) равен

$$d_1 = e^{-\frac{2b}{\sigma^2}N} \frac{1}{N^2} \int_{r_1}^{N} t^2 e^{\frac{2b}{\sigma^2}t}\,dt.$$

Интегрируя, заключаем, что при всех $N > r_0$ и некоторой постоянной c_2

$$d_1 \ge c_2 > 0. \tag{3.29}$$

Аналогичные соображения показывают, что при всех N и некоторой постоянной $c_3 > 0$

$$\frac{1}{z_1(r_1)} \int_{r_1}^{N} z_1(t)\,dt \leq e^{\frac{2b}{\sigma^2} r_1} \int_{r_1}^{N} e^{-\frac{2b}{\sigma^2} t}\,dt \leq c_3$$

$$\frac{1}{z(r_1)} \int_{r}^{r_1} z(t)\,dt \leq r_1^2 e^{-\frac{2b}{\sigma^2} r_1} \int_{r}^{r_1} \frac{1}{t^2} e^{\frac{2b}{\sigma^2} t}\,dt \leq c_3 . \tag{3.30}$$

Из (3.29), (3.30), (3.27) следует (3.28). Тем самым справедливость соотношения (3.26) установлена. В силу (3.26), (3.25) на любом ограниченном множестве изменения x

$$\lim_{N\to\infty} V_N(x) = \lim_{r_1\to\infty} \frac{2}{\sigma^2} \int_{r}^{|x|} z(t)\,dt \int_{t}^{r_1} \frac{dt_1}{z(t_1)} .$$

Подставляя сюда вместо функции $Z(t)$ её выражение (3.24), получаем

$$\lim_{N\to\infty} V_N(x) = V_0(x) = \frac{1}{b}\left[|x| - r + \frac{\varrho^2}{b} \ln \frac{|x|}{r} + \frac{\sigma^4}{2b^2}\left(\frac{1}{r} - \frac{1}{|x|}\right)\right]. \tag{3.31}$$

Используя представление (3.31) для $V_0(x)$, найдем управление $u_0(x)$, доставляющее минимум по u, $|u| \leq b$ выражению

$$u_1 \frac{\partial V_0}{\partial x_1} + u_2 \frac{\partial V_0}{\partial x_2} + u_3 \frac{\partial V_0}{\partial x_3} .$$

Имеем

$$u_0(x) = -bx\,|x|^{-1} . \tag{3.32}$$

Для доказательства оптимальности управления (3.32) достаточно, ввиду следствия 2 к теореме, проверить, что управление (3.32) допустимо, т. е. проверить, что при этом управлении существует решение системы (3.20). Однако это немедленно следует из (3.32), (3.31) теоремы 3.4.1 ([5]) и того факта, что $L_{u_0} V_0 = -1$.

Заметим, что при $\sigma = 0$ выражение (3.31) установлено ранее в [19] для более общего случая детерминированных систем с инвариантной нормой, отличным от изложенного методом.

В связи с примерами 3—5, в которых найдено оптимальное управление, отметим, что в них оптимальное стохастическое управление совпадает с оптимальным управлением для соответствующей детерминированной системы при $\sigma = 0$.

Выражаю благодарность Ф. Л. Черноусько за внимание к работе и ценное обсуждение, способствовавшее улучшению работы.

Литература

1. *Гихман И. И., Скороход А. В.* Введение в теорию случайных процессов. М., «Наука», 1965.
2. *Понтрягин Л. С., Болтянский В. Г., Гамкрелидзе Р. В., Мищенко Е. Ф.* Математическая теория оптимальных процессов. М., Физматгиз, 1969.
3. *Dorato, P., Chang-Ming Hsieh, Robinson, P.*, Optimal bang-bang control of linear stochastic systems with a small noise parameter, IEEE Trans. Automatic Control, AC–12, 1967.
4. *Красовский Н. Н.* Об оптимальном регулировании при случайных возмущениях. ПММ, вып. 1, т. 24, 1960.
5. *Хасьминский Р. З.* Устойчивость систем дифференциальных уравнений при случайных возмущениях их параметров. М., «Наука», 1969.
6. *Дынкин Е. Б., Юшкевич А. А.* Теоремы и задачи о процессах Маркова. М., «Наука», 1967.
7. *Тихонов А. Н., Самарский А. А.* Уравнения математической физики. М., «Наука», 1966.
8. *Ладыженская О. А., Уральцева Н. Н.* Линейные и квазилинейные уравнения эллиптического типа. М., «Наука», 1973.
9. *Fleming, W. H.*, Some Markovian optimization problems. J. Math. Mech., 12, 1963.
10. *Наймарк М. А.* Нормированные кольца. М., «Наука», 1968.
11. *Миранда К.* Уравнения с частными производными эллиптического типа. М., Изд-во иностр. литературы, 1957.
12. *Дынкин Е. Б.* Марковские процессы. Физматгиз, М., 1963.
13. *Люстерник Л. А., Соболев В. И.* Элементы функционального анализа. М., «Наука», 1965.
14. *Fleming, W. H.*, Duality and a priori estimates in Markovian optimization problems. J. of Math. Anal. and Applic. Vol. 16, No. 2, 1966.
15. *Ито К., Нисио М.* Стационарные решения стохастических дифференциальных уравнений. Математика (сборник переводов), II, 5, 1967.
16. *Гихман И. И., Скороход А. В.* Стохастические дифференциальные уравнения. «Наукова Думка», 1968.
17. *Ли Э. Б., Маркус П.* Основы теории оптимального управления. М., «Наука», 1972.
18. *Фихтенгольц Г. М.* Курс дифференциального и интегрального исчисления. М., Физматгиз, т. 2, 1959.
19. *Атанс М., Фалб П.* Оптимальное управление. М, Машиностроение, 1968.

Some problems of time optimal control for stochastic systems

V. B. KOLMANOVSKY

(Moscow)

The paper is devoted to some problems of time optimal stochastic systems. The motion differential equation (1.1) and the functional (1.3) were given. First of all the problem of control existence such that finite functional (1.3) is investigated in Section 2 and some examples are also given.

Theorems 3.1 and 3.2 are the main results of the paper. These theorems give also algorithm for optimal control and for the optimal value of constructing functional (1.3).

There are two mechanical examples in which optimal control is found with the help of Theorems 3.1 and 3.2.

В. Б. Колмановский
Институт проблем механики АН СССР
СССР, Москва А—40,
Ленинградский проспект, д. 40.

Problems of Control and Information Theory, Vol. 4 (4), pp. 369–371 (1975)

INTERNATIONAL IFAC SYMPOSIUM ON OPTIMIZATION METHODS

VARNA (BULGARIA), OCTOBER 8–11, 1974

The lectures delivered at the Symposium represented the wide-spreading research work and the last restults achieved in practice with optimization methods and their applications in the field. The Symposium comprised the classical topics of the IFAC:

- models of the concrete, continuous-time and discrete-time systems in chemical industry, metallurgy, cellulose and paper industry;
- synthesis of the optimal and quasi-optimal control systems, taking into account the non-complete information of the objects and interfering effects. In this field an effort is distinguished to produce a connection required between the most up-to-date theory and the practical realization of the optimal control.

Important problems of the optimization of the socio-economic "large" systems — developed jointly by the IFORS and IFIP in the last years — were also extensively discussed. The participants were deeply interested in these themes because of the solution of the socio-economic development is very significant for all those countries whose scientists attended the Symposium.

250 researchers from 23 countries participated at the Symposium. The list of the lectures accepted by the International Program Committee and presented actually at the Symposium is shown in the table.

It is to be noted that from the very interesting lectures presented by the Soviet scientists only four studies were read since the Sixth All-Union Conference on Control Problems was held at the same time as the Varna IFAC Symposium To the participants' greatest regret Academician L. S. Pontryagin, the great Soviet authority on optimal control theory because of his illness could not deliver his lecture on "Linear differential games". The main topics of the symposium were as follows:

- linear and nonlinear programming
- dynamic programming

Country	Lectures	
	accepted	delivered
Bulgaria	14	11
England	2	1
Finland	3	3
France	4	4
FRG	2	1
Hungary	2	2
Italy	2	1
Japan	2	1
Roumania	1	—
Spain	1	1
Sweden	2	2
USA	1	1
USSR	13	4
Yugoslavia	3	3
total	53	36

— optimization methods in the case of non-complete information
— numerical methods
— quasi-optimal control methods
— automatic optimization methods
— optimal control of large systems
— optimization of hierarchical control systems
— decomposition methods of optimal control
— analysis of the sensitivity of optimal systems
— optimization of differential games
— optimization of biological and economical systems.

The studies were arranged into three sessions

1. Applied mathematical programming
2. Large systems
3. Synthesis of feedback optimal systems.

Both the lectures and discussions dealt with the problems of the extension of the instrumentation of the methods of mathematical programming and algorithms. We have to emphasize the importance of those lectures which were devoted to the development of the automatic control theory, especially those which treated the variable-structure systems opening new ways in the field of optimization (Soviet Union). Very interesting studies dealt with the formation of the strongly convergent algorithms (England); scientists from the USA reported the results obtained by differential games in dynamic systems. Development could be stated in the field of the development and

new applications of the classical variation methods (France, Sweden); in dynamic programming (Spain) and in linear programming (Bulgaria).

At the Plenary Session Professor R. Lunderstadt (FRG) delivered a lecture entitled "Numerical Methods of Optimization) showing a real picture of the methods, their present state and future applications.

Acknowledgments are due to IFAC and IFORS, first of all to Professor M. Thomas, for the successful organization of the Symposium.

The members of the International Program Committee offered valuable help to the Bulgarian National Organizing Committee in the selection of the presented lectures and elaboration of the final program.

Prof. N. NAPLATANOV
(Sofia)

BOOK REVIEW

R. KRAUSHAAR, L. JAKOB, D. GOTH: **Datenfernverarbeitung** (Data Teleprocessing). Siemens Aktiengesellschaft, Berlin and Munich, 1974, 192 pp., 88 figs. (in German)

In recent years a rapid development can be expected in data teleprocessing. More and more institutions have to employ experts in this field. This book gives a brief survey for these experts.

The total data teleprocessing system is treated from the data stations until data teleprocessing equipments — including the possible methods of data transfers. The book does not touch the applications in process control.

The book deals with the questions at a simple level. It does not require special technical qualification. In this way the readers — who are interested in general engineering and system organization questions — may obtain a good insight to the problems. Though the Siemens equipments are mentioned as examples, the principles of application are valid for other equipments, too.

The book consists of six chapters. The first chapter is a preface. The second chapter is an introduction to data teleprocessing: the organization preconditions and the ways of solutions are reviewed. The third chapter deals with the indirect data teleprocessing: its scope is the formulation of the organization task, the methods of solution, instructions for the planning and the methods of data collection, input and output. The fourth chapter deals with the direct data teleprocessing. Besides surveying the organization task, the ways of solutions and the instructions for planning, this chapter describes the organization of the central data processing system, the data terminals and data stations and the preliminary data teleprocessing. The fifth chapter gives an account on the data networks: on the public telegraph systems, on the private telegraph systems, on the selection of the data transfer channels and on data insurance. The sixth chapter gives a review on the codes and procedures.

In addition the book contains a collection of words and terminologies introduced in the text, as well as a reference list and an index.

L. Lakatos

JULIUS T. TOU, RAFAEL C. GONZALES: **Pattern Recognition Principles.** Addison-Wesley Publishing Company, Advanced Book Program, London, 1974, 377 pp.

This book is an easy-to-read and well-presented survey of the basic methods and ideas of pattern recognition theory. It was written to provide engineers, scientists, and students with a comprehensive, well-organized, and up-to-date account of basic principles.

This volume consists of eight chapters. In the first chapter the authors introduce the basic concepts of pattern recognition and the fundamental problems in pattern recognition system design.

In the second chapter the principal functions, i.e. the decision or discriminant functions are described. At the end of this chapter the orthogonal and orthonormal functions (e.g. the Legendre, the Laguerre polynomials etc.) are also given.

In the third chapter the idea of pattern classification by distance functions and the nearest neighbour classification rule are treated.

In the fourth chapter the authors "begin a study of the statistical approach to pattern recognition". One of the most important classifier is the so-called Bayes classifier.

It minimizes the total expected loss and from the statistical point of view it repreesnts the optimum measure of performance In this chapter we may find several basic approaches to obtain estimate of the probability density function.

In the fifth chapter the reader can begin the study of trainable pattern classifiers. Here gradient method and least-mean-square-error (LMSE) algorithms are presented.

In the next chapter there are some methods for estimation of the densities $p(w_i/\mathbf{x})$. First of all some stochastic approximation methods are presented.

In the seventh chapter the authors consider various feature extraction and selection methods. These are the following: feature selection through entropy minimization, f.s. through orthogonal expansion (e.g. Karhunen–Loéve expansion), f.s. through functional approximation etc.

To bridge the gap between the mathematical approach discussed thus far and the linguistic pattern recognition in the last chapter the authors give a survey into the syntactic pattern recognition methods.

One of the main advantages of this book is to present several illustrative examples. Let us emphasize once again that this book is very useful for engineers, scientists, and students. It is very easy to read and understand it because of the excellent style, the correct proof and the illustrative examples.

M. Vajta, Jr.

E. D. GILLES: **Systeme mit verteilten Parametern** (Distributed Parameter Systems). R. Oldenbourg Verlag, Munich—Wien, 1973, 185 pp. (in German)

Not long ago there were only a few papers dealing with Distributed Parameter Systems (dps), but recently this situation has profoundly changed. Nowadays the dps have become an organic part of control technics. This change was brought about by the need, even the constraint of being able of construct such new process models which describe the ramifying chemical technologies and the processes in the atomic power stations with an adequate accuracy. In these processes the parameters are dependent not only on time but on spatial position as well, consequently the process behaviour can be characterized by Partial Differential Equations (pde).

It is a well-known fact that applying the pde demands a more advanced knowledge of mathematics than is taught at most universities, or can be expected of the majority of industrial even research engineers. So their studying this field meets severe difficulties.

At the same time the rapid spreading of process computers has already established the precondition of practical employment, so there is urgent need of a book that treats the dps starting from its mathematical basis, that sums up the different results otherwise difficult to survey.

This book aims to fulfil this task. It consists of 13 chapters.

Chapter 1 presents the characteristic types of the dps, the behaviour of which can be described mathematically with the help of the pde. We shall find here the phenomena of heat convection, heat conduction, wave propagation and the properties of vibration of elastomechanical systems, which result the pde of 1st, 2nd and 4th order.

Chapters 2–6 contain the mathematical bases needed later on when discussing the dps. The attitude of the book is concise and clear, corresponding to engineer's mentality. The author focuses the attention primarily on the relation between the infinite dimensional vector space and the function space. Further on he places the inversion of differential operators and by this Green's function in the centre.

The chapters and their main topics are as follows:

Chapter 2. Vectors and Functions (Infinite Dimensional Spaces, Convergence of Vectors, Orthogonalization of Vectors, Linear Transformations, Function Space, Completeness of Orthogonal Systems, Relation between Function Space and Hilbert Space)

Chapter 3. Functional Operators (Integral and Differential Operators and their Adjuncts, their Existence and Uniqueness)

Chapter 4. Inversion of Differential-operators (Green's function, Green's matrix and their Existence)

MTA Könyvtára
Periodika 3802/19 /75 sz.

Chapters 5 and 6 present controlled plants of 2nd and 4th order and enumerates the most important ways how they can be described.

Chapter 5 investigates a three-dimensional controlled plant of 2nd order giving its pde and determining Green's function and the block diagram. It determines the eigenfunction of the controlled plant and Green's function in spectral form, then on the basis of them gives the self-adjoint of the controlled plant in spectral form.

These equations are then transformed into an algebraic form in order to present the state-space method as well. Finally it will be shown how Green's function has to be modified if there is a flow too in the system.

Chapter 6 investigates an elastic distributed parameter system of 4th order, which shows a close connection with ships, aeroplanes, rockets, antennas and mirror-telescopes.

Chapter 7 gives two other possibilities for describing the dps. Applying the theory of variations the system is given by a set of the pde's in canonical form and we can get acquainted also with the state-space method.

Chapter 8 extends Lyapunov stability criterion to the dps. Generalizing state-space representation, Lyapunov functional is initiated too, and with the help of this method stability investigation is carried out on three control circuits which can practically be of high importance.

In Chapter 9 the observability and controllability of the dps is briefly treated.

Chapter 10 examines the cooperation of a distributed parameter controller and a controlled plant of the same type.

In Chapter 11 only one discrete value of the controlled variable will be controlled at a certain place. Then the stability of the control circuit is investigated by means of root-locus method, Nyquist stability criterion and D-separation method. Finally a procedure is shown how the structure of controller can be determined.

Chapters 12 and 13 present the available possibilities (modal control and point-by-point control) for modelling control circuits in case of space-dependent controlled variables

D. Takács

Ya. Z. TSYPKIN: **Automatic Relay Control Systems.** Izd. Nauka, Moscow, 1974, 575 pp. (in Russian)

This book is based on the author's former book on the theory of relay control systems published in 1955.

The new possibilities to realize contactless "relay" effects offered by the development of electronics give new actuality to the theory and practice of automatic relay control systems.

Relay systems are applied in several technical areas, firstly because of their simplicity. In many cases they yield also better properties then other systems. In their contactless realization the convenience in changing the parameters offers new possibilities in the system construction.

The relay systems are essentially nonlinear, nevertheless some concepts and methods of linear system theory (such as the time and frequency characteristics) may be retained in their study.

The book gives the theory of relay control systems based on these concepts and follows up the operation of the relay systems in the simplest way illustrative also from physical aspects.

The volume is devoted to the general phemomena and properties of the relay systems and their study (e.g. limit cycle, sliding, oscillation, etc.). The theory has developed since the preceding variation of the book. This new book deals also with the new results, discusses the problems of stability, the analysis of subharmonic oscillations, the theory of pulse-relay system, the effects of negative hysteresis, the problems of multivariable relay systems, relay extremum systems, etc.

The book consists of 14 chapters.

In the introduction the author describes the general construction of control systems and points out the place of relay control systems.

Chapter 1 deals with the properties of the relay systems and their classification with several examples for illustration.

Chapter 2 discusses the equations of the relay systems. The author investigates here first separately the linear part of the system with its time and frequency characteristics and the equations of the relay part, and then considers the system as a whole with examples for better understanding.

Chapter 3 describes the properties of the relay systems, their various modes of operation, the construction method of their transition functions.

Chapter 4–10 are devoted to the problems of stability and various types of oscillations. The effects of inner loops are also studied. Several stability criteria are given.

Chapter 11 deals with the determination of the forms of periodical operation modes.

Chapter 12 discusses the method of lineraization and extends also to the sliding effect.

Chapter 13 describes the problems of pulsed-data relay systems.

Chapter 14 treats the extremum searching relay systems.

The Appendix summarizes the mathematical foundations in short.

The book is fundamental in this topic. It is characterized by clarity. Besides the mathematical treatment the physical aspects are always taken into consideration. The theory is illustrated by numerous examples.

The book may be of great help of engineers designing or working with relay-type control systems.

R. Bors

L. B. Berkovitz: **Optimal Control Theory.** In Applied Mathematical Sciences; Vol. 12, Springer-Verlag, New York – Heidelberg – Berlin, 1974, 304 pp.

The book is an introduction to the mathematical theory of optimal control processes governed by ordinary differential equations.

The volume consists of six chapters.

Chapter 1 presents some short examples of control problems drawn from different areas of application.

Chapter 2 deals with the formulation of the control problems and discusses the relationship between the methods in the calculus of variations and specific control problems.

Chapter 3 contains the basic existence theorems for optimal problems in which a certain convexity condition is present. Here the control variables are measurable functions and there are no restrictions on the rate at which the control variables can be varied.

Chapter 4 investigates such systems in which certain sets need not to be convex. Section 2 of this chapter is devoted to the inertial controllers, whereas Section 3 to the so-called relaxed problem.

Chapter 5 deals with the presentation of maximum principle and with some of its applications. The dynamic programming argument is used to derive the maximum principle. The last section of this chapter presents the so-called LQ problem, that is, the linear plant with quadratic cost criterion problem. It is shown that the necessary conditions are sufficient and the standard synthesis method of the optimal control is also derived.

Chapter 6 is devoted the to detailed proof of the maximum principle. Actually a theory is proved that is more general than the maximum principle and the maximum principle is obtained as a special case of this theorem.

The book is of mathematical character, practical problems serve only as illustrations for the theoretical investigations. Consequently it is recommended for students and experts in mathematics and in areas of application who want a concise and coherent introduction to the subject and to its relationship with applications.

Tivadar Kovács

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

PROBLEMS OF CONTROL AND INFORMATION THEORY, VOL. 4, NO. 4
ENGLISH TRANSLATIONS OF THE COMMUNICATIONS IN RUSSIAN

ON THE COST SEQUENCE TRANSMISSION

G. L. KHODAK
(Omsk)

Message transmission through the noiseless channel is considered. Message source is an arbitrary discrete stochastic process, and coding is any measurable mapping of the set of input sequences into output ones. Transmission cost is defined as the average number of output letters required to decode one letter of a message. This definition differs from the traditional one. For finite-automaton coding, the new cost is proved to be equal to the traditional one if coding is decodable, and to be infinite otherwise. The fundamental theorem for noiseless channels appears to be generalized to the widest class of sources and codings.

1. Introduction

The fundamental theorem for noiseless channels as it was formulated by C. Shannon [1], states that transmission cost (inverse variable is speed) cannot exceed the source entropy divided by the logarithm of the number of letters in the output alphabet, but it can be made arbitrarily close to this value. The negative part of this statement is proved in [1] for coding of any source by an automaton substituting letter for letter; in [2] for block prefix coding of Bernoulli source; in [3] for arbitrary block coding of Bernoulli source (these results are generalized to the stationary source in [4]); in [5] for finite-automaton coding of Markov source; in [6] for stationary coding of stationary source; and in [7] for a rather wide class of sources and mappings of input words into output ones. All these papers regard coding as decodable. The positive part of the theorem is proved in [1] for the stationary source. Codings with cost equal to the lower boundary were found in [8] for the stationary source, and in [7] for a wider class of sources.

In the papers mentioned above, tranmission cost is understood as an average number of output alphabet letters given by the coder in response to one input letter. This is coder's operation index. A situation is possible where cost is finite (speed is nonzero), but message cannot be restored with any code. Since coding is not completely characterized by transmission cost, it should have certain decodability, i.e. possibility to restore message by its code. Decodability does not ensure, however, that decoding speed is the same as that of coding.

The present paper regards as the transmission cost the average number of output alphabet letters required for decoding of one input alphabet letter. It is clear from the definition that the cost reflects operation of the whole of the communication system rather than of the coder only. Since the cost is defined through decoding, no decodability condition is imposed on coding.

Below, any discrete stochastic process whose sample functions (messages) are sequences of finite or denumerable input alphabet letters, is understood as message source; and measurable mapping of input sequences into output ones is understood as coding. The most close to the above definition is formulated in [6, 8] where either stationary, or monotone mapping of sequences is regarded as coding. In the general case, even block coding is known not to be coding in this sense. In [1–5, 8] coding is one or another mapping of a set of words in the input alphabet into a set of words in the output one. If decodable, word mapping induces the only measurable mapping of sequences [7], and, therefore, it may be regarded as a special case of coding defined in the present paper. At the same time, not every sequence mapping may be induced by word mapping. For noiseless channels, only finite output alphabet is of interest, since, otherwise, capacity is infinite. It follows from output alphabet finiteness that the input alphabet cannot be more than denumerable. Word probability at coder output goes hand in hand with measurability of sequence mapping (coding). Therefore, the class of sources and codings under consideration is the widest one.

The conventional, traditional cost and that introduced in this paper are proved (Section 6) to coincide, in particular, under decodable finite-automaton Markov source coding (under slight restrictions, Theorem 1), letter-wise decodable coding of Bernoulli sources, and "decodable" [7] coding.

Decodability in conventional terms is shown to follow from the cost finiteness (Section 7). With finite cost, coding is nearly everwhere one-to-one (Theorem 2), i.e. decodable in the sense of [6]. Under finite-automaton coding of Markov source (under slight restrictions), cost is finite if and only if coding is decodable (Theorem 3).

The fundamental theorem is generalized (Section 8) to the class of sources and codings described above. To deduce results obtained by other authors from the theorem proved below, one has to establish equality of costs to leading to them results of [1–4,7] for "decodable" coding, and of [5] under slight restrictions. The positive part of the theorem is proved in a similar way [7].

2. Notations

Let $\mathfrak{A}$ be some alphabet; $\mathfrak{A}^*$ be a set of all words in alphabet $\mathfrak{A}$; Λ be an empty word; $|A|$ be the number of letters in A; and $\mathfrak{A}^\infty$ be a set of all sequences of form $\{a_i\}_{i=1}^{\infty}$ ($a_i \in \mathfrak{A}$: $i = 1, 2, \ldots$). If $A_1 \in \mathfrak{A}^*$ and $A_2 \in \mathfrak{A}^* \cup \mathfrak{A}^\infty$, then

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

$A_1 \cdot A_2$ results from adding A_2 to the right of A_1. If $A \in \mathfrak{A}^*$, $V \subset \mathfrak{A}^*$, and n is a natural number, then $A^n = A \cdot A \ldots A$ (n times); $A^0 = \Lambda$, $V^n = \{A_1 \ldots A_n\colon A_1, \ldots, A_n \in V\}$; $V^0 = \{\Lambda\}$; $A^\infty = A \cdot A \ldots$; $V^\infty = \{A_1 \cdot A_2 \ldots\colon A_1, A_2, \ldots \in V\}$. Notation $A_1 \prec A_2$ where $A_1 \in \mathfrak{A}^*$ and $A_2 \in \mathfrak{A}^* \cup \mathfrak{A}^\infty$ means that A_1 is a prefix (beginning) of A_2. Set $V \subset \mathfrak{A}^*$ is referred to as prefix set if for any two different words from V one cannot be the beginning of another. If $A \in \mathfrak{A}^*$, $V \subset \mathfrak{A}^* \cup \mathfrak{A}^\infty$, then $A \prec V$ means that word A is prefix of all elements of set V. $\| V \|$ is the cardinality of set V; $\varphi^{-1}(V)$ is a complete prototype of V under mapping φ.

Below, $\mathfrak{A}$, is always a finite or denumerable input alphabet, $\| \mathfrak{A} \| \geq 2$; $\mathfrak{B}$ is a finite output alphabet of cardinality m, $2 \leq m < \infty$; $\sigma(\mathfrak{A})$ is some σ-algebra of subsets of $\mathfrak{A}^\infty$ involving all finite-dimensional cylinders, i.e. $A\mathfrak{A}^\infty \in \sigma(\mathfrak{A})$ for any word $A \in \mathfrak{A}^*$; $\sigma_0(\mathfrak{B})$ is σ-algebra of subsets of $\mathfrak{B}^\infty$ generated by all finite-dimensional cylinders; p is a probabilistic measure over $\sigma(\mathfrak{A})$.

3. Source of messages and its entropy

By definition, probabilistic space $\{\mathfrak{A}^\infty, \sigma(\mathfrak{A}), p\}$ will be referred to as a source. A sequence in alphabet $\mathfrak{A}$, i.e. an elementary event of probabilistic space $\{\mathfrak{A}^\infty, \sigma(\mathfrak{A}), p\}$, ois called message. Denote measure of a set of form $A \cdot \mathfrak{A}^\infty$ (where $A \in \mathfrak{A}^*$) by $p(A)$ and call it probability of word A. A source is called positive if probabilities of all words are positive.

Define, in the conventional fashion, entropy of prefixes of length n ($n = 1, 2, \ldots$.) as $H(n) = -\sum\limits_{A \in \mathfrak{A}^n} p(A) \ln p(A)$*. Function $H(\ )$ is monotone non-decreasing. Always existing $\overline{H} = \overline{\lim\limits_{n \to \infty}} \dfrac{1}{n} H(n)$ and $\underline{H} = \underline{\lim\limits_{n \to \infty}} \dfrac{1}{n} H(n)$ will be, respectively, referred to as upper and lower entropies of the source. If $\overline{H} = \underline{H}$ the source is said [5, 9] to have entropy $H = \overline{H} = \underline{H}$

4. Coding

Mapping $\varphi\colon \mathfrak{A}^\infty \to \mathfrak{B}^\infty$ measurable with respect to $\sigma(\mathfrak{A})$ anf $\sigma_0(\mathfrak{B})$, i.e. $\forall V \in \sigma_0(\mathfrak{B})(\varphi^{-1}(V) \in \sigma(\mathfrak{A}))$ will be referred to as coding. Codings coinciding almost everywhere are termed as equivalent. Let $\sigma(\mathfrak{B}) = \{V \subset \mathfrak{B}^\infty\colon \varphi^{-1}(V) \in \sigma(\mathfrak{A})\}$. $\sigma(\mathfrak{B})$ is known [9] to be a σ-algebra. $\sigma_0(\mathfrak{B}) \subset \sigma(\mathfrak{B})$ by the definition of coding. Coding transfers measure p from $\sigma(\mathfrak{A})$ to $\sigma(\mathfrak{B})$ [9]. By definition, measure of set $V \in \sigma(\mathfrak{B})$ is equal to that of its complete prototype, i.e. $p(\varphi^{-1}(V))$.

* Product $p \cdot \ln p$ is considered as equal to zero under $p = 0$. The sum is always definite because the number of summands is not greater than denumerable, and all of them have the same sign. $H(n)$ may be equal to ∞.

5. Transmission cost

It is natural to believe that prefix A of some message A' is decoded with respect to prefix B of sequence $\varphi(A')$ if almost everywhere messages having codes beginning with B begin with A, i.e. almost everywhere $A \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty)$. This definition of decodability corresponds to a case where message code letters enter decoder evenly, without gaps between the words. Therefore, for codings of type $\psi : \mathfrak{A}^* \to \mathfrak{B}^*$, it is not known, when word B is received, whether it is a code of some word A $(B = \psi(A))$ or a prefix of another word A′ for which $A \not\prec A'$ and $B < \psi(A')$.

For $n = 1, 2, \ldots$ and any sequence $B' \in \varphi(\mathfrak{A}^\infty)$, assume that $t(n, B')$ is equal to the minimal length of sequence B' prefix with respect to which some word of length n may be decoded; if there is no such a prefix, then $t(n, B') = +\infty$. Assume formally that

$$t(n, B') = \begin{cases} \min\limits_{B \in V(n, B')} |B| & \text{if } V(n, B') = \varnothing; \\ +\infty & \text{if } V(n, B') = \varnothing, \end{cases}$$

where $V(n, B') = \{B \in \mathfrak{B}^*: B < B' \,\&\, \exists A \in \mathfrak{A}^*(|A| = n \,\&\, A \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty)$ almost everywhere)$\}$.

It may be readily verified that for any natural n function $t(n, \ )$ is defined almost everywhere over $\mathfrak{B}^\infty$ and measurable with respect to $\sigma(\mathfrak{B})$. If codings φ_1 and φ_2 are equivalent, corresponding to them functions t_{φ_1}, and t_{φ_2} are equivalent.

Denote by $C(n)$ the average of output alphabet letters required to decode first n letters of a message: $C(n) = \int\limits_{\mathfrak{B}^\infty} t(n,)\, dp = \int\limits_{\mathfrak{A}^\infty} t(n, \varphi(\ ))\, dp$ $(n = 1, 2, \ldots)$*

Function $C(\ )$ is monotone non-decreasing. $\overline{C} = \overline{\lim\limits_{n \to \infty}} \dfrac{1}{n} C(n)$ and $\underline{C} = \underline{\lim\limits_{n \to \infty}} \dfrac{1}{2} C(n)$ will be referred to as upper and lower costs of transmission, respectively. These limits always exist. If codings φ_1 and φ_2 are equivalent, then corresponding to them $\overline{C}_{\varphi_1}$, $\overline{C}_{\varphi_2}$ and $\underline{C}_{\varphi_1}$, $\underline{C}_{\varphi_2}$ are pairwise equal. If $\overline{C} = \underline{C}$, $C = \overline{C} = \underline{C}$ will be referred to transmission cost.

In discussing codings of type $\psi : \mathfrak{A}^* \to \mathfrak{B}^*$, usually $L = \lim\limits_{n \to \infty} \dfrac{1}{n} \sum\limits_{A \in \mathfrak{A}_n} p(A) \,|\, \psi(A) \,|$ is referred to as transmission cost (limit is assumed to exist) [5, 7]. In simpler cases, the explicit expression of the limit is taken as definition [1–4].

* The integral is definite even if $t(n, \ ) = \infty$ over a set of positive measure because $t \geq 0$. In htis case $C(n) = +\infty$.

Illustrate cost definition. Let $\mathfrak{A} = \{\alpha, \beta, \gamma\}$, $\mathfrak{B} = \{0, 1\}$, $\sigma(\mathfrak{A}) = \sigma_0(\mathfrak{A})$, source be Bernoullian $p(\alpha) = p(\beta) = p(\gamma) = \frac{1}{3}$, and coding be letter-wise $\varphi(\alpha) = 0$, $\varphi(\beta) = 01$, $\varphi(\gamma) = 11$, $\varphi(a_1, a_2 \ldots) = \varphi(a_1) \cdot \varphi(a_2) \ldots$. Let $B' = 0^{k_1} \cdot 1^{k_1} \cdot 0^{k_3}$ be a sequence from $\varphi(\mathfrak{A}^\infty)$. If $k_1 = 0$, $k_2 \neq 0$, then $V(1,B') = \{1, 1^2, \ldots, 1^{k_2}, 1^{k_2}0, \ldots\}$ and $t(1, B') = 1$ because $\gamma \prec \varphi^{-1}(1, \mathfrak{B}^\infty)$, and $\bar{1}$ is a minimal prefix of B' with respect to which some letter is decoded. If $k_1 \geq 2$, $V(1, B') = \{00{,}000, \ldots, 0^{k_1}, 0^{k_1}1, \ldots\}$, $t(1, B') = 2$ because $\alpha \prec \varphi^{-1}(00, \mathfrak{B}^\infty)$ and $\varphi^{-1}(0, \mathfrak{B}^\infty) = \alpha \cdot \mathfrak{A}^\infty \cup \beta \cdot \mathfrak{A}^\infty$ (not a single letter can be decoded with respect to prefix 0). If $k_1 = 1$, and k_2 is even, $V(1, B') = \{0 \cdot 1^{k_2}0, 01^{k_2}00, \ldots\}$ and $t(1, B') = k_2 + 2$ because $\alpha \prec \varphi^{-1}(0 \cdot 1^{k_2} \cdot 0 \cdot \mathfrak{B}^\infty)$ and $\varphi^{-1}(0 \cdot 1^{k_2} \cdot \mathfrak{B}^\infty) = \alpha \cdot \gamma^{k_2/2} \cdot \mathfrak{A}^\infty \cup \beta\gamma^{k_2/2} \cdot \mathfrak{A}^\infty$. If $k_1 = 1$ and k_2 is odd, similarly, $V(1, B') = \{0 \cdot 1^{k_2} \cdot 0, \ldots\}$ and $t(1, B') = k_2 + 2$. If $B = 0 \cdot 1^\infty$, $V(1, B') = \emptyset$ and $t(1, B') = \infty$. For $k = 1, 2, \ldots$ $p(0 \cdot 1^{2k} \cdot 0 \cdot \mathfrak{B}^\infty) = p(\alpha \cdot \gamma^k \cdot \alpha \cdot \mathfrak{A}^\infty \cup \alpha \cdot \gamma^k \cdot \beta \cdot \mathfrak{A}^\infty) = \frac{2}{3^{k+2}} p(0 \cdot 1^{2k-1} \cdot 0 \cdot \mathfrak{B}^\infty) = p(\beta \cdot \gamma^{k-1} \cdot \alpha \cdot \mathfrak{A}^\infty \cup \beta \cdot \gamma^{k-1} \cdot \beta \cdot \mathfrak{A}^\infty) = \frac{2}{3^{k+1}}$, we have $p(1 \cdot \mathfrak{B}^\infty) = p(\varphi^{-1}(1 \cdot \mathfrak{B}^\infty)) = p(\gamma \cdot \mathfrak{A}^\infty) = \frac{1}{3} p(00, \mathfrak{B}^\infty) = p(\alpha^2 \cdot \mathfrak{A}^\infty \cup \alpha \cdot \beta \cdot \mathfrak{A}^\infty) = \frac{2}{9}$. $C(1) = 1 \cdot \frac{1}{3} + 2 \cdot \frac{2}{9} + \sum_{k=1}^{\infty} (2k+2) \frac{2}{3^{k+2}} + \sum_{k=1}^{\infty} (2k+1) \frac{2}{3^{k+1}} + \infty \cdot 0 = \frac{16}{9}$.

To decode the first letter of a message, on the average $\frac{16}{9}$ of output letters will be required. The input unit will give in response to the first input $L = \frac{1}{3} + \frac{2}{3} + \frac{2}{3} = \frac{5}{3}$ which is less that the value above. This discrepancy is accounted for by the fact that, knowing only the code of letter α, one cannot tell which letter message begins with. It is easy to verify that for any natural n, word $A \in \mathfrak{A}^n$ and sequence $B' = \varphi(A)B'' \in \varphi(A \cdot \mathfrak{A}^\infty)$, $t(n+1, B') = |\varphi(A)| + t(1, B'')$. Therefore $C(n+1) = nL + C(1)$.

6. Relation between various definitions of cost

Lemma 1. If the source is positive and cost L is defined, $C \geq L$. The lemma statement follows from the fact that no word $a_1 \ldots a_n \in \mathfrak{A}^*$ can be decoded with respect to $\psi(a_1 \ldots a_{n-1})$.

It is just proved that if coding is "decodable" in the sense of [7], then $\bar{C} \leq L$. If, besides, the source is positive, $C = L$ in virtue of Lemma 1.

In Ref. [5] it has shown that there is cost L for decodable finite-sautomaton coding of Markov sources. Definitions related to the finite-automaton coding which are standard and listed in a publication by the author [10], are not repeated here.

Theorem 1. If the source is Markovian, of arbitrary finite order of connectivity and positive; automaton coding is decodable; and automaton is strongly connected, cost C exists and is equal to cost L.

Proof. Let all conditions of the theorem be satisfied, $\varphi : \mathfrak{A}^* \cup \mathfrak{A}^\infty \to \to \mathfrak{B}^* \cup \mathfrak{B}^\infty$ be automaton coding, and U be a universal decoding word whose existence is proved in Lemma 5 [10]. Take

$$P = \min_{A \in \mathfrak{A}^*} \frac{p(A \cdot U)}{p(A)} = \min_{A \in \mathfrak{A} \cup \{\Lambda\}} \frac{p(A \cdot U)}{p(A)} > 0.^* \tag{6.1}$$

Fix arbitrarily natural number n. Represent each sequence $a_1 a_2 \ldots \in \mathfrak{A}^\infty$ as $a_1 a_2 \ldots a_n A_1 \cdot A_2 \ldots$ where $|A_1| = |A_2| = \ldots = |U|$. If for some i (such i will be found with probability 1) $A_i = U$, then $a_1 \ldots a_n$ (and even $a_1 \ldots a_n A_1 \ldots A_{i-1}$) is decoded with respect to $\varphi(a_1 \ldots a_n A_1 \ldots A_i)$ and therefore

$$t(n, \alpha(A)) \leq |\varphi(a_1 \ldots a_n \cdot A_1 \ldots A_i)| \leq |\varphi(a_1 \ldots a_n)| + M \cdot i \cdot |U| \tag{6.2}$$

where M is the length of a maximal word assigned to one letter.

Integrate termwise inequality (6.2)

$$\begin{aligned} C(n) &\leq \sum_{A \in \mathfrak{A}^n} p(A)\,|\varphi(A)| + M \cdot |U| \cdot \sum_{i=0}^{\infty} (i+1)(1-p)^i = \\ &= \sum_{A \in \mathfrak{A}^n} p(A)\,|\varphi(A)| + M \cdot |U| \cdot p^{-2}. \end{aligned} \tag{6.3}$$

Divide both sides of inequality (6.3) by n and pass to limit under $n \to \infty$, obtain

$$\bar{C} \leq L. \tag{6.4}$$

The statement of the theorem follows from (6.4) and Lemma 1.

Corollary 1. If the source is Bernoullian (not necessarily positive), the input alphabet is finite, and coding is letter-wise and decodable in the sense of [3], then cost C exists and is equal to cost L.

To prove this, it is sufficient to consider coding over a sub-alphabet of letters with non-zero probabilitites.

* In discussing any finite-automaton coding, alphabet $\mathfrak{A}$ is assumed to be finite. Assume that $U \neq \Lambda$, otherwise any non-empty word can be taken as U.

7. Cost finiteness and decodability

Theorem 2. If the lower cost is finite, then contraction of coding to some set of full measure is one-to-one.

Proof outline. It follows from the definition of cost and from the theorem that one can select set $V \subset \mathfrak{A}^\infty$ so that for all $A = a_1 a_2 \ldots \in V$, all $n = 1, 2, \ldots$ and all $B \in V(n, \varphi(A))$ the following relations hold: $t(n, \varphi(A)) < \infty$; $a_1 \ldots, \ldots, a_n \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty) \cap V$ (everywhere); $p(V) = 1$. Mapping $\varphi : V \to \varphi(V)$ will be one-to-one.

Coding φ: $\mathfrak{A}^\infty \to \mathfrak{B}^\infty$ was called by Billingsley [6] as decodable if it is almost everywhere one-to-one.

Corollary 2. If the lower cost is finite, there is an equivalent coding which is almost everywhere one-to-one (decodable in the sense of [6]).

Proof. Mapping $\varphi : V \to \varphi(V)$ is one-to-one in virtue of the above theorem. If $V = \mathfrak{A}^\infty$, the corollary holds. Let $V \neq \mathfrak{A}^\infty$. Change coding to $\mathfrak{A}^\infty \setminus V$. There will always be sequence $B \in \mathfrak{B}^\infty$ of zero probability. Assume that $\varphi(A) = B$ for all $A \in \mathfrak{A}^\infty \setminus V$. Than any sequence from $V \setminus \{\varphi^{-1}(B)\}$ is determined one-to-one by its code, and $p(V \setminus \{\varphi^{-1}(B)\}) = 1$. Corollary is proved.

Theorem 3. If a source is Markovian, of arbitrary finite order of connectivity and positive, coding is done by a strongly connected automaton, then it is decodable, if and only if cost C is finite (definitions are given in [10]).

Proof. Let all conditions of the theorem be satisfied. In virtue of Theorem 1, $C = L < \infty$ for decodable coding. Show that $C = \infty$ for non-decodable coding. Let $t(A)$ be final state of an automaton in response to input word $A \in \mathfrak{A}^*$. If coding is not decodable, there will be words $A_1 A_2 \in \mathfrak{A}^*$ for which

$$A_1 \neq A_2, \ \varphi(A_1) = \varphi(A_2), \ t(A_1) = t(A_2), \tag{7.1}$$

It follows from (18) that for any sequence $A \in \mathfrak{A}^\infty$

$$\varphi(A_1 \cdot A) = \varphi(A_2 \cdot A). \tag{7.2}$$

It follows from (19) and positiveness of probabilitites $p(A_1)$ and $p(A_2)$ that there is no one-to-one contraction of coding to a set of full measure. Therefore, in virtue of Theorem 2, $C = \underline{C} = \infty$. Q. E. D.

Corollary 3. If a source is Bernoullian, the input alphabet is finite, and coding is letter-wise, it is decodable in the sense of [3] if and only if cost C is finite.

To prove this, it is sufficient to consider coding contraction to a subalphabet consisting of letters with nonzero probabilities.

8. The fundamental theorem for noiseless channels

Theorem 4A. For any source and coding, $\overline{C} \geq \overline{H} \cdot \ln^{-1} m$ and $\underline{C} \geq \underline{H} \cdot \ln^{-1} m$.

Proof. Fix arbitrarily source and coding. Assume that

$$\overline{C} < \infty. \tag{8.1}$$

Otherwise, statement of the theorem is trival. Take

$$V_0(n) = \bigcup_{B \in \varphi(\mathfrak{A}^\infty)} V(n, B) \tag{8.2}$$

$$V(n) = \{B \in V_0(n) \colon \forall B_1 \in V_0(n)(B = B_1 V B_1 \not\prec B)\} \ (n = 1, 2, \ldots) \tag{8.3}$$

where $V(n, B)$ is the set defined in Section 5. Set $V(n)$ involves all prefixes of message codes with respect to which one can first decode any message of length n, this set is a prefix one. If sequence $B' \in \varphi(\mathfrak{A}^\infty)$ belongs to $B \cdot \mathfrak{B}^\infty$ under some $B \in V(n)$, then the minimal length of sequence B' prefix with respect to which one can decode some word of length n, is

$$t(n, B') = |B|. \tag{8.4}$$

If B' does not belong to $\bigcup_{B \in V(n)} B \cdot \mathfrak{B}^\infty$, then

$$t(n, B') = \infty. \tag{8.5}$$

It follows from (8.4) and (8.5) that

$$C(n) = \sum_{B \in V(n)} p(B)\,|B| + \infty \cdot p\Big(\mathfrak{B}^\infty \setminus \bigcup_{B \in V(n)} B \cdot \mathfrak{B}^\infty\Big) * \ (n = 1, 2, \ldots). \tag{8.6}$$

Fix arbitrarily a natural number n. For $A \in \mathfrak{A}^n$, assume that

$$V(A) = \{B \in V(n) \colon A \prec \varphi^{-1}(B \cdot \mathfrak{B}^\infty) \text{ almost everywhere}\}. \tag{8.7}$$

From (8.2), (8.3) and (8.7), obtain

$$\bigcup_{A \in \mathfrak{A}^n} V(A) = V(n). \tag{8.8}$$

It follows from (8.7) that $A \cdot \mathfrak{A}^\infty \supset \bigcup_{B \in V(A)} \varphi^{-1}(B \cdot \mathfrak{B}^\infty)$, and therefore

$$p(A) \geq \sum_{B \in V(A)} p(B) \qquad (A \in \mathfrak{A}^n). \tag{8.9}$$

* By definition,

$$\infty \cdot x = \begin{cases} \infty & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases}$$

We have from (8.1), (8.6): $p(B^\infty \setminus \bigcup_{B \in V(n)} \mathfrak{B} \cdot \mathfrak{B}^\infty) = 0$, therefore

$$\sum_{B \in V(n)} p(B) = 1 \tag{8.10}$$

It follows ıfom (8.6) that, if $p(B) > 0$ $(B \in V(n))$, then $B \in V(A)$ only for one $A \in \mathfrak{A}^n$. Therefore, from (8.9), (8.10) and the fact that $\sum_{A \in \mathfrak{A}^n} p(A) = 1$, it follows that

$$p(A) = \sum_{B \in V(A)} p(B) \qquad (A \in \mathfrak{A}^n). \tag{8.11}$$

Entropy [4] leads to

$$-\sum_{B \in V(n)} p(B) \ln p(B) \geq -\sum_{A \in \mathfrak{A}^n} p(V(A) \cdot \mathfrak{B}^\infty) \ln p\left(V(A) \cdot \mathfrak{B}^\infty\right) =$$

$$= -\sum_{A \in \mathfrak{A}^n} p(A) \ln p(A) = H(n) \qquad (n = 1, 2, \ldots). \tag{8.12}$$

It follows from (8.12) and prefixness of set $V(n)$ that

$$C(n) = \sum_{B \in V(n)} p(B) \mid B \mid \geq \ln^{-1} m \cdot \sum_{B \in V(n)} p(B) \ln p(B) \geq \ln^{-1} m \cdot H(n). \tag{8.13}$$

Divide (8.13) by n and pass to limit under $n \to \infty$. The statement of the theorem results.

Illustrate the proof with the same example as in Section 5. It follows from what was said there that $V_0(1) = \{1, 00, 010, 0110, 01110, \ldots\}$. For $n = 1$, $V(\alpha) = \{00, 0110, 011110, \ldots\}$, $V(\beta) = \{010, 01110, \ldots\}$, $V(\gamma) = \{1\}$.

We have $p(\alpha) = p(00) + p(0110) + \ldots = \sum_{k=0}^{\infty} \frac{2}{3^{k+2}} = \frac{1}{3}$

$$p(\beta) = p(010) + \ldots = \sum_{k=1}^{\infty} \frac{2}{3^{k+1}} = \frac{1}{3}$$

$$p(\gamma) = p(1) = \frac{1}{3}.$$

At $n > 1$ $V_0(n) = \varphi(\mathfrak{A}^{n-1})\ V_0(1)$ and any word $A = A_1 \cdot a \in \mathfrak{A}^n$ $V(A) = \varphi(A_1) \cdot (Va)$.

Theorem 4B. For any source, there is coding for which $\overline{C} = \overline{H} \cdot \ln^{-1} m$, $\underline{C} = \underline{H} \cdot \ln^{-1} m$.

Give outline of the proof. Each message $A \in \mathfrak{A}^\infty$ is represented in the form of $A_1(A) \cdot A_2(A)$ where $|A_1(A)| = 1, \ldots |A_n(A)| = n^2$. For all messages with common beginning $A_1, \ldots A_{n-1}$ such that $p(A_1 \ldots A_{n-1}) > 0$, words

$A_n(A)$ are coded by a prefix code similar to that of C. Shannon [1] and constructed with respect to conditional probability distribution

$$p_{A_1 \ldots A_{n-1}}(A) = \frac{p(A_1 \ldots A_{n-1} \cdot A)}{p(A_1 \ldots A_{n-1})}.$$

For this coding, $C\left(\sum_{i=1}^{n} i^2\right) \leq \ln^{-1} m \; H\left(\sum_{i=1}^{n} i^2\right) + n$ under any $n = 1, 2, \ldots$.

By passing to limit with allowance for monotony of $C(\quad)$ and $H(\quad)$, obtain the statement of the theorem.

Conclusion

The results above show that the definition of cost given in the present paper is not only intuitively natural, but also is in agreement with the traditional definition and enables comparatively simple generalization of the preceeding results, in particular, the fundamental theorem for noiseless channels.

References

1. *Shannon, C.*, A mathematical theory of communication. Bell System Techn. J., **27** (1948), No. 3, pp. 379—423; **27** (1948), No. 4, pp. 623—656.
2. *Feinstein, Amiel*, Foundations of information theory. McGrow-Hill Book Company, Inc., 1958.
3. *McMillan, B.*, Two inequalities implied by unique decipherability, IRE Trans., IT-2, No. 4 (1956), pp. 115—116.
4. *Fano, Robert M.*, Transmission of information. Published jointly by MIT Press and John Wiley & Sons, Inc., 1961.
5. *Sidel'nikov, V. M.*, On statistical properties of transformations done by finite automata. Kibernetika (1965), 6, pp. 1—14 (in Russian).
6. *Billingsley, P.*, Ergodic theory and information.
7. *Csiszár, I.*, On noiseless channels. Problemi peredachi informatsii (1970), **4**, No. 6, pp. 3—15 (in Russian).
8. *Billingsley, P.*, On the coding theorem for the noiseless channel, Ann. Math. Statistics. (1961), **32**, No. 2, pp. 594—601.
9. *Neveu, Jacques*, Bases Mathématiques du calcul des probabilités. Masson et Cie, Paris, 1964.
10. *Khodak, G. L.*, On the average delay of decipherable code decoding. Dokladi Akademii Nauk SSSR (1972), **207**, No. 6, pp. 1284—1287 (in Russian).

SOME PROBLEMS OF TIME OPTIMAL CONTROL FOR STOCHASTIC SYSTEMS

V. B. KOLMANOVSKY
(Moscow)

The problem of time optimal control in stochastic systems is considered. Some existence conditions for optimal and allowable control are established. A way to approximate solution of the problem is suggested. Specific examples are demonstrated.

1. Let a controlled system has the following form

$$\dot{x}(t) = f(x(t)) + Bu(x(t)) + \sigma(x(t))\,\dot{\xi}(t),\ t \geq 0$$

$$x(0) = x_0 \tag{1.1}$$

where β is a given constant matrix; $x(t)$ is a vector of phase coordinates from Euclidean space E_n; function f and matrix σ satisfy conditions providing existence of solution of Eq. (1.1) at $u = 0$, i.e. [1] at a constant $c \geq 0$ and $x_1,\ x_2 \in E_n$.

$$| f(x_1) - f(x_2) | + | \sigma(x_1) - \sigma(x_2) | \leq c\, | x_1 - x_2 |.$$

Finally, $\xi(t)$ is a Wiener process; vector $x \in E_n$ is given; and $u(x) \in E_m$ is control where for all x

$$u(x) \subset U \tag{1.2}$$

U being a given convex limited set.

Let in phase space E_n some set Q be also given (limited or not). Boundary of Q is denoted below as Q_1, and the moment when process (1.1) with initial condition $x(0) = x$ (arbitrary initial point $x \in E_n$) reaches for the first time set Q under control u, is denoted as $\tau_x\,(u)$.

If some realization of process (1.1) do not reach Q during any finite time, stochastic value of $\tau_x\,(u)$ for these realizations is assumed to be equal to infinity.

Problem. Choose control u satisfying restraint (1.2) and minimizing functional

$$M\tau_x(u) \tag{1.3}$$

where M is the sign of mathematical expectation.

In the theory of optimal control of determinate systems (systems (1.1) at $\sigma = 0$) the problem of time occupies a special place [2]. Rather complete and finished results have been obtained for this problem; in particular, theorems of existence and uniqueness of time optimal control were proved, the optimal form of control was established, algorithms for construction of approximate control were suggested, various cases were considered.

At the same time, in the formulated above problem of minimization of the mean time required by a controlled stochastic process to reach a given set, known complete results (see, e.g., [3]) reduce to the fact that under certain conditions of smoothness of minimal value $V(x)$ of functional (1.3), it will satisfy Bellman's partial differential equation:

$$\min_{u \in U} (L_u(V(x) + 1) = 0, \qquad x \in Q_2$$

$$V(x) = 0, \quad x \in Q_1$$

$$L_u V = (f + BU)^1 \frac{\partial V}{\partial x} + \frac{1}{2} \operatorname{Tr} \sigma\sigma^2 \frac{\partial^2 V}{\partial x^2}, \qquad V(x) = \min_{u \in U} M\tau_x(n),$$

where Q_2 is complement of $Q \cap Q_1$ with respect to E_n, stroke is the sign of transposition; vector $\partial V / \partial x$ has coordinates $\partial V / \partial x_i$; matrix $\partial^2 V / \partial x^2$ has components $\partial^2 V / \partial x_i \, \partial x_j$, $(i, = 1, 2, \ldots, n)$ and finally, symbol Tr stands for matrix trace.

It should be stressed, however, that problem (1.4) does not have, generally speaking, a unique solution. Therefore, even if $V(x)$ is a sufficiently smooth function, it is not clear in advance which solution of Bellman equation (1.4) coincides with it.

Explain the above-said by means of an example.

Example 1. Let system (1.1) be scalar and have the form of

$$\dot{x}(t) = u(t) + \sqrt{2}\, \dot{\xi}(t), \; | u(t) | \leq 1, \quad x(0) = x. \tag{1.5}$$

Set Q is interval $(-\infty, \alpha]$ at some constant α. For $x < \alpha$, Bellman equation (1.4) has the form of

$$\frac{d^2 V(x)}{dx^2} - \left| \frac{dV(x)}{dx} \right| + 1 = 0 , \quad x > \alpha$$

$$V(\alpha) = 0$$

any function of form $x - \alpha + c(\exp(x - \alpha) - 1)$ (where c is an arbitrary non-negative constant) is its solution. Note, that for this example the optimal control is -1. Indeed, if $x(t, u)$ is solution of (1.5) under control u, then

$x(t_1 - 1) \leq x(t, u)$ with probability 1 under any control u, $|u| \leq 1$. In doing so, the optimal transition time under control $u = -1$ is shown below to be $x - \alpha$ for any $x \geq \alpha$.

Another generalization of some problems of optimal time to linear stochastic systems is proposed in [4] which considers the problem of minimization of the total probability of a controlled stochastic process being outside a given vicinity of zero during infinite time. Ref [4] uses the Lyapunov function method to solve this problem, and, moreover, discusses ways to its approximate solution.

In Section 2 of this paper some existence conditions for allowable control are established, formulated in terms of existence of Lyapunov functions of system trajectories. In Section 3 conditions for existence of optimal control are discussed, particular solution of Bellman equation (1.4) is found coinciding with the minimal value of functional (1.3), and examples are presented. Remind that process (1.1) under control u is referred to as recurrent if $\tau_\alpha(u) < \infty$ with probability 1, and positive if $M\tau_\alpha(u) < \infty$ for any finite x.

2. This section considers examples of existence of allowable control.

Control u will be referred to as allowable if under this control for any finite x, $M\tau_\alpha(u) < \infty$. and if there is a strong solution of (1.1), i.e. [1] there is a stochastic process $x(t)$ measurable with respect to $\xi(s)$, $0 \leq \varrho \leq t$, and satisfying with probability 1 the following integral identity.

$$x(t) - x = \int_0^t \left[f(x(s)) + Bu(x(s))\right] ds + \int_0^t \sigma(x(s))\, d\xi(s).$$

The lemma below follows from [5, p. 130]:

Lemma 2.1. Let in Q exist function $u(x)$ and twice continuously differentiable function $V(x) \geq 0$ such that under control $u(x)$ there is solution of system (1.1) and

$$L_u V(x) \leq -c, \quad x \in Q_2 \tag{2.1}$$

where c is some positive constant. Then $u(x)$ is allowable control, and also $M\tau_x(u) \leq c^{-1} V(x)$.

Example 2. Let a point move along a straight line under action of control u, $|u| \leq 1$, and a stochastic disturbing force $\sigma\dot{\xi}$ where σ is a given constant. Movement equations have the following form:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = u + \sigma\dot{\xi}. \tag{2.2}$$

It is required to get as soon as possible into axis x_2, i.e. in this case Q is a straight line $x_1 = 0$. Demonstrate that under any σ allowable control is

$$u = 1, \; x_1 < 0$$

$$u = -1, \; x_1 < 0. \tag{2.3}$$

To this end, fix some constant $k_1 < \sigma^2/2$ and select k_2 so that for all x_1, x_2 function V be non-negative where V has the following form:

$$V = k_1 x_2 + \frac{1}{2} x_2^2 + k_2 + x_1, \quad x_1 > 0,$$

$$V = k_1 x_2 + \frac{1}{2} x_2^2 + k_2 - x_1, \quad x_1 < 0.$$

With allowance for (2.3) obtain that everywhere outside the straight line $x_1 = 0$,

$$L_u V = -k_1 + \frac{\sigma^2}{2}.$$

Hence, on the basis of Lemma 2.1, process (2.2) under control (2.3) is positive. It may be shown further that control (2.3) is optimal. Indeed, at $x_1(0) > 0$ and any control $u_1(x) \geq 1$ inequality

$$M\tau_{u_1}(x) < M\tau_{-1}(x)$$

does not hold because

$$x_1(t_1 - 1) \leq x_1(t_1 u_1)$$

with probability 1 where $x_1(t, u)$ denotes solution $x_1(t)$ of system (2.2) under control u. In the same manner, we see that control $u = 1$ in domain $x_1 < 0$ is optimal.

Remark. In studying existence of allowable control, requirements of Lemma 2.1, in certain situations, may be somewhat relaxed. To put it more precisely, let there be in Q_2 function $u(x)$ and continuous function $\Delta(x) \geq 0$ such that under control $u(x)$ there is solution of system (1.1), and for any points $t_1 \leq t_2$ in trajectories $x(t, u)$ of system (1.1)

$$MV(x(t_2, u)) - MV(x(t_1, u)) \leq -c(t_2 - t_1),$$

where C is a positive constant. Then $u(x)$ is an allowable control. In particular, it is sufficient to require from function V that

$$\lim_{\Delta \to +0} \frac{1}{\Delta} [MV(x(\Delta, u)) - V(x_0)] \leq -1. \tag{2.4}$$

In this case, if V satisfies the requirements of Lemma 2.1, limit in (2.4) is $L_u V(x_0)$. Validity of these assertions follows from the results of [5, p. 130].

3. In this section existence of a solution of the Bellman equation and ways to find it are discussed. It is assumed below that Q is a bounded domain in phase space E_n with smooth boundary Q of Lyapunov type [7, p. 346].

Denote by R_N a sequence of spheres of radius N in space E_n such that $Q < R_N$. Boundary R_N demote by V_N. Define sequence of scalar functions $V_N(x)$ from the following boundary problem:

$$\begin{gathered} \min_{u \in U} \left(L_u V_N(x) + 1\right) = 0\,, \quad x \in Q_2 \cap R_n\,, \\ V_N(x) = 0\,, \quad x \in Q_1\,, \quad x \in V_N\,. \end{gathered} \tag{3.1}$$

Here expression $L_u V_N(x)$ is defined by formula (1.4). List, in the form of a lemma, some of the properties of functions $V_N(x)$ following from the results of [8–10].

Lemma 3.1. Assume that coefficients of Eq. (3.1) satisfy requirements of Section 1, matrix $\sigma(x)$ is non-degenerate at any $x \in E_N$, and bounded domain Q has Lyapunov-type boundary Q_1. Then there is a unique solution $V_N(x)$ of (3.1), $V_N(x) \geq 0$, and for any $N_2 \geq N_1$

$$V_{N_2}(x) \geq V_{N_1}(x), \; x \in Q_2 \cap R_{N_1}\,.$$

Proof. Existence of the unique solution of the boundary problem (3.1) follows from Theorem 4.10.2 of [8], or from Theorem 3 of [9]. Further, by virtue of [10, p. 603], there is measurable function $u_N(x) \in \cap$ providing minimum in Eq. (3.1). Function $V_N(x)$, thus, satisfies linear equation

$$L_{u_N} V_N(x) + 1 = 0, \quad x \in Q_2 \cap R_N\,.$$

Hence, with allowance for zero boundary values (3.1) for $V_N(x)$ and [8, p. 160], it follows that $V_N(x) \geq 0$. Validity of the latter statement of Lemma 3.1 follows from the non-negativeness of all $V_N(x)$ and from Theorem 40.1 of [11].

Theorem 3.1. Let all requirements of Lemma 3.1 be valid and let there be allowable control $\omega(x)$. Then

1. there is minimal positive solution of external boundary problem (3.1) denoted as $V_0(x)$;

2. over any limited set of variations of x, sequence $V_N(x)$ converges evently to $V_0(x)$ at $N \to \infty$ without decreasing monotonically;

3. there is no allowable control giving value of functional (1.3) less than $V_0(x)$;

4. if under control $u_0(x)$ defined by condition

$$\min_{u \in U} L_u V_0(x) = L_{u_0} V_0(x)$$

there is solution of system (1.1), then $u_0(x)$ is optimal.

Proof. Assume that

$$W(x) = M\tau_x(\omega).$$

First show that for all N

$$V_N(x) \leq W(x). \tag{3.2}$$

Denote by $\tau_x^N(u)$ the moment when process (1.1) reaches for the first time boundaries of set $Q_2 \cap R_N$ under control u and $x \in Q_2 \cap R_N$, and by $x(t, u)$ solution of system (1.1) under control u.

Since it is evident that

$$M\tau_x^N(\omega) \leq W(x), \tag{3.3}$$

to prove (3.2) it is sufficient to establish that

$$V_N(x) \leq M\tau_x^N(\omega) \tag{3.4}$$

and, to this end, apply Dynkin's formula [12, p. 191] to function $V_N(x)$ and Markov moment $\tau_N(\omega)$, this being possible on the basis of the definition of $V_x(x)$, Lemma 3.1 and finiteness of mathematical expectation $\tau_x^N(\omega)$ by virtue of (3.3). With allowance for (3.1) at $x \in Q_2 \cap R_N$ we have

$$V_N\big(x(\tau_x^N(\omega), \omega) - V_N(x) = M \int_0^{\tau_x^N(\omega)} L_\omega V_N(x(t, \omega))\, dt \geq$$

$$\geq M \int_0^{\tau_x^N(\omega)} \min_{u \in U} L_u V_N(x(t, \omega))\, dt = -M\tau_x^N(\omega)\,.$$

Thus, inequality (3.4) and, together with it, inequality (3.2) are established.

Fix arbitrary bounded domain $Q_3 \subset E_n$ with boundary of the type like $Q_1, Q < Q_3$, and establish convergence of $V_N(x)$, $x \in Q_2 \cap Q_3$ at $N \to \infty$. To this end, note that on the basis of conditions of the theorem and inequalities (3.2) and with allowance for the *a priori* estimate of [8, page 316] at any N

$$\max_x \left\| \frac{\partial V_N(x)}{\partial x} \right\| \leq c_1\,, \quad x \in Q_3 \cap Q_2\,. \tag{3.5}$$

Here and below C_i are some non-negative constants.

On the basis of (3.5), (3.2), the non-degeneracy of matrix $\sigma(x)$ and validity for it of Lipschitz' condition, conclude using estimate of second derivatives of solutions of equations (3.1) [8, Theorem 4.6.3], that

$$\max_x \left\| \frac{\partial^2 V_N(x)}{\partial x^2} \right\| \leq c_2, \quad x \in Q_3 \cap Q_2. \tag{3.6}$$

Estimates (3.2), (3.5) and (3.6) mean that over any limited set $Q_3 \cap Q_2$ sequence $V_N(x)$ is limited evenly together with its two derivatives. Therefore [13, p. 240], sequence $(V_N(x), \partial V_N(x) | \in x)$ is compact in the space of continuous functions for $x \in Q_3 \cap Q_3$. As usual, assume that the sequence $(V_N(x), \partial V_N(x) | \partial x)$ itself converges evenly. Denote limit $V_N(x)$ under $N \to \infty$ by $V_0(x)$. Then limit $V_0(x)$ is solution of (1.4) on the basis of [11, p. 176, Theorem 42.9] and the arbitrariness of domain Q_3.

Establish now that $V_0(x)$ is the minimal positive solution of problem (1.4). Assume the oppposite, i.e. that there would be another positive solution $V(x)$ of (1.4) and that

$$0 \leq V(x) \leq V_0(x). \tag{3.7}$$

Being solution of (1.4), function $V(x)$ satisfies Eq. (3.1) at any N inside domain $x \in Q_2 \cap R_N$ and, moreover,

$$\begin{aligned} V(x) &= V_N(x) = 0, \ x \in Q_1 \\ V(x) &\geq V_N(x) = 0, \ x \in r_N. \end{aligned} \tag{3.8}$$

On the basis of (3.7), (3.8) and the maximum principle for quasi-linear elliptical equations (3.1) [11, p. 162, Theorem 40.1] conclude that for all $x \in Q_2 \cap R_N$ the following holds:

$$V_N(x) \leq V(x). \tag{3.9}$$

Equality $V(x) = V_0(x)$ follows from (3.7), (3.9) and established above convergence of $V_N(x)$ to $V_0(x)$. Thus, statements 1. and 2. of the theorem are proven.

To prove Statement 3 of the theorem, assume that under some allowable control value $W_0(x)$ of functional (1.3) is obtained less than $V_0(x)$. Then, in a way similar to that of estimate (3.2), the fact is established that for any N

$$V_N(x) \leq W_0(x).$$

This fact and convergence of $V_N(x)$ to $V_0(x)$ lead us to the conclusion that the opposite holds, i.e. $W_0(x) \geq V_0(x)$. Finally, to prove Statement 4. of the theorem, take in the conditions of Lemma 2.1 that $V(x) = V_0(x)$ and $u(x) =$

$= u_0(x)$. Since V_0 and u_0 satisfy Bellman equation (1.4), inequality (2.1) holds for them, and on the basis of Lemma 2.1

$$M\tau_x(u_0) \leq V_0(x).$$

On the other hand, by virtue of Statement 3 of the theorem

$$M\tau_x(u_0) \geq V_0(x).$$

Hence, Statement 4 of the theorem is correct.

Q.E.D.

Remark. Let $\beta = 0$ in (1.1) (i.e. system is uncontrollable) and process (1.1) under $\beta = 0$ be positive. Then assertion of Theorem 3.1 that attainment time is the minimal positive solution of Eq. (1.4) at $\beta = 0$, follows from the results of [5, 6]. The theorem above allows a number of generalizations. Some of them will be formulated as corollaries.

Corollary 1. Let all conditions of the theorem be satisfied, but function $f(x)$ in (1.1) grow not faster than a certain power $||x||$ at $||x|| \to \infty$ and satisfy Lipschitz' condition over each bounded set, i.e. at $||x_1|| \leq N$, $||x_2|| \leq N$, $x_1, x_2 \in E_n$

$$||f(x_1) - f(x_2)|| \leq c(N)\,||x_1 - x_2||, \tag{3.10}$$

where N is an arbitrary number, and $c(N)$ is a given constant depending, generally speaking, on N.

Then all the assertions of the theorem hold, and validity of Corollary 1 is established by repetition of the proof of the theorem.

Corollary 2. As it may be seen from the proof of the theorem, the requirement of existence of allowable control was used only for justification of estimate (3.2). Thus, the statements of the theorem hold if either all of its conditions, or those of Corollary 1 are satisfied except for the requirement of existence of allowable control which is substituted by any of the following assumptions:

a) there is non-negative solution W_0 of problem (1.4);

b) sequence $V_N(x)$ is evenly limited over any bounded set of variations of x; and

c) sequence $V_N(x)$ converges pointwise to the solution of problem (1.4).

To prove this, it is sufficient to demonstrate that under any of Conditions a) to c) estimate of the type (3.2) holds. Indeed, under Condition a), sequence $V_N(x) \leq W_0(x)$ for any N by virtue of [11, page 162], and Condition

a) follows from Condition b). These facts give a technique for finding the time optimal control in the stochastic problem. To do this, one has either to determine the minimal positive solution of problem (1.4), or to construct sequence $V_N(x)$ from (3.1) and to pass to the limit at $N \to \infty$. In any of these cases, if minimum in (1.4) is attained over the allowable control, the latter will be optimal.

Remark. Let $V(x)$ be positive solution of Bellman equation (1.4) and the requirements of Section 1 be correct. Then for any $1 > \varepsilon > 0$ there will be allowable control $u_0(x)$ such that

$$M\tau_x(u_0) \leq \frac{1}{1-\varepsilon} V(x). \tag{3.11}$$

To substantiate (3.11), note that from Ref. [14, proofs of Lemmas 2.1, 2.2] an estimate follows:

$$u_0' \frac{\partial V}{\partial x} \leq \min_{u \in U} u' \frac{\partial V}{\partial x} + \varepsilon \tag{3.12}$$

and also $u_0(x) \in U$ and satisfies local Lipschitz' condition (3.10). This means that in view of boundedness of U under control $u_0(x)$ there is [15] a unique solution of the system (1.1). From (3.12) and Bellman equation for V we see that

$$L_{u_0} V \leq -1 + \varepsilon.$$

Thus, on the basis of Lemma 2.1, control u_0 is allowable, and inequality (3.11) holds. This leads to the following:

Theorem 3.2. Let assumptions of Lemma 3.1 be correct. Then

1. if there is the minimal positive solution V_0 of problem (1.4), then for any fixed initial point x_0 and any $\varepsilon > 0$ there will be allowable control u such that

$$M\tau_{x_0}(u) \leq V_0(x_0) + \varepsilon. \tag{3.13}$$

2. if the problem of time optimal control is solvable and minimal time $V_0(x)$ is twice continuously differentiable, V_0 is the minimal positive solution of Bellman equation (1.4). Indeed, estimate (3.13) follows from (3.11). To prove the second statement, note that minimal time $V_0(x) \geq 0$ is solution of (1.4). Therefore, by virtue of Corollary 2, a), there is minimal positive solution W of problem (1.4). There had been $W(x_0) < V_0(x_0)$ in some point x_0, then there would have been, in view of Theorem 3.2, 1, such an allowable control u that $M\tau_{x_0}(u) < V_0(x_0)$, this being in contradiction with the minimality of time $V_0(x_0)$.

Note that Theorem 3.1 enables also approximate or numerical determination of the required time through constructing an appropriate sequence $V_N(x)$.

Example 3. Let system (1.1) be scalar and have the form of

$$\dot{x}(t) = f(x(t)) + u(t) + \sigma\dot{\xi}(t). \tag{3.14}$$

Here $\sigma \neq 0$ is a constant; control $u(t)$ complies with the requirement that $|u(t)| \leq 1$; and, finally, set Q, like in Example 1, is interval $(-\infty, \alpha]$.

Require that

$$\int_{\alpha}^{\infty} \frac{1}{b(s)}\, ds < \infty. \tag{3.15}$$

Here

$$b(t) = \exp\left\{-\int_{0}^{t} \frac{2}{\sigma^2}(f(t) - 1)\, ds\right\}. \tag{3.16}$$

Requirement (3.15) is satisfied, for example, if $\sup_{x \geq \alpha} f(x) < 1$.

Show that the minimal transition time in this example is as follows:

$$W(x) = \frac{2}{\sigma^2}\int_{\alpha}^{x} b(t)\, dt \int_{t}^{\infty} \frac{ds}{b(s)}, \qquad x \geq \alpha. \tag{3.17}$$

It follows from Eqs. (3.14)–(3.17) that $L_{-1}W = -1$. Hence control $u = -1$ is allowable and, therefore, optimal, because, on the strength of the comparison theorem of [16, p. 120], $\alpha(t, -1) \leq x(t, u)$ with probability 1 under any allowable control u. This and Theorem 3.1 lead to the conclusion that optimal transition time $V(x)$ corresponding to control -1 is defined by the following relation

$$V(x) = \lim_{N \to \infty} V_N(x), \tag{3.18}$$

where functions $V_N(x)$ are solutions of the following boundary problem

$$\frac{\sigma^2}{2}\frac{d^2V_N(x)}{dx^2} + (f(x) - 1)\frac{dV_N(x)}{dx} + 1 = 0, \qquad x > \alpha$$

$$V_N(x) = 0, \; x = \alpha, \; x = N.$$

After calculations obtain

$$V_N(x) = \frac{2}{\sigma^2}\left[\int_{\alpha}^{N} b(t)\,dt\right]^{-1} \int_{\alpha}^{x} b(t_1)\,dt_1 \int_{x}^{N} b(t_2)\,dt_2 \int_{t_1}^{t} \frac{dt_3}{b(t_3)}. \tag{3.19}$$

But under control $u = -1$ process (3.14) is recurrent. Therefore [5, p. 138]

$$\lim_{N\to\infty} \int_{\alpha}^{N} b(t)\,dt = \infty$$

Hence, by passing to limit at $N \to \infty$ in (3.18), and by allowing for (3.15), (3.17) and the L'Hospitale rule, come to the conclusion that the optimal transient time $V(x) = W(x)$.

Note that in this example under $u = 0$ process (3.14) may even be non-recurrent, for instance, by virtue of [5] if

$$f(x) = \frac{1}{2} \text{ at } x \geq \alpha .$$

Example 4. Consider controlled motion of a rigid body subject to stochastic disturbances. Let x_i, $i = 1, 2, 3$, be components of the kinetic moment with respect to a body-fixed reference frame whose axes coincide with the principal central body axes. Euler's motion equations have the form of

$$\begin{aligned}
\dot{x}_1 &= x_2 x_3 (a_2 - a_3)\,(a_3 a_2)^{-1} + u_2 + \sigma\xi_1 \\
\dot{x}_2 &= x_1 x_3 (-a_1 + a_3)\,(a_1 a_3)^{-1} + u_2 + \sigma\xi_2 \\
\dot{x}_3 &= x_1 x_2 (a_1 - a_2)\,(a_1 a_2)^{-1} + u_3 + \sigma\xi_3
\end{aligned} \tag{3.20}$$

where a_t are principal central moments of inertia; σ is a constant; control $u = (u_1, u_2, u_3)$ at constant $b > 0$ satisfies restriction

$$|\, u \,| = (u_1^2 + u_2^2 + u_2^3)^{1/2} \leq b. \tag{3.21}$$

Note that system (3.20) together with restriction (3.21) may be interpreted, for instance, as follows [17]: there is a pair of controlled gas jets creating torques; gas nozzles may be positioned at any angle with respect to the rigid body; and control action on the system is the same in both directions.

For system (3.20), the problem is formulated as that of coming to sphere $Q = |\, x \,| \leq r$ as quick as possible, in terms of functional (1.3), where r is a given positive number.

Solution of this problem is obtained by Corollary 2 of the theorem which dictates that, first, sequence of boundary problem (3.1) solutions V_N should be found; second, limit V_0 of sequence V_N at $N \to \infty$ should be determined,

and, finally, optimal control established using V_0. Render relations (3.1) more concrete to fit the case under consideration. For $V < |x| < N$ we have

$$\frac{\sigma^2}{2}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial x_3^2}\right)V_N(x)+\min_{u,|u|\le b}\left(u_1\frac{\partial V_N(x)}{\partial x_1}+u_2\frac{\partial V_N(x)}{\partial x_2}+u_3\frac{\partial V_N(x)}{\partial x_3}\right)+$$

$$+x_2x_3\frac{a_2-a_3}{a_3a_1}\frac{\partial V_N(x)}{\partial x_1}+x_1x_3\frac{a_3-a_2}{a_1a_3}\frac{\partial V_N(x)}{\partial x_2}+$$

$$+x_1x_2\frac{a_1-a_2}{a_1a_2}\frac{\partial V_N(x)}{\partial x_3}=-1,\quad V_N(x)=0,\quad |x|=r,\quad |x|=N. \tag{3.22}$$

Define number $r_1=r_1(N)$ as the root of equation

$$\int_r^r z(t)\,dt\int_t^r\frac{du}{z(u)}=\int_{r_1}^N z_1(t)\,dt\int_{r_1}^t\frac{du}{z_1(u)}. \tag{3.23}$$

The following notations are introduced here

$$z(t)=\left(\frac{r}{t}\right)^2 e^{\frac{2b}{\sigma^2}(t-r)}$$

$$z_1(t)=\left(\frac{r}{t}\right)^2\exp\left(\frac{2b}{\sigma^2}(r-t)\right). \tag{3.24}$$

The left side of (3.23) is zero at $r_1=r$ and, taking into account (3.24), grows monotonically. The right side of (3.23) is zero at $r_1=N$ and decreases monotonically at $r\le r_1\le N$. This implies that Eq. (3.23) for any fixed N has a unique solution $r_1(N)$ and

$$r<r_1(N)<N. \tag{3.25}$$

By substituting into (3.22) make sure that solution of (3.22) is defined by formulae

$$V_N(x)=\frac{2}{\sigma^2}\int_r^{|x|}z_1(t)\,dt\int_t^{r_1}\frac{dt_1}{z_1(t_1)}\quad \text{at } r\le(x)\le r_1,$$

$$V_N(x)=\frac{2}{\sigma^2}\int_{|x|}^N z_1(t)\,dt\int_{r_1}^t\frac{dt_1}{z_1(t_1)}\quad \text{at } r_1\le|x|\le N.$$

Show now that

$$\lim_{N\to\infty}r_1(N)=\infty \tag{3.26}$$

Compute derivative $dr_1(N) \mid dN$ existing by virtue of [18, p. 452] and, to this end, differentiate, with respect to N, both sides of equality (3.23). Obtain

$$\frac{dr_1(N)}{dN} = \left[z_1(N) \int_{r_1}^{N} \frac{dt}{z_1(t)} \right] \cdot$$

$$\left[\frac{1}{z_1(r_1)} \int_{r_1}^{r} z_1(t)\, dt + \frac{1}{z(r_1)} \int_{r}^{r_1} z(t)\, dt \right]^{-1} . \tag{3.27}$$

This and (3.25) suggest that sequence $r_1(N)$ grows monotonically, i.e. it has limit at $N \to \infty$. Assume that equality (3.26) is not correct. This means that there is number r_0, $r < r_0 < \infty$, such that $r_1(N) \leq r_0$ for all N. Show, however, that in this case there will be constant $c_1 > 0$ such that for all $N > r_0$ estimate

$$\frac{dr_1}{dN} \geq c_1 \tag{3.28}$$

holds which is in evident contradiction to inequality $r_1(N) \leq r_0$.

By virtue of (3.24), the first term in the right side of (3.27) is

$$d_1 = e^{-\frac{2b}{\sigma^2} N} \frac{1}{N^2} \int_{r_1}^{N} t^2 e^{\frac{2b}{\sigma^2} t}\, dt .$$

After integration see that at all $N > r_0$ and some constant c_2

$$d_1 \geq c_2 < 0. \tag{3.29}$$

Similar reasoning shows that at all N and some constant $c_3 > 0$

$$\frac{1}{z_1(r_1)} \int_{r_1}^{N} z_1(t)\, dt \leq e^{\frac{2b}{\sigma^2} r_1} \int_{r_1}^{N} e^{-\frac{2b}{\sigma^2} t}\, dt \leq c_3$$

$$\frac{1}{z(r_1)} \int_{r}^{r_1} z(t)\, dt \leq r_1^2\, e^{-\frac{2b}{\sigma^2} r_1} \int_{r}^{r_1} \frac{1}{t^2} e^{\frac{2b}{\sigma^2} t}\, dt \leq c_3 . \tag{3.30}$$

(3.28) follows from (3.29), (3.30) and (3.27). This establishes validity of (3.26). By virtue of (3.26) and (3.25), over any bounded set of variations of x the following holds:

$$\lim_{N \to \infty} V_N(x) = \lim_{r_1 \to \infty} \frac{2}{\sigma^2} \int_{r}^{|x|} z(t)\, dt \int_{t}^{r} \frac{dt_1}{z(t_1)} .$$

Substituting (3.24) for function $z(t)$, obtain $\lim_{N\to\infty} V_N(x) =$

$$V_0(x) = \frac{1}{b}\left[|x| - r + \frac{\sigma^2}{b}\ln\frac{|x|}{r} + \frac{\sigma^4}{2b^2}\left(\frac{1}{r} - \frac{1}{|x|}\right)\right].$$

Using representation (3.31) for $V_0(x)$, obtain control $u_0(x)$ providing minimum with respect to u, $|u| \leq b$, in expression

$$u_1\frac{\partial V_0}{\partial x_1} + u_2\frac{\partial V_0}{\partial x_2} + u_3\frac{\partial V_0}{\partial x_3}.$$

Obtain

$$u_0(x) = -bx\,|x|^{-1}. \tag{3.32}$$

On account of Corollary 2 of the theorem, to prove the optimality of control (3.32), it is sufficient to check whether control (3.32) is allowable, i.e. to make sure that under this control there is a solution of (3.20). This follows directly from (3.32), (3.31), Theorem 3.4.1 of [5] and the fact that $L_{u_0}V_0 = -1$.

Note that for $\sigma = 0$ expression (3.31) was derived previously in [19] for a more general case of determinate systems with invariant norm by using a different method.

In connection with Examples 2, 3, 4 for which optimal controls were determined, note that in them the optimal stochastic control coincides with the optimal control for the corresponding determinate system, i.e. the system at $\sigma = 0$.

Acknowledgement. My thanks are due to F. L. Chernous'ko for interest and a helpful discussion which was beneficial for the paper.

References

1. *Gikhman, I. I., Skorokhod, A. V.*, Introduction into the stochastic process theory· Nauka Publ. Moscow, 1965 (in Russian).
2. *Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., Mishchenko, E. F.*, Mathematical theory of optimal processes, Fizmatgiz Publ., Moscow, 1969 (in Russian).
3. *Dorato, P., Chand-Ming Hsieh, Robinson, P.*, Optimal bang-bang control of linear stochastic systems with small noise parameter, IEEE Trans. Automatic Control, AC–12, 1967.
4. *Krasovskii, N. N.*, On optimal control under stochastic disturbances, Prikladnaya Mekhanika i Matematika, Vol. 14, No. 1, 1960 (in Russian).
5. *Khasminskii, R. Z.*, Stability of systems of differential equations under stochastic disturbances of their parameters, Nauka Publ., Moscow, 1969 (in Russian).
6. *Dynkin, E. B., Yushkevich, A. A.*, Theorems and problems concerning Markov processes, Nauka Publ., Moscow, 1967 (in Russian).
7. *Tikhonov, A. N., Samarskii, A. A.*, Equations of mathematical physics, Nauka Publ., Moscow, 1966 (in Russian).

8. *Ladyzhenskaya, O. A., Ural'tseva, N. N.*, Linear and quasi-linear equations of elliptical type, Nauka Publ., Moscow, 1973 (in Russian).
9. *Fleming, W. H.*, Some Markovian optimization problems, J. Math. Mech., 12, 1963.
10. *Naimark, M. A.*, Normed rings, Nauka Publ., Moscow, 1968 (in Russian).
11. *Miranda, K.*, Elliptical partial differential equations. Izdatelstvo innostrannoi literatury, Moscow.
12. *Dynkin, E. B.*, Markovian processes, Fizmatgiz Publ., Moscow, 1963 (in Russian).
13. *Lyusternik, L. A., Sobolev, V. I.*, Elements of the functional analysis, Nauka Publ., Moscow, 1965 (in Russian).
14. *Fleming, W. H.*, Duality and a priori estimates in markovian optimization problems. J. of Math. Anal. and Applic., Vol. 16, No. 2, 1966.
15. *Ito, K.*, Steady-state solution of stochastic differential equations. Matematika (Sbornik perevodov) II, 5, 1697.
16. *Gikhman, I. I., Shorokhod, A. V.*, Stochastic differential equations, Naukova Dumka Publ. Kiev, 1968, (in Russian).
17. *Li, E. B.*, Introduction to optimal control theory. Nauka Publ., Moscow, 1972,
18. *Fikhtengolts, G. M.*, Course of differential and integral calculus, Fizmatgiz Publ., Moscow, Vol. 2, 1959 (in Russian).
19. *Atans, M.*, Optimal control theory. Mashinostroenie, Moscow, 1968.

MAGYAR
TUDOMANYOS AKADEMIA
KONYVTARA

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 гранок.

Статье должна предшествовать аннотация объемом до 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 высылаются гранки статьи, которые они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Voronov, A. A., Petrushinin, E. P.: Control optimization of multiple operation resources (*Воронов А. А., Петрушинин Е. П.* Оптимизация управления ресурсами многократного действия) 275

Marton, K.: On the rate distortion function of stationary sources (*Мартон К.* Об ε-энтропии стационарных источников) 289

Ходак Г. Л. О стоимости передачи последовательностей (*Khodak G. L.:* On the cost sequence transmission) 299

Gerencsér, L.: The extension of the extrapolation technique in the SUMT method (*Геренчер Л.* Развитие экстраполяционных алгоритмов в методе штрафных функций) 311

Dashevsky, M. L.: A semiinvariant method of closing the equations for moments in statistical analysis of nonlinear systems (*Дашевский М. Л.* Семиинвариантный метод замыкания уравнений для моментов в задачах статистического анализа нелинейных систем) 317

Bliznakov, J.: The efficiency of complex hierarchical multilevel systems witd special regard to reliability (*Близнаков Й.* Эффективность сложных иерархических многоуровневых систем с учетом надежности) 327

Breuer, P., Vajta, M., Jr.: Structural character recognition by forming projections (*Брейер П., Вайта М.* Структурное распознавание букв путем проектирования) 339

Колмановский В. Б. О некоторых вопросах оптимального быстродействия в стохастических системах (*Kolmanovsky, V. B.:* Some problems of time optimal control for stochastic systems) 353

Naplatanov, N.: International IFAC Symposium on Optimization Methods 369

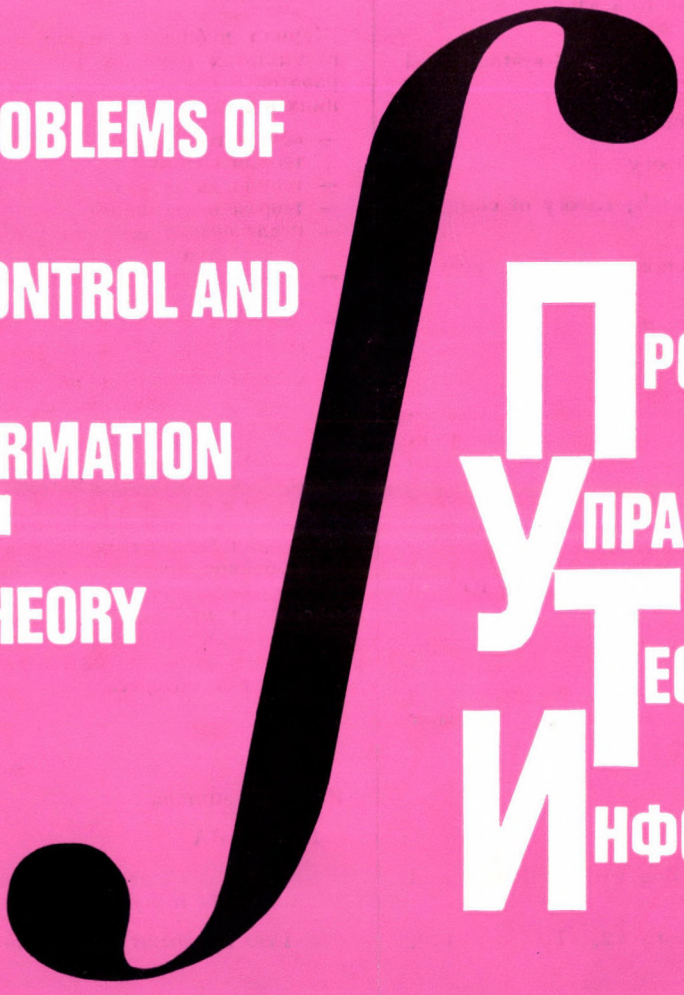
Book Review 373

316.920

Jubilee Number
Юбилейный 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

1975

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 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

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, Jubilee Number, pp. 3—11 (1975)

THE HUNGARIAN–SOVIET SCIENTIFIC AND TECHNICAL COOPERATION

G. BOGNÁR
(Budapest)

(Received April 7, 1975)

On the occasion of the 25th anniversary of signing the Hungarian–Soviet agreement on scientific-technical cooperation the significance of this cooperation and the results obtained in the fields of both science and industry in the course of the cooperation were commemorated all over the country at festive sessions. Scientific and technical lectures appreciated the results attained so far by the scientific-technical cooperation agreement, pointing out the further extension on the basis of the past experiences. In addition, beyond these conferences and consultations the concrete industrial results were illustrated by exhibitions.

The Hungarian Academy of Sciences organized on this occasion a scientific session, with Soviet scientists of world fame, lecturing on the most recent results of Soviet science and on the themes of the cooperation, discussing the advance and the future prospects in several fields of research.

The aim of the session was not alone festive commemoration, but also an opportunity for the Soviet and the Hungarian scientists to survey in the form of a direct exchange of experiences the obtained results of the scientific cooperation, to confirm its contents and to lay the foundations of the future extension of this cooperation.

Surveying the complete scientific cooperation is the task of the scientific session. In this paper we mention merely those fields of cooperation — with no intention to diminish by far the importance and the results of other ones — whose successful cultivation in Hungary would not have been possible without the efficient support of the Soviet Union. In these fields the Soviet Union attained outstanding results by an immense intellectual and material concentration. At the same time, their independent cultivation by a small country like Hungary could not have been carried out with success.

In the age of the scientific-technical revolution, in the process of science turning into a direct productive force, the concentration of huge intellectual and material capacities for solving individual problems of outstanding importance is characteristic. The dimensions and the installed equipments of the

research institutes are comparable with those of industrial objects and the activities of the network of these great size research institutions are controlled by a central special purpose research program. So the characteristic features are the high concentration of the intellectual and material forces and the orientation of the research activities to predetermined aims.

In the framework of big-scale programs like these are realized the research for the utilization of nuclear power, the space research, the development of rocket technics and automatic control, the intensive enhancement of computer capacities and the extension of their application, the utilization of solar energy, the research of new structural materials capable to withstand extraordinarily high loads, the realization of a worldwide communication network, etc.

The age of the scientific-technical revolution is characterized by the development not being controlled by the direct needs, but the very high demands, justified by the attainment of peaks. For an example let us mention that the degree of miniaturization ensured by the integrated circuits containing 10 000 elements in a few square mm area is not required by radio and television program receivers, whose dimensions are fixed by the dimensions of the video-tube, the loudspeaker and the human hand, operating the control knobs. Nevertheless the integrated circuits designed for computers mounted in rockets are ever more extensively applied also in these apparatuses, simply by the fact that the technology elaborated for the purposes of rockets is available; the mass production results in acceptable prices, so the capacity and the reliability of the program receivers may also be perfected by the use of these technics.

The utilization of the specific structural materials applied in rocket technics permitted the use of supersonic aircraft for passenger traffic.

The daily utilization of long-distance telephone calls and TV broadcasting with the help of space communication systems elaborated for rocket control is generally known.

In Hungary there is no possibility to concentrate intellectual and material capactities in a degree permitting the attainment of the aims of a peak program influencing the whole technical progress, with any probability. This is why the efforts of Hungary's participation in the work of the scientific programs carried out by the Soviet Union in the field of the utilization of nuclear power, space research, computer technics aimed at concrete goals are of an extraordinary importance.

The cooperation in the field of utilizing nuclear power for peaceful purposes started in 1955 with our experts signing an agreement in the Soviet Union for settling a well-proved research reactor in Hungary.

This research reactor started to operate in 1959 and since it fulfilled with no breakdown the tasks demanded by the research activities.

During the past 15 years the research reactor became indispensable both for the scientific research and numerous branches of our industry. It created the possibillity of starting the development of nuclear physics, nuclear chemistry, reactor technics, several modern fields of solid-state physics, etc. in our country. The production of isotopes, activation analytics, biological irradiation research, etc. attained an ever increasing importance in many branches of our national economy.

One of the most important effects of the research reactor may be measured by the fact that during the past 15 years internationally acknowledged experts of reactor technics and reactor operation were raised, who joined already in the international work distribution organized in the framework of CMEA and in the solution of the technical-scientific tasks connected with the erection of the nuclear power station at Paks.

The wide-range propagation of radioactive isotopes was made possible in our country by the operation of the research reactor. Methods of producing short-life isotopes were elaborated and numerous isotope preparations for special purposes were produced. A very great number of home and foreign utilization demands were initiated and satisfied later at a high level.

All this shows that the research reactor realized in the framework of the Hungarian–Soviet scientific-technical cooperation contributed significantly to the development of the whole Hungarian technical and natural-scientific research. Our experts, who learned very much from the Soviet scientists, became subsequently actively participating partners in the solution of the common tasks, capable for cooperation in numerous fields.

It should be especially mentioned the cooperation developed in the field of reactor research with the Kurchatov Atomic Energy Institute and the Institute for Physical Energetics. Soviet experts participated in our Central Research Institute for Physics in the initiation of the so-called zero-reactor experiments, whose results contributed significantly to the suggestion of creating a temporary international research collective in the Central Research Institute for Physics for studying the neutron-physical properties of the reactor zones of 1000 MW capacity compressed water nuclear power plants, as well as for controlling and perfecting the calculation methods. The Committee for Atomic Energy of CMEA decided in 1972 — on the basis of the approval of the competent government organs — to establish the research collective involved in these tasks indeed in Budapest, in the Central Research Institute for Physics. The period, which passed since that date, was devoted to intensive work leading to important results, as reported by the head of the international

research collective in Moscow at the November 1973 session of the Committee for Atomic Energy of CMEA.

One of the most important Soviet contributions to the home program of the peaceful utilization of nuclear power is the Nuclear Power Plant for producing electric power, to be erected at the Paks area. The preparations for operating the nuclear power plant and the solution of the scientific-technical problems connected with the investment project would even be unimaginable without the profitable effects of the assistance supplied by the coworkers and managers of Soviet institutes, planning bureaus and state control organs to Hungary in order to promote the peaceful utilization of nuclear power.

Our country is member of the Dubna United Nuclear Research Institute and participates in the work since its foundation. We participate in projecting and realizing the yearly, medium- and long-range plans of the Institute. Our researchers took and take part in the management of the Institute at all levels, as well as in the daily research activities. A number of our experts obtained scientific degrees in Dubna, or at home by utilizing their Dubna results.

The institutes of our Academy (KFKI, ATOMKI) established connections extending year-by-year with the United Dubna Institute, whose great equipments represent in several research themes the only experimental basis. At present about 40 themes are investigated in cooperation between the Hungarian institutes and the Dubna Institute in the field of the physics of elementary particles, nuclear physics and the physics of condensed systems. The Hungarian research in electronics contributes significantly to the automation of the equipments and the measurements in Dubna.

The themes of the cooperation form an organic part of our home research plans, so the results of our research workers attained in Dubna are added to those reached at home and enhance their value.

For developing and rendering more efficient the Hungarian–Soviet cooperation in the field of solid-state research a joint committee was established in 1971 for the promotion and coordination of the common cultivation of solid-state research. The tasks of the committee are as follows:

1. Organizing and developing the cooperation in certain fields of solid-state research qualified as important for both parties, by the exchange of informations, research coordination and carrying out common research activities;

2. Determining the main directions of the cooperation, evaluating the development of the selected research fields, elaborating proposals for common research themes and analyzing multiply the experiences of the cooperation;

3. Rendering mutual information on the long-range conceptions going beyond the scope of the mixed committee's tasks, but involving fundamentally the future development of solid-state research;

4. Initiating propositions for the practical utilization of solid-state research appearing as useful for other branches.

The suggestion for creating the Joint Committee for Solid-State Research was accepted by the presidiums of both Academies. The founding session of the Joint Committee was held in Budapest in 1973, where the parties mutually outlined their fundamental research conceptions, determined the themes of cooperation and agreed in devoting the work of the Joint Committee mainly to discuss scientific problems.

The Joint Committee held sessions in the spring of 1974 in Moscow and Leningrad.

According to the experiences gained so far, the Joint Committee is promoting the development of Soviet–Hungarian connections in the field of solid-state research with a high efficiency and managed indeed to enrich our home research program by revealing numerous new possibilities, of which those appearing in the field of laser and crystal physics, magnetic and semiconductor research deserve special attention.

The Computer and Automation Research Institute of the Hungarian Academy of Sciences has been cooperating for years with the institutes of the Academy of Sciences of the Soviet Union in the framework of the problem "The scientific problems of computation technics". The Committee TEKMO coordinating and directing the theoretical research in the field of computation technics performed by the institutes of the Academies of the socialist countries was established in 1967. Besides the Computer and Automation Research Institute the Central Research Institute for Physics participates also in this cooperation.

Mutually useful cooperation has been going on for years between the Computer and Automation Research Institute and the Institute for Mathematics and Computation Technics of the Academy of Sciences of the SSR of Latvia in the theme "Computer-aided design of logic elements constructed of great and medium integrated circuits" in the framework of the problem "Construction methods of optimum computerized control systems and the development of modern means of automation". In the framework of the same problem the Computer and Automation Research Institute performs common research with the Institute of Control Problems of the Academy of Sciences of the Soviet Union in the themes "Adapting search methods of complex systems" and "Research and elaboration of the synthesis of pneumatic pulse units and control

devices based on them". The tasks of the elaboration and application of the modern fluid and pneumatic elements are solved in common by both institutes.

The Computer and Automation Research Institute concluded also a cooperation agreement with the Institute for Physics and Mathematics of the Academy of Sciences of the SSR of Lithuania in the following themes: "Statistical methods in the field of computation technics", "The multidimensional analysis and statistics of stochastic processes", "Statistical methods in the field of information processing and pattern recognition".

In connection with the cooperation themes listed above 30—40 research workers visit yearly the Soviet Union for participating in the exchange of experiences, on conferences and for study tours and about the same number of Soviet guests is accepted by the Institute. In addition several aspirants to scientific degrees from Hungary attend courses in the institutes of the Academy of Sciences of the USSR.

Space research is carried out in the framework of the Space Research Cooperation of the Socialist Countries. The main research fields of the theme are as follows:

1. Research in cosmic physics
2. Research in cosmic meteorology
3. Space communication research
4. Medico-biological research

In the framework of these research activities the observation of sputniks (satellites) and upper atmospheric research by using data supplied by sputniks were carried out.

In cosmic meteorology the distribution of the clouds and air contamination were the subjects of studies. These observations were evaluated by computer programs devised for this purpose. Electronic apparatuses were constructed for future instrumental observations, designed for space flight. Also the rare earth contents of the moon rock sample supplied by the LUNA-16 experiments were determined.

The erection of a Hungarian space communication terrestrial station with the help of Soviet experts is under way.

The biological effects of irradiations appearing under space flight conditions were studied.

Let me be allowed to give a more detailed information on two themes of the scientific cooperation with the Soviet Union — beyond simply listing the tasks — all the more so, as I take part directly with my coworkers in these themes, so I am most familiar with them. These themes were not suggested

for filling up the framework of official agreements, but were adopted in the mutual knowledge of the scientific work carried on by the cooperating parties, through their personal connections, weighing up the possibilities of cooperation and determining concrete aims on this basis.

More than 25 years ago I came into contact during a common work of nearly half a year duration through an international conference on radio engineering with Academician V. I. Siforov, Director of the Institute of Information Theory of the Soviet Academy of Sciences. The problem at that time was the determination of the principles and the scientific bases of short-wave frequency distribution among the countries. About 10 years later at a meeting in Moscow the idea was raised that for solving the concrete practical problems of communication and for their development a well-founded theoretical basis should be created. In this period the possibility of the application of space communication was already expected. The application of the methods and the latest results of information theory in addition to the classical theoretical methods and the subsequent stochastic methods became necessary. In the institute of Academician V. I. Siforov the whole complexity of information transfer and processing problems is studied. The idea of the "whole complexity" will be better understood, if we consider that for the engineering construction of the means of communication a deep understanding of the appearing problems and a solid theoretical mathematical basis are necessary. As both the transmission and the reception of the informations occur finally through the sensory organs of man, obviously also the eye and the ear involved in information processing must be studied. And if we think of the fact that the information supplied by the sensory organs is processed after passing the nervous system in the human brain, it is clear that the complete information chain extends from the technical means over the sensory organs up to the brain.

There are problems in communication which require the study of this complete chain in the age of the ever accelerating development. Studies of this nature may lead to revealing the functioning of the nervous system of higher order, whose better knowledge means a release for humanity, just as mechanization made man free of heavy physical work, automation from complex mechanical intellectual load and computerization from mechanical routine activities. According to our present prognostics about 15—20 years are necessary for reaching more significant results in this field, with the help of properly directed scientific research.

Out of this immense complex problematics we cooperate in the research of the electrocardiographical diagnostics of the disorders of the heart and vascular system and the application of stochastic approximation and feedback channels in pattern recognition problems by exchanging researchers, by orga-

nizing seminaries and finally by the practical elaboration of the concrete system. The participants in this cooperation are the Research Institute for Telecommunication and the Computer and Automation Research Institute.

In the field of microwave radio communication a cooperation is realized between the institute NIIR of the Ministry of Posts of the Soviet Union and the Hungarian Research Institute for Telecommunication, started in 1958 by mutual visits of the directors of both institutes and the assessment of the mutual possibilities. The first objective of the cooperation was the interconnection of the radio relay systems developed independently in the Soviet Union and in Hungary to permit their unified, interconnected functioning, even on applying them over the same geographical areas.

This coordination was successful: the Soviet and Hungarian microwave systems developed at that time are functioning in their interconnectiou undisturbed over a distance of several times 10 000 km.

Over the difficulties of harmonizing the parameters of the independently developed Soviet and Hungarian systems we came to the conclusion that it is more advised to coordinate the systems in the phase of their development. Accordingly, after preliminary experiments and theoretical considerations a common work program was elaborated — under an appropriate work distribution — for the development of a new, high-capacity radio relay system to be realized in the years 1963—1967. The volume and the difficulty of this work are characterized by the fact that this was the period of changing over from the first to the second generation in the field of the microwave radio-relay systems.

In the last years a grandious conception of a unified automatic communication network was evolved in the Soviet Union, with the determination of its requirements including cable, microwave and space communication, as well as conducted electromagnetic waves (also laser rays) with the necessary switching systems conceived as a unified communication system, taking into consideration the ever increasingly complex functions of telephone communication, televison image broadcasting and in addition the requirements of information transmission.

The fundamental aim of our present cooperation is the development of microwave systems forming organic parts of this unified automatic communication network, meeting the requirements of the ever increasing demands of service. The space communication systems are developed by the Soviet party; in this program the Hungarian party undertakes partial tasks. The elaboration of the terrestrial communications were distributed according to frequency bands between the partners. The 2 and 4 GHz systems are developed by the Soviet party, while the 6 and 8 GHz ones by the Hungarian party. As a result of the

cooperation the introduction of the the 6 GHz system into production is already under way.

The perfection of the semiconductor devices and the further development of the technology of integrated circuits open new perspectives before the development of the microwave systems. Our future cooperation is based on prognoses made in terms of these prespectives.

The scientific cooperation with the Soviet Union is efficiently promoted by the common journals of the Soviet and the Hungarian Academies of Sciences publishing our scientific results fast in Russian and in English. These are: the "Problems of Control and Information Theory", "Reaction Kinetics and Catalysis Letters" and "Analysis Mathematica".

Our scientific connections are assisted by the students graduating at Soviet universities, the candidatures of Hungarian scientific workers in the Soviet Union, who join in this way in the work of the scientific schools. The scientific long-range fellowship grants are also of high importance in this respect.

The fields of cooperation listed and emphasized above show that the cooperation does not follow a predetermined pattern, but develops in multiple ways depending on the actual scientific fields and possibilities.

The common research tasks determined by agreements, the work distribution in various research themes, the common scientific seminaries, the exchange of research workers, the formation of cooperation fields on the basis of personal connections, the common journals — and the list could be still continued — are all well proved forms of scientific cooperation, giving a framevork in which it is continually expanding and becomes more and more efficient.

G. Bognár
Research Institute for Telecommunication
H–1026 Budapest, II. Gábor Áron u. 65
Hungary

Problems of Control and Information Theory, Jubilee Number, pp. 13–32 (1975)

A SURVEY OF STATISTICAL CONTROL THEORY

V. S. PUGACHEV
(Moscow)

(Received April 23, 1975)

1. Introduction

The use and development of statistical methods in control theory were prepared by the works of physicists in their studies of Brownian motion and other fluctuation phenomena (A. Einstein, A. Campbell, A. D. Fokker, M. Planck) and by subsequent development of probability theory at the end of twenties and in thirties (E. E. Slutsky, A. N. Kolmogorov, N. Wiener, W. Feller, A. Ya. Khinchin).

The first result which could really be used to solve some special problems of control system accuracy was obtained by A. Campbell who derived the expression for the steady-state variance of the output of a stable stationary linear system excited by a Poisson flow of unit pulses [1]. His formula in effect is valid for any white noise at the input.

A further step was made by A. A. Andronov, A. A. Witt and L. S. Pontryagin [2] who used the Fokker–Planck–Kolmogorov equation to find a stationary distribution in the state-plane of a simple second-order stationary system paving thus a way for the solution of some statistical problems of control theory.

The general theory for statistical studies of any systems described by stochastic differential equations was developed in [3] (see also [4]). This theory is based on the equation satisfied by the one-dimensional characteristic function of a process defined by a stochastic differential equation. If the process $X(t)$ is defined by the equation

$$\dot{X} = \varphi(t, X) + \psi(t, X)\, V(t), \tag{1.1}$$

$V(t)$ being a white noise and $\varphi(t, X), \psi(t, X)$ some definite functions, then the one-dimensional characteristic function $g(\lambda; t)$ of $X(t)$ satisfies the equation

$$\frac{\partial g(\lambda; t)}{\partial t} = M e^{i\lambda^T X(t)} \left\{ (i\lambda^T \varphi(t, X(t)) + \chi(\lambda; t, X(t)) \right\}, \tag{1.2}$$

where all the vectors are represented as matrix-columns, T indicates transposed matrices, M means expectation, and $\chi(\lambda; t, x)$ represents the value at $s = t$ of

the partial derivative of the characteristic function $\theta(\lambda\,;\,t,\,s,\,x)$ of the integral

$$W(s) - W(t) = \int_t^s \psi(\tau, x(\tau))\, V(\tau)\, d\tau \tag{1.3}$$

with respect to s:

$$\chi(\lambda; t, x) = \left[\frac{\partial \theta(\lambda\,;\,t, s, x)}{\partial s}\right]_{s=t}. \tag{1.4}$$

Equation (1.2) determines $g(\lambda; t)$, certainly, only if the function $\chi(\lambda; t, x)$ in (1.4) and the expectation in the right member of (1.2) exist for any t. These are the only restrictions on φ, ψ and $V(t)$ necessary for Eq. (1.2) to be valid. If this condition is satisfied, then (1.2) represents in fact the linear integro-differential equation in $g(\lambda; t)$. In the special case in which $\varphi(t, x)$ is a polynomial in x and $\psi(t, x)$ is independent of x, (1.2) represents the linear equation in partial derivatives whose order coincides with the degree of $\varphi(t, x)$ in x.

If (1.2) has a solution $g(\lambda; t)$ which has derivatives of all orders, then using well-known relations between the characteristic function and the moments of a random variable (1.2) yields the infinite-order set of ordinary differential equations for the moments of the process $X(t)$ [3, 4].

2. Linear system analysis

In the special case of a linear equation (1.1)

$$\dot{X} = a(t)X + b(t) + c(t)V(t) \tag{2.1}$$

$\chi(\lambda; t, x)$ is independent of x, $\chi(\lambda; t, x) \equiv \chi(\lambda; t)$, and Eq. (1.2) becomes

$$\frac{\partial g(\lambda\,;\,t)}{\partial t} = \lambda^T a(t) \frac{\partial g(\lambda\,;\,t)}{\partial \lambda} + [i\lambda^T b(t) + \chi(\lambda\,;\,t)]\, g(\lambda\,;\,t), \tag{2.2}$$

$\partial/\partial\lambda$ being the column-matrix with elements $\partial/\partial\lambda_1, \ldots, \partial/\partial\lambda_n$. This is the linear first-order equation in partial derivatives which is easily integrated by well-known standard techniques yielding

$$g(\lambda\,;\,t) = \tag{2.3}$$

$$= g_0(\Phi^{-1}(t)\lambda) \exp\left\{\int_{t_0}^{t} \left[i\lambda^T \Phi^{-1}(t)^T \Phi(\tau)^T b(\tau) + \chi(\Phi(\tau)\Phi^{-1}(t)\lambda\,;\,\tau)\right] d\tau\right\},$$

$g_0(\mu)$ being an arbitrary characteristic function, and $\Phi(t)$ the solution of the adjoint equation

$$\dot{y} = -a(t)^T y \tag{2.4}$$

satisfying $\Phi(t_0) = I$.

In the special case of normal white noise $V(t)$

$$\chi(\lambda\,;t) = -\frac{1}{2}\,\mu^T\,\nu(t)\,\mu \tag{2.5}$$

where $\nu(t)$ is the intensity of the white noise $V(t)$, and $\mu = c(t)^T\lambda$. Then assuming $X(t_0) = 0$ (2.3) becomes

$$g(\lambda\,;t) = \exp\left\{i\lambda^T\int_{t_0}^{t} w_1(t,\tau)\,b(\tau)\,d\tau - \frac{1}{2}\,\lambda^T\left[\int_{t_0}^{t} w_2(t,\tau)\,\nu(\tau)\,w_2(t,\tau)^T\,d\tau\right]\lambda\right\}, \tag{2.6}$$

$$w_1(t,\tau) = \Phi^{-1}(t)^T\Phi(\tau)^T \quad\text{and}\quad w_2(t,\tau) = w_1(t,\tau)\,c(\tau)$$

being the weighting functions of the system with respect to the inputs $b(t)$ and $V(t)$. Formula (2.6) shows that the process $X(t)$ is normally distributed with the expectation and the correlation matrix given by

$$m_x(t) = \int_{t_0}^{t} w_1(t,\tau)\,b(\tau)\,d\tau, \tag{2.7}$$

$$\Theta_x(t) = \int_{t_0}^{t} w_2(t,\tau)\,\nu(\tau)\,w_2(t,\tau)^T\,d\tau\,. \tag{2.8}$$

These formulae now widely used are thus consequences of the expression (2.3) for the characteristic function of $X(t)$ published firstly in 1944 in [3].

In the special case of a linear system the infinite set of ordinary differential equations for the moments of the process $X(t)$ obtained in [3] is partitioned into independent sets of finite-order equations for moments of various orders. In particular we obtain the following equations for the moments of the first and the second orders m_x and Θ_x

$$\dot{m}_x = a(t)\,m_x + b(t)\,, \tag{2.9}$$

$$\dot{\Theta}_x = a(t)\,\Theta_x + \Theta_x a(t)^T + c(t)\,\nu(t)\,c(t)^T. \tag{2.10}$$

These equations were independently obtained later by D. B. Duncan [5].

The general statistical theory of linear systems was given in [6]. This theory which is applicable to any linear system contains two approaches to the statistical study of linear systems.

One method is based on general formulae for the expectation and the correlation function of the output $Y(t)$:

$$m_y(t) = \int_{t_0}^{t} g(t, \tau)\, m_x(\tau)\, d\tau, \tag{2.11}$$

$$K_y(t, t') = \int_{t_0}^{t}\int_{t_0}^{t'} g(t, \tau)\, K_x(\tau, \tau')\, g(t', \tau')^T\, d\tau\, d\tau', \tag{2.12}$$

$m_x(t)$ and $K_x(t, t')$ being the expectation and correlation function of the input $X(t)$, and $g(t, \tau)$ weighting function of the system. These formulae represent now conventional techniques for the statistical study of linear systems.

Another method is based on the theory of canonical expansions of random functions. If we represent the input to the system $X(t)$ by some canonical expansion

$$X(t) = m_x(t) + \sum_{\nu=1}^{\infty} V_\nu\, x_\nu(t) \tag{2.13}$$

or by an integral canonical expansion

$$X(t) = m_x(t) + \int x(t, \lambda)\, Z(d\lambda) \tag{2.14}$$

$V_1, V_2, \ldots$ being uncorrelated random variables, $Z(A)$ a stochastic measure, and $x_\nu(t)$, $x(t, \lambda)$ the respective coordinate functions, we obtain the similar canonical expansion for the output

$$Y(t) = m_y(t) + \sum_{\nu=1}^{\infty} V_\nu\, y_\nu(t) \tag{2.15}$$

or

$$Y(t) = m_y(t) + \int y(t, \lambda)\, Z(d\lambda) \tag{2.16}$$

the functions $m_y(t)$, $y_\nu(t)$, $y(t, \lambda)$ representing the responses of the system to the inputs $m_x(t)$, $x_\nu(t)$, $x(t, \lambda)$ respectively. They may be calculated by using expressions of the form (2.11), or by simulating the system, or even by using the system itself feeding it by the respective inputs $m_x(t)$, $x_\nu(t)$, $x(t, \lambda)$.

The latter techniques involve as a special case the use of spectral expansions of stationary random functions for studying stable stationary linear systems [7].*

* See [8, 9] and [10], Ch. 13, and respective references there. See also [11], Ch. 7.

3. Nonlinear system analysis

The generality and simplicity of the statistical theory of linear systems lead to a natural tendency to use it for the approximate study of nonlinear systems. So various types of linearization techniques have been employed. The simplest linearization technique is based on the Taylor-series expansion of nonlinearities [6, 10, 12]. Yet this type of linearization is not applicable to some typical nonlinearities encountered in control systems such as discontinuities, saturations, etc. Statistical linearization techniques are generally used in such cases. This method was proposed independently by R. C. Booton [13] and I. E. Kazakov [14, 15] and developed in many subsequent papers which cannot be mentioned here individually. So we confine ourselves by mentioning a recent survey of statistical linearization methods by I. N. Sinitsyn [16] where numerous references may be found. We mention also the excellent book by F. Csáki [17] devoted to nonlinear problems of modern control theory where the exposition of statistical linearization techniques with detailed bibliography is given.

The idea of the method of statistical linearization is to replace the actual input-output relation for a memoryless nonlinearity $Y = \varphi(X)$ by an approximate expression

$$Y \approx k_0 m_x + k_1(X - m_x), \tag{3.1}$$

k_0 and k_1 being some coefficients (matrices in general) depending on the expectation m_x and the correlation matrix K_x of the input. The difference between Kazakov's and Booton's approaches is that Kazakov uses different gains k_0, k_1 for the expectation m_x (which is supposed to be a signal carrying information) and for the noise $X - m_x$, while Booton takes $k_0 = k_1$. The gains k_0 and k_1 are generally chosen so as to minimize the mean-square error of the representation (3.1). Using the expression of the form (3.1) for all nonlinearities involved, we linearize the system with respect to the fluctuations only, and it remains nonlinear with respect to signals carrying information, preserving hence all specific features of nonlinear systems.

To take into account the deformation of the spectrum of the signal by nonlinearity the term of the form $k_2(\dot{X} - \dot{m}_x)$ may be added to the right member of (3.1) [18].

The most general statistical linearization technique is based on expressions of the form (3.1), m_x and X being considered as elements of a function space and k_0, k_1 as integral operators.

Another approach to approximate statistical study of nonlinear systems is based on equations for moments of the output. The simplest way to obtain such equations is the use of the statistical linearization [19]. More general

approaches are based on Pugachev's Eq. (1.2) for the characteristic function $g(\lambda; t)$ of the process $X(t)$.

Let $f(x; t)$ be the (unknown) probability density of the instantaneous value of the process $X(t)$. Assuming a definite functional form $f(x; t, \alpha)$ for $f(x; t)$ depending on some unknown finite-dimensional vector α we may always suppose that the components of α represent the moments of $X(t)$ up to a given order. Using then well-known relations between the moments and the characteristic function of a random variable we obtain from (1.2) the closed finite set of equations

$$\dot{\alpha}_{k_1, \ldots, k_n} = \tag{3.2}$$
$$= \int \frac{\partial^{k_1 + \cdots + k_n}}{\partial(i\lambda_1)^{k_1} \ldots \partial(i\lambda_n)^{k_n}} \left\{ e^{i\lambda^T x} \left[i\lambda^T \varphi(t, x) + \chi\left(\psi(t, x)^T \lambda; t\right) \right] \right\} \Big|_{\lambda=0} f(x; t, \alpha)\, dx .$$

This method was proposed by V. I. Demukh [20]. Assuming normal $f(x; t, \alpha)$ with $\alpha = \{m_x, \Theta_x\}$ the equations for m_x and Θ_x derived by the statistical linearization method in [19], are obtained as a special case. So Demukh's approach leads to more general and more precise results than those obtained by statistical linearization.

A similar equations may be obtained for the cumulants of the process $X(t)$. Assuming a definite functional form $f(x; t, \beta)$ for the unknown probability density $f(x; t)$ depending on the set of unknown cumulants β up to a given order and wrighting $g(\lambda; t, \beta)$ for the respective function $g(\lambda; t)$ we obtain

$$\dot{\beta}_{k_1, \ldots, k_n} = \tag{3.3}$$
$$= \int \frac{\partial^{k_1 + \cdots + k_n}}{\partial(i\lambda_1)^{k_1} \ldots \partial(i\lambda_n)^{k_n}} \left\{ \frac{e^{i\lambda^T x}}{g(\lambda; t, \beta)} \left[i\lambda^T \varphi(t, x) + \chi\left(\psi(t, x)^T \lambda; t\right) \right] \right\} \Bigg|_{\lambda=0} f(x; t, \beta)\, dx .$$

This method was proposed by M. L. Dashevskii and R. S. Liptser [21, 22]. Equations (3.3) are sometimes preferable as compared with Eqs. (3.2) because the cumulants of higher order are small. Yet Eqs. (3.3) are more difficult to obtain in an explicit form than Eqs. (3.2). Thus the combined method is perhaps the best one. This combined method is based on using such $f(x; t, \alpha)$ in (3.2) for which all the cumulants of order higher than a given number are zeros. This is equivalent to the closing a truncated set of equations for the moments of [3, 4] by the relations obtained by equating to zero the cumulants of higher orders [23].

Other approaches to the approximate solution of Eq. (1.2) were developed in [24, 25].

Other group of methods for the approximate statistical study of nonlinear systems — the so-called methods of equivalent disturbances — was initiated by B. G. Dostupov [26] and developed in various works, especially in [27, 28].

There is yet a wide class of nonlinear systems the precise statistical study of which can be accomplished by using the statistical theory of linear systems. This is the class of systems reducible to linear ones [8, 9, 29].

A system is called reducible to linear one if its input-output relation has the form

$$y(t) = A\varphi(t, \tau_1, \ldots, \tau_n, x(\tau_1), \ldots, x(\tau_n)), \tag{3.4}$$

$\varphi(t, \tau_1, \ldots, \tau_n, x_1, \ldots, x_n)$ being a finite-dimensional vector function of the arguments indicated, and A a linear operator in the space of respective vector functions of n variables $\tau_1, \ldots, \tau_n$. This class of systems covers systems described by truncated Volterra series, $\varphi(t, \tau_1, \ldots, \tau_n, x_1, \ldots, x_n)$ being the vector whose components are the products $x_1, x_1x_2, \ldots, x_1x_2 \ldots x_n$ and A the integral operator with the kernel

$$g(t, \tau_1, \ldots, \tau_n) = \begin{bmatrix} g_{11}(t, \tau_1) \ldots g_{1n}(t, \tau_1, \ldots, \tau_n) \\ g_{21}(t, \tau_1) \ldots g_{2n}(t, \tau_1, \ldots, \tau_n) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ g_{m1}(t, \tau_1) \ldots g_{mn}(t, \tau_1, \ldots, \tau_n) \end{bmatrix}. \tag{3.5}$$

The system described by Eq. (3.4) can be reduced to the linear system putting

$$z(t, \tau_1, \ldots, \tau_n) = \varphi(t, \tau_1, \ldots, \tau_n, x(\tau_1), \ldots, x(\tau_n)). \tag{3.6}$$

If the $2n$-dimensional probability distribution of the random input $X(t)$ is given, then using well-known formulae of probability theory the moments of the first and second orders for the random function

$$Z(t, \tau_1, \ldots, \tau_n) = \varphi(t, \tau_1, \ldots, \tau_n, X(\tau_1), \ldots, X(\tau_n)) \tag{3.7}$$

can be calculated after which the standard techniques of the linear theory yields the expectation and correlation function of the output $Y(t)$. Some special systems of this kind, i.e. systems described by Hammerstein operators were studied by L. A. Zadeh [30, 31]. The general statistical theory of this class of systems was given in [29].* The systems described by truncated Volterra series were previously studied in [8, 9]. Further development of this theory was given in [32–37] and in many other works which cannot be cited here.

* See also [10], §§ 106, 107 (105–107 in the 1st Russian edition, 1957).

4. Linear system synthesis

A mathematical tool for solving optimization problems for discrete-time stationary linear systems using the mean-square error criterion was first given by A. N. Kolmogorov [38, 39], and independently some time later by N. Wiener who extended his theory also to continuous-time stationary linear systems [40]. Some further steps were made in [41–45]. The general statistical theory of optimal linear systems using the mean-square criterion was developed in [46–49]. The general solutions of optimization problems for any linear systems having any number of inputs and outputs were obtained in these papers using unified method of canonical expansions of random functions [6, 12] (see also [10], Chs. 9, 10, 17; [11], Ch. 14, and [8]).*

All the optimization problems for linear systems are reduced to an equation of the form

$$L_\tau K(\tau, s) = \varphi(t, s), \tag{4.1}$$

L_τ being unknown linear operator on the space of functions of τ mapping this space into the space of functions of t, $K(\tau, s)$ the correlation function of some random function $X(t)$, and $\varphi(t, s)$ some known function. Representing $X(t)$ by some canonical expansion of the form (2.13) we find at the same time the set of linear functionals $\{f_\nu\}$ satisfying together with the coordinate functions $x_\nu(t)$ the biorthogonality condition $f_\nu x_\mu = \delta_{\nu\mu}$. Then the solution of Eq. (4.1) is given by

$$L_\tau = \sum_{\nu=1}^{\infty} \frac{f_{\nu s}\varphi}{D_\nu} f_{\nu\tau}, \tag{4.2}$$

subscripts s and τ indicating that $f_{\nu s}$ acts on φ as function of s, and $f_{\nu\tau}$ acts on functions of the variable τ. In special case of sampled-data system $f_\nu\psi$ (and consequently L_τ) represents the linear combination of the values of the function $\psi(t)$ at given values $t_1, \ldots, t_n$ of t. In the case of continuous-time systems $f_\nu\psi$ (and consequently L_τ) represents the integral $\int f_\nu(t)\,\psi(t)\,dt$.

The expression similar to (4.2) may also be obtained by using an integral canonical expansion of $X(t)$ of the form (2.14). As it was pointed out in [47, 48], all existing techniques for the solution of equations of the form (4.1) in a closed form are based either in explicit or implicit form on integral canonical representations of random functions of the form (2.14). It was also indicated in [47] that the theory developed solves also the optimization problems for linear distributed-parameter systems and for systems reducible to linear ones (see also [29]).

* Note that the well-known Karhunen–Loève expansion is a special case of Pugachev's canonical expansion.

The extension of linear optimization theory to a class of complex statistical criteria different from the mean-square one was given by N. I. Andreev [50] (see also [10], Chs. 16, 17).*

An essentially new approach to linear filtering and extrapolation theory was started by R. Kalman and R. Bucy [51, 52] who developed the method leading to an easily instrumentable solution. The Kalman–Bucy theory was extended to the cases of coloured noise in observations [53, 54], and to distributed-parameter systems [55–57].

Following Kalman and Bucy we consider the case in which the signal $W(t)$ to be filtered (estimated) is determined by a linear stochastic differential equation of the form

$$\dot{W} = aW + V \tag{4.3}$$

and the input to the system, the observed signal $Z(t)$ is determined by

$$Z = bW + V_1, \tag{4.4}$$

a and b being some definite matrices (which may depend on the time t), and V, V_1 being uncorrelated white noises. The optimal system with infinite memory is described in such a case by the differential equation

$$\dot{\hat{W}} = (a - kb)\hat{W} + kZ\,, \tag{4.5}$$

where $\hat{W}(t)$ is the output of the optimal system (i.e. the best least-square estimate of the signal $W(t)$), and $k = Hb^T v_1^{-1}$, H being the covariance matrix of the error $E(t) = \hat{W}(t) - W(t)$ of the optimal system, and v_1 the intensity of the white noise $V_1(t)$ (it is essential that v_1 be invertible, as is always the case when the components of $Z(t)$ are measured independently). The matrix H is determined by the Riccati matrix equation

$$\dot{H} = aH + Ha^T - Hb^T v_1^{-1} bH + v\,, \tag{4.6}$$

where v is the intensity of the white noise $V(t)$. Equation (4.5) shows that the optimal system is obtained by closing the system forming the signal $W(t)$ by the feedback with the matrix gain b and using the amplifier with the gain k in the direct line.

Similar results are easily obtained in the case of sampled-data systems if the signal $W(t)$ is determined by a linear stochastic difference equation.

The extension of Kalman's theory to the case of the noise $V_1(t)$ in observed signal different from white noise, but shaped from a white noise by a linear

* At the same time this extension was given independently by V. S. Rybakov who did not publish his work.

filter described by differential equations, was obtained in [54] by using the system inverse with respect to the shaping filter for the noise $V_1(t)$ reducing thus the proplem to the case of Kalman and Bucy.

The most convenient way to extend Kalman's theory to distributed-parameter systems is to convert the equation determining the signal $W(t)$ into the form of Eq. (4.3) in a suitably chosen Hilbert space [57], a, b, k, v and v_1 being linear operators on this Hilbert space in such a case.

5. Nonlinear system synthesis

The first nonlinear system optimization problems were naturally raised in communication theory as problems of minimizing the probability of an error in signal detection. The first results in this were obtained by V. A. Kotelnikov in his theory of potential noise-stability of radio-receivers [58, 59]. Further results and the development of the general theory of signal detection were given in numerous works of which we mention here only [60–63].

Some nonlinear signal estimation problems were solved in [62–66]. In [62, 63] a statistical decision theory approach to signal detection and estimation was also initiated.

To formulate the general problem of optimizing a system let $r(z, \hat{w})$ be the average risk (the cost function) associated with the input z and output $\hat{w}$. It is often used in statistical decision theory the utility function $u(z, \hat{w})$ instead of the average risk $r(z, \hat{w})$. One of these cases can always be reduced to another by taking $u(z, \hat{w}) = r_0 - r(z, \hat{w})$, r_0 being any number chosen so as to make $u(z, \hat{w})$ non-negative for any $z, \hat{w}$. In the case of an automatic system $r(z, \hat{w})$ (and consequently $u(z, \hat{w})$) may easily be evaluated if a loss function is chosen characterizing the cost of deviations of system output $\hat{w}$ from the desired output w and the conditional distribution of the desired output for any given input z is known. Denoting the loss function $l(w, \hat{w})$ we obtain the following expression for the average risk

$$r(z, \hat{w}) = \int l(w, \hat{w})\, p(dw \mid z), \tag{5.1}$$

where $p(w \mid z)$ is the conditional distribution (the regular conditional probability measure) of the random desired output W for a given input z.*

Denoting $\delta(w \mid z)$ the *decision function* of the system, i.e. the conditional distribution (necessarily regular) of the random system output $\hat{W}$ for a given

* For simplicity we write $p(w \mid z)$ instead of more rigorous symbol $p(\Delta w \mid z)$, Δw being a measurable set in the space of all possible outputs (i.e. an element of the appropriate σ-algebra of sets on which $p(\Delta w \mid z)$ is defined). The domains of integration of all the integrals in the sequel represent the whole respective spaces.

input z, the average risk associated with a given decision function δ and a given input z is determined by

$$R(\delta \mid z) = \int r(z, \hat{w})\, \delta(d\hat{w} \mid z). \tag{5.2}$$

The problem of optimizing the system (or the decision rule) is then reduced to finding an optimal $\delta(\hat{w} \mid z)$.

If there exists a non-empty set D_z of values of $\hat{w}$ for which $r(z, \hat{w}) = \inf_y r(z, y)$:

$$D_z = \{\hat{w}: r(z, \hat{w}) = \inf_y r(z, y)\}$$

for any possible input z, then an arbitrary probability distribution

$$\delta_0(\hat{w} \mid z) \text{ on } D_z, \quad \delta_0(D_z \mid z) = 1$$

represents the optimal decision function and $R(\delta_0 \mid z) = \inf_y r(z, y)$. In the special case often encountered in problems of practice where D_z consists of a single point $\hat{w} = \hat{w}(z)$ for any z, the optimal system is deterministic and not stochastic. So statistical optimum system theory often leads to deterministic optimal systems.

Another situation arises when there exist no values of $\hat{w}$ minimizing the average risk $r(z, \hat{w})$ so that $\inf_y r(z, y)$ is not achieved for any $\hat{w}$. In such a situation for any $\varepsilon > 0$ exist always values of $\hat{w}$ for which $r(z, \hat{w}) - \inf_y r(z, y) < \varepsilon$. An arbitrary probability distribution $\delta_\varepsilon(\hat{w} \mid z)$ on the set

$$D_z^\varepsilon = \{\hat{w}: r(z, \hat{w}) < \inf_y r(z, y) + \varepsilon\},$$

$\delta_\varepsilon(D_z^\varepsilon \mid z) = 1$, represents then an ε-optimal decision function, and $R(\delta_\varepsilon \mid z) < \inf_y r(z, y) + \varepsilon$. Hence ε-optimal decision functions always exist and cannot be unique for any $\varepsilon > 0$. And, as one never can indicate exactly an admissible upper bound for ε, the admissible set of decision functions which make the system arbitrarily close to optimum is in fact much wider.

We see thus that modern statistical decision theory permits in principle to find nearly optimal systems in any situations.

The optimization procedure just outlined is based on the assumption that $r(z, \hat{w})$ is known. But it is not usually and completely known. It is often known only up to certain unknown parameters, say is a definite function of z, $\hat{w}$ and some unknown finite-dimensional vector u, $r(z, \hat{w}, u)$. Thus the optimization problem with some degree of uncertainty arises. Assuming that the unknown parameter u represents a realization of certain random vector U

with some prior distribution $\alpha(u)$ ($\alpha(u)$ may be taken as Lebesgue measure in the case where all the components of U are unknown non-random quantities which can assume any values [68, 70, 10, § 144]) we may average $r(z, \hat{w}, u)$ with respect to the posterior distribution $\omega(u \mid z)$; i.e. to replace $r(z, \hat{w})$ in (5.2) by

$$\bar{r}(z, \hat{w}) = \int r(z, \hat{w}, u)\, \omega(du \mid z). \tag{5.3}$$

The uncertainty may in particular arise owing to dependence of the random input Z on some finite-dimensional random vector U, $p(w \mid z)$ depending on u in such a case. General methods of finding optimal systems in the cases where the conditional distribution of Z is normal for any given value u of U and the correlation functions of Z, W are independent of u were developed for any loss functions $l(w, \hat{w})$ in [67–72]. These methods were extended to the cases where not only conditional expectations of Z, W but also their conditional correlation functions depend on u in [73–75]. The general theory developed in [67–75] covers all the statistical problems of signal estimation and signal detection and recognition as well. It is also applicable to a lot of similar problems in other fields such as, for instance, the problems of nonlinear filtering of random fields, etc. All these general results were obtained by means of the method of canonical expansions of random functions (see also [10], Ch. 18; [11], Ch. 15). The extension of this theory to a more general class of signals was given in [76] on the basis of the simultaneous canonical expansions of two random functions with the same set of coordinate functions the theory of which was previously developed in [12] (see also [77]).

The general statistical theory for optimizing systems reducible to linear ones and, in particular, systems described by truncated Volterra series was given in [29] (see also [10], § 137). Further development of the optimization theory for such systems was given in [33–36, 78–81] and in many other papers.

Another technique for solving nonlinear filtering problems was proposed by R. L. Stratonovich [82, 83] who used for this purpose his theory of conditional Markov processes.

An extension of the Kalman–Bucy filtering theory to nonlinear systems was given by A. N. Shiryaev and R. S. Liptser [84–91]. They have noticed that if the signal $W(t)$ to be estimated is a subvector of the vector Markov process determined by stochastic differential equations of the form

$$\left.\begin{aligned} \dot{W} &= a_1(t, W_1)W + a_0(t, W_1) + V, \\ \dot{W}_1 &= b_1(t, W_1)W + b_0(t, W_1) + V_1 \end{aligned}\right\} \tag{5.4}$$

and the input to the system $Z(t) = \dot{W}_1(t)$, then the process $W(t)$ is conditionally normal, i.e. its conditional distribution for a given realization $z(t)$ of the input $Z(t)$ (or, as is the same, for a given realization $w_1(t)$ of $W_1(t)$) is normal. This

fact permits to obtain the closed equations similar to (4.5) and (4.6) for the first and second conditional moments of $W(t)$. These equations represent an extention of the Kalman–Bucy equations describing optimal system because the conditional expectation of $W(t)$ is just the optimal estimate $W(t)$ of the signal $\widehat{W}(t)$.

A special direction in statistical optimization theory is formed by works extending Pontryagin's maximum principle to stochastic optimum control problems. We confine ourselves here to mention only the books [92–97] where numerous references to the original papers are given.

6. Learning system theory

The first scientific approach to processes of learning, self-learning and using experience in automatic systems appears to be made by A. Špaček and his numerous pupils [98–107]. Their approach is based on using statistical estimation of unknown probability distributions of the signals involved for designing algorithms of learning and self-learning. Rather general statistical statement of the problem of self-learning of an automatic system was given in [108], where also the way of its solution using nonparametric statistical estimation was indicated. A Bayesian approach to one problem of pattern recognition seems to be given at first in [109]. The general statistical theory of learning and self-learning of automatic systems was developed in [110–114, 75]. In these papers a unified approach to learning and self-learning of automatic control systems, information processing systems and recognizing systems based on statistical decision theory was given. The concept of the real teacher, i.e. the teacher performing teaching processes with random errors, was introduced here and some general properties of optimal Bayesian learning and self-learning procedures were studied. There was also given in [75, 114] the theory of learning and self-learning processes in randomly varying environment, the effects of teaching information aging in such situations were studied. Finally some approaches to estimate the performance of a taught or self-taught system were proposed.

The possibility of using learning, or self-learning of a system arises when some parameters of a control or information processing problem are unknown and either have constant values or vary slowly enough that their estimates were available with sufficient accuracy to be used in control or information processing algorithms. If, for instance, the parameter u in (5.3) (or some of its components) remains constant or is varying slowly enough during some time of system operating (or some cycles of system operating) then the posterior distribution of u may be estimated more precisely if some additional informa-

tion, except the usual system input is used. Denoting the additional information ξ we may replace the posterior distribution $\omega(u \mid z)$ when the input z only is known, by the posterior distribution $\omega(u \mid z, \xi)$ when z and ξ are known. Then (5.3) takes the form

$$\bar{r}(z, \xi, \hat{w}) = \int r(z, \hat{w}, u)\, \omega(du \mid z, \xi). \tag{6.1}$$

The additional information ξ is called teaching information, the system using $\bar{r}(z, \xi, \hat{w})$ instead of $r(z, \hat{w})$ in (5.2) is called learning system, and the source of the additional information is termed the teacher. If there is no external source of additional information and all the informations ξ consist of past system inputs, then the system is self-learning. In the case of a varying parameter u, $\omega(u \mid z, \xi)$ represents the conditional distribution of its value at a given time instant or at a given cycle of system operating.

The posterior distribution $\omega(u \mid z, \xi)$ was evaluated for the case of a conditionally normal input Z (i.e. conditional distribution of Z with respect to u being normal) in [111–113] for constant u and in [75, 114] for varying u.

Some problems of self-learning of automatic systems using empirical Bayesian approach were considered in [115–119]. Other applications of nonparametric statistical estimation to learning and self-learning systems were developed in [120–123].

An important direction in the statistical theory of learning systems is formed by recursive alogrithms of learning and self-learning. This direction was initiated by A. Špaček and his pupils [102, 104–107, 124–126] and continued in [127–138] and in numerous other works which cannot be cited here individually.

7. Stochastic systems

Modern control problems especially those concerning large-scale systems cannot be solved by usual techniques of automatic control theory. In particular more general models for such systems than those considered in conventional control theory are necessary.

Usual deterministic models are insufficient and must give the way to more general indeterministic models. So a stochastic system seems to be an adequate model for real complex systems encountered in modern control problems.

A general definition of a stochastic system in terms of probability distributions in the space of outputs was given and some general properties of the class of stochastic systems were studied in [139, 140]. Stochastic models may serve to solve many problems of large-scale system control as well as many operation research and system analysis problems.

In view of the complexity of a general mathematical description of a stochastic system and the lack of available information for designing general stochastic models for real complex systems the necessity of using some simpler stochastic models arises. The simplest stochastic model for a real system is represented by a normal stochastic system, i.e. such stochastic system whose input and output have normal joint probability distribution if the input is normally distributed [141]. The notion of the normal system is a natural generalization of that of the linear system. The problem of normalizing arbitrary stochastic systems for purposes of approximate study and control-system design naturally arises. The first steps in this direction were made in [142, 143] where a general approach to the problem was proposed and techniques for normalizing queueing systems were developed.

Numerous new statistical problems arise in stochastic system theory. Some of these problems represent a generalization of similar problems arising in conventional automatic control theory, and some are quite new. The generalization of methods of automatic control theory to more general situations arising in stochastic system theory may be based on respective generalization of the theory of the canonical expansions of random functions [144].

One of the main problems of stochastic system theory is the problem of designing an adequate stochastic model for a real complex (e.g. large-scale) system. This problem is a generalization of ordinary identification problems. An important special class of stochastic models is the class of models described by stochastic differential equations. The problem of identification of systems described by definite linear stochastic differential equations by adjusting some unknown parameters in the expressions of the coefficients of these equations was solved in [145]. But this problem is in effect much more difficult in stochastic system theory because not only some parameters in a definite stochastic differential equation, but an adequate order and form of this equation must also be estimated using operating data.

We cannot give here a detailed survey of works in the important field of identification. A very good textbook on the problems of identification by P. Eykhoff [146] permits us to confine ourselves here by mentioning this book as well as the popular book by N. S. Raibman and V. M. Chadeev [147].

One of the most important problems of stochastic system theory is the problem of decision-making in control and management. Thus the statistical theory of decision-making processes is now one of the main parts of stochastic system theory.

The procedures for finding optimal decisions outlined in Section 6 are also applicable to problems of elaborating decisions made by men. It is only necessary to complete these procedures by some techniques for estimating personal

subjective probabilities and utility functions which characterize personal experience and activity of each manager. But this is a separate problem to be discussed. So we confine ourselves by mentioning only the book [148] and the surveys [149, 150] where further references may be found.

Concluding remarks

We outlined here a brief survey of the main directions of development of statistical control theory. It is easily seen from this survey that the Soviet scientists as well as their collegues from Hungary and Czechoslovakia have brought an essential contribution to this important branch of modern science which covers now the great majority of modern control problems and is in fact very promising for future studies in many other fields.

The author begs to apologize him for not mentioning many interesting and valuable works. It was quite impossible to pay attention to all important contributions to this vast field in a brief paper.

In writing this survey I used reviews of various parts of statistical control theory prepared by A. V. Avtonomova, R. A. Batarshin, A. V. Dobrovidov, I. D. Siluyanova, M. E. Shaikin and L. P. Sysoev. To all them I express grateful thanks. I thank also E. A. Pukhalskii and I. V. Sinitsyna for assistance.

References

1. *Campbell, A.*, Proc. Cambridge Phil. Soc., **15**, 117—136, 310—328 (1909).
2. *Andronov, A. A., Witt, A. A., Pontryagin, L. S.*, Journ. Exper. and Theoretic Phys., **3**, *3*, 165—180 (1933) (Russian).
3. *Pugachev, V. S.*, Trudy Akademii Zhukovskogo, *118* (1944) (Russian).
4. *Pugachev, V. S.*, Izvestiya Akademii Nauk SSSR, Energetika i Avtomatika, *3*, 46–57 (1961) (Russian).
5. *Duncan, D. B.*, Journ. Appl. Phys., **24**, *5*, 609—611(1953).
6. *Pugachev, V. S.*, Foundations of General Random Function Theory. Moscow, 1948 (Russian).
7. *James, H. M., Nichols, N. B., Phillips, R. S.* (eds.), Theory of Servomechanisms. Mass. Inst. of Technol., Radiation Lab. Series, **25**, 1947.
8. *Pugachev, V. S.*, Izvestiya Akademii Nauk SSSR, ser. math., **17**, *5*, 401—420 (1953) (Russian).
9. *Pugachev, V. S.*, Trudy 2nd All-Union Conference on Automatic Control Theory (Moscow, Oct. 1953), **2**, 403—424 (1955) (Russian).
10. *Pugachev, V. S.*, Theory of Random Functions and its Application to Control Problems, Pergamon Press, 1965 (Translation of the Russian book, published by Fizmatgiz, Moscow, 1st edition 1957, 2nd edition 1960, 3rd edition 1962).
11. *Pugachev, V. S.* (ed.), Fundamentals of Automatic Control, Moscow, 1963, 1968, 1974 (Russian).
12. *Pugachev, V. S.*, Foundations of General Random Function Theory, Moscow, 1952 (Russian).
13. *Booton*, R. C., IRE Trans., **CT-1**, 9—18 (1954).
14. *Kazakov, I. E.*, Trudy Akademii Zhukovskogo, *394* (1954) (Russian).
15. *Kazakov, I. E.*, Avtomatika i Telemekhanika, **17**, *5*, 385—409 (1956) (Russian).
16. *Sinitsyn, I. N.*, Avtomatika i Telemekhanika, **5**, 82—94 (1974) (Russian).

17. *Csáki, F.*, Modern Control Theories. Budapest, 1972.
18. *Pugachev, V. S.*, Statistical Methods in Automatic Control. Proc. 2nd IFAC Congress (Basel, Switzerland, August 1963). Nonlinear System Theory. Butterworths, London, 1964, 1—13.
19. *Aleksandrov, V. M., Batkov, A. M., Staroverov, A. N., Shchukin, B. A.*, Avtomatika i Telemekhanika, **26**, *3*, 492—499 (1965) (Russian).
20. *Demukh, V. I.*, Avtomatika i Telemekhanika **26**, *6*, 1021—1025 (1965) (Russian).
21. *Dashevskii, M. L., Liptser, R. S.*, Avtomatika i Telemekhanika, *8*, 32—43 (1967) (Russian).
22. *Dashevskii, M. L.*, Avtomatika i Telemekhanika, *11*, 62—81 (1967); *10*, 63—71 (1968) (Russian).
23. *Dashevskii, M. L.*, Problems of Control and Information Theory, **4**, *4*, 311—317 (1975).
24. *Kazakov, I. E.*, Avtomatika i Telemekhanika, **26**, *11*, 1926—1937 (1965) (Russian).
25. *Malchikov, S. V.*, Avtomatika i Telemekhanika, *5*, 43—51 (1970); *10*, 33—38, *11*, 16—21 (1973) (Russian).
26. *Dostupov, B. G.*, Avtomatika i Telemekhanika, **18**, *11*, 999—1009 (1957) (Russian).
27. *Andrienko, A. Ya.*, in "Tekhnicheskaya Kibernetika". "Science", Moscow, 1965, 3—16 (Russian).
28. *Chernetskii, V. I.*, Analysis of Precision of Non-Linear Control Systems, Moscow, 1965 (Russian).
29. *Pugachev, V. S.*, IRE Trans., **CT-7**, *4*, 506—512 (1960).
30. *Zadeh, L. A.*, Journ. Appl. Phys., **24**, *4*, 396—404 (1953).
31. *Zadeh, L. A.*, Journ. of the Franklin Inst., **255**, *5*, 387—408 (1953).
32. *Wiener, N.*, Non-linear Random Theory. John Wiley, 1958.
33. *Lubbock, J. K.*, Proc. IEE, **107**, Part C, *11*, 60—74 (1960).
34. *Lubbock, J. K.*, Proc. 1st IFAC Congress (Moscow, 1960) Butterworths, London, 1 (1961).
35. *Prasad, T.*, Proc. 1st IFAC Congress (Moscow, 1960), Butterworths, London, **1**, 437—445 (1961).
36. *Van Trees, H. L.*, Synthesis of Optimum Non-Linear Control Systems. The MIT Press, 1962.
37. *Popkov, Yu. S.*, Avtomatika i Telemekhanika, *10*, 89—108 (1967); *3*, 55—64 (1970) (Russian).
38. *Kolmogorov, A. N.*, Bull. Moscow State University, **2**, *6*, 1—40 (1941) (Russian).
39. *Kolmogorov, A. N.*, Izvestiya Akademii Nauk SSSR, ser. math., **5**, *1*, 3—14 (1941) (Russian).
40. *Wiener, N.*, Extrapolation, Interpolation and Smoothing of Stationary Time Series, Wiley, 1949.
41. *Zadeh, L. A., Ragazzini, J. R.*, Journ. Appl. Phys., **21**, *7*, 645—655 (1950).
42. *Booton, R. C.*, Proc. IRE, **40**, *8*, 977—981 (1952).
43. *Davis, R. C.*, Journ. Appl. Phys., **23**, *9*, 1047—1053 (1952).
44. *Dolph, C. L., Woodbury, M. A.*, Trans. Amer. Math. Soc., **72**, *3*, 519—550 (1952).
45. *Semenov, V. M.*, Sborn. Nauchnykh Trudov Akademii Zhukovskogo, **1**, 67—94 (1954) (Russian).
46. *Pugachev, V. S.*, Sborn. Nauchnykh Trudov Akademii Zhukovskogo, **1**, 33—45, 47—66 (1954) (Russian).
47. *Pugachev, V. S.*, Avtomatika i Telemekhanika, **17**, *4*, 289—295; *6*, 489—499; *7*, 585—589 (1956); **18**, *11*, 971—984 (1957) (Russian).
48. *Pugachev, V. S.*, Prikladnaya Matematika i Mekhanika, **23**, *1*, 3—14; *3*, 527—533 (1959) (Russian).
49. *Pugachev, V. S.*, Trudy 6th All-Union Conference on Probability Theory and Math. Statistics (Vilnius, Palanga, Sept. 1960). Vilnius, 1962, 233—237 (Russian).
50. *Andreev, N. I.*, Avtomatika i Telemekhanika, **18**, *7*, 615—619 (1957); **19**, *12*, 1077—1090 (1958); **20**, *7*, 833—838 (1959) (Russian).
51. *Kalman, R. E.*, Trans ASME, Journ. Basic Eng., 82D, 35—45 (1960).
52. *Kalman, R. E., Bucy, R. S.*, Trans. ASME, Journ. Basic Eng., 83D, 95—100 (1961).
53. *Bryson, A. E., Johansen, D. E.*, IEEE Trans., **AC-10**, *1*, 4—10 (1965).
54. *Gulko, F. B., Novoseltseva, J. A.*, Avtomatika i Telemekhanika, **27**, *4*, 122—141, *10*, 153—168 (1966) (Russian).

55. *Falb, P.*, Information and Control, **11**, *1/2*, 102—137 (1967).
56. *Kuhr, D.*, Regelungstechn. und Prozess Datenverarb., **18**, *11*, 506—507 (1970).
57. *Dubenko, T. I.*, Avtomatika i Telemekhanika, *12*, 37—40 (1972) (Russian).
58. *Kotelnikov, V. A.*, Materialy 1st All-Union Conference on Problems of Reconstruction of Communication, Upravleniye Svyazi RKKA (1933) (Russian).
59. *Kotelnikov, V. A.*, Theory of Potential Noise-Stability (Theoriya Potentsialnoi Pomekhoustoichivosti), Gosenergoizdat, 1956 (Russian).
60. *Middleton, D.*, IRE Trans., **IT-3**, 26—51 (1954).
61. *Middleton, D.*, Journ. Appl. Phys., **24**, *4*, 371—391 (1953).
62. *Van Meter, D., Middleton, D.*, IRE Trans., **IT-4**, 119—445 (1954).
63. *Middleton, D., Van Meter, D.*, Journ. Soc. Indust. Appl. Math., **3**, 192—253 (1955); **4**, 86—119 (1956).
64. *Slepian, D.*, IRE Trans., **IT-3**, 68—89 (1954).
65. *Joula, D. C.*, IRE Trans. **IT-3**, 90—105 (1954).
66. *Lanning, J. H., Battin, R. H.*, Random Processes in Automatic Control, McGraw-Hill, 1956.
67. *Pugachev, V. S.*, Avtomatika i Telemekhanika, **19**, *6*, 519—539 (1958) (Russian).
68. *Pugachev, V. S.*, Izvestiya Akademii Nauk SSSR, Energetika i Avtomatika, **2**, 83—97 (1960) (Russian).
69. *Pugachev, V. S.*, IRE Trans., **IT-6**, *1*, 4—7 (1960).
70. *Pugachev, V. S.*, IRE Trans., **CT-7**, *4*, 491—505 (1960).
71. *Sysoev, L. P.*, in "Tekhnicheskaya kibernetika", "Science", Moscow, 1965, 47—55 (Russian).
72. *Sysoev, L. P.*, in „Avtomaticheskoye regulirovaniye in upravleniye", Akademiya Nauk SSSR, Moscow, 1961, 392—398 (Russian).
73. *Sysoev, L. P.*, Avtomatika i Telemekhanika, **26**, *10*, 1709—1719 (1965) (Russian).
74. *Sysoev, L. P.*, Trudy 3rd All-Union Conference on Automatic Control (Odessa, Sept.,
1965), Optimal Systems. Statistical Methods, "Nauka", Moscow, 1967, 247—255 (Russian).
75. *Pugachev, V. S.*, Avtomatika i Telemekhanika, *10*, 163—167 (1967) (Russian).
76. *Pugachev, V. S.*, Izvestiya Akademii Nauk SSSR, Energetika i Avtomatika, **5**, 123—135 (1961) (Russian).
77. *Pugachev, V. S.*, Stochastic Systems, *12*, "Science", Moscow, 1974 (Russian).
78. *Kichatov, Yu. F.*, Avtomatika i Telemekhanika, **26**, *8*, 1348—1356 (1965) (Russian).
79. *Popkov, Yu. S.*, Trudy 3rd All-Union Conference on Automatic Control (Odessa, Sept., 1965) Optimal Systems. Statistical Methods. "Nauka", Moscow, 1967, 382—396 (Russian).
80. *Popkov, Yu. S.*, Trudy 3rd International Congress of IFAC (London, June, 1966), Optimal Systems. Statistical Methods, 278—279, "Nauka", Moscow, 1971 (Russian).
81. *Popkov, Yu. S., Kiselev, O. N.*, Trudy 1st All-Union Symposium on Statistical Methods in Automatic Control (Moscow, Febr., 1967), 64—74, "Nauka", Moscow, 1971 (Russian).
82. *Stratonovich, R. L.*, Teoriya Veroyatnostei i eye Primeneniya, *2*, 172—195 (1960) (Russian).
83. *Stratonovich, R. L.*, Conditional Markov Processes and Their Application to Optimum Control Theory. Moscow State University Press, 1966 (Russian).
84. *Liptser, R. S.*, Teoriya Veroyatnostei i eye Primeneniya, **11**, *3*, 528—533 (1966); **12**, *4*, 764—765 (1967) (Russian).
85. *Liptser, R. S.*, Kibernetika, *3*, 63—70, *6*, 70—76 (1968) (Russian).
86. *Liptser, R. S.*, Avtomatika i Telemekhanika, *1*, 35—41 (1974) (Russian).
87. *Liptser, R. S.*, Problemy Peredachi Informatsii, *2*, 75—94 (1974) (Russian).
88. *Liptser, R. S., Shiryaev, A. N.*, Trudy Matematicheskogo Instituta im. V. A. Steklova Akademii Nauk SSSR, **104**, 135—180 (1968) (Russian).
89. *Liptser, R. S., Shiryaev, A. N.*, Teoriya Veroyatnostei i eye Primeneniya **13**, *1*, 17—38; *3*, 569—571; *4*, 602—620 (1968) (Russian).
90. *Liptser, R. S., Shiryaev, A. N.*, Proc. of 6th Berkeley Symp. on Math. Statistics and Probability (1970), Univ. of Calif. Press, II, 389—422 (1972).
91. *Liptser, R. S., Shiryaev, A. N.*, Statistics of Random Processes (non-linear filtering and related topics), Moscow, 1973 (Russian).

92. *Boguslavskii, I. A.*, "Stochastic Optimal Control Synthesis", Ch. 7 in „Modern Techniques of Design of Automatic Control Systems", "Mashinostroyeniye", 1967 (Russian).
93. *Bryson, A. E., Yu-Chi Ho*, Applied Optimal Control, Blaisdell Publ. Co., 1969.
94. *Moiseev, N. N.*, Optimization Techniques, Vychislitelnyi Centr Akademii Nauk SSSR, Moscow, 1969 (Russian).
95. *Åström, K. J.*, Introduction to Stochastic Control Theory, Acad. Press, 1970.
96. *Kushner, H. J.*, Stochastic Stability and Control, Acad. Press, 1967.
97. *Balakrishnan, A. V.*, Stochastic Differential Systems I. Lecture Notes in Economics and Math. Systems, 84, Springer-Verlag, 1973.
98. *Fabian, V., Špaček, A.*, Czechoslovak Math. Journ., **6** (1956) 190—194.
99. *Špaček, A.*, Trans. 1st Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague, 1957, 253—258.
100. *Driml, M., Špaček, A.*, Trans. 1st Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague, 1957, 43—60.
101. *Winkelbauer, K.*, Trans. 1st Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague, 1957, 297—354.
102. *Hanš, O.*, Trans. 1st Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague, 1957, 105—125.
103. *Prouza, L.*, Trans. 1st Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1956). Prague, 1957, 37—41.
104. *Driml, M., Hanš, O.*, Trans. 2nd Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague, 1960, 93—111, 113—122.
105. *Driml, M., Nedoma, J.*, Trans. 2nd Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague, 1960, 145—158.
106. *Hanš, O., Špaček, A.*, Trans. 2nd Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague, 1960, 203—213.
107. *Šefl, O.*, Trans. 2nd Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1959). Prague, 1960, 597—608.
108. *Pugachev, V. S.*, Automatic and Remote Control. Proc. 1st Intern. Congress of IFAC (Moscow, June, 1960) Butterworths, London, 1961, **2**, 702—706.
109. *Shaikin, M. E.*, Trudy 3rd All-Union Conference on Automatic Control (Odessa, Sept., 1965). Optimal Systems. Statistical Methods, Moscow, 1967, 262—270 (Russian).
110. *Pugachev, V. S.*, Trudy 3rd All-Union Conference on Automatic Control (Odessa, Sept., 1965), Self-Adjusting Systems. Pattern Recognition, Moscow, 1967, 192—202 (Russian).
111. *Pugachev, V. S.*, Preprints of papers of 3rd IFAC Congress (London, June, 1966), pp. 14A.1—14A.6. London, 1966.
112. *Pugachev, V. S.*, Doklady Akademii Nauk SSSR, **172**, *5*, 1039—1042, **175**, *5*, 1022—1025 (1967) (Russian).
113. *Pugachev, V. S.*, Izvestiya Akademii Nauk SSSR. Tekhnicheskaya Kibernetika, *6*, 26—42 (1967) (Russian).
114. *Pugachev, V. S.*, Kybernetika, **7**, *5*, 347—376 (1971).
115. *Robbins, H.*, Proc. 3rd Berkeley Symposium on Math. Statistics and Probability, 1, 157—164 (1955).
116. *Robbins, H.*, Review of the International Statistical Institute, **31**, *2*, 195—208 (1963).
117. *Robbins, H.*, Annals of Math. Statistics, **35**, *1*, 1—20 (1964).
118. *Johns, H. V.*, Annals of Math. Statistics, **28**, 649—669 (1957).
119. *Johns, H. V.*, Studies in Item Analysis and Prediction (ed. H. Solomon). Stanford University Press, 1961, 221—232.
120. *Sysoev, L. P.*, Avtomatika i Telemekhanika, *12*, 27—38 (1969); *11*, 19—28 (1970); *3*, 69—82 (1973) (Russian).
121. *Dobrovidov, A. V.*, Doklady 2nd All-Union Conference on Statistical Methods in Control Theory (Tashkent, Apr., 1971), Detection, Recognition, Planning of Experiments, "Nauka", Moscow (1970), 55—61 (Russian).
122. *Dobrovidov, A. V.*, Avtomatika i Telemekhanika, *2*, 88—99; *10*, 51—56 (1971) (Russian).

123. *Dobrovidov, A. V.*, Problems of Control and Information Theory, **1**, *2*, 163—176 (1972) (Russian).
124. *Fabian, V.*, Aplikace Matematiky, *6*, 162—183 (1961); *7*, 426—440 (1962) (Czech).
125. *Fabian, V.*, Trans. 3rd Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1962). Prague, 1964, 85—105.
126. *Šefl, O.*, Trans. 3rd Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1962). Prague, 1964, 611—620 (Russian).
127. *Gardner, L. A.*, Trans. 3rd Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (1962). Prague, 1964, 123—192.
128. *Cooper, D. V.*, IEEE Trans. **EC-13**, *3*, 306—307 (1964).
129. *Tsypkin, Ya. Z.*, Avtomatika i Telemekhanika, **27**, *1*, 23—61 (1966); *8*, 15—27 (1968); *1*, 42—50; *4*, 12—21 (1970) (Russian).
130. *Tsypkin, Ya. Z.*, Adaptation and Learning in Automatic Systems. Academic Press, 1971.
131. *Tsypkin, Ya. Z.*, Foundations of the Theory of Learning System. Academic Press, 1973.
132. *Csibi, S.*, Proc. of the Colloquium on Information Theory. Bolyai Math. Soc., Debrecen, 1967, 89—100.
133. *Csibi, S.*, Problems of Control and Information Theory, **1**, *1*, 37—50 (1972).
134. *Gulyás, O.*, Problems of Control and Information Theory, **1**, *1*, 51—64 (1972).
135. *Győrfi, L.*, Colloquia Series. Research Institute for Telecommunication, Budapest, 1970 (preprint, Hungarian).
136. *Győrfi, L.*, Problems of Control and Information Theory, **1**, *3—4*, 247—265 (1972); **3**, *1*, 11—17 (1974).
137. *Rejtő, L., Révész, P.*, Problems of Control and Information Theory, **2**, *1*, 67—80 (1973).
138. *Fritz, J.*, Problems of Control and Information Theory, **2**, *1*, 81—93 (1973).
139. *Pugachev, V. S.*, Doklady Akademii Nauk SSSR, **197**, *6*, 1288—1290 (1971) (Russian).
140. *Pugachev, V. S.*, Problems of Control and Information Theory, **1**, *1*, 65—75 (1972).
141. *Pugachev, V. S.*, Doklady Akademii Nauk SSSR, **208**, *3*, 566—569 (1973) (Russian).
142. *Pugachev, V. S.*, Avtomatika i Telemekhanika, *5*, 74—81 (1974) (Russian).
143. *Pugachev, V. S.*, Preprints of papers of 6th IFAC Congress (Boston, USA, August, 1975).
144. *Pugachev, V. S.*, Proc. of IFAC Stochastic Control Symposium (Budapest, Sept. 1974). General Lectures. Budapest, 1974.
145. *Balakrishnan, A. V.*, Tezisy Dokladov 3rd All-Union Conference on Statistical Methods in Control Processes (Vilnius, Sept., 1973), Pt. 1, Moscow 1973, 118—119.
146. *Eykhoff, P.*, System Identification. Parameter and State Estimation. J. Wiley, 1974.
147. *Raibman, N. S., Chadeev, V. M.*, What is Identification. Moscow, 1970. (Russian).
148. *Morris, W. T.*, Management Science. A Bayesian Introduction. Prentice-Hall, 1968.
149. *Emelyanov, S. V., Ozernoi, V. M., Larichev, O. I.*, Problems and Methods of Decision Making (a survey). International Center of Scientific and Technical Information, 1973 (Russian).
150. *Emelyanov, S. V., Ozernoi, V. M.*, Decision Making in Multi-Objective Problems: A Survey. This issue, 51—64.

V. S. Pugachev
Institute of Control Problems
Profsoyuznaya 81, V-485, Moscow,
USSR

Problems of Control and Information Theory, Jubilee Number, pp. 33—49 (1975)

THE DEVELOPMENT OF THE WORK CONNECTIONS BETWEEN THE SOVIET AND THE HUNGARIAN RESEARCH SCHOOLS IN THE FIELD OF CONTROL AND INFORMATION THEORY

F. CSÁKI
(Budapest)

(Received April 7, 1975)

1. The significance of cooperation

It is very difficult to tell, even if restricted to a narrow research field, how much the technical-scientific cooperation with the socialist countries and especially with the Soviet Union meant and means for us.

We call our age by right the age of the scientific-technical revolution, with new important scientific disciplines developing, with the scientific discoveries going over within a relatively short time into practice forming the indispensable basis of modern production processes.

Automation (open- and closed loop controls), measurement technique, information transfer and processing, computation sciences form an essential part of the scientific-technical revolution. Though some of their solutions were met with before, sporadically, the named disciplines and their engineering realizations, developed during the last quarter of a century and started with the introduction of transistors and followed by the integrated circuits, brought about revolutionary changes, which could not have been even followed up without the efficient technical-scientific cooperation with the socialist countries and mainly with the Soviet Union. The results of the technical-scientific cooperation permitted our country not to be merely a passive observer, but to become an active participant in the revolutionary changes brought about among others in the field of automation, information processing and digital computation.

2. Some results of the multiple socialist cooperation

The socialist countries may be proud by right of having managed to develop by their technical-scientific cooperation a modern, uniform computer system, (UCS) called ESR. The family of machines ranging from R-10 to R-60 has been designed, developed, realized and manufactured in series production with the participation of the Soviet Union, Poland, the German Democratic Republic, the Bulgarian People's Republic and the Hungarian People's Republic, whose experts continue its development and application promotion

in the framework of numerous consultative and work groups of the Intergovernmental Committee for Computation Technics (see Fig. 1).

Our age may be regarded rightly also the age of space research and space navigation. In these fields the Soviet Union and the United States play the leading role, but certain research tasks were made available for the scientists

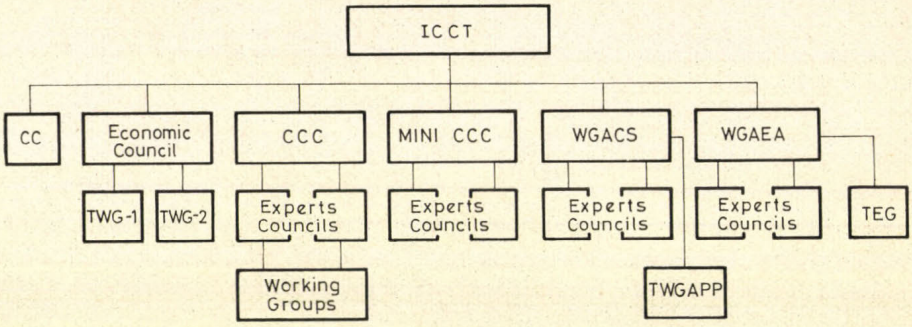


Fig. 1.

ICCT	— Intergovernmental Committee for Computation Technics
CC	— Coordination Centre
CCC	— Council of Chief Constructors
MINI CCC	— Council of Chief Constructors of Mini-Computers
WGACS	— Work Group of Automatic Control Systems
WGAEA	— Work Group of Automatized Engineering Activities
TWG	— Temporary Working Group
TEG	— Temporary Experts Group
TWGAPP	— Temporary Working Group for Application Program Packages

of the socialist countries as well by INTERCOSMOS. This union is directed by Academician B. N. Petrov, whose activity will be mentioned in the following also in other respects.

Several new forms of cooperation are practised by the CMEA member countries in the framework of the integration program. Undertakings of a high importance are the international economical associated enterprises. E.g. INTERATOMINSTRUMENT, founded by Bulgaria, Hungary, the German Democratic Republic, the Soviet Union and Czechoslovakia, carries out nuclear scientific research, experiments, designs in nuclear-technical activities, INTERETALONPRIBOR supplies special measuring means, and standards for the research and planning institutes of the socialist countries.

The experts of the Soviet Union and Hungary cooperate closely also in the framework of the international scientific organizations. With no claim to completeness we mention IFAC, the International Federation for Automatic Control, IMEKO, the International Congress for Measurement Technics, IFIP, the International Federation for Information Processing and the activities of the International Institute of System Theory in Laxenburg, Austria.

3. Some results of the bilateral technical-scientific cooperation

Concerning Hungary the bilateral technical-scientific cooperation developed most closely with the Soviet Union. In the following only a few important results of this form of cooperation bearing on automation, communication and metrological aspects are mentioned.

The Research Institute for Telecommunication developed in cooperation with the Moscow institute NIIR the microwave chain *Druzhba* covering two continents. Over the 8-track wide channel of the chain 1920 telephone lines can simultaneously be used, suitable for transmitting black-white and colour television, telex and picture telegraph signals. In the development works G. Bognár, Hungarian Academician and his coworkers played an important role.

With a similar pride and satisfaction we may mention that the development and the delivery of the automatic control systems serving the Soviet oil lines supplying the socialist countries have been realized with Hungarian cooperation. We have supplied also numerous fully automatized equipments for the food preserving industries. In the field of vehicle production we participate by the supply of instrument panels and car radio receivers produced in long series, in the production of the "*Zhiguli*" cars.

The Soviet Union took an important part in the design and the supply of the equipments of the Hungarian Danube Iron Works and Danube-shore Power Plant. Automation plays an important role in both industrial objects. We may mention similarly the Tisza-shore Chemical Combinate, the Alumina Factory of Almásfüzitő, the Aluminium Blast Furnace Plant of Ajka and the Danube Rock Oil Enterprise.

The National Bureau for Water Management develops automated irrigation systems in close cooperation with the Soviet counterpart Bureau, with the extensive utilization of computation technics.

The Hungarian lamp factory United Incandescent realized the fully automatic bulb production line with Soviet cooperation. In the construction of the METRO underground lines and the line automation we relied on Soviet experiences. Mention must be made also of the development of the black-white and coloured TV broadcasting stations, the high capacity radio transmitters and the erection of the terrestrial follow-up stations of the artificial earth satellites.

4. Cooperation in the research

The Soviet–Hungarian technical-scientific cooperation contributed to the development of the network of Hungarian research institutes. The Soviet and the Hungarian brother-institutes carry out common research in numerous

themes. So the Soviet IPU (Institute for Control Problems) and its predecessor IAT (Institute of Automation and Telemechanics) and the Hungarian SZTAKI (Computer and Automation Research Institute) and its predecessor AKI (Research Institute for Automation) have maintained connections for several years. The director of the institute IPU, Academician V. A. Trapeznikov is Honorary Member of the Technical Department VI of the Hungarian Academy of Sciences.

An especially close connection exists between the institutes IPPU (Institute of Information Transfer) and TÁKI (Research Institute for Telecommunication), both of which organize common seminars in alternate years in Moscow and Budapest, respectively. The director of IPPU, Academician V. I. Siforov is Honorary Member of the Technical Department VI of the Hungarian Academy of Sciences.

Academician M. V. Keldish, the president of the Academy of Sciences of the USSR at that time, Honorary Member of the Technical Department VI of the Hungarian Academy of Sciences visited some years ago the scientific institutions of our country. On this occasion Academician T. Erdey-Grúz, Honorary Member of the Academy of Sciences of the USSR suggested that the Soviet–Hungarian scientific cooperation could be made still closer by publishing common journals. Since that time three common journals were started. The first was PUTI (Problemy Upravleniya i Teorii Informatsii) — PCIT (Problems of Control and Informatiom Theory) in Russian and English, whose Soviet chief editor, Academician B. N. Petrov is Honorary Member of the Technical Department VI of the Hungarian Academy of Sciences. The journal presently appearing four times annually, entered the fourth year of its publication and soonly will appear six times a year (see the Editorial Board on the internal title page). As a complementary volume, the proceedings of the 2nd International Symposium on Information Theory, held in Tsahkadsor, was published in bookform in 1973. The sessions of the Editorial Board, held twice yearly in Moscow and in Budapest, respectively, serve as occasions for discussing actual scientific problems in connection with the evaluation of the received papers.

From among the multiplicity of the scientific connections let us mention only that numerous Hungarian research workers contributed papers to conferences, and symposia organized by the Academy of Sciences of the USSR and we have also opportunity to greet at our scientific events organized by the Hungarian Academy of Sciences numerous outstanding Soviet scientists (e.g. on the Vth Microwave Colloquium, or the festive session on the occasion of the jubilee of the Soviet–Hungarian scientific-technical cooperation).

We mention here that the activities of the Soviet research workers are well known in our country also by their published books and papers. Several

Soviet works dealing with our professional field were translated into Hungarian, while the works of the Hungarian scientists published in foreign languages are known in the Soviet Union and some of these were translated into Russian as well.

5. Cooperation in the field of education

On the basis of the Hungarian–Soviet technical-scientific agreement a great volume of high-level scientific and technical knowledge keeps flowing into our country, which we are able to return only in a lesser measure. In addition to handing over documentations, an important contribution is to be noted in the education of experts as well. A great number of Hungarian students are learning at Soviet universities and colleges, also in the fields of automation, communiation and digital computation. Many times also possibilities were made available for specific education, which we could have realized only with great difficulties, or not at all, by ourselves. The acquisition of the professional linguistic knowledge is also notable, as it promotes the possibility of future technical-scientific cooperation.

The mutual connections between the Soviet and the Hungarian universities are confirmed by numerous socialist cooperation contracts. We may emphasize especially the close contact of over a decade between the Moscow University for Energetics (Moskovskii Energeticheskii Institute, MEI) and the Technical University of Budapest permitting not alone the exchange of experiences in the field of teaching and education, but also personal connections by mutual visits to the professional counterparts, advancing the scientific cooperation greatly. Hungarian lecturers participate regularly at scientific sessions organized in Moscow, while Soviet lecturers contribute papers to scientific conferences held in Budapest.

The guidance of diploma planning students has been introduced and is becoming ever more frequent. The exchange of guest lecturers started as well.

In addition to educating engineers, the Soviet lecturers and research workers undertook a considerable role also in training Hungarian aspirants (post-graduate students) to scientific degrees and we may send candidates for doctorate as well. Some Soviet postgraduates are being trained in Hungary.

6. Contributions by Soviet research workers to the development of control theory

My narrower research field, control engineering and its basis, control theory look back only to a past of a few decades, but during this time it developed immensely and represents even today one of the most rapidly changing disciplines. The Soviet research scientists played a highly important role in

this fast development. E.g. the central theme of our days is optimum control, in which the maximum-principle developed by Academician Pontryagin (Honorary Member of Department III of the Hungarian Academy of Sciences) and his coworkers plays a decisive role. Part of the activities of the Soviet research workers relies on earlier results of Russian research. An example of this is the stability test method of nonlinear systems elaborated by the Russian Academician Lyapunov (in 1892) and perfected since considerably by Soviet scientists.

In the framework of a short paper it is impossible to give even a sketchy outline of the results attained by the Soviet research workers, therefore merely a few of the most important methods may be presented mosaic-like here.

6.1. *Some results involving linear systems*

One of the most important range of problems in linear systems of constant, concentrated parameters is stability. Mikhailov suggested in 1938 the following stability criterion. Let us start from the characteristic polynomial

$$K(s) = a_0 s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n; \quad (a_0 > 0).$$

Substituting the value $s = j\omega$ into this characteristic polynomial and varying the radian frequency ω between the limits $0 = \omega < \infty$, the system is stable if and only if the locus curve $K(j\omega)$ passes exactly n plane quadrants in the complex plane (that is, surrounding the origin just over the n quadrants). If $K(j\omega_c) = 0$, for a certain critical radian frequency ω_c, then with this radian frequency a remanent oscillation arises in the system.

Another important stability testing method is the so-called D-separation permitting to resolve not alone the stability, but also to determine the permissible range of some parameter, usually the loop gain.

This method was initiated by Neumark and propagated by Meierov. Let $D(\mu, \iota, \nu)$ denote a set of the coefficients, or the system parameters, to which number μ of negative real part, number ι imaginary and number ν of positive real part roots of the characteristic equation of n-order ($\mu + \iota + \nu = n$) belongs. For deciding upon stability it must be assessed, if a set $D(n, 0, 0)$ exists, and if so, which one is it, i.e. which coefficient, or to which parameter ranges it may be assigned.

The stability test is only simple, if a single coefficient, or parameter P may vary freely. In this case the characteristic equation

$$K(s) = 0$$

is written in the form

$$Q(s) = PR(s) = 0$$

or

$$P = -\frac{Q(s)}{R(s)}.$$

Although the parameter P may usually be only real, it is permitted to assume complex values as well. By substituting $s = j\omega$, we have:

$$P(j\omega) = -\frac{Q(j\omega)}{R(j\omega)}.$$

While the radian frequency varies between the limits $-\infty < \omega < \infty$, essentially the imaginary axis of the complex plane s is being transformed by conformal mapping on the curve $P(j\omega)$ of the complex plane $P(s)$. As on running over the imaginary axis in the direction of the increasing ω the left half plane s falls left from the axis, therefore the left side of the curve $P(j\omega)$, i.e. the innermost left side ranges formed by the curve $P(j\omega)$ may only claim the title of set $D(n, 0, 0)$. Whether the named ranges determine indeed the values ensuring the stability of the parameter P, or we have merely to do with the set $D(\mu, \iota, \nu)$ ensuring the highest number of left side roots μ, may be decided by assuming appropriate P values and by other tests.

From the stability range enveloped by the curve $P(j\omega)$ also the permissible maximum and minimum values of parameter P may be assessed.

Meierov studied in 1947 the structural stability of multiloop systems. He established e.g. in a system with rigid feedback and consisting of first order lag elements, whose transfer function is $K_i/(1 + sT_i)$; $(i = 1, 2, \ldots, n)$, that only the intermediate feedback of two first order lag elements may be realized by a stabilizing element of the transfer function $sT/(1 + sT)$, if the resultant gain of both involved elements is to be increased arbitrarily. Naturally other conditions must also be met with for ensuring the stability. Meierov studied several systems in detail and gave the stability conditions, as found in his book, which is translated also into Hungarian.

Tsypkin and Bromberg introduced in 1945 the conception of the degree of "stability". They established on the basis of the stability criterion of Routh–Hurwitz the condition, at which none of the left side roots of the characteristic equation $K(s) = 0$ approximates the imaginary axis by more, than the value δ, meaning that even the most slowly damped component tends to zero, as $e^{-\delta t}$, or in other words even the highest time constant of the system is lower, than $T = 1/\delta$.

The above range of problems leads already to the quality investigation of the linear systems. In this connection the trapezoid rule for calculating the unit step response from the frequency function given by Solodovnikov (published in 1945, 1948) is of a high importance. Let us approximate the real part Re $W(j\omega)$ of the frequency characteristic curve $W(j\omega)$ by trapezoids. Be the equation of the ith trapezoid approximation

$$\text{Re } W(j\omega) = \lambda_i; \quad 0 \leqq \omega \leqq \omega_i - \Delta_i,$$

$$\text{Re } W(j\omega) = \lambda_1\left[1 - \frac{\omega - \omega_i + \Delta_i}{2\Delta_i}\right], \quad \omega_i - \Delta_i \leqq \omega \leqq \omega_i + \Delta_i .$$

Then the ith component of the unit step response may be expressed in the following form:

$$v_i(t) = \frac{2}{\pi}\lambda_i\left\{\text{Si}(\varkappa t) + \frac{1}{1-\varkappa}\left[\text{Si}(t) - \text{Si}(\varkappa t) + \frac{\cos t - \cos \varkappa t}{t}\right]\right\}$$

where $\omega_i + \Delta_i = 1$ is assumed, $\varkappa = (\omega_i - \Delta_i)/(\omega_i + \Delta_i) = \omega_i = \Delta_i$ and $\text{Si}(\varkappa t)$ is the so-called integral sinus function:

$$\text{Si}(\varkappa t) = \int_0^{\varkappa} \frac{\sin \omega t}{\omega}\, d\omega .$$

Voronov (1952) uses instead of the trapezoid approximation a triangular approximation; in this case $\varkappa = 0$ and the ith component of the unit step response is

$$v_i(t) = \frac{2}{\pi}\lambda_i\left[\text{Si}(t) - \frac{1 - \cos t}{t}\right].$$

The quality test and the optimization is performed often on the basis of integral criteria of the form

$$I = \int_0^{\infty} F(x(t), t)\, \mathrm{d}t = \min.$$

Kharkevich, Feldbaum, Krasovskii suggested subsequently linear and general quadratic integral criteria weighted by time. Letov attained notable results in the analytical design of controllers based on integral criteria.

Recently the sampled-data, discrete-time and digital systems play an important role in control systems. In this respect especially Tsypkin's oeuvre is most notable, who suggested the introduction of a transformation similar to the z-transformation and reached numerous excellent results.

The statistical analysis and synthesis methods gain also ever more ground. The activities of Solodovnikov, Pugachev and Kazakov may be pointed out in this respect from among the great number of Soviet research workers. (Academician Pugachev is deputy editor of the common journal PCIT.) Raibman attained considerable results in the field of statistical identification.

A special Soviet school was formed for the elaboration of the so-called invariance principle. This method is used to study how certain disturbances could be eliminated totally from the control systems. In this field the works of Academician Kulebakin, Academician Petrov, Kukhtenko and Chinaev may be distinguished.

In connection with the varying-parameter systems Solodov's name have to be mentioned, while Butkovski and Egorov performed pioneering work in the field of distributed-parameter systems.

6.2. *Some results of the Lyapunov stability test method*

The most important problem of control systems is stability. The method of the Russian Academician Lyapunov developed at the end of the last century may be applied to nonlinear systems.

With the Lyapunov method the stability and the asymptotic stability, respectively, of a system is evaluated by constructing e.g. a positive definite Lyapunov function and on the basis of the negative semi-definiteness and negative definiteness, respectively, of the derived function.

Stability theorem. If for the nonlinear system of n-order written up in the form of the vector differential equation

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{x} = \mathbf{f}(\mathbf{x}), \quad \mathbf{f} = (f_1, \ldots, f_n)^{\mathbf{T}}, \quad \mathbf{x} = (x_1, \ldots, x_n)^{\mathbf{T}} \quad (\mathbf{f} = \mathbf{0}, \text{ when } \mathbf{x} = \mathbf{0})$$

a continuous and derivable definite function $V = V(\mathbf{x})$ may be selected in a way that its derivate with respect to the time

$$W(\mathbf{x}) = \frac{\mathrm{d}V(\mathbf{x})}{\mathrm{d}t} = \dot{V}(\mathbf{x}) = (\nabla_{\mathbf{x}} V)^{\mathbf{T}} \dot{\mathbf{x}} = (\nabla_{\mathbf{x}} V)^{\mathbf{T}} \mathbf{f}$$

where

$$\nabla_{\mathbf{x}} = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_m} \right)^{\mathbf{T}}$$

is also definite, but of the opposite sign, then the system is asymptotically stable in a small region around the origin. If the function $W(\mathbf{x})$ is only semi-

definite, then the system is stable, but not asymptotically stable (however — as Barbashin and Krasovskii demonstrated — if $W(\mathbf{x})$ is semi-definite and $\dot{V}(\mathbf{x}) = 0$ is no trajectory of the differential equation $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, then the solution is asymptotically stable). The global stability, or asymptotical stability means that the conditions are satisfied in the whole phase space.

Unstability theorem. If for the nonlinear system of n-order a Lyapunov function $V(\mathbf{x})$ may be selected, which is continuous and whose derivate with respect to the time $W(\mathbf{x})$ is negative definite, then

a) the system is unstable in the finite range, where $V(\mathbf{x})$ is non-positive semi-definite,

b) with the extension of time the deflection of the system increases beyond every limit, if $V(\mathbf{x})$ is globally non-positive semi-definite.

Numerous variations of the method are known for evaluating the stability in the small and in the large. The greatest problem is even the construction of the Lyapunov functions. Several procedures for autonomous systems are known in this respect. Relatively few theorems apply to non-autonomous systems. The literature of the Lyapunov method is very extensive. Some outstanding names in connection with the Lyapunov method are Aizerman, Barbashin, Chetaev, Duboshin, Erugin, Krasovskii, Letov, Lur'e, Malkin, Moiseev, Nemytskii, Persidskii, Razumikhin, Rumyantsev, Zubov. For linear autonomous systems, when $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, the selection of the Lyapunov function is relatively simple. In this case we select the positive definite quadratic form of

$$V(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{P}\mathbf{x}$$

as Lyapunov function, where $\mathbf{P}$ is the positive definite symmetric matrix. The derivate of $V(\mathbf{x})$ may be brought into the form of

$$W(\mathbf{x}) = -\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}$$

where

$$-\mathbf{Q} = \mathbf{A}^{\mathrm{T}}\mathbf{P} + \mathbf{P}\mathbf{A}$$

$W(\mathbf{x})$ is negative definite, if and only if $\mathbf{Q}$ is a positive definite symmetric matrix.

Perhaps the most general theorem for nonlinear $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ systems is the theorem of Krasovskii. Here the positive definite quadratic form

$$V(\mathbf{f}) = \mathbf{f}^{\mathrm{T}}\mathbf{B}\mathbf{f}$$

is selected as Lyapunov function, whose derivate is the quadratic form

$$W(\mathbf{f}) = -\mathbf{f}^{\mathrm{T}}\mathbf{C}\mathbf{f}$$

where

$$-\mathbf{C} = \mathbf{J}^{\mathrm{T}}\mathbf{B} + \mathbf{B}\mathbf{J},$$

furthermore $\mathbf{J} = \mathbf{f}(\nabla_{\mathbf{x}})^{\mathrm{T}}$ is the Jacobi matrix. Here

$$(\nabla_{\mathbf{x}})^{\mathrm{T}} = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n}\right).$$

The quadratic form $W(\mathbf{f})$ is negative definite if, and only if $\mathbf{C}$ is a positive definite matrix.

The method of Aizerman may also be applied often. If the differential equation of the nonlinear system is

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{g}(\mathbf{x})$$

and a $\mathbf{K}$ matrix may be found, where

$$\mathbf{g}(\mathbf{x}) \leqq \mathbf{K}\mathbf{x}$$

for each component, then the linear system

$$\mathbf{x} = (\mathbf{A} + \mathbf{K})\mathbf{x}$$

may be considered with the help of the method presented for linear autonomous systems.

In the stability test of nonlinear systems the first canonical form of Lur'e plays an important role. System equations of the form

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{b}g(\sigma)$$
$$\sigma = \mathbf{c}^{\mathrm{T}}\mathbf{x}$$

may be transformed by the coordinate transformation

$$\mathbf{x} = \mathbf{L}\mathbf{z}$$

in the case of distinct eigenvalues into the first canonical form of

$$\dot{\mathbf{z}} = \mathbf{S}\mathbf{z} + \mathbf{u}g(\sigma)$$
$$\sigma = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{z}.$$

Here the transformation matrix $\mathbf{L}$ is to be selected in a way to have on the one hand

$$\mathbf{L}^{-1}\mathbf{A}\mathbf{L} = \mathbf{S} = \mathbf{diag}\,[s_1, s_2, \ldots, s_n]$$

and on the other hand

$$\mathbf{L}^{-1}\mathbf{b} = \mathbf{u}; \qquad \mathbf{u} = (1, 1, \ldots, 1)^{\mathbf{T}}$$

where $s_1, s_2, \ldots, s_n$ are the distinct, nonzero roots of the characteristic equation $|\mathbf{A} - s\mathbf{I}| = 0$.

Finally

$$\boldsymbol{\alpha}^{\mathbf{T}} = \mathbf{c}^{\mathbf{T}}\mathbf{L}.$$

In this case the Lyapunov function may be selected (with neglecting a term, which may be rendered arbitrarily small) in the form of

$$\mathbf{V} = \mathbf{z}^{\mathbf{T}}\mathbf{P}\mathbf{z}$$

where the p_{ik} element of the matrix $\mathbf{P}$ is

$$p_{ik} = \frac{a_i a_k}{s_i + s_j}.$$

Here a_i, a_k $(i, k = 1, 2, \ldots, n)$ are at present undetermined coefficients. The derivate of the Lyapunov function may be brought, with due consideration to the canonical form, into the form

$$\dot{V} = \mathbf{z}^{\mathbf{T}}(\mathbf{S}^{\mathbf{T}}\mathbf{P} + \mathbf{P}\mathbf{S})\,\mathbf{z} - \sigma g(\sigma) + [\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{z} + 2\mathbf{u}^{\mathbf{T}}\mathbf{P}\mathbf{z}]\,g(\sigma).$$

For $\dot{V}$ to be negative definite, the multiplier factor of $g(\sigma)$, that is, the expression in the square brackets must be zero. From this we obtain the system of quadratic equations

$$\alpha_i - 2a_i \sum_{k=1}^{n} \frac{a_k}{s_i + s_k} = 0; \qquad (i = 1, 2, \ldots, n).$$

The sufficient condition of the global asymptotic stability is that the a_i solutions belonging to the real (complex) s_i roots are also real (complex). Lur'e and Letov derived several other, so-called secondary stability criteria as well, which are similar to the above.

6.3. *Results in the field of optimum control*

Pioneering work was done in the field of optimum control especially by Feldbaum, Pontryagin and his coworkers, Boltyanskii, Gamkrelidze and Mishenko.

The discussion of this topic in any detail would lead very far, so we merely point out here the formulation of the maximum principle of Pontryagin and the similarity of the Hamiltonian equations in mechanics.

Be the differential equation of the $(n+1)$-order of the dynamic system

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}, \mathbf{u})$$

where $\mathbf{f}$ and $\mathbf{x}$ are $(n+1)$-component, $\mathbf{u}$ is an r-component column vector.

Let us formulate the Hamiltonian function H:

$$H(\mathbf{p}, \mathbf{x}, \mathbf{u}) = \mathbf{p}^{\mathbf{T}}\mathbf{f}(\mathbf{x}, \mathbf{u})$$

where

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \nabla_{\mathbf{p}} H; \qquad \nabla_{\mathbf{p}} = \left(\frac{\partial}{\partial p_0}, \ldots, \frac{\partial}{\partial p_n}\right)^{\mathbf{T}}$$

$$\frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\nabla_{\mathbf{x}} H; \qquad \nabla_{\mathbf{x}} = \left(\frac{\partial}{\partial x_0}, \ldots, \frac{\partial}{\partial x_n}\right)^{\mathbf{T}}$$

where superscript $\mathbf{T}$ refers to the transposition.

Be $\mathbf{u}(t)$, $0 \leq t \leq T$ a permissible control vector ensuring the arrival of the trajectory $\mathbf{x}(t)$ from point $\mathbf{x}(0)$ into the point $\mathbf{x}(t)$. The necessary condition of $\mathbf{u}(t)$ and $\mathbf{x}(t)$ to be optimum is the existence of a non-zero vector function $\mathbf{p}(t)$ of $(n+1)$ continuous components, which is also a function of $\mathbf{u}(t)$ and $\mathbf{x}(t)$ and the Hamiltonian function $H(\mathbf{p}, \mathbf{x}, \mathbf{u})$ for

(i) should assume in each time t the maximum value $(0 \leq t \leq T)$ at $\mathbf{u} \in U$

$$H[\mathbf{p}(t), \mathbf{x}(t), \mathbf{u}(t)] = M[\mathbf{p}(t), \mathbf{x}(t)]; \text{ and}$$

(ii) in the final instant T

$$p_0(T) \leq 0; \quad M[\mathbf{p}(T), \mathbf{x}(T)] \geq 0$$

should be satisfied (generally condition (ii) is satisfied not alone for the instant T, but for any $0 \leq t \leq T$ instant as well).

Remark. In the above vector equations the optimization condition

$$J = \int f_0[\mathbf{x}(t), \mathbf{u}(t)]\,\mathrm{d}t = x_0 = \min$$

appears in the form:

$$\frac{\mathrm{d}x_0}{\mathrm{d}t} = f_0[x_1, \ldots, x_n, \mathbf{u}(t)].$$

E.g. if $f_0 \equiv 1$, then the shortest conversion period is to be found

$$J = T = \min.$$

The Pontryagin principle gives a very general formulation of the optimum problems in control systems, involving possibly not alone systems of optimum transition periods, but also other optimum systems minimizing some integral criterion, or any other systems.

The Pontryagin principle is a relatively new principle, but the number of scientific publications related to it amounts already to several hundreds, dealing mainly with the study of optimum time systems and only in a lesser measure with other optimum systems. The greatest problem seems to be how far the so-called switching hyper-surface can be also given in a closed form, when the expression of the function is relatively complex, rendering the physical realizability questionable. Part of the studies sets the aim of approximating these switching hyper-surfaces to practically simpler realizable planes, or surfaces.

Numerous papers were published in the literature pointing out the relationships between the maximum principle, the classical calculus of variation and dynamic programming (in this respect the series of articles by Rosenauer deserves to be emphasized). The maximum principle and the dynamic programming form the basis of one of the most important chapters in control systems and we may count rightfully on the appearance of many more publications and papers in the years to come, related to this range of problems.

6.4. *A few other results*

In nonlinear control systems several methods were developed showing a certain affinity to nonlinear oscillation theory, but attaining their full development specifically in control systems. Such is e.g. the harmonic linearization method, initiated by Krylov and Bogolyubov. Harmonic linearization, known also under the name of the "method of the describing functions" assumes sinusoid oscillations in the input of the nonlinear element and takes only fundamental on the output of the nonlinear element, neglecting the upper harmonics. As the other linear parts of the control system are generally of a low-pass filter character, this approximation leads often to very good results, especially in the evaluation of the stability, or the limit cycle, respectively. The spreading of harmonic linearization is due to the fact that it permits essentially the calculation with the equivalent linear system, i.e. the application of certain well-known methods of the linear control systems to nonlinear systems.

Be the function of the nonlinear element $y = f(u)$, where y is the output signal, u the input signal. If the latter is a harmonic signal:

$$u(t) = B \sin \omega t,$$

then the output signal is

$$y(t) = A_0 + \sum_{n=1}^{\infty} (A_n \cos n\omega t + B_n \sin \omega t)$$

where A_0; A_n, B_n $(n = 1, 2, \ldots)$ are the coefficients of the Fourier series. In the harmonic linearization only the constant coefficient and the fundamental is taken into account. So the output signal is approximated by the relationship

$$y(t) \approx A_0 + q(B)\, B \sin \omega t + q'(B)\, B \cos \omega t.$$

Here

$$A_0 = \frac{1}{2\pi} \int_0^{2\pi} f(B \sin \omega t)\, \mathrm{d}\omega t\,,$$

$$q(B) = \frac{1}{\pi B} \int_0^{2\pi} f(B \sin \omega t) \sin \omega t\, \mathrm{d}\omega t\,,$$

$$q'(B) = \frac{1}{\pi B} \int_0^{2\pi} f(B \sin \omega t) \cos \omega t\, \mathrm{d}\omega t\,.$$

The describing function itself is:

$$N(B) = q(B) + jq'(B).$$

The literature dealing with the method of the describing functions is very extensive and even in recent years there appear new and new papers discussing individual variations of the method (see e.g. the book of Popov and Paltov).

A widespread method of the linearization in the time domain similarly to the harmonic linearization is the so-called statistical linearization. With this method a statistical input signal is assumed, regarded generally to be of the normal Gaussian distribution. Also the output signal is of normal distribution and on the basis of both signals e.g. an equivalent linear gain may be assigned to the nonlinear element. Naturally other characteristics may additionally be defined as well, but most frequently the linearized gain is used.

Be (in the simplest case) the function $y = f(u)$ of the nonlinear element an odd equivalent function and the input signal $u(t)$ a stationary stochastic signal with zero expected value:

$$\mathbf{M}[u(t)] = 0.$$

The real stochastic signal $y(t)$ may be approximated by an ideal signal $i(t)$ rendering the mean square of the difference of both signals minimum:

$$\mathbf{M}\{[y(t) - i(t)]^2\} = \min.$$

According to the method of the statistical linearization

$$i(t) = K_S u(t),$$

and the equivalent gain is

$$K_S = \frac{\int_{-\infty}^{\infty} uf(u)\, p(u)\, \mathrm{d}u}{\int_{-\infty}^{\infty} u^2 p(u)\, \mathrm{d}u}$$

where in the case of a stationary process the probability density function $p(u)$ of the input signal $u(t)$ is time-independent. In connection with the statistical linearization the name of Kazakov must be mentioned in the first place.

Another range of themes in the field of nonlinear control systems developing greatly is extremum control, or optimum search method, as it is also called. Here essentially the finding of the extremum of a nonlinear single-valued function is involved. The tasks may be classified in two groups: on the one hand to various methods for determining the gradients and on the other hand to steps towards the extremum value. These latter tasks are often called also "hill climbing method". The most frequently used processes are the Gauss–Seidel method, the gradient method and the method of the steepest ascent or descent. This range of problems has an ample literature extending not alone to theoretical considerations, but also to the possibilities of practical realization. In this connection we refer to the work of Feldman, Krasovskii and others.

The methods of optimum search form a subgroup of the adaptive systems of varying structures. This latter range of problems is in the centre of the attention of the control engineers, as shown also by the great number of publications related to the adaptive systems. From among the Soviet scientists Tsypkin attained excellent results, especially in the field of the stochastic approximation. Academician Petrov and his school attained remarkable results in the field of the adaptive control of flying bodies.

Concerning the systems of varying structure the activities of Emelyanov are outstanding.

An extensive research work is pursued also in the field of computer-aided design and computerized production control. In the former field Academician Gavrilov, in the latter the name of Emelianov is to be mentioned.

The recently deceased Academican Sotskov deserved undying merits in the elaboration of apparatuses based on new principles in physics.

7. Conclusions

The preceding could be by necessity merely a schematic survey of some of the more important results of socialist and specifically of the Soviet–Hungarian technical-scientific cooperation, the forms of the various level connections and some methods in individual fields of automation, instrument technics, communication and computation sciences.

In the age of the technical-scientific revolution, where the time factor is of decisive importance, the perfection of the technical-scientific cooperation, the development of new forms and modes of the same is an indispensable necessity.

On the basis of the experiences gained in cooperation over 25 years it may be rightfully expected that the Soviet–Hungarian technical and scientific cooperation will develop further, the connections become ever closer, the exchange of experiences and the common research of the Soviet and the Hungarian research workers will promote the growth of the scientific, technical and economical potential of our countries.

We believe that the Academies of Sciences of the Soviet Union and Hungary may play an important role in developing the technical-scientific cooperation further and in creating closer and still more active mutual connections.

F. Csáki
Department of Automation
Technical University of Budapest
H–1111 Budapest XI., Goldmann Gy. tér 3.
Hungary

Problems of Control and Information Theory, Jubilee Number, pp. 51–64 (1975)

DECISION-MAKING IN MULTI-OBJECTIVE PROBLEMS: A SURVEY

S. V. EMELYANOV, V. M. OZERNOI
(Moscow)

(Received February 25, 1975)

The state-of-art and major problems of the multi-objective decision-making theory and methodology are discussed. A classification is proposed of multi-objective problems and methods of their solution. Major difficulties involved into justification of decision methods for multi-objective problems are treated. Various approaches to construction of decision rules, their advantages and disadvantages are analyzed. Consideration is given to new approaches to the development of decision procedures for multi-objective problems.

Introduction

It is a well-established fact now that for the majority of practically significant decision problems the quality of decision cannot be estimated by a single, scalar-valued parameter. In such situations one has to estimate and compare alternative courses of action with respect to quite a few criteria, i.e. one meets with multi-objective decision problems (decision-making under a vector-valued criterion) [1—6].

Necessity to solve complicated practical multi-objective problems and interest in the theoretical aspects involved into them led to extensive development of studies in this field. Methods were proposed to solve particular problems based on various assumptions and various kinds of information; comprehensive mathematical literature also exists dealing with multi-objective problems.

However, in spite of the great number of publications on multi-objective problems, presently there is no formalized decision method which could be applied to a wide variety of practical tasks. Major difficulties involved into development of such methods are due to the fact that multi-objective decision problems are ill-structured [7]; and to formalize them successfully, one has to allow for many factors from the measurement theory, sociology, psychology and other disciplines. If one neglects these factors in formalizing decision problems, this would inevitably say on the decision-makers' attitude towards recommendations given on the basis of these formalized methods, thus devaluating them.

A great number of works was published recently discussing various theoretical, methodological and practical aspects of decision-making in multi-objective problems. The volume of publications grows dramatically, the percentage of theoretical works is much greater than that of announcements about successful applications of the proposed methods. Many publications on solution of multi-objective problems do not indicate appropriate conditions and areas of application. Still more and more new methods are proposed dealing with particular and specific situations and, thus, making it impossible to draw some conclusions and make generalizations characterizing problems/classes of problems under consideration.

An essential part of publications on multi-objective problems is discussed in surveys [4—6, 8—10]. The papers analyzed, as well as the surveys themselves, however, do not give a well-defined test for complete and consistent classification of decision methods for multi-objective problems. The present paper attempts

— to propose such a classification;
— to discuss within its framework some typical methods;
— to estimate their possibilities, advantages and disadvantages; and
— to contemplate possible ways to improve them.

Classification of decision methods for multi-objective problems

The multi-objective decision problems considered in the literature belong usually to one of the following three categories:

1) identification of the set of non-inferior elements (Pareto set);
2) identification of the most preferable alternatives;
3) ordering of the feasible alternative set in the decreasing order of preferences.

The basic elements of multi-objective problems which in the majority of publications are taken into consideration to some degree of detailing and justification are as follows:

— set of feasible alternatives;
— set of criteria;
— preference judgements of the decision-maker; and
— decision rule.

In accordance to the above, the multi-objective decision problems may be classified as follows:*

* Our references do not comprise all known publications, but are of rather illustrative nature.

— problems with a prescribed set of feasible alternatives [11] and with feasible alternatives generated during decision-making [12];

— with finite [13] or infinite [14] set of feasible alternatives;

— with prescribed set of criteria [15] and with criteria generated during decision-making [16];

— with analytical [17] and non-analytical [18] criteria;

— with decision-maker's preference judgements expressed as numerical estimates [19] or in some other form [20]; and

— with a given decision rule [21] and with that constructed during decision-making [22].

A very great diversity of multi-objective problems stems from the possibility of having various combinations of classifications. The theory and approaches to problems of various kinds are elaborated to different degrees.

It is reasonable, in analyzing the present status and major problems involved into the development of the decision theory and methods for multi-objective problems, to use as basic attributes for classification the principles and procedures underlying decision rule construction. With this approach, classification will be carried out with respect to:

1) *construction principles:*

— axiomatic principle of decision-rule construction where either decision rules, or some assumptions enabling their construction are postulated [23];

— heuristic principle of construction based on non-formal considerations and peculiarities of particular problems, without any explicit axiomatics [24];

2) *construction procedures:*

— one-step methods built around single application of a postulated or constructed decision rule [15];

— multi-step methods based on multiple application of a decision rule(s) generated as problem is being solved [22].

One-step methods

Among multi-objective decision problems discussed most frequently there is that of identifying a set of non-inferior alternatives (Pareto set) [15], first formulated in economics [25] for comparison of preferences given to various sets of goods.

The problem of identifying the non-inferior alternative set was posed in the operation research theory [1] and in the optimal control theory [3] owing

to the emergence of multi-objective problems and the criticism levelled against concepts based on scalar-valued criteria.

In identifying a non-inferior alternative set, it is usually assumed that alternatives is characterized by n-dimensional vector $x = (x_1, x_2, \ldots, x_n)$ whose real components are values of n parameters. Alternative ϱ is estimated by m-dimensional vector

$$f(x) = (f_1(x), f_2(x), \ldots, f_m(x)),$$

where f_i, $i = 1,2, \ldots, m$, are given real functions of x. Alternative $x_0 \in Y$ (where Y is subset of m-dimensional Euclidean space R^m) is non-inferior with respect to $f(x)$ if and only if there is no alternative $\bar{x}$ in Y such that $f_i(\bar{x}) \geq f_i(x^0)$, $i = 1, 2, \ldots, m$, at least one of the inequalities being strict, i.e. there is no alternative more prefereable than x^0. Other terms standing for "non-inferior" are "non-improvable" [17] and "effective" [26].

The notion of optimality under vector-valued criterion is introduced in [3]: alternative $x^0 \in Y$ is referred to as optimal in Y with respect to $f(x)$ if $f_i(x^0) \geq f_i(x)$, $i = 1, 2, \ldots, m$ for all $x \in Y$. It follows from the optimality of x^0 that it is non-inferior. The reverse statement, generally speaking, is not valid. However, if $f(x)$ is a scalar function, these notions coincide.

Zadeh [3] indicates that the set of all non-inferior points lying in the boundary of Y may be obtained by determining optimal points in Y under weighted scalar-valued criterion

$$f(x, \alpha) = \sum_{i=1}^{m} \alpha_i f_i,$$

where α_i are non-negative real numbers.

It is shown in [26], however, that coefficients α_i should be strictly positive; otherwise, "false" non-inferior points might be obtained.

The problem of finding all non-inferior (effective) alternatives was termed the vector maximization problem [27]. Some of the papers regard vector maximization (or optimization) as implying also separation of the most preferable alternative from a set of non-inferior ones [15, 17]. Vector optimization has the most natural mathematical formulation, in contrast to other multi-objective problems necessarily requiring allowance for non-formal, meaningful considerations.

The vector maximization problem is formulated in [27] as follows: vector function $\mathcal{Y}(x)$ over R^n is considered under restrictions $\mathcal{F}(x) \geq 0$, $x \in Y$ where Y is an isolated and convex set. Components of functions $\mathcal{Y}(x)$ and

$\mathcal{F}(x)$ are assumed to be concave functions of x. Vector x^0 satisfying restrictions (feasible vector) is an effective decision if there is no other feasible vector x for which $\mathcal{Y}(x) \geq \mathcal{Y}(x^0)$. The well-known Karlin lemma [27] establishes relation between the vector maximization problem and the parametric one: if x^0 is a non-inferior (effective) point, there exists vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ having components $\alpha_i \geq 0$ and satisfying equality $\sum^m \alpha_i = 1$ such that over the set of all x satisfying restrictions $\mathcal{F}(x) \geq 0$, $x \in Y$, function $g(x) = (\alpha, \mathcal{Y}(x))$ has maximum in $x = x^0$. Necessary and sufficient existence conditions for non-inferior points are formulated in a theorem based on this lemma. As it was shown in [28], $x^0 \in Y$ is an effective solution of the linear vector maximization problem if and only if x^0 is the maximal solution of a corresponding linear parametric problem for some vector parameter $t > 0$. By means of this theorem, the linear vector optimization problem is reduced to a parametric problem of linear programming. This enables the study of the complete set of non-inferior alternatives and characterization of each of them by a corresponding parameter vector.

Identification of the non-inferior alternative set is usually regarded as the first step towards identification of the most preferable alternative [3, 15, 21]. Zadeh [3] draws attention to the fact that this might be done either by introducing a scalar-valued criterion over the set of non-inferior alternatives, and not necessarily over Y (such a criterion is used, e.g., in [29]), or by the random choice. However, if we decide upon the random choice of the most preferable alternative that is based on the assumption of the decision-maker's indifference to the choice between non-inferior alternatives, this would imply putting the end to any further analysis of the problem. Actually, in situations where additional information about non-inferior alternatives is lacking, it is more reasonable to entrust the decision-maker with the final choice of the most preferable one. For the two-criteria case, [8] recommends to plot curves enabling decision-maker to make the final choice.

It ought to be noted here that identification of the non-inferior alternative set makes it possible to identify the most preferable alternative, but does not enable linear ordering of the set of feasible alternatives. With one-step methods, ordering of feasible alternative set is based, as a rule, on introducing global criteria defining perfect-order relation over this set. Such criteria enable also identification of the most preferable alternative but information used to this end may turn out to be redundant.

Both in problems of identification of the most preferable course of action and in those of linear ordering of feasible alternative set, global criteria have the form of functions of estimates with respect to the scales of all criteria

("criteria convolution" [30, 31]):

$$u(x) = u(f_1(x), f_2(x), \ldots, f_m(x)) . \tag{1}$$

The most exhaustive list of all global criteria ("compromise schemes") is contained in [32].

The following function is a typical example of global criterion:

$$\sum_{j=1}^{m} [\alpha_j(x_j^* - x_{ij})]^S \tag{2}$$

where x_j^* is "ideal" [15], "optimal" [14] or "required" [33] estimate with respect to the scale of jth criterion;

x_{ij} is normalized estimate of ith alternative with respect to the scale of jth criterion;

α_j is relative importance factor ("weight") of jth criterion; and

S is an index enabling generation of different decision rules.

It is assumed that by taking corresponding values of α_j and S one can obtain different decision rules and choose from them those corresponding to a problem at hand. In fact, the choice and justification of these values are, in the majority of problems, as difficult for the decision-maker as the direct choice of the most preferable alternative. In a number of papers [15, 33], a situation is discussed where the decision-maker has to choose from a set of decision rules one which best corresponds to his understanding of the problem. In doing so, "arbitrary rule is transferred from one instance to another", as in [34] is noted.

With application of a global criterion, one meets with the problem of choosing a principle for reducing scales of all criteria to a single scale ("normalization of criteria") and with the problem of choosing a way to allow for their relative importance. Borisov [15] gives some recommendations contributing to the solution of these problems. However, the methodological difficulties involved into the application of global criteria are exteremely great.

Another group of global-criteria-oriented methods is based on the utility theory [23, 35, 36]. This theory has shown that if a decision-maker acts on the basis of certain assumptions, over the set of feasible alternative outcomes a utility function may be constructed mapping this set in that of real numbers. Stated differently, the utility function is a way to consistent description of decision-maker's preferences by means of the real number scale, provided that his preference judgements are consistent with the assumptions of the utility theory.

The utility theory permits to represent the global criterion in the following form:

$$u(x) = u(u_1[f_1(x)], u_2[f_2(x)], \ldots, u_m[f_m(x)]), \tag{3}$$

where u is a utility function over the vector estimate set, and $u_1, u_2, \ldots, u_m$ are utility functions of criteria $f_1, f_2, \ldots, f_m$, respectively.

The most important notions of the utility theory as applied to multi-objective decision-making are those of criterion independence and additivity [37–44]. The criterion independence requirement boils down to the fact that estimates with respect to the scale of each criterion may be ordered by the decision-maker independently of the estimates with respect to the scales of other criteria. Hence, if for $m = 2$ decision-maker prefers estimate (x_1, x_2) to (x_1, y_2), the independence condition requires that estimate (y_1, x_2) should be preferred to (y_1, y_2) for any y_1.

Fishburn [37] gives definitions of criterion independence and utility function additivity based on the decision-maker's indifference judgements about the choice between specially composed lotteries. Some tests for additivity are given in [37–40] and generalized in [41].

The following additive function

$$u(x) = \sum_{i=1}^{n} k_i u_i[f_i(x)] \tag{4}$$

is the simplest form of the utility function where k_i are factors of criterion relative importance.

For cases where the assumption about utility function additivity does not hold, a method for comparison of alternatives using quasi-separable (quasi-additive) utility functions is proposed [42]. For the two-dimensional case, quasi-additive utility function $u(x_1, x_2)$ is proved to have, under a number of assumptions, the following form

$$u(x_1, x_2) = u_1(x_1) + u_2(x_2) + k u_1(x_1) u_2(x_2) \tag{5}$$

where constant k should be determined empirically. Existence assumptions for this utility function are less restrictive than those for additivity. The additive utility functions may be easily shown to be a special case of the quasi-additive function.

Multiplicative utility functions are discussed in [43, 44]. The latter paper points out that in a number of particular problems multiplicative utility functions enable approximation of experts' utility estimates with accuracy close to that of additive functions.

Sometimes, multi-objective decision-making reduces to linear [14], quadratic [45], discrete [46] or geometric [47] programming. Complexity of the mathematical apparatus used in these problems does not reduce difficulties involved into the substantiation of the global criterion, choice of criterion relative importance factors, etc. For example, in [14] the squared distance between an arbitrary point characterizing an estimated alternative and some "optimal" point (quadratic function) is taken as the global criterion. The problem boils down to finding an alternative minimizing this distance. The global criterion in [14] is chosen, however, for mathematical convenience of studying this problem and has no evident meaningful justification.

Multi-step methods

The disadvantages of one-step decision-making methods led to the development of new approaches based on more detailed analysis of multi-objective problems. The most important element of such an analysis is gathering and use of additional information required to substantiate decision rules coordinated with decision-maker's preference judgements and assumptions made. With one-step methods, such an analysis is carried out only partially, if any.

Presently, the multi-step decision-making methods are weighted mostly towards combining mathematical models with information obtained from the decision-maker. Such an approach enables one to develop iterative procedures using obtained information for construction and further updating of the model.

One of the first and most interesting approaches in the field of multi-step methods is the successive concession method suggested in [1] and developed in [48–50].

With application of the successive concession method, all criteria $f_1, f_2, \ldots, f_m$ are placed in decreasing order of importance. For simplicity, it is assumed that estimates with respect to all criteria should be made maximal.

The search algorithm for finding the most preferable alternative is as follows. First, alternative x_1, optimal with respect to criterion f_1, is looked for. Next, a "concession" Δf_1 is made defining losses tolerable with respect to this criterion and leading to a better result with respect to the scale of the second important criterion f_2. Then alternative x_2 is found, optimal with respect to criterion f_2 under additional restriction

$$f_1(x_2) \geq f_1^* - \Delta f_1,$$

where f_1^* is the most preferable (maximal) estimate with respect to criterion f_1 scale.

Next, "concession" Δf_2 is made defining losses with respect to the criterion f_2 scale that may be allowed in order to obtain the most preferable estimate with respect to criterion f_3 scale, etc.

This method, thus, reduces to a sequence of m problems:

(1) find $\max f_1(x) = f_1^*$, $x \in Y$;
(2) find $\max f_2(x) = f_2^*$ under conditions $x \in Y$, $f_1(x) \geq f_1^* - \Delta_1$;

. . . . . .

(m) find $\max f_m(x) = f_m^*$ under conditions $x \in Y$, $f_i(x) \geq f_i^* - \Delta_i$, $i = 1, 2, \ldots, m-1$

where Δ_i is the magnitude of allowable abatement (concession) with respect to criterion f_i scale, $\Delta_i \geq 0$.

As it is pointed out in [49], any solution of problem (m) is not effective, but at least one of them is. In this case, for any effective solution x^0 there are concessions Δ_i^0 such that x^0 is the unique solution of problem (m).

Another direction of studies in the field of multi-step decision-making methods is related to the method ELECTRE [51] and its modifications [9] enabling one to identify various non-inferior alternative sets on the basis of some coefficients ("concordance index", "discordance index", and a parameter characterizing the ordinal number of "discordance interval"). Correctness of ELECTRE is called in question by its numerous drawbacks (possibility of losing transitivity, requirement to choose and justify three parameters, use of criterion weights, etc.). However, underlying this method, idea of successive changes in the non-inferior alternative set deserves further improvement and development.

The ideas of the adaptation theory to the solution of multi-objective problems which are regarded as those of optimization under a special type of uncertainty, i.e. uncertainty in choosing the optimality criterion are attempted to be applied [52–55]. This uncertainty may be eliminated in two ways. On the one hand, one may carry out experiments giving information sufficient to overcome this uncertainty.

In [53, 54] such experiments involve checking for certain conditions which may be satisfied in a number of particular multi-objective problems. For cases where these conditions are met, theorems are formulated defining the form of the optimality criterion.

However, one cannot accumulate this information until it is enough to eliminate the initial uncertainty, but can change or correct decision successively in accordance with this information as it comes. In doing so, its form, means of processing and use are defined by the structure of employed adaptive

algorithm. A general scheme for forming such algorithms is given in [52], and in [55] is shown how these alogirthms may be applied to the solution of a multi-objective problem formulated as that of learning.

Computers make it possible to construct iterative dialogue (man-computer) procedures for multi-step problems [56]. A typical man-computer procedure for solution of multi-objective problem [57] is as follows:

$$\max_{x \in X} u(f(x))$$

where f is an r-dimensional vector-valued criterion function, x is an n-dimensional vector of variables, $X \subset R^n$ is a domain of allowable values of x, and u is a utility function defined over the domain of values of f. u and each component of f are assumed to be convex and differentiable, with u increasing with respect to each f_i; and set X is assumed to be convex and compact. The iterative dialogue procedure for problem solution bears resemblance to the procedure for solving nonlinear programming problems. Two problems are solved at each step: choice of next step direction and determination of its length [56]. In order to choose direction for kth step, $r - 1$ questions are put to the decision-maker: "To what decrease of Δf_i^k with respect to criterion i would you agree to get increase Δf_1^k, with respect to criterion 1, provided that estimates with respect to the scales of the balance of criteria $f_j^k, j \neq 1, i$, are not changed?" Values of Δf_i^k, $i = 2, \ldots, r$ determined through answers allow to make approximate computation of coordinates

$$w_i^k = \frac{\partial u / \partial f_i \,|\, f = f(x^k)}{\partial u / \partial f_1 \,|\, f = f(x^k)} \approx \frac{\Delta f_{i,1}^k}{\Delta f_i^k}, \qquad i = 1, 2, \ldots, r \qquad (6)$$

of a vector collinear to the vector of gradient $\nabla_f u(f(x^k))$. This enables solution of the direction problem. The step length choice is the problem of optimization on the straight line which is solved in collaboration with the decision-maker.

At each step the following problem is solved:

$$\max_{y \in X} W^{k\mathrm{T}} f'(k^k)\, y \qquad (7)$$

where w^k is obtained from (6), and $f'(x^k)$ is value of the Jacobi matrix $r \times n$ in point x^k. Direction of the next step is determined by vector $d^k = y^k - x^k$ where y^k is the solution of (7). It set X is given by linear restrictions, (7) is a linear programming problem.

To obtain w_i^k, $i \neq 1$, computer prints out vector f^k denoted as $\mathcal{A}$ and vector $(f_1^k + \Delta f_1^k, f_2, \ldots, f_i^k - \Delta f_r^k, \ldots, f_r^k)$ denoted as $\mathcal{B}$.

If $\mathcal{B}$ is more preferable to the decision-maker than $\mathcal{A}$, Δf_i^k is doubled and the question is repeated until the decision-maker prefers $\mathcal{A}$. Further search

of the value of Δf_i^k for which $\mathcal{A}$ and $\mathcal{B}$ become indifferent to the decision-maker is done by dichotomy. Next, $w_i^k \approx \frac{\Delta f_1^k}{\Delta f_i^k}$ is computed and the procedure is repeated for all $i = 2, \ldots, r$. The step length is chosen in the following way. Computer prints out a number of points lying in a segment of the straight line connecting points x^k and y^k. The decision-maker picks the most preferable point, thus defining the step length and the initial point for the next step.

Conclusion

The major difficulties involved into the justification of decision-making methods for multi-objective problems are due to the necessity of solving in each particular case the following set of problems:

- identification of the problem type;
- generation of a set of feasible alternatives;
- generation of a set of criteria characterizing each alternatives;
- development of criterion scales;
- determination of estimates of alternatives with respect to criterion scales;
- identification of decision-maker's preference judgements;
- generation of a decision rule; and
- ordering of feasible alternatives in terms of the decision rule.

Analysis of the literature shows that none of the publications treats exhaustively all of the above-mentioned problems. The emphasis is made on the decision rule generation and on the use of the rule for obtaining the ordering required by a problem. In doing so, in one-step methods one generally postulates the decision rule as a global criterion, and decision-maker's preference judgements are expressed as some factors ("criterion weights"). The possibility of justified choice of these factors, however, is doubtful, thus not allowing in particular problems to establish correspondence of global criteria to decision-makers' preference judgements. Global criteria enable linear ordering of the feasible alternative set; in problems where linear ordering is not required, this leads to redundant (and often unreliable) information concerning decision-maker's preferences. At the same time, some types of multi-objective problems (e.g. identification of a given number of non-inferior alternatives) cannot, in the general case, be solved by one-step methods.

Mathematical decision-making methods enable combination of mathematical model study with revision of decision-makers' preference judgements. Presently, however, solely those procedures are most widespread where the decision-maker answers questions aimed at elucidating preferences of a certain

type. This approach is related to the use of a certain type of decision rule and to successive revision of the model parameters; on the other hand, this does not exclude discrepancies in information related to preference judgements. Development of methods involving construction of a system (sequence) of coordinated models appears to be more promising.

The quest for use of formalized methods and procedures in multi-objective decision-making requires development of involved models which would merge capacity to solve complicated problems by means of formalized methods and computers with the contribution of decision-maker's experience, intuition and common sense.

There does not exist the "only correct" model for each multi-objective problem whose results should be accepted and used. Since every practical problem may be characterized by a great number of criteria, the same situation may be represented by various models depending on purpose in view and on possibilities of getting information about decision-maker's preference judgements. Development of formalized models and decision-making methods require special organization of this activity and interaction of experts on decision-making with the decision-maker [57, 58]. Use of a model developed and revised by the analysis expert in dialogue with the decision-maker is seemingly the most general approach in the decision-making methodology.

References

1. *Venttsel, K. S.*, Introduction into the operations research. Sovietskoe Radio Publ., Moscow, 1964, 388 pp. (in Russian).
2. *Venttsel, E. S.*, Operation research. Sovietskoe Radio Publ., Moscow, 1972, 552 pp. (in Russian).
3. *Zadeh, L. A.*, Optimality and non-scalar-valued performance criteria. IEEE Transactions on Automatic Control, **AC-8**, *1*, pp. 59—60.
4. *Ozernoi, V. M.*, Decision-making: A survey. Automation and Remote Control, 1971, **32**.
5. *Larichev, O. I.*, Man-computer decision-making procedures: A survey. Automation and Remote Control, 1971, **32**.
6. *Emelyanov, S. V. et al.*, Problems and methods of decision-making: A survey. International Center of Scientific and Technological Information. Moscow, 1973, 44 pp. (in Russian).
7. *Optner, S.*, Systems analysis for business and industrial problem solving. Prentice-Hall, Inc., 1965.
8. *Roy, B.*, Décisions avec critères multiple et méthodes. Metra, **XI**, *I*, 1972, pp. 121—151.
9. *Bernard, G., Besson, M. L.*, Douze méthodes d'analyse multicritères. Revue Française d'Informatique et de Recherche opérationnelle, *V-3*, 1971.
10. *Volkovich, V. L.*, Methods for decision-making with respect to a set of optimality criteria: A survey. Proceedings of the seminar "Complex Control Systems". Naukova Dumka Publ., Kiev, Vol. 1, pp. 100—112 (in Russian).
11. *Alexandrov, A. D. et al.*, Quantitative comparison of alternative versions or strategies in control problems, in "Fifth All-Union Conference on Control Problems (Moscow, 1971). Abstracts of papers. Part 1". "Nauka" Publ., Moscow, 1971, pp. 268—270 (in Russian).

12. *Artobolevskii, I. I. et al.*, Looking for compromise alternative in selecting machine parameters. Dokladi Akademii Nauk SSSR, 1974, **219,** *1,* pp. 53—56 (in Russian).
13. *Larichev, O. I.*, A method for estimating alternative courses of action in applied research and development. Automation and Remote Control, 1972, *8.*
14. *Salukvadze, M. E.*, On a linear programming problem with vector-valued quality criterion. Automation and Remote Control, 1972, *5.*
15. *Borisov, V. I.*, Vector optimization problems, in "Operations research: Methodological aspects", Nauka Publ., Moscow, 1972, pp. 72—91 (in Russian).
16. *Emelyanov, S. V. et al.*, Multi-objective estimation of technological process local control systems (preprint). Order of Lenin Institute of Control Sciences, Moscow, 1974, 40 pp. (in Russian).
17. *Gorokhovik, V. V.*, On vector optimization problem. Izvestiya AN SSSR. Tekhnicheskaya kibernetika. 1972, *6,* pp. 63—70 (in Russian).
18. *Emelyanov, S. V., et al.*, Selection of rational versions of mine process flows with respect to multiple criteria. Izvestiya vysshikh uchebnykh zavedenii. Gornyi zhurnal 1972, *5,* pp. 8—14 (in Russian).
19. *Kuzmin, I. V. et al.*, Decision-making principles for multi-alternative problems with vector criterion, in "Second All-Union Conference on Problems of Scientific Organization of Control in the Socialist Industry". Vol. 7, Moscow, 1972, VINITI Publ., pp. 168—179 (in Russian).
20. *Kurilov, A. E.*, Construction of additive criterion on the basis of expertise using order scales. "Trudy MFTI, 1971", "Astrophysics and applied mathematics", MFTI Publ., Dolgoprudnyi, 1972 (in Russian).
21. *Borisov, V. I.*, Vector optimization of systems, in "Systems study. (Proceedings of the All-Union Symposium)" VINITI Publ., Moscow, 1971, pp. 106—114 (in Russian).
22. *Burchakov, A. S. et al.*, "A procedure for selecting a set of preferable (rational) alternatives of coal mine process flows" (Manual for the course in "Basics of coal mine design"). Ministry of Higher and Specialized Secondary Education. Moscow Red Banner of Labor Mining Institute, Moscow, 1972, 58 pp. (in Russian).
23. *Makarov, I. M. et al.*, Making decision upon an alternative version of an involved automatic control system. Automation and Remote Control 1971, *3.*
24. *Glotov, V. A. et al.*, A method for estimating the quality of journal publications, in "VI-th Symposium on Cybernetics. Part III."; Institute of Cybernetics, Georgian Academy of Sciences, Tbilisi, 1972, pp. 58—61 (in Russian).
25. *Pareto, V.*, Cours d'Economie Politique. Lausanne, Rouge, 1896.
26. *Klinger, A.*, Vector-Valued Performance Criteria. IEEE Transactions on Automatic Control **AC-9,** *1,* Jan. 1964, pp. 117—118.
27. *Karlin, S.*, Mathematical methods and theory in games, programming and economics. Pergamon Press, 1959.
28. *Focke, J.*, Vektormaximumproblem und parametrische Optimierung. Mathematische Operationsforschung und Statistik **4,** *5,* 1973.
29. *Kazakov, M. F. et al.*, Some questions in the multi-objective decision-making, in "IV th Symposium on Cybernetics. Part III", Institute of Cybernetics, Georgian Academy of Sciences, Tbilisi, 1972, pp. 91—92 (in Russian).
30. *Germeier, Yu. B.*, Introduction into the operations research theory. Nauka Publ., Moscow, 1971, 384 pp. (in Russian).
31. *Germeier, Yu. B.*, On vector-valued criteria of effectiveness convolution into a single criterion under uncertainty in the convolution parameters, in "Cybernetics serves communism", collection of papers edited by A. I. Berg, Vol. 6, Energiya Publ., 1971 (in Russian).
32. *Emelyanov, S. V. et al.*, Models and methods of vector optimization, in "Results of science and technology", series "Technical cybernetics", Vol. 5, VINITI Publ., Moscow, 1973, pp. 386—448 (in Russia).
33. *Krukhmalev, G. A.*, A method for comparative estimation of the quality of ship automatic devices. Collection of papers on automatic control of ship equipment, Vol. 4, Sudostroenie Publ., 1970, pp. 84—88 (in Russian).
34. *Venttsel, E. S.*, Speech at a discussion at the symposium on "Operations research and analysis of scientific progress" (Moscow, 1967), in "Operations research. Methodological aspects", Nauka Publ., Moscow, 1972, pp. 106—111 (in Russian).
35. *Ozernoi, V. M.*, Application of the utility theory to decision-making, in "Results of

science", series "Organization of control", Vol. 1, VINITI Publ., Moscow, 1971, pp. 74—97 (in Russian).
36. *Vilkas, E. I.*, The utility theory and decision-making, in "Abstracts of papers presented at the I-st All-Union Conference on Operations Research", Minsk, 1972, pp. 68—76 (in Russian).
37. *Fishburn, P. C.*, Independence in utility theory with whole product sets. "Operations Research", 1965, Jan.-Febr., **13**, *1*.
38. *Adams, E. W., Fagot, R. F.*, A model of riskless choice. Behavioral Science, 1959, Vol. 4.
39. *Fishburn, P. C.*, Independence, trade-offs, and transformations in bivariate utility functions. Management Science, Series A, 1965, July, **11**, *9*.
40. *Luce, R. D., Tukey, J. W.*, Simultaneous conjoint measurement: A new type of fundamental measurement. J. of Mathematical psycology, 1964, **1**.
41. *Fishburn, P. C.*, A study of independence in multivariate utility theory. Econometrica, 1969, Jan., **37**, *1*.
42. *Keeney, R. L.*, Evaluating multidimensional situations using quasi-separable utility function. IEEE Transactions on Man-Machine Systems, 1968, June, **MMS-9**, *2*.
43. *Keeney, R. L.*, Multiplicative utility functions. Operations Research, 1974, Jan.-Febr., **22**, *1*, pp. 22—34.
44. *Huber, G. P.*, Multi-attribute utility models: A review of field and field-like studies. Management Science, 1974, Apr., **20**, *10*, pp. 1393—1402.
45. *Kafarov, V. V. et al.*, Method for solving multi-criteria problems of control of complex chemical system. Doklady Akademii Nauk, 1971, **198**, pp. 62—63 (in Russian).
46. *Yampolskii, V. Z. et al.*, Method of complicated decision-making based on pair-wise comparison of alternatives, in "VI-th Symposium on cybernetics. Part III", Institute of cybernetics, Georgian Academy of Sciences, Tbilisi, 1972, pp. 202—207 (in Russian).
47. *Meshcheryakov, G. A. et al.*, On formalized decision-making in control with respect to a single aim, ibid., pp. 131—133 (in Russian).
48. *Meleshko, V. I.*, Method of concessions in multi-objective problem, ibid., pp. 125—130 (in Russian).
49. *Podinovskii, V. V. et al.*, Study of the successive concession method, in "Methods for analysis and reconstruction of complex systems". Abstracts of papers", Zinatne Publ., Riga, 1972, pp. 61—63 (in Russian).
50. *Gromov, G. G.*, On a model of choosing an alternative of designed system with respect to two optimality criteria, ibid., pp. 71—72 (in Russian).
51. *Benayoun, R., et al.*, ELECTRE: une méthode pour guider le choix en présence de points de vue multiple. SEMA (Metra International), Direction Scientifique, note de travail, *49*, 1966.
52. *Kaplinskii, A. I. et al.*, Randomization and smoothing in problems and algorithms of adaptation. Automation and Remote Control., 1974, *6*.
53. *Krasnenker, A. S.*, On adaptive approach to decision-making with respect to several criteria, in "Questions of optimal programming in industrial problems", NIEI VGU Publ., Voronezh, 1972, pp. 26—36 (in Russian).
54. *Krasnenker, A. S.*, On adaptation in decision-making under uncertainty in choosing the aim. Proc. of VI-th Winter Workshop on Mathematical Programming and Related Problems. Drogobytch, 1973, TsEMI AN SSSR, 1973 (in Russian).
55. *Kaplinskii, A. I. et al.*, On construction of stochastic approximation algorithms. Preprint of the Industrial economics institute, Ukranian Academy of Sciences. Donetsk, 1974. (in Russian).
56. *Geoffrion, A. et al.*, An interactive approach for multi-criterion optimization with an application to the operation of an academic department. Management Science, 1972, Dec., **19**, *4*, pp. 357—368.
57. *Dyer, D.*, Time-sharing compüter program for the solution of the multiple criteria problem. Management Science, 1973, Aug., **19**, *12*, pp. 1379—1383.
58. *Larichev, O. I.*, Multiple-criteria decision-making method based on preference judgements of the decision-maker. "Proc. of the Conference on System Analysis and Long-term Planning". Computer Center of the Academy of Sciences, USSR, Moscow, 1973, pp. 189—196 (in Russian).
59. *Ozernoi, V. M.*, Expert estimates in decision-making with vector-valued criterion, ibid., pp. 180—188 (in Russian).

Problems of Control and Information Theory, Jubilee Number, pp. 65–76 (1975)

RESULTS OF CO-WORK IN INFORMATION THEORY AND RECOGNITION PROBLEMS, ADVANCES IN THE DEVELOPMENT OF COMPUTER ASSISTANCE TO HEART AND VASCULAR SYSTEM EXAMINATIONS*

S. CSIBI
(Budapest)

(Received April 21, 1975)

1. Introduction

At this Jubilee Session of the 25th Anniversary of the Soviet–Hungarian Scientific-Engineering Cooperation Academicians F. Csáki and V. I. Siforov already gave an overview of (a) the work relations of the Soviet and the Hungarian research schools in the field of Control and Information Theory and (b) the experience of co-work, specifically in the field of Information Theory [1] and the development of a service system relying upon recognition techniques [2]. Academician Siforov pointed out that efforts are made, specifically in the latter area, to assist mass-wise examinations in cardiology by computer.

We add, in what follows, to these lectures some comments concerning specifically project No. 10 of the joint research program of the Scientific Academies of the USSR and Hungary, viz. the research in Information Theory and recognition problem, and the development of the afore-mentioned service system.

2. Goals

The purpose of this work is (a) to ensure manysided exchange of views between Soviet and Hungarian researchers pursuing mathematical topics in this field and (b) the development of an R-10 based mass service of immediate practical significance and of interest also in a more general context concerning recognition techniques.

As it was pointed out already by Academician Siforov, series of annual interinstitutional seminars and international conferences serve, since 1968, for the regular exchange of ideas between Soviet and Hungarian researchers

* Presented at the Scientific Session of Department 6, Hungarian Academy of Sciences, Budapest, September 19, 1974, during the Festive Meeting arranged on the occasion of the 25th Anniversary of the Hungarian–Soviet Scientific-Engineering Cooperation.

in the field within Information Theory and Statistical Information Processing. These activities got a new publication forum, the common journal of the Academies of the USSR and Hungary: the "Problems of Control and Information Theory".

This manysided joint activity has, of course, an essential impact on the techniques and way of thinking of the participating Hungarian institutes, in research as well as in actual practice.

The Soviet and the Hungarian participants of this common scientific activity are listed in Table 1.

Table 1

Interinstitutional seminars, from 1968

Soviet institutes	Hungarian institutes
IPPI occasionally: Steklov Mathematical Institute; IPU (Academy of Sciences of the USSR) MGU	Telecommunication Research Institute (TKI); Mathematical Research Institute of the Hungarian Academy of Sciences; Computer and Automation Institute (SZTAKI) of the Hungarian Academy of Sciences; József Attila University of Sciences (Szeged), Cybernetics Laboratory; Technical University of Budapest, Department of Automation; Technical University of Budapest, Institute of Communication Electronics

On the other hand, we have been developing since 1969 a specific task (a "pilot project", in the usual sense of international organizations), which might be also instructive, in a broader context for computer-based mass service, for assisting mass examinations, by the computer R-10, in cardiology.

The participants of this latter activity of immediate practical interest are listed in Table 2.

Table 2

Computer assistance to cardiological examinations, from 1968

Soviet institutes	Hungarian institutes
IPPI, Academy of Sciences of the USSR	Hungarian Cardiology Institute
MONIKI COLIU	Telecommunication Research Institute (TKI)

3. On selecting the topics

The fundamental idea of this twofold activity came up still in 1968, during the first exchange of views between Academicians G. Bognár and V. I. Siforov and the intention was, already from the very outset, at the same time a lasting, manysided, joint scientific life, concerning these two Academies, and the joint realization of a complex actual endeavour.

The attention turned to the computer R-10, as this country was joining the Unified Computer System and, within this mainly, the development of minicomputers. Dr. I. Váradi, General Director of the Telecommunication Research Institute, had a decisive role in combining academic and industrial aims. This kind of topic selection was supported by the very definite intention of the Hungarian Academy of Sciences to offer further possibilities for increasing scientific activities within the country for developing basic approaches to computer applications.

Of course, all actual and personal experiences gathered during the previous decade, in the co-work of the Telecommunication Research Institute and the NIIR (Research Institute for Radio Science, Moscow) and personally of the Director of the latter, Professor Fortushenko and Academician G. Bognár, and his coworkers — among them the present author — in the field of broadband radio relay systems, especially during the joint development of the radio relay system *Druzhba* [3], had an essential impact forming these well-defined actual efforts.

Setting the specific tasks precisely was preceded by a study of the available technical literature and solving several pilot tasks, and searching for such recognition type mass services which are related to sufficiently frequent and widespread applications and appear, at the same time, appropriately instructive also for other kind of frequently appearing actual jobs.

In this respect the role of the IPPI, and personally that of Academician Siforov and Associate Director, Dr. Ovseevich, was particularly relevant, as they presented the pattern recognition research at the IPPI, during the first Moscow seminar, for a prestudy. In our final choice we have found enthusiastic partners at the Hungarian Cardiology Institute (OKI), viz., personally in Professor Gy. Gábor and his coworkers.

Finally, we have arrived at the assistance to mass examinations in cardiology, as a specific system project, for the following two reasons: (a) because of the increasing social significance of Public Health Information Systems and the anticipated gradual expansion of these, and (b) also because of the frequency at which heart and vascular system diseases actually occur in mass health care.

4. Co-work in information theory

Today, at this session of Department 6, of the Hungarian Academy of Sciences, I should like to point out, by a few specific facts the actual significance, I guess, our close relations with the research schools of the Academy of Sciences of the USSR have, relying mainly upon the Institute of Information Transmission Problems (IPPI) — particularly in the field of Information Theory (meaning by this in the present context, probabilistic approaches related to the study of the information moments).

In this case, information theory in concerned particularly with the principal possibilities of information transmission and processing, as a matter of fact with revealing, how efficient the available actual techniques are with respect to the principally attainable possibilities.

Information theoretic studies are of fundamental interest (a) in obtaining a clear overview of information transmission and processing principles, and (b) in various specific research problems.

With the development of microelectronics and the fast decrease of relative implementation costs — both in information processing and transmission — the scope of those topics, concerning everyday design, are gradually extending in which the good, or bad exploitation of the available possibilities, done by the designer, leads to striking differences.

Accordingly, more and more information theory topics come closer to everyday design practice. This trend gives a particular significance to the information theory aspects of this co-work, viz., to the close contacts with the research groups in information theory at the IPPI, as well as the furthering of the existing traditions in this country, in the field of information theory.

The survey by Academician Siforov pointed already out that these traditions started, here in Hungary, by Academician A. Rényi, the founder of the information theoretical research in our country, who, however, unfortunately deceased four and a half year ago.

Table 3 enlists the main field, the contributors and the actual significance of the topics, viz., those which appear, at least, for the present author, as most significant, concerning the Soviet–Hungarian co-work in information theory.

Table 3, obviously, includes not all the internationally relevant results of Soviet information theoretic schools, missing those, to which no lasting activities have been related, at least in the framework of the present co-work (due to the limited range of our activities here, or possibly to our one-sidedness only). The survey does not extend, naturally to everything, concerning activities within the country in this field either.

Table 3

Co-work in information theory

Results	Significance
— Principial possibilities of information transmission under constraints close to real life conditions (Rate distortion, coding theorems) (Dobrushin, Pinsker, Marton)	Possibility of studying fidelity criteria; of interest in analog information transmission
— Principal possibilities of information transmission with feedback and efficient algorithms (Pinsker, Dobrushin, Khasminskii, Zigangirov)	in teleprocessing
— Principal possibilities of information transmission under complexity constraints (Pinsker, Dobrushin and others)	in teleprocessing
— Source coding and coding theorems for noiseless channels (Csiszár, Marton)	in teleprocessing
— Simultaneous servicing of several sources (Körner)	in computer communication
— Estimation efficiency (Ibraginov, Khasminskii)	in information processing
— Information moments for statistical purposes, estimations (Rényi, Csiszár, Perez, Nemetz, Vajda, Fritz, Győrfi)	in information processing

For further orientation, concerning this topic, I refer to the Siforov lecture [2], to our state-of-art report, written up early in 1972 together with I. Csiszár and G. Gordos [7], a review on the Pinsker Seminar Series, held in Budapest in 1974 [5] and the survey by A. Wyner, which appeared early in 1974 [6].

5. Co-work in recognition problems

In this part of our joint activities the attention was mainly directed to the development of the techniques which guarantee in advance that an algorithm will learn, from masses of labelled individual cases, classification and interpretation rules. We are concerned, in this sense, mainly with statistical learning processes.

In this respect we were mainly interested in (a) recurrent estimations, (b) computational constraints, (c) opening procedures, stopping rules and (d) the study of real life conditions, e.g., that of the interdependence between samples. These were the main topics in our work, done so far.

Referring to some related comments in the Siforov lecture, we point out, this time again, some features together with the contributors (Table 4).

This work is directly related to the design of the statistical learning algorithms, and to the very service outlined in Section 6. This is the theoretical

Table 4

Recognition problems

Results	Significance
-- Algorithms with preguaranteeable learning capability, for independent samples (Aizerman, Tsypkin, Braverman, Rosonoer, Khasminskii, Nevelson, Gulyás, Győrfi)	Fundamental algorithms for recurrent learning
— Improving density function estimation (Révész, Rejtő, Major)	Consistent estimates
— Lyapunov-type theorems (Braverman, Rosonoer, Csibi)	Stability of processes
— Simultaneous estimation of error probability, optimum approximations in the sense of error probability (Győrfi, Fritz)	Stopping problems, meaningful approximations
-- Exploitation of the dependence of the samples, recurrent learning from weakly dependent samples, domain of attraction for learning capabilities (Révész, Csibi)	Extending the range of learning algorithms, considering computational constraints

background, on which a program library has already been successfully used for learning interpretation rules by the computer in various pilot tasks. We intend to apply (and develop) this also within the experimental service to be described in Section 6.

An additional insight into this work is given by the afore-mentioned state-of-art report [7], a previous survey by the Telecommunication Research Institute [8], several related articles published in "Problemy Peredachi Informatsii" ("Information Transmission Problems"), "Avtomatika i Telemekhanika" ("Automation and Telemechanics"), "Problems of Control and Information Theory", and several preprints of the Mathematical Research Institute of the Hungarian Academy of Sciences and of the Telecommunication Research Institute, which appeared in the meantime.

Within this activity our work has been definitely restricted to the statistical topics of recognition tasks. The activity, concerning the structural methods of recognition is, of course, also growing, within the Academy of Sciences of the USSR as well as the Hungarian Academy of Sciences, viz. in image processing, scene analysis and in other aspects. But we are, for the time being, still not much pursuing (perhaps just because of excessive one-sidedness) the structural aspects, within the presently reported co-work of the two Academies.

In information theory, as well as in the study of recognition topics, our work is closely related to the topics in probability theory and statistics, jointly

reported at this present Jubilee Session by Academician Yu. V. Prokhorov, Associate Director of the Steklov Institute of the Academy of Sciences of the USSR and M. Arató, Associate Director of the Computer and Automation Institute of the Hungarian Academy of Sciences.

6. Assisting examinations in cardiology by computers

Computer assistance to examinations in cardiology has, for the time being, already an extensive literature, from which it is obvious that — besides significant results — the approaches of such kind of services are still at a fairly developing stage. The improvement of these approaches opens up, however, obviously new possibilities before the cardiologist and raise new open problems as well. The implementation of such services, within each new community of medical doctors, as a matter of fact, takes a long time and calls for regular work in common, including cardiologists and system designers.

The development of this project is directed at the following: (a) the posterior analysis and interpretation of the data masses obtained in the course of screening and ability tests, (b) the assistance to cardiologist examinations in the presence of the patient, (c) consultation, (d) teaching, (e) assistance to the analysis of archive stuffs.

We wish, of course, in accordance with the spirit of academic research, to offer further investigation possibilities to the cardiologist, through practical as well as mathematical approaches.

Figure 1 shows a CII 10010-based experimental system, which we have been using at the Hungarian Cardiology Institute of Budapest and the Telecommunication Research Institute, since late 1972.

This system may be operated in two ways: (a) three-lead ECG signals are fed into an analog magnetic tape unit (Fig. 2), additional data of cardiologist examinations are recorded on a standard format sheet and these two kinds of data are introduced into the computer at a later date; (b) the afore-mentioned ECG record is conveyed in the form of an analog signal, and the alphanumerical data (by using a keyboard) over a switched telephone network into the center. (In the latter case the terminal may be switched over simply into one of the following modes: (i) speach, (ii) alphanumerical information exchange, (iii) ECG signal input.)

By pressing a pushbutton a 10-sec long observation is made of the heart current activity of the patient and fed into the computer, together with a calibration square wave of 1 mV peak-to-peak amplitude (Fig. 3). The computer center draws from each input signal 300 samples/sec, and rounds off the analog samples obtained in this way to about 128 levels.

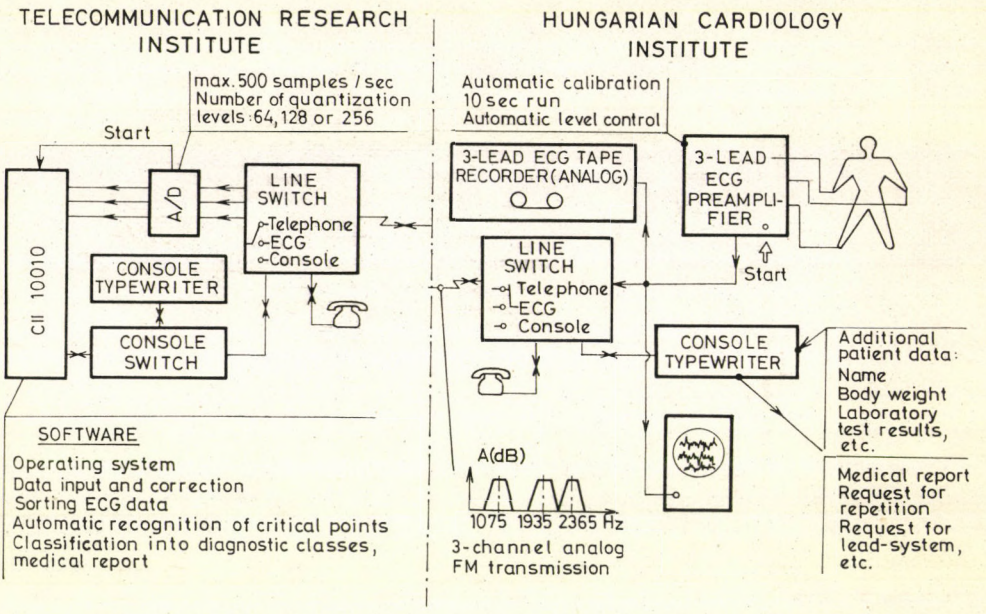


Fig. 1. The CII 10010-based experimental system of OKI and TKI

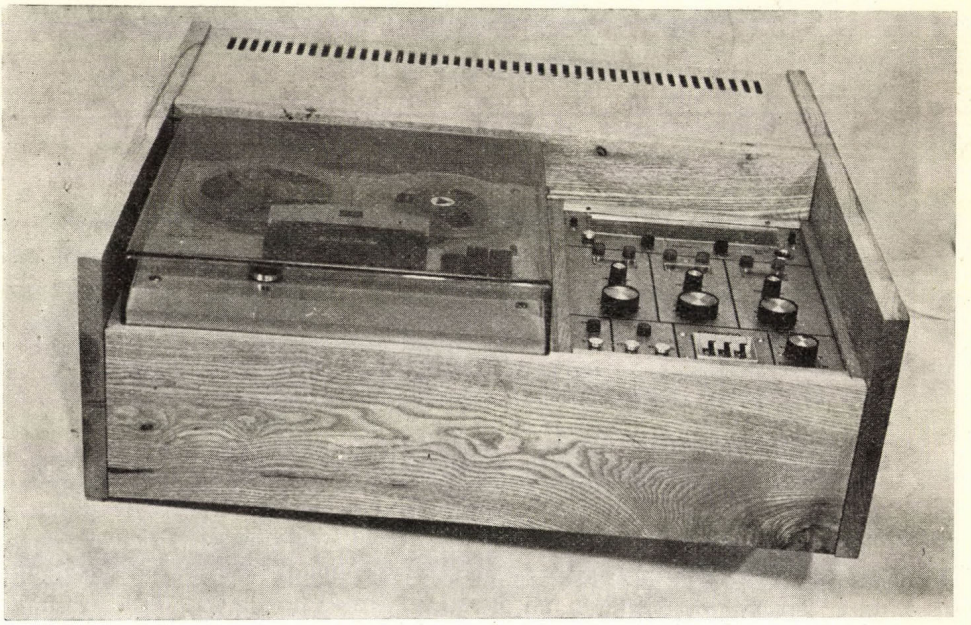


Fig. 2. Magnetic tape unit

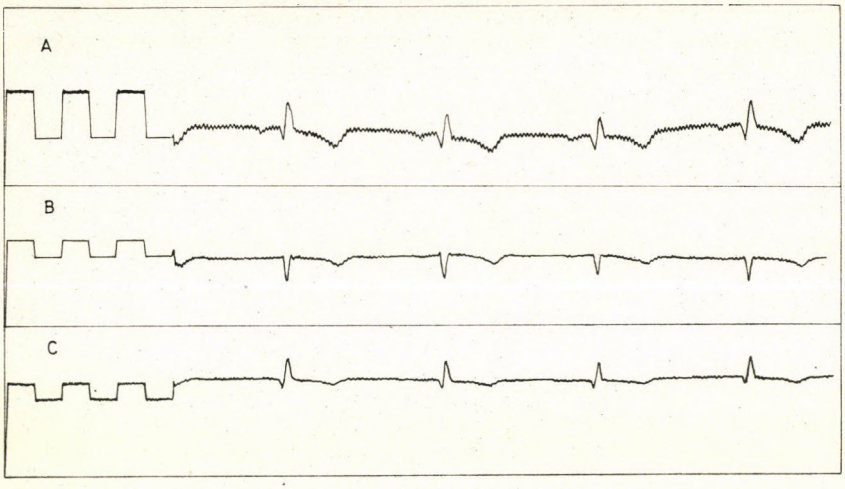


Fig. 3. Record drawn back from digital data by an automatic plotter

We have tried out, during 1972 in the first version of this system, all the functions of the entire program system up to the medical report, but the measurement results could not be well reproduced for noisy actual patient record. At present a program of better noise protection is operated only up to the vector-cardiographic description. A type of medical report program, connected to this, is for the time being under development.

Next to semantic noise protection, calibration, base line determination, coordinate and vector peak recognition and segmentation the program supplies, in rhythmical cases, three coordinate curves and the vector model curve of a single selected patient-period for further study. Otherwise it reports arrhythmia.

At the present stage of development we can gather reliably standard archive material. Up to now we have registered two-hundred carefully collected archive cases, supplied with the corresponding medical diagnoses. This is, of course, still insufficient for the computer, for learning classification rules. From the next year on a faster collection of archive stuffs is planned by introducing addition M. D. workplaces, supplied with magnetic tapes, both in Budapest and Moscow.

By the end of 1973 an R-10 configuration was implemented in the Telecommunication Research Institute, in order to co-work also with the Hungarian Cardiology Institute in the second half of 1975 in an experimental service, essentially equivalent to the one, available by the end of this year on the present CII 10010 configuration.

In the R-10 based service, the present algorithm will, however, run — according to estimations — within one minute up to the vector-cardiographic description, instead of the present 8 min run time on the CII 10010 configuration. In addition, we intend to replace the presently used teleprinter terminal as well for an alphanumerical display, supplied with a keyboard. The curve-depicting possibilities of the alphanumerical display are presented in Table 4. This also seems to be already useful for experimental cardiological examinations. For a routine, however, we are planning to apply a quasi-graphic display, to be developed in VIDEOTON, and a fine point-grid displaying of the curves investigated.

To sum up, the outlined project is at an advanced stage of automation, but for cardiologist procedures will be really fit only next to further advances.

The co-working cardiologists have done, up to know, instructive investigations on vector-cardiographic curve-complexes, reproduced by a plotter in analog form for the automatic recognition and interpretation of physiologically critical points. In addition we have, up to now, a learning program package, tested partly on school examples, and partly on some simple actual tasks, taken, however, from other medical areas. By means of these and further programs we intend to enable the learning and revision of classification rules on the basis of the total archive material and various partial data of the same

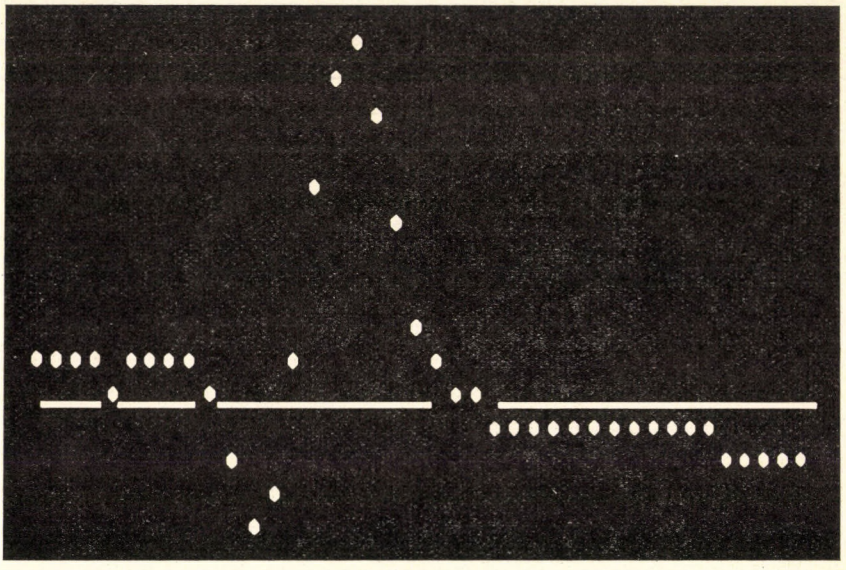


Fig. 4. A segment of one of the lead-signals, depicted of an alphanumeric display, for experimental purposes

The contributors to this work were on the part of the Hungarian Cardiology Institute, Drs. Gy. Lamm, K. Ghyczy, J. Németh and occasionally other cardiologists on duty, and also two medical assistants, and on the part of the Telecommunication Research Institue (in addition to the co-authors of Ref. [8]) A. B.-Nagy, B. Balog, F. Csánki, I. Kerekes, Mrs. Éva Szilléri, T. Wolf, Mrs. Mária Szélesi, a programming technician and a medical-computer assistant. The manager of the minicomputer center was formerly L. Ungvári and recently E. Neményi.

V. Shakin, C. Sc. (Math.-Phys.) is contributing, on the part of IPPI, co-principal investigator of the project, to research and system design, here in Budapest. He spent already at this place one month early in 1972, and is acting now, from the beginning of September 1974, here as a Keldish Fellow, for a period of 9 months at the Telecommunication Research Institute.

The plans for the actual service system have been presented to the Moscow Research Institute for Clinical Research (MONIKI). The delegations of MONIKI, IPPI and the COLIU are expected to come next year, in the first quarter, for the coordination of the actual service objectives to be supplied by the R-10-based system. We intend to consult this project also from further aspects with the Ministries of Public Health of the USSR and the Hungarian People's Republic, and follow as for as possible, also the "Minimum Operation Requirements" to be discussed by the TC IV, IFIP.

References

1. *Csáki, F.*, "The development of the work connections between the Soviet and the Hungarian research schools in the field of control and information theory". Scientific Session of Department VI of the Hungarian Academy of Sciences, 19 September, 1974, at the Meeting of "25 years of the Hungarian–Soviet scientific-engineering cooperation" (in this issue).
2. *Siforov, V. I.*, "Experiences and prospects of the Soviet–Hungarian cooperation in the field of information theory and its applications" (at this session).
3. *Váradi, I., Bognár, G., Fortushenko, A. D.*, lectures, Technical Day on Telecommunication and Communication Technology, House of Engineering, 23 September, 1973, at this meeting.
4. *Arató, M., Prokhorov, Yu. V.*, lectures, session of Department 3 of the Hungarian Academy of Sciences, 18 September, 1974, at this meeting.
5. "M. S. Pinsker lecturing in Budapest", Problems of Control and Information Theory, **3**, *3*, 231—232 (1974).

6. *Wyner, A.*, Review IEEE Trans. IT, Jan. 1974.
7. *Gordos, G., Csiszár, I., Csibi, S.*, "State-of-art in information theory", Committee for Telecommunication, Department VI of the Hungarian Academy of Sciences, 1972. Telecommunication Research Institute, Transactions, Vol. XVIII, Special issue, 1974, in Hungarian
8. *Bak, J., Csibi, S., Dévai, F., Gulyás, O., Győrfi, L., Kobzos, L., Molnár, L., Székely, E.*, "The mathematical and software questions of the machine interpretation of observations", Yearbook, Telecommunication Research Institute, 1973, 285—294.

S. Csibi
Telecommunication Research Institute
H–1026 Budapest II., Gábor Áron u. 65
and
HEI, Technical University of Budapest
H–1111 Budapest XI., Stoczek u. 2.
Hungary

Problems of Control and Information Theory, Jubilee Number pp, 77—96 (1975)

DEVELOPMENT OF THE FINITE AUTOMATA THEORY IN THE DISCRETE CIRCUIT DESIGN

M. A. GAVRILOV
(Moscow)

(Received June 2, 1975)

The paper discusses new problems posed to the finite automata theory by the discrete circuit design automation. Along with a general discussion of the development of the theory, some new results by the author are presented.

The beginning of the finite automata theory is due to the progress of the discrete technology. A Russian physicist P. S. Erenfest [1] was first to recognize the possibility of substituting objective mathematical methods for intuitive design of control circuits. He supposed that circuits controlling telephone switching networks might be described in terms of the Boolean algebra. The finite automata theory as a mathematical model of discrete circuits was developed in the works by W. McCulloch and W. Pitts [2] and S. Kleene [3] who were the first to introduce the notion of finite automaton. Among the first works where the Boolean algebra was applied to the design of relay circuits, those by C. Shannon [4], V. Shestakov [5], and A. Nakashima and M. Hanzava [6] may be mentioned.

The book by M. A. Gavrilov [7] was the first monograph dedicated to the theoretical and practical aspects of Boolean algebra application to design of discrete circuits.

Further development of this field is due to international contributions. References [8–11] contain rather detailed reviews of this initial period in the history of the finite automata theory, that of the "classical" formulation of the discrete circuit design.

Conditionally, the history of the finite automata theory may be broken down into the following periods:

a) Years 1938 to 1960 when the basic methods for minimization of Boolean functions written in the canonical form were developed and extended to unilateral conductivity elements, contactless ones, etc.

b) The next period covers studies of the sequential circuit theory. Although the first problems (the so called methods of "transition

tables") were considered as early as 1948 to 1950 [12], this area has been extensively explored only since 1958 [13]* and involves now stability studies initiated by the same paper and works on fault-tolerance started in 1960 [14].

c) The following period is characterized by extension of the finite automata theory to the so called "non-determinate" and "probabilistic" automata [15], development of the "regular event" theory [16] and "event algebra" [17], application of the finite automata theory to fault detection and diagnosis tests [18], etc.

d) The present trend to practical application of the finite automata theory has led to the development of machine methods for synthesis of its structure.

Let us dwell upon the problems which emerged during this period.

Rather well developed "classical" methods of determining the structure of discrete circuits are based upon the model of finite automaton as mentioned above.

This model is characterized by a quintuple of sets:

$$A = (X, Y, Z, \varphi, \psi)$$

where X is the input set of the automaton; Y is the internal state set; Z is the output set; φ is the function defining transition from one state to another $(y_{t+1} = \varphi(x_t, y_t))$, and ψ is the output function $(z = \psi(x_t, y_t))$.**

The so-called transition tables [13] and graphs are used in practical problems to describe this model.

Input (x) and internal element (y) states at time t are assigned, respectively, to columns and rows of the transition table whose entries contain internal element states at time $t + 1$ and output z.

Functional algorithms of combinational automata are described by means of the so-called state tables where columns are assigned to input variables and outputs, and rows contain appropriate output states assigned to each input state.

Although transition and state table-based methods for synthesis of discrete circuits are sufficiently developed, they have a number of essential drawbacks standing in the way to their application in problems of such dimensionality as met in practice.

One of the major drawbacks is due to the fact that the present-day methods of the classical finite automata theory require consideration of every input, output and internal state, and that to obtain minimal discrete unit in

* Here and below, references are made mostly to fundamental works in a given field of study.

** t is a discrete time.

terms of given primitives one has to perform, in principle, a complete throw-over which even with computers becomes unpractical when the number of input variables is about 10 to 12.

The only way out is, first, to develop methods that would be based on sets of states rather than on individual states, and, second, to work out approximate optimization methods based on some criterions reducing selection or eliminating it at all.

Among the first attempts to make use of these ideas Refs. [19, 20, 21] should be quoted.

For classical finite automata, the most compact form of presenting the related information is the parenthesis form of Boolean functions.

Each form of this type, however, limits, in a way, the domain where the optimal realization of an automaton is looked for because it is only the state table or the initial transition one that contain all possible realizations.

Thus, the problem arises of finding analytical methods enabling determination of all the data required to find the closest to the optimal realizations through given function but without analyzing separately each of required states. In essence, this problem boils down to find through given Boolean functions the power and compositions of state subsets satisfying given requirements. Presently, this is one of the major problems in the development of modern methods of optimal realization of finite automata.

Another related problem is that of extracting subfunctions involving states with required properties from given (in any form) Boolean function. Solution of all the modifications of this problem arising in implementation of discrete devices is based upon the notion of the so called "derivative" of Boolean functions which was first introduced in [22] but only recently has been applied to this field.*

Basically, all the problems arising at all stages of logical synthesis of discrete devices may be reduced to compositions or decompositions of finite automata. For instance, the classical problem of synthesizing the structure of one-output discrete device given in the following form

$$F = f(x_1, x_2, \ldots, x_n)$$

(where $x_1, x_2, \ldots, x_n$ are input variables) on the basis of components given also by Boolean functions $\varphi = \psi(y_1, y_2, \ldots, y_q)$ (where y_1, y_2, y_q are component's input variables) is to represent function F as

$$F^* = f^*(\varphi_1, \varphi_2, \ldots, \varphi_m).$$

* Solution of both problems will be considered below.

Generally, this may be represented as decomposition of an automaton described by function F into sub-automata described by functions φ.

Analysis of the discrete device block-diagram is, in essence, restoration of the original function F through F^*.

It should be noted that the analysis entails some specific problems. For diagnosis or reliability analysis, for example, one needs complete information about the structure, i.e. all the data about components involved into the circuit and their interconnections are required. With the analysis methods of the classical finite automata theory where one returns to the original function F of only input variables, this information vanishes. To preserve it, a model is required differing from the classical finite automaton model. Such a model was proposed in [23].

The scope of this paper does not allow exhaustive discussion of all problems involved into the design automation and related to the finite automata theory. As an example, composition and decomposition of combinational automata is presented below. New problems arising in the finite automata theory because of application of methods for directed search of optimal implementations of discrete circuits and due to difficulties generated by high dimensionality are most prominent there.

Discuss first composition of combinational circuits. The problem is posed as follows: a multi-output combinational automaton is given as a set of incompletely specified functions $F_1, F_2, \ldots, F_k$ given in arbitrary form; the functions are to be defined so as to decrease "distance" between them and to determine the general realization layout with complete fusion of separate functions wherever possible.

Existing methods for implementation of multi-output incomplete combinational automata are built around determination of the so called "multi-output" implicants [24, 25, 26, 27] or around using outputs of components implementing one of output functions for implementation of other output functions [28].

The shortcoming of these methods is that they ignore some function metrics characterized by the fact that "closer" functions are most connected and sometimes (for certain ways be completing them) may be fused with each other.

Let two combinational automata A_i and A_j be given described by functions F_i^1, F_i^0 and F_j^1 F_j^0 characterizing their permit and forbid states.*

* States giving "1" at the automaton output will be referred to as "permit" states, those giving "0" will be referred to as "forbid" ones.

Distance between these elements will be defined as cardinality of subsets of input states which are permit states in one automaton, and forbid ones in another. Denote them, respectively, as M_i^{r1}, M_i^{r0} and M_j^{r1}, M_j^{r0}. By definition,

$$M_i^{r1} = M_j^{r0} \quad \text{and} \quad M_i^{r0} = M_j^{r1}.$$

Denote functions characterizing these states as F_i^{r0}, F_i^{r1} and F_j^{r1}, F_j^{r0}, respectively, $F_i^{r1} = F_j^{r0}$ and $F_i^{r0} = F_j^{r1}$. Evidently, $F_t^{r1} = F_i^1 F_j^0$ and $F_i^{r0} = F_i^0 F_j^1$.

Let α_i^{r1}, α_i^{r0}, α_j^1 and α_j^{r0} be powers of appropriate state subsets having identical indexes. Assume that the lower is powers of state subsets union $M_i^{r1} \cup M_i^{r0}$ or $M_j^{r1} \cup M_j^{r0}$, the closer are automata A_i and A_j.

Identify in addition the following subsets:

Subset M_{ij}^{s1} of permit states similar in A_i and A_j;

Subset M_{ij}^{s0} of forbid states similar in A_i and A_j;

Subset M_i^{t1} of permit states in A_i which are absent in A_j;

Subset M_i^{t0} of forbid states in A_i which are absent in A_j;

Subset M_j^{t1} of permit states in A_j which are absent in A_i;

Subset M_j^{t0} of forbid states in A_j which are absent in A_i;

Subset $M^{\sim}$ of the so called "do not care" states, i.e. states for which the output state is of no importance or which are known to be impossible.

Functions characterizing these states are as follows:

$$F_{ij}^{s1} = F_i^1 F\,; \; F_{ij}^{s0} = F_i^0 F_j^0$$

$$M_i^{t1} = F_i^1 F^{\sim} = F_i^1 \bar{F}_j^1 \bar{F}_j^0$$

$$M_i^{t0} = F_i^0 F_j^{\sim} = F_i^0 \bar{F}_j^1 \bar{F}_j^0$$

$$M_j^{t1} = F_j^1 F^{\sim} = F_j^1 \bar{F}_i^1 \bar{F}_i^0$$

$$M_j^{t0} = F_j^0 F^{\sim} = F_j^0 \bar{F}_i^1 \bar{F}_i^0.$$

Demonstrate that if $F_i^{r1} = F_i^{r0} = 0$ or $F_{ij}^{s1} = F_{ij}^{s0} = 0$ i.e. if distance between automata A_i and A_j is zero or maximal, the automata may be implemented by the same structure.

Represent state sets F^1 and F^0 as unions of the subsets above and write the automaton functions as

$$A_i = (M_{ij}^{s1} \cup M_i^{r1} \cup M_i^{t1}),\ (M_{ij}^{s0} \cup M_i^{r0} \cup M_i^{t0})$$

$$A_j = (M_{ij}^{s1} \cup M_j^{r1} \cup M_j^{t1}),\ (M_{ij}^{s0} \cup M_j^{r0} \cup M_j^{t0}).$$

If $M_i^{r1} = M_i^{r0} = M_j^{r1} = M_j^{r0}$ are empty, then

$$A_i = (M_{ij}^{s1} \cup M_i^{t1}),\ (M_{ij}^{s0} \cup M_i^{t0})$$

and

$$A_j = (M_{ij}^{s1} \cup M_j^{t1}),\ (M_{ij}^{s0} \cup M_j^{t0}).$$

Complete A_i and A_j as follows:

Separate from $M_i^{\sim}$ of A_i some state subset $\Delta M_i^{\sim} = M_j^{t1}$ and some state subset $\Delta M_i^{\sim} = M_j^{t0}$ and add them to M_i^1 and M_i^0, respectively. Similar operation is done in A_j. As a result obtain

$$A_i = (M_{ij}^{s1} \cup M_i^{t1} \cup M_j^{t1}),\ (M_{ij}^{s0} \cup M_i^{t0} \cup M_j^{t0})$$

$$A_j = (M_{ij}^{s1} \cup M_j^{t1} \cup M_i^{t1}),\ (M_{ij}^{t0} \cup M_j^{t0} \cup M_j^{t0}).$$

Composition of state subsets in both automata is the same; therefore, they may be implemented by the same structure. In this case automata A_i and A_j will be referred to as completely connected.

A similar extension could be done if $M_{ij}^{s1} = M_{ij}^{s0}$ are empty. In this case completion of automata A_i and A_j is done as follows:

$$A_i = (M_i^{r1} \cup M_i^{t1} \cup M_r^{t0}),\ (M_i^{r0} \cup M_i^{t0} \cup M_j^{t1})$$

$$A_j = (M_j^{r1} \cup M_j^{t1} \cup M_j^{t0}),\ (M_j^{t0} \cup M_j^{t0} \cup M_j^{t1}).$$

It is easy to see that under such a completion $A_j = \bar{A}_i$ and both automata may be implemented by the same structure with the only difference that the output of one of them should be connected to inverter. Such automata will be referred to as inversely completely connected. Using some estimates (those of Sholomov [6], for instance), one can readily show that joint and separate implementations of completely or inversely completely connected functions are of similar complexity (to within criterion used) only in the extreme case where subsets M_{ij}^{s1} and M_{ij}^{s0} are empty and completion of F_i and F_j is maximal, in all other cases connected implementation is more expedient.

Totality of output functions of a multi-output net that are completely (or inversely completely) connected with each other will be referred to as connectivity "node". It stands to reason that in the general case an expedient in terms of minimality structure of output function may be represented as several node functions and related to them functions which are most close in terms powers of subsets M_{i1}^r and M_j^{r1}, and which it is reasonable to implement (from the viewpoint of simplicity) by adding some more circuits to those realizing the node functions (see e.g. [7]). In doing so, "profitability" of connected or

disconnected realizations of these functions should be established by means of appropriate criteria.

Determination of power of sets M^1 and M^0, subsets of classes S, r and t, etc. is done on the basis of their functions which may be written in an arbitrary parenthetic or some other forms. Given function is reduced to some "canonical" form where the number of states realizable by the function is representable as a sum of states realizable by its separate parts.

Parts of a function connected by the sign of disjunction/conjunction will be referred to as disjunctive/conjunctive terms. Then, according to [8], a form where all terms included into disjunctive terms are mutually orthogonal, and those included into conjunctive ones do not contain the same variables is the "canonical" form satisfying the condition above.

The number of states realizable by a disjunctive term is

$$\alpha^{\vee} = \sum_{i=1}^{i=m} \alpha_i^{\wedge}, \tag{4}$$

and that realizable by the conjunctive one is

$$\alpha^{\wedge} = 2^{n-k}\, \beta_1, \beta_2, \ldots, \beta_s$$

where m is the number of conjunctive terms in the disjunctive term under consideration, n is the total number of variables, k is the number of variables in the conjunctive term, s is the number of disjunctive terms in it, and $\beta_1, \beta_2, \ldots, \beta_s$ are numbers of states realizable by each of these disjunctive terms.

Preliminary completion of functions with the aim of their composition is sometimes very effective and, as some examples demonstrate, results in 35 to 40 % less number of elements as compared with implementations of incompletely specified multi-output discrete nets.

Consider now decomposition of combinational automata. One of the methods used here will be demonstrated through realization of a function given in the minimal normal form.* The classical theory of combinational automata tackles this problem by determination of the set of the so called "minterms" ("simple implicants")** and selecting, in a way, some subset realizing all permit and forbid states of the given Boolean function [32–34]. In [21,

* This solution of the decomposition problem is of interest for the design of large scale integrated circuits.

** A minimal conjunction of letters realizing non-contradictory one or more states in the permit and forbid parts of the state table and such that omission of at least one letter would result in a contradictory realization (where at least one permit and one forbid state are realized simultaneously) is referred to as minterm or simple implicant.

25–27] this problem is solved through state tables which enab essential reduction of problem dimensionality as it was indicated leabove.

Consider a more general approach to this problem based upon transformation of Boolean functions given in arbitrary form and determination of Boolean function derivatives and their parts with the view to identify subfunctions characterizing state sets that have certain features and to make a count of their powers in order to apply tests for directed search of optimal realizations.

Let an incomplete function F_i be given by its functions F_i^1 and F_i^0 characterizing permit and forbid states. The first step in the synthesis of combinational automaton net is elimination of unessential variables from these functions, i.e. variables which are not essential for non-contradictory and complete realization of given functions.

Unessential variables are determined by means of the so called "obligatory" letters, i.e. letters present in two states of which one is in the permit set and another is in the forbid set, and which differ only in value of this variable. Elimination of such a variable leads to inherent contradiction, this is the reason why such variables should be left.

Variables without obligatory letters, thus, are unessential.

To identify unessential variables, introduce a notion of derivative Boolean function:

$$\frac{dF_i}{dx_i} = F^1_{i(x_i=1)}\, F^0_{i(x_i=0)} + F^1_{i(x_i=0)}\, F^0_{i\,x_i=1)}.$$

Its first term defines states where variable x_i is obligatory and is 1, the second term defines the states where variable x_i is obligatory as well but is 0.

Elimination of unessential variables is done one by one in the order of their closeness to constants which is determined by the following test:

$$t_{x_i} = \alpha^1_1 x_i \alpha^0_{0x_i} + \alpha^1_{0x_i} \alpha^0_{1x_i}$$

where $\alpha^1_{1x_i}$ and $\alpha^0_{1x_i}$ are numbers of states in the permit part where variable x_i is, respectively, 1 or 0, $\alpha^0_{1x_i}$ and $\alpha^0_{0x_i}$ are numbers of forbid states where variable x_i is, respectively, 0 or 1. The lower is the value of test t_{x_i},* the closer to constant is variable x_i.

* It was proved in [35] that in the extreme case where a function is realized by one letter, criterion t_{x_i} will have the maximal value.

Consider an example.

Let some function F be given by appropriate functions characterizing its permit and do not care states:

$$F^1 = x_5 + \bar{x}_5\bar{x}_1x_6 + x_1(\bar{x}_5x_6 + x_2\bar{x}_5x_6x_{10}x_{11}) + x_1x_2(\bar{x}_3 + \bar{x}_4 + \bar{x}_7 + \bar{x}_8 + \bar{x}_9),$$

$$F^{\sim} = x_3x_4 + \bar{x}_8\bar{x}_9 .$$

Then

$$F^{1,\sim} = x_5 + \bar{x}_5\bar{x}_1x_6 + x_1(\bar{x}_5x_6 + x_2\bar{x}_5\bar{x}_6x_{10}x_{11}) +$$

$$+ x_1x_2(\bar{x}_3 + \bar{x}_4 + x_7 + x_8 + x_9)\,[x_3x_4 + \bar{x}_8\bar{x}_9],\ *$$

$$F^{0,\sim} = [x_5 + \bar{x}_5\bar{x}_1x_6 + x_1(\bar{x}_5 - x_6 + x_2\bar{x}_5\bar{x}_6x_{10}x_{11}) +$$

$$+ x_1x_2(\bar{x}_3 + \bar{x}_4 + \bar{x}_7 + x_8 + x_9)]\,[x_3x_4 + \bar{x}_8\bar{x}_9].$$

After decomposition with respect to variables x_1, x_2, x_3 and x_4, canonical forms of these functions will be as follows:

$$F^{1,\sim} = ({}_1({}_2x_1x_2x_3\bar{x}_4 + x_1x_2\bar{x}_3 + ({}_3x_1\bar{x}_2 + \bar{x}_1)_3 \cdot$$

$$\cdot\,({}_3x_5 + \bar{x}_5 - x_6)_2\,({}_3\bar{x}_3 + x_3\bar{x}_4)_3)_2\,[{}_2\bar{x}_8\bar{x}_9]_2$$

and

$$F^{0,\sim} = ({}_1({}_2x_1x_2 + \bar{x}_1)_2\,[{}_2x_5 + \bar{x}_5x_6]_2\,({}_2\bar{x}_3 + x_3\bar{x}_4)_2\,[{}_2x_8x_9]_2)_1.$$

Values of t_{x_i} are listed in Table 1.

Table 1

	x_1	x_2	x_3	x_4	x_5	x_6	x_8	x_9
α_1^1	63	63	39	39	72	72	78	78
α_0^1	54	54	78	78	45	45	39	39
α_1^0	9	9	9	9	0	0	18	18
α_0^0	18	18	18	18	27	27	9	9
t_{x_i}	1727	1727	1402	1402	1945	1945	1402	1402

* The conventional inversion bar is substituted here by square brackets.

The lowest values are obtained by the test for variables x_3, x_4, x_8 and x_9. Consider first variable x_0:

$$\frac{dF}{dx_3} = \big(x_1x_2\bar{x}_4 + (x_1\bar{x}_2 + \bar{x}_1)(x_5 + \bar{x}_5x_6)\big)\bar{x}_4[\bar{x}_8\bar{x}_9] \cdot$$
$$\cdot (x_1\bar{x}_2 + \bar{x}_1)[x_5 + \bar{x}_5x_6][\bar{x}_8\bar{x}_9] + \big(x_1x_2 + (x_1\bar{x}_2 + \bar{x}_1)(x_5 + \bar{x}_5x_6)\big) \cdot$$
$$\cdot [\bar{x}_8\bar{x}_9](x_1\bar{x}_2 + \bar{x}_1)[x_5 + \bar{x}_5x_6]\bar{x}_4[\bar{x}_8\bar{x}_9] = 0\,.$$

Variable x_3 is unessential. Eliminate it from expressions for $F^{1,\sim}$ and $F^{0,\sim}$, compute t_{x_i} again and establish whether next variable is essential or not. Functions $F^{1,\sim}$ and $F^{0,\sim}$ are simplified in this manner until only essential variables are left.

The final form of $F^{1,\sim}$ and $F^{0,\sim}$ is as follows:

$$F^{1,\sim} = x_1x_2 + (x_1\bar{x}_2 + \bar{x}_1)(x_5 + \bar{x}_5x_6)$$
$$F^{0,\sim} = (x_1\bar{x}_2 + \bar{x}_1)[x_5 + \bar{x}_5x_6] = (x_1\bar{x}_2 + \bar{x}_1)\bar{x}_5\bar{x}_6\,.$$

Next stage of synthesis is to separate the kernel minterms and the so called "insufficient minterms". Minterms are conjunctions consisting of obligatory letters and giving non-contradictory realization of permit and forbid states. Conjunctions of obligatory letters realizing concurrently both permit and forbid states and therefore requiring specification to be completed in order to avoid contradictory realization are referred to as insufficent minterms.

Therefore, the first step in implementation of the functions obtained is to determine obligatory letters and consisting of them conjunctions.

Note that analysis of unessential variables gives also for all essential variables both lists of essential letters and expressions characterizing states into which they are involved.

Expressions obtained above as the result of determination of essential variables are shown in Table 2.

Table 2

Obligatory letter	Function characterizing states in which the letter is involved
$x_1 = 1$	$x_1x_2\bar{x}_5\bar{x}_6$
$x_2 = 1$	$x_1x_2\bar{x}_5\bar{x}_6$
$x_5 = 1$	$(x_1\bar{x}_2 + \bar{x}_1)x_5\bar{x}_6$
$x_6 = 1$	$(x_1\bar{x}_2 + \bar{x}_1)\bar{x}_5x_6$

To obtain conjunctions of obligatory letters, compare pairwise the expression above and determine the number of states realized by a function which is conjunction of these expressions. If for a pair of expressions their conjunction is not zero, this means that corresponding obligatory letters form conjunction and that the expression of the function into which they are involved will enable both determination of the number of states which they realize and the function characterizing these states.

In the case under consideration, conjunction of obligatory letters is formed only by x_1 and x_2 because their expressions in Table 2 intersect and expressions for x_5 and x_6 are orthogonal both to each other and to expressions for x_1 and x_2.

Verify now which of obligatory letter conjunctions and separate obligatory letters are minterms of the kernel or insufficient minterms and whether they give complete realization of $F^{1,\sim}$. Denote disjunction of all minterms as

$$\varphi = x_1x_2 + x_5 + x_6$$

and determine expressions

$$F_0 \cdot \varphi \quad \text{and} \quad F_1 \cdot \bar{\varphi}.$$

If the first is equal to zero it means that implementation is non-contradictory; if the second is zero, implementation is complete.

$$F_0 \cdot \varphi = (x_1\bar{x}_2 + \bar{x}_1)\,\bar{x}_5\bar{x}_6(x_1x_2 + x_5 + x_6) = 0$$

$$F_1 \cdot \bar{\varphi} = (x_1x_2 + (x_1\bar{x}_2 + \bar{x}_1)\,(x_5 + \bar{x}_5x_6))\,[x_1x_2 + x_5 + x_6] = 0.$$

Thus, obtained above functions $F^{1,\sim}$ and $F^{0,\sim}$ are non-contradictorily realized as the following canonical sum-of-products

$$\varphi = x_1x_2 + x_5 + x_6$$

and all terms involved are minimal kernel terms.

Consider a case where in determination of obligatory letters defective minterms occur along with kernel minterms.

Let the following functions of six variables result from elimination of unessential variables:

$$F^1 = \bar{x}_1\bar{x}_2(\bar{x}_3x_4x_5 + x_3(x_4\bar{x}_5x_6 + \bar{x}_4)\,(x_5\bar{x}_6 + \bar{x}_5x_6)) +$$
$$+ \bar{x}_5\bar{x}_6x_4(x_3(x_1\bar{x}_2 + \bar{x}_1x_2) + \bar{x}_1\bar{x}_2\bar{x}_3) + x_2x_6 \cdot (\bar{x}_1\bar{x}_3x_4\bar{x}_5 + x_1x_3\bar{x}_4x_5)$$

$$F^0 = x_1x_3\bar{x}_5\bar{x}_6(\bar{x}_2\bar{x}_4 + x_2x_4) + x_6(x_3(x_4\bar{x}_5(x_1\bar{x}_2 + \bar{x}_1x_2) +$$
$$+ x_5(\bar{x}_1\bar{x}_2\bar{x}_4 + x_1x_2x_3)) + x_5\bar{x}_4\bar{x}_3\bar{x}_2x_1).$$

Functions containing obligatory letters are given in Table 3.

Table 3

Obligatory letter	Function characterizing states in which the letter is involved
$x_1 = 0$	$x_3x_4\bar{x}_5(\bar{x}_1\bar{x}_2x_6 + \bar{x}_1x_2\bar{x}_6)$
$x_2 = 0$	$x_3x_4\bar{x}_5(\bar{x}_1\bar{x}_2\bar{x}_6 + x_1\bar{x}_2\bar{x}_6)$
$x_3 = 0$	$\bar{x}_1x_2\bar{x}_3x_4\bar{x}_5x_6$
$x_4 = 1$	$x_1\bar{x}_2x_3x_4\bar{x}_5\bar{x}_6$
$x_4 = 0$	$x_1x_2x_3\bar{x}_4x_5x_6$
$x_5 = 0$	$x_1x_2x_3\bar{x}_4\bar{x}_5x_6$
$x_6 = 0$	$\bar{x}_1(x_3\bar{x}_6(x_2x_4\bar{x}_5 + \bar{x}_2\bar{x}_4x_5) + x_2x_3x_4\bar{x}_5\bar{x}_6$

Compare these functions pairwise and construct a triangular matrix where 1 and 0 stand for intersection or its absence, respectively.

Table 4

$x_2 = 0$	1					
$x_3 = 0$	0	0				
$x_4 = 1$	0	1	0			
$x_4 = 0$	0	0	0	0		
$x_5 = 0$	0	0	0	0	0	
$x_6 = 0$	1	1	0	1	0	0
	$x_1 = 0$	$x_2 = 0$	$x_3 = 0$	$x_4 = 1$	$x_4 = 0$	$x_5 = 0$

It follows from the matrix that the obligatory letters form the following function

$$\varphi = \bar{x}_1\bar{x}_2 + \bar{x}_1\bar{x}_6 + \bar{x}_2x_4\bar{x}_6 + \bar{x}_3 + \bar{x}_4 + \bar{x}_5 .$$

As it may be seen from $F^0\varphi$, conjunctions $\bar{x}_1x_6$ and $\bar{x}_2x_4\bar{x}_6$ are kernel minterms, and $\bar{x}_1\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$ and $\bar{x}_5$ are defective minterms.

$$\varphi_\vartheta = \bar{x}_1\bar{x}_6 + \bar{x}_2x_4\bar{x}_6$$

$$\varphi_H = \bar{x}_1\bar{x}_2 + \bar{x}_3 + \bar{x}_4 + \bar{x}_5 = \varphi_{1,H} + \varphi_{2,H} + \varphi_{3,H} + \varphi_{4,H}.$$

According to [35], the next step is selection from the set of insufficient minterms an optimal (having minimal number of minimal-length conjunctions) subset completely realizing the residuum of the function which is not realizable by the kernel minterms.

First, determine this residuum:

$$\Delta F^1 = F^1[\varphi_\vartheta] = \underset{1\,2}{(}\underset{3}{(}\bar{x}_1\bar{x}_2(\bar{x}_3x_4x_5 + x_3(x_4\bar{x}_5x_6 + \underset{4}{\bar{x}_4}\underset{5}{(}x_5\bar{x}_6 + \bar{x}_5x_6)\underset{5\,4\,3}{)}\underset{}{)} +$$

$$+ x_5\bar{x}_6x_4(\underset{3\;4}{x_3}(x_1\bar{x}_2 + \underset{4}{\bar{x}_1}x_2) + \bar{x}_1\bar{x}_2\underset{3}{\bar{x}_3}) + x_2x_6(\underset{3}{\bar{x}_1}\bar{x}_3x_4\bar{x}_5 +$$

$$+ x_1x_3\bar{x}_4x_5))\,[\underset{3\,2\;2}{\bar{x}_1}x_6 + \underset{2\,1}{\bar{x}_2}x_4\bar{x}_6]) = \bar{x}_1\bar{x}_2(x_3x_4x_5x_6 + x_3(x_4\bar{x}_5x_6 +$$

$$+ \bar{x}_4\bar{x}_5x_6) + x_2x_6(\bar{x}_1\bar{x}_3x_4\bar{x}_5 + x_1x_3\bar{x}_4x_5).$$

Insufficient minterms may lead to redundant realization. Therefore, the next step is to select among them a subset which would completely and most effectively realize all the states without remainder.

Take the maximal number of permit states realized along with the minimal number of forbid ones as a test for selection of the most effective terms. According to Sholomov's test, this would result in the simplest realization of completing functions. To determine these parameters, one has to make a count of the number of states characterized by functions resulting from substitution into ΔF^1 and F_0 of values of obligatory letters which are included into insufficient minterms.

For example, for insufficient minimal term $\bar{x}_1\bar{x}_2$ the following holds:

$$\Delta F^1_{x_1=0,\,x_2=0} = x_3x_4x_5x_6 + x_3(x_4\bar{x}_5x_6 + \bar{x}_4\bar{x}_5x_6)$$

$$F^0_{x_1=0,\,x_2=0} = \bar{x}_4x_3x_5x_6$$

Hence $\alpha_{\Delta F^1} = 3$, $\alpha_{F^0} = 1$.

Values obtained for all the functions are tabulated in Table 5.

Table 5

Defective minimal term		Function	Number of states
$\bar{x}_1\bar{x}_2$	ΔF^1	$\bar{x}_3x_4x_5x_6 + x_3(x_4\bar{x}_5x_6 + \bar{x}_4\bar{x}_5x_6)$	3
	ΔF^0	$x_3\bar{x}_4x_5x_6$	1
$\bar{x}_3$	ΔF^1	$\bar{x}_1\bar{x}_2x_4x_5x_6 + \bar{x}_1x_2x_4\bar{x}_5x_6$	2
	ΔF^0	$x_1\bar{x}_2\bar{x}_4x_5x_6$	1
$\bar{x}_4$	ΔF^1	$\bar{x}_1\bar{x}_2x_3\bar{x}_5x_6 + x_1x_2x_3x_5x_6$	2
	ΔF^0	$x_1\bar{x}_2\bar{x}_5\bar{x}_6 + x_6(\bar{x}_1\bar{x}_2x_3x_5 + x_1\bar{x}_2\bar{x}_3x_5)$	3
$\bar{x}_5$	ΔF^1	$\bar{x}_1\bar{x}_2x_3x_6 + \bar{x}_1x_2\bar{x}_3x_6$	3
	ΔF_0	$x_1x_3\bar{x}_6(\bar{x}_2\bar{x}_4 + x_2x_4) + x_3x_4x_6(x_1\bar{x}_2 + \bar{x}_1x_2)$	4

The functions tabulated above ΔF^1 and ΔF^0 define at the same time for appropriate incomplete terms completing functions eliminating contradictions introduced by the terms.

Within the subset of effective insufficient minterms, $\bar{x}_1\bar{x}_2$ is taken as the first one on the ground of values of α.

To avoid redundant realization in selection of next minimal terms, states realized by insufficient term $\bar{x}_1\bar{x}_2$ are eliminated from the function for insufficient terms $\bar{x}_3$, $\bar{x}_4$, $\bar{x}_5$ be means of formulae

$$\bar{x}_3\Delta F^1_{\bar{x}_3} = \bar{x}_3\,F^1_{\bar{x}_3}[\bar{x}_1\bar{x}_2\,F^1_{\bar{x}_1\bar{x}_2}]; \quad x_3\Delta F^0_{\bar{x}_3} = \bar{x}_3\,F^0_{\bar{x}_3}[\bar{x}_1\bar{x}_2\,F^0_{\bar{x}_1\bar{x}_2}].$$

As the formulae show, these states are eliminated from ΔF^1 and ΔF^0 simultaneously with the aim of obtaining in F_0 a more compact group of states which, in compliance with Sholomov's test, would result in better realization of completing functions.

Then values of α^1 and α^0 are computed once again, and the next insufficient closest to the optimal term is selected. Continue this procedure until the selected defective minterms realize all permit states, i.e. until the following equality is reached:

$$\varphi_{i,H}\Delta F^1_{\varphi_{i,H}} = \varphi_{i,H}\,F^1_{\varphi_{i,H}}[\varphi_{i-1,H}\,F^1_{\varphi_{i-1,H}}] = 0.$$

A number of defective minterms might turn out to be redundant and would not be considered.

In the example under consideration, it is x_5 that is such a defective term.

For the rest of insufficient terms, the following completing functions are obtained:

$$\text{for } \bar{x}_1\bar{x}_2 \quad F^1_{\bar{x}_1,\bar{x}_2} = \bar{x}_3x_4x_5x_6 + x_3(x_4\bar{x}_5x_6 + \bar{x}_4\bar{x}_5x_6),$$

$$F^0_{\bar{x}_1,\bar{x}_2} = x_3\bar{x}_4x_5x_6,$$

$$\text{for } \bar{x}_3 \quad F^1_{\bar{x}_3} = x_1x_2x_4\bar{x}_5x_6,$$

$$F^0_{\bar{x}_3} = x_1\bar{x}_2\bar{x}_4x_5x_6,$$

$$\text{for } \bar{x}_4 \quad F^1_{\bar{x}_4} = x_1x_2x_3x_5x_6,$$

$$F^0_{\bar{x}_4} = x_1\bar{x}_2x_3\bar{x}_5x_6 + (\bar{x}_1\bar{x}_2x_3 + x_1\bar{x}_2\bar{x}_3)\,x_5x_6.$$

These functions are realized similarly to the realization of initial functions F^1 and F^0. The process begins with the function giving the simplest realization in terms of some test (that of Sholomov, for example).

Sometimes, the completing function of the last among all insufficient minterms turns out to be such a function because from it all states are excluded realizable by formerly selected terms.

The completing functions are realized either in the order of increasing complexity (with respect to appropriate tests), or in the order reverse to that of selecting insufficient minterms.

In a case where several realizations of some completing function exist, these versions are checked for realization of states of other completing functions, and the most effective function in this sense is selected. Corresponding states are eliminated from these functions to avoid redundancy. For example, for defective minterm x_3, completing functions may be realized also as follows:

$$\bar{x}_1\bar{x}_3,\ x_2\bar{x}_3,\ x_3\bar{x}_4,\ x_3\bar{x}_5.$$

From these versions either $\bar{x}_1x_3$, or $x_3\bar{x}_4$ should be selected because they realize one of the states in the completing function of defective minterm $\bar{x}_1\bar{x}_2$.

After realization of all completing functions, the following realization of given function F^1 results:

$$F^1 = \bar{x}_2x_4\bar{x}_6 + \bar{x}_1\bar{x}_6 + x_2\bar{x}_4 + \bar{x}_1\bar{x}_3 + \bar{x}_1\bar{x}_2\bar{x}_5x_6.$$

This method of decomposition is applicable to all functional components with symmetrical inputs and parameter $\tau_\varphi = 1$,* as well as to those components whose properties may be reduced so as to comply with these conditions [35].

* τ_φ is a parameter of a component, determining its structural properties [35].

With various components, application of the method above differs only in realizations of given functions with respect to either permit or forbid states, or alternately with respect to both, like in the case of NAND's and NOR's.

Synthesis for components with non-symmetrical inputs is more complicated due to a number of new problems arising here such as determination of the optimal sequence of filling inputs of components, and selection of optimal variables fed into them; determination of the optimal order of constructing input completing functions and of procedures to determine them with as great as possible allowance for the structural properties of components; consideration of constraints on the number of component's inputs, depth, fan-outs, etc.

Lack of constraints on both form of functions describing automaton operation and on the structural properties of components is of prime importance in methods for solving these problems from the viewpoint of both application and better realizations.

The corresponding methods are oriented to functions given in any form (canonical or parenthesis of any depth) and to arbitrary components given by any kind of Boolean function as well.

The differential Boolean algebra, various heuristic tests and arbitrary Boolean function transformation techniques find wide application to these problems. One of such methods for optimal realization of combinational automata is the so-called "transition table" method discussed first in [28] and later elaborated in [23] and [36].

In the transition table columns are assigned to inputs and rows are assigned to permit and forbid states of a given function. Assume that a function given by permit F^1 and forbid F° states is to be realized by an element whose structural properties are given by components φ^0 and φ^0 characterizing its permit and forbid states, respectively.

Through some heuristic test most appropriate variables are selected to be fed into element inputs [36]. This leads to some output functions $f(h)$ and $f(g)$ characterizing, respectively, permit and forbid states of the element resulting after substitution in φ^1 and φ^0 of input variable symbols for those of element inputs.

If $F \cdot \overline{f(h)} = 0$ and $F^0 \cdot \overline{f(g)} = 0$, a state of $f(h)$ will correspond to each permit state of F^1 in the state transition table, thus resulting in 1 at the element output, and a state from $f(g)$ will correspond to each forbid state, of F^0 thus giving 0 at the output. Function F will be realized non-contradictorily by a single output element.

Otherwise, to eliminate contradiction, input variables should be replaced by complementing functions.

Determination of these functions is rather a complicated task which is a good illusration to the problem of separating from the set of states characterizing a given function a subset having some preassigned properties as it was mentioned in the beginning of this paper.

If one considers a transition table column corresponding, from the one hand, to some element input y_i, and, from the other hand, to input variable x_i, the following groups of states may be identified with respect to the column:

a) states where $x_i = 1$ or $x_i = 0$ and the given function is realized at the output correctly, with the change of x_i's value changing the output and making it contradictory. This value, obviously, should be preserved in the completing function replacing x_i; and

b) states where x_i is also 1 or 0 but where function of x_i is realized at the output incorrectly, with the change of x_i's value leading to elimination of the contradiction. Like in the case above, letter x_i is obligatory, and, to avoid contradiction, this state should be replaced by a state in the opposite part of the table.

If the state under consideration differs from those in the other part of the table, changing x_i value by means of a completing function might prove useful because it decreases the "distance" between the given state and that correctly realizing the given function, thereby simplifying realization of completing functions at other inputs of the element.

With analytical representation of function and element, one has to find a function characterizing all these groups of states in order to obtain expressions for completing function.

As it was demonstrated in [23], permit and forbid states of the completing function for input y_i are determined in this case as follows:

$$F^1_{y_i} = F^1(\partial(h)_{y_i} + S(g)_{\bar{y}_i}) + F^0(\partial(h)_{\bar{y}_i} + S(h)_{\bar{y}_i})$$

$$F^0_{y_i} = F^1(\partial(h)_{\bar{y}_i} + S(g)_{y_i}) + F(\partial(h)_{y_i} + S(h)_{y_i}),$$

where $F^1_{y_i}$ and $F^0_{y_i}$ are functions characterizing permit and forbid states of the completing function for the input y_i of the output element;

h and g are functions characterizing permit and forbid states of the function realized at the element output after feeding variables at its inputs;

$\partial(h)_{y_i} = h(y_i = 1)\, g\,(y_i = 0)$ is a so called "partial" first-order derivative of function h with respect to y_i with symbol y_i substituted for variable fed into y_i's input;

$\partial(h)_{\overline{y}_i} = h(y_i = 0)\, g(y_i = 1)$ is partial first-oder derivative of function h with respect to $\overline{y}_i$ under the same conditions;

$\partial(g)_{y_i}$ and $\partial(g)_{\overline{y}_i}$ are first-order derivatives of function g with respect to y_i and $\overline{y}_i$ defined in the similar manner;

$S(h)_{y_i} = \bigvee_{j=2}^{q} \partial^j(h)_{y_i}$ is disjunction of derivatives of function h with respect to y_i from 2-nd to qth order; and

$S(h)_{\overline{y}_i}$, $S(g)_{y_i}$ and $S(g)_{\overline{y}_i}$ are similar disjunctions of derivatives of h with respect to $\overline{y}_i$, and of g with respect to y_i and $\overline{y}_i$.

Recently, the above-described results of the finite automata theory have found application to the design of sequential automata [37–39]. They are also used in other problems such as analysis of statistical and dynamical races, synthesis of reliable discrete circuits, design of modular discrete circuits, construction of dynamical and checking tests, etc. (Methods for solving these problems and other problems of modern finite automata theory are reviewed in [40].)

References

1. *Erenfest, P.*, Critique of the book by L. Kutyura "Algebra of Logic", ZhRFKhO, Phys. Sect., 1910, **42**, *10*.
2. *McCulloch, W., Pitts, W.*, A logical calculus of the ideas immanent in nervous activity, Bull. of Math., Biophysics. **5**, 1943.
3. *Kleene, S. C.*, Representation of events in nerve sets, Automata Studies, 1953, pp. 3—41.
4. *Shannon, C.*, A symbolic analysis of relay and switching circuits, AJEE Trans, 1938, **57**, p. 713.
5. *Shestakov, V.*, Algebra dvukhpolyusnykh skhem postroennykh iskluchitel'no iz dvukhpolyusnikov (Algebra A-skhem) (Algebra of dipole circuits constructed of dipoles only: A-circuit Algebra), Avtomatika i Telemekhanika, 1941, *2*.
6. *Nakashima, A., Hanzava, M.*, Theory of equivalent transformation of simple partial paths in relay circuit, Nippon Electr. Comm. Engineers, 1938, *9*.
7. *Gavrilov, M. A.*, Teoriya releino-kontaktnikh skhem (Theory of relay networks), Academy of Sciences Publ., 1950.
8. *Gavrilov, M. A.*, A survey of research in the theory of relay networks in the USSR, Harvard Univ. Computation Laboratory Annals, 1959, **28**.
9. *Gavrilov, M.*, Sovremennoe sostoyanie teorii releinykh ustroistv i konechnykh avtomatov (State-of-art in the theory of relay networks and finite automata). Trudy Mezhdunarodnogo Symposiuma po teorii releinykh ustroistv i konechnykh avtomatov (IFAC), Teoriya konechnykh i veroyatnostnykh avtomatov, Nauka Publ., 1965.
10. *Kautz, W.*, IEEE Transactions on Electronic Computers, Apr. 1966, **EC-15, 2**.
11. *Moisil, G.*, Razvitie teorii tranzistornykh i drugikh beskontaknykh skhem v Rumynskoi Narodnoi Respublike (Development of the theory of transistor and other contactless networks in the Roumanian Peoples Republic), Trudy Mezhdunarodnogo Simposiuma po teorii releinykh ustroistv i konechnykh avtomatov (IFAC), Teoriya konechnykh veroyatnostnykh avtomatov. Nauka Publ., 1965.
12. *Gavrilov, M.*, Metody sinteza releino-kontaktnykh skhem (Methods for synthesis of relay circuits), Elektrichestvo, 1946, *2*.
13. *Caldwell, S.*, Switching Circuits and Logical Design, Willey, 1958.

14. *Gavrilov, M.*, Structural redundancy and reliability of relay circuits, Proc. of the First Intern. Congress of IFAC, Butterworth, London.
15. *Rabin, M. O.*, Probabilistic automata, Inf. and Control, 1963, **6**, *3*.
16. *Rabin, M., Scott, D.*, Finite automata and their decision problems, IBM J. Res. Develop., 1959, **3**, pp. 114—125.
17. *Brzozowski, J. A.*, A survey of regular expressions and their application, IRE Trans, 1962, **EC-11**, *3*, pp. 324—335.
18. *Gill, A.*, Cascades of finite-state machines, IRE Trans. on Electrical Computers, Sept., 1961, **EC-10**, *3*.
19. *Gavrilov, M.*, Sintez tablits perekhodov metodom obobchennykh sostoyanii vykhodov (Synthesis of transition tables by generalized input state method), Avtomatika i telemekhanika 1967, *1*.
20. *Parkhomenko, P.*, Sintez struktur releinykh ustroistv metodom zameni vykhodnykh peremennykh (Synthesis of relay networks by input variable replacement), Avtomatika i telemekhanika, 1967, *1*.
21. *Gavrilov, M.*, Problemy napravlennogo poiska minimal'nykh struktur releinykh ustroistv (Problems in the directed search of minimal relay networks) in "Vychislitel'nye Sistemy", Trudy I Vsesoyuznoi konferentsii po vychislitel'nym sistemam, Publ. by Inst. of Mathematics, Siberian Section of Academy of Sciences, USSR.
22. *Akers, S.*, On a theory of Boolean function, Journ. Soc. Ind. and Appl. Math., 1959, **7**, *4*.
23. *Bazarbaeva, T. et al.*, Proektirovanie struktur releinykh ustroistv s uchetom ikh nadezhnosti i nekotorykh osobennostei strukturnoi realizatsii (Relay network design allowing for its reliability and some properties of structural implementation) "Diskretnye Sistemy", vol. 2, Zinatne Publ., 1974.
24. *Naughton, R., Mitchell, B.*, The minimality of rectifier nets with multiple incompletely specified. Journ. of the Franklin Inst., Dec. 1957, **246**, pp. 457—480.
25. *Toropov, N. R.*, Priblizhonnyi algoritm minimizatsii sistem slaboopredelennykh Bulevykh funktsii (An approximate algorithm for minimization of sets of weakly defined Boolean functions), Izvestiya AN SSSR, Tekhnicheskaya Kibernetika, 1969, *1*.
26. *Pupyrev, E. I.*, Dve programmy minimizatsii sistem Bulevykh funktsii (Two programs for minimization of Boolean function sets), Informatsionnye Materialy, 1971, **7** (54).
27. *Markovskii, A. V.*, Minimizatsiya sistem Bulevykh funktsii v klasse DNF (Minimization of Boolean function sets in the DNF class) in "Abstraktnaya i strukturnaya teoriya releinykh ustroistv", Nauka Publ., Moscow, 1972.
28. *Gavrilov, M. A., Kopylenko, V. M.*, Metod perekhodnykh tablits sinteza mnogovykhodnykh kombinatsionnykh struktur na proizvol'nykh elementakh (Transition table method for synthesis of conbinational networks with arbitrary elements), Published by the Order of Lenin Inst. of Control Sciences, 1970.
29. *Sholomov, L. A.*, Kriterii slozhnosti Bulevykh funktsii (Complexity tests for Boolean Functions), in Problemy Kibernetiki, NAUKA Publ., 1966, **17**.
30. *Gavrilov, M. A.*, Problemy poiska minimal'nukh reshenii pri sinteze struktury releinykh ustroistv (Problems in the search of minimal solutions for synthesis of relay networks), Trudy II Vsesoyuznogo soveschaniya po avtomatischeskomu upravleniyu (tekhnicheskoi kibernetike), Nauka Publ., 1965.
31. *Kopylenko, V. M.*, Sintez struktur releinykh ustroistv pri zadanii iskhodnykh logicheskikh funktsii v proizvol'noi skobochnoi forme (Relay network synthesis with original functions presented in arbitrary parenthesis form), in Proektirovanie informatsionno-logicheskikh ustroistv dlya irrigatsii, ILIM, Frunze, 1973.
32. *McClusue, E.*, The minimization of Boolean functions, Bell Syst. Tech. J., 1956, **35**, pp. 1467.
33. *W. Quine*, A way to simplifying truth function, American Mathematic Monthly, Nov. 1955, **LXA**, *9*.
34. *Roth, J.*, Algebraic topological methods. Trans. of Amer. Math. Society, July 1958, **88, 2**.
35. *Gavrilov, M. A., Kopylenko, V. M.*, Sintez struktury beskontaktnykh releinykh ustroistv (Synthesis of contactless relay networks) Bulleten Mathem. de la Societe des Science Mathematical de la Republic Social de Rumanie. 1967, **10** (68), *3*.

36. *Gavrilov, M. A.*, Dekompozitsiya kombinatsionnykh avtomatov (Combinational automata decomposition) in Doklady seminara "Struktura i metodologicheskie rekomendatsii po sozdaniyu sistem avtomatizirovannogo proektirovaniya diskretnykh ustroistv". Published by the of Computer and Automation of Research Institute the Hungarian Academy of Sciences, 1973.
37. *Devyatkov, V. V., Potekhin, A. I.*, Intervalnii metod analiza povedeniya avtomata, zadannogo funktsiyami vozbugdeniya i vikhodov (Interval method for analysis of behaviour of automaton given a set of excitation and output functions), Avtomatika i telemekhanika, 1972, **5**.
38. *Lipatnikov, V. I., Devyatkov, V. V.*, Minimizatsiya nepolnost'yu opredelonnogo avtomata zadannogo funktsiyami vozbuzhdeniya i vykhoda s ispol'zovaniem interval'nogo zadaniya otnosheniya razlichimosti (Minimization of incompletely defined automaton given by its excitation and outpüt functions through interval representation of descernability relation) in Avtomatizatsiya proektirovaniya tsifrovykh upravlyayushchikh ustroistv, Politekhnicheskii Institut, Chelyabinsk, 1973.
39. *Lipatnikov, V. I., Devyatkov, V. V.*, Ispol'zovanie Bulevykh funktsii dlya analiza diskretnykh ustroistv pri avtomatizirovannom proektirovanii (Application of Boolean functions to analysis of discrete networks in automated design) in Diskretnye sistemy, vol. 3, "Avtomatizatsiya sinteza", Zinatne Publ., Riga, 1974.
40. Diskretnye sistemy (Discrete systems). International Symposium of IFAC. vols. 1, 2, 3, 4, 5, Zinatne Publ., Riga, 1974.

M. A. Gavrilov
Institute of Control Problems
Profsoyuznaya 81, V-485, Moscow,
USSR

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

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

Bognár, G.: The Hungarian–Soviet scientific and technical cooperation 3

Pugachev, V. S.: A survey of statistical control theory 13

Csáki, F.: The development of the work connections between the Soviet and the Hungarian research schools in the field of control and information theory 33

Emelyanov, S. V., Ozernoi, V. M.: Decision-making in multi-objective problems: A survey 51

Csibi, S.: Results of co-work in information theory and recognition problems, advances in the development of computer assistance to heart and vascular system examinations 65

Gavrilov, M. A.: Development of the finite automata theory in the discrete circuit design 77