316.920

VOL. 10 * NUMBER 1
TOM ТОМ НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1981

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

Is an international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Roumania, Viet-Nam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H–1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford 0X3 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1980 Subscription Rate US Dollars 94.00 per annum including postage and insurance

316.920

PROBLEMS OF CONTROL AND INFORMATION THEORY VOL. 10, 1981

SUBJECT INDEX

Bakule, L.: On a new hierarchical approach to the large-scale system solution. Vol. **10,** No. *5*, pp. 353–361

Bíla, J.: Modeling of morphological and functional structure in the symbolical system for the automatic synthesis of technical object. Vol. **10,** No. *3*, pp. 191–198

Bokor, J., Pető, J.: Minimal realisation procedure from partial function representation of a rational transfer matrix. Vol. **10,** No. *6*, pp. 403–415

Boltyansky, V. G.: Methods of optimization applied to the relativistic theory. Vol. **10,** No. *4*, pp. 219–235

Dumer, I. I., Zinov'ev, V. A., Zyablov, V. V.: Cascaded decoding according to minimal generalized distance. Vol. **10,** No. *1*, pp. 3–19

Dyachkov, A. G., Rykov, V. V.: On lower bound of the length of B_4-code. Vol. **10,** No. *5*, pp. 301–307

Dynkin, V. N., Katsman, G. L., Kurdyukov, A. P.: Lower bound of the region of acceptable code rates fro a summary channel with correlated sources. Vol. **10,** No. *5*, pp. 309–316

Galló, V.: An experimental recognition system using pattern grammars. Vol. **10,** No. *1*, pp. 21–31

Grim, J.: An algorithm for maximizing a finite sum of positive functions and its application to cluster analysis. Vol. **10,** No. *6*, pp. 427–437

Györfi, L.: Recent results on nonparametric regression estimate and multiple classification. Vol. **10,** No. *1*, pp. 43–52

Györfi, Z., Molnár, L., Osváth, L., Pomozi, A.: Real-time digital filtering of electrocardigraphic signals by means of an INTEL 8080 microprocessor. Vol. **10,** No. *2*, pp 137–147

Hinkel, H.: An application of adaptive Kalman filtering to short-term load forecasting for power systems. Vol. **10,** No. *5*, pp. 329–339

Jereb, L., Szabó, Cs.: A simple asynchronous reservation protocol for a multiple access channel. Vol. **10,** No. *6*, pp. 417–425

Kowalewski, A., Kotarski, W.: On the application of conical approximations to an optimal control problem for systems described by partial differential equations of parabolic type with time delay. Vol. **10,** No. *5*, pp. 341–351

Kučera, V., Šebek, M.: Linear optimal dynamic regulators. Vol. **10,** No. *3*, pp. 181–189

Kurzhanskii, A. B.: Dynamic control system estimation under uncertainty conditions, II. Vol. **10,** No. *1*, pp. 33–42

Kusolitsch, N: On a re-sampling problem in the theory of information. Vol. **10,** No. *2*, pp. 129–136

Kuznetsov, N. A., Yashin, A. I.: On consistent estimation in adaptive filtering. Vol. **10,** No. *5*, pp. 317–327

Outrata, J. V., Schindler, Z.: An augmented Lagrangian method for a class of convex continuous optimal control problems. Vol. **10,** No. *2*, pp. 67–81

Outrata, J. V., Jarusek, J., Beran, Z.: An application of the augmented Lagrangian approach to the optimal control of a biharmonic system with state-space constraints. Vol. **10,** No. *6*, pp. 363–373

Petkovski, D., Rakić, M.: The compensation of unmeasurable disturbances via proportional-plus-integral regulators with a prescribed degree of stability. Vol. **10,** No. *3*, pp. 199–212

Pík, J.: Syntactic recognition of nonstylized handwritten numerals. Vol. **10,** No. *3*, pp. 149–160

Pugachev, V. S.: The finite-dimensional distribution of a random process determined by a stochastic differential equation and their application to control problems. Vol. **10,** No. *2*, pp. 95–114

Sinha, N. K., Rózsa, P.: Algorithm for transformation of a polynomial matrix to canonical form. Vol. **10,** No. *3*, pp. 175–180

Sinha, N. K., Rózsa, P.: Some system theory applications of Egerváry's algorithm for transformation of a matrix to the Hermite normal form. Vol. **10,** No. *4*, pp. 267–277

Spal, J.: Modified fuzzy sets. Vol. **10**, No. *6*, pp. 375–386

Stepanov, S. N.: Probabilistic characteristics of an incompletely accessible multi-phase service system with several types of repeated calls. Vol. **10,** No. *6*, pp. 375–386

Tuzar, A., Beran, Z.: Optimization of the discrete description for continuous systems. Vol. **10,** No. *2*, pp. 83–94

Visek, J. Á.: Bounds of the asymptotic rate of convergence of the error probabilities of a robust test. Vol. **10,** No. *4*, pp. 253–266

Wertz, W.: A remark on the estimation of functionals of a probability density. Vol. **10,** No. *4*, pp. 279–285

Yemelyanov, S. V., Korovin, S. K.: Development of feedback types and their application to design of closed-loop dynamic systems. Vol. **10,** No. *3*, pp. 161–174

Yemelyanov, S. V., Korovin, S. K., Sizikov, V. I.: Use of coordinate-parametric feedback in design of control systems. Vol. **10,** No. *4*, pp. 237–251

Zilinkas, A. G.: On statistical models of multimodal functions and their application for the construction of the minimization algorithms. Vol. **10,** No. *1*, pp. 53–65

AUTHOR INDEX

Bakule, L. vol. **10,** No. *5*, pp. 353–361
Beran, Z. vol. **10,** No. *6*, pp. 363–373
Bíla, J. vol. **10,** No. *3*, pp. 191–198
Bokor, J. vol. **10,** No. *6*, pp. 403–415
Boltyansky, V. G. vol. **10,** No. *4*, pp. 219–235

Dumer, I. I. vol. **10,** No. *1*, pp. 3–19
Dyachkov, A. G. vol. **10,** No. *5*, pp. 301–307
Dynkin, V. N. vol. **10,** No. *5*, pp. 309–316

Galló, V. vol. **10,** No. *1*, pp. 21–31
Grim, J. vol. **10,** No. *6*, pp. 427–437
Györfi, L. vol. **10,** No. *1*, pp. 43–52
Györfi, Z. vol. **10,** No. *2*, pp. 137–147

Hinkel, H. vol. **10,** No. *5*, pp. 329–339

Jarusek, J. vol. **10,** No. *6*, pp. 363–373
Jereb, L. vol. **10,** No. *6*, pp. 417–425

Katsman, G. L. vol. **10,** No. *5*, pp. 309–316
Korovin, S. K. vol. **10,** No. *3*, pp. 161–174; vol. **10,** No. *4*, pp. 267–277
Kotarski, W. vol. **10,** No. *5*, pp. 341–351
Kowalewski, A. vol. **10,** No. *5*, pp. 341–351
Kučera, V. vol **10,** No. *3*, pp. 181–189
Kurzhanskii, A. B. vol. **10,** No. *1*, pp. 33–42
Kusolitsch, N. vol. **10,** No. *2*, pp. 129–136
Kuznetsov, N. A. vol. **10,** No. *5*, pp. 317–327

Osváth, L. vol. **10,** No. *2*, pp. 137–147
Outrata, J. V. vol. **10,** No. *2*, pp. 67-81; vol. **10,** No. *6*, pp. 363–373

Petkovski, D. vol. **10,** No. *3*, pp. 199–212
Pik. J. vol. **10,** No. *3*, pp. 149–160
Pomozi, A. vol. **10,** No. *2*, pp. 137–147
Pugachev, V. S. vol. **10,** No. *2*, pp. 95–114

Rakic, M. vol. **10,** No. *3*, pp. 199–212
Rózsa, P. vol. **10,** No. *3*, pp. 175–180; vol. **10,** No. *4*, pp. 267–277
Rykov, V. V. vol. **10,** No. *5*, pp. 301–307

Schindler, Z. vol. **10,** No. *2*, pp. 67–81
Šebek, M. vol. **10,** No. *3*, pp. 181–189
Sinha, N. K. vol. **10,** No. *3*, pp. 175–180; vol. **10,** No. *4*, pp. 267–277
Spal, J. vol. **10,** No. *6*, pp. 375–386
Stepanov, S. N. vol. **10,** No. *6*, pp. 387–401
Szabó, Cs. vol. **10,** No. *6*, pp. 417–425

Tuzer, A. vol. **10,** No. *2*, pp. 83–94

Visek, J. Á. vol. **10,** No. *4*, pp. 253–266

Wertz, W. vol. **10,** No. *4*, pp. 279–285

Yashin, A. I. vol. **10,** No. *5*, pp. 317–327
Yemelyanov, S. V. vol. **10,** No. *3*, pp. 161–174; vol. **10,** No. *4*, pp. 267–277

Zilinkas, A. G. vol. **10,** No. *1*, pp. 53–65
Zinov'ev, V. A. vol. **10,** No. *1*, pp. 3–19
Zyablov, V. V. vol. **10,** No. *1*, pp. 3–19

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)

G. BOGNÁR (Budapest)

COORDINATING EDITORS

S. V. EMELYANOV (Moscow)
N. N. KRASOVSKII (Sverdlovsk)
E. P. POPOV (Moscow)
V. S. PUGACHEV (Moscow
V. I. SIFOROV (Moscow)
E. D. TERYAEV (Moscow)
T. VÁMOS (Budapest)
L. VARGA (Budapest)
A. PRÉKOPA (Budapest)
S. CSIBI (Budapest)
I. CSISZÁR (Budapest)
J. KOCSIS (Budapest)
J. BENEŠ (Prague)
V. STREJC (Prague)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)

Г. БОГНАР (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

С. В. ЕМЕЛЬЯНОВ (Москва)
Н. Н. КРАСОВСКИЙ (Свердловск)
Е. П. ПОПОВ (Москва)
В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Е. Д. ТЕРЯЕВ (Москва)
Т. ВАМОШ (Будапешт)
Л. ВАРГА (Будапешт)
А. ПРЕКОПА (Будапешт)
Ш. ЧИБИ (Будапешт)
И. ЧИСАР (Будапешт)
Я. КОЧИШ (Будапешт)
И. БЕНЕШ (Прага)
В. СТРЕЙЦ (Прага)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 10 (1), pp. 3—19 (1981)

КАСКАДНОЕ ДЕКОДИРОВАНИЕ ПО МИНИМУМУ ОБОБЩЕННОГО РАССТОЯНИЯ

И. И. ДУМЕР, В. А. ЗИНОВЬЕВ, В. В. ЗЯБЛОВ

(Москва)

(Поступила в редакцию 30 июля 1979 г.)

Рассмотрено декодирование двоичных обобщенных каскадных кодов в канале с непрерывным выходом. Предложен алгоритм каскадного декодирования этих кодов по минимуму обобщенного расстояния. Исследованы реализуемые корректирующие свойства для различных систем декодирующих порогов. При наличии достаточного числа порогов предложенный метод реализует гарантированное кодовое расстояние обобщенных каскадных кодов. Полученные результаты могут быть перенесены на дискретные каналы и дают новый алгоритм декодирования обобщенных каскадных кодов.

1. Введение

Рассмотрим декодирование по минимуму обобщенного расстояния (МОР) [1] при передаче двоичных символов по каналу с аддитивным непрерывным шумом. На приемном конце после демодулятора символы принятого слова $y = (y_1, \ldots, y_n)$ поступают на детектор правдоподобия, который на выходе выдает для каждого символа y_i не пару условных вероятностей $P(y_i \mid 1)$ и $P(y_i \mid 0)$, а логарифм отношения правдоподобия f_i:

$$f_i = \ln \frac{P(y_i \mid 1)}{P(y_i \mid 0)},$$

где $P(y_i \mid a)$, $a = 0, 1$ — вероятность символа y_i при передаче a. Пусть (a, b) обозначает скалярное произведение векторов a и b в поле действительных чисел (в дальнейшем всегда будем считать, что векторы двоичного кода состоят из символов -1 и $+1$; 0 заменяется на $+1$, а 1 заменяется на -1). Как известно [1], при передаче кодового слова c по каналу без памяти вероятность $P(y \mid c)$ максимальна, если и только если максимально значение скалярного произведения (f, c), где $f = (f_1, \ldots, f_n)$. Следующий за детектором ограничитель имеет характеристику $q(x)$,

$$q(x) = \begin{cases} 1, & x \geq T, \\ \dfrac{x}{T}, & -T \leq x \leq T, \\ -1, & x \leq -T, \end{cases}$$

где T — некоторый порог. В результате получаем вектор z равный $(q(f_1), \ldots, q(f_n))$. Ограничение $-1 \leq q(x) \leq 1$ позволяет доказать теорему, на которой основано декодирование по критерию МОР.

Теорема 1 (Форни [1]). В коде длины n с минимальным расстоянием d существует не более одного кодового вектора x, для которого

$$(z, x) > n - d, \tag{1.1}$$

где z — слово, принятое после ограничения.

Доказательство теоремы 1 основано на следующем факте, который нам будет нужен в дальнейшем.

Лемма 1. Пусть для некоторого вектора x_1 из кода C с расстоянием d

$$(z, x_1) > n - m,$$

где z — произвольный вектор, полученный после ограничителя, а $m \leq d$. Тогда для любого вектора $x_2 \in C$, $x_2 \neq x_1$,

$$(z, x_2) < n - 2d + m.$$

Доказательство. Хотя доказательство этого факта, по существу, имеется у Форни [1; стр. 70], приведем его, так как эти рассуждения нам понадобятся в дальнейшем. Пусть x_2 — кодовое слово, отличающееся от x_1 в d позициях, то есть

$$x_{1,i} \neq x_{2,i}, \quad i \in S, \quad |S| = d.$$

Тогда

$$(z, x_1) = \sum_{i \notin S} z_i x_{1,i} + \sum_{i \in S} z_i x_{1,i} = m_1 + m_2,$$

$$(z, x_2) = \sum_{i \notin S} z_i x_{2,i} + \sum_{i \in S} z_i x_{2,i} = m_1 - m_2.$$

Но по условию

$$m_1 + m_2 > n - m \quad \text{и} \quad m_1 \leq n - d,$$

откуда и следует утверждение леммы.

Теорема 1 верна при любой характеристике ограничителя $-1 \leq q(x) \leq 1$. В частности, она справедлива в случае ошибок и стираний, которые соответствуют следующей характеристике ограничителя:

$$q(x) = \begin{cases} +1, & x \geq T, \\ 0, & -T < x < T, \\ -1, & x \leq -T. \end{cases}$$

Оказалось, что декодирование с исправлением ошибок и стираний можно использовать для декодирования по МОР, если справедливо условие (1.1). Пусть для любого $z = (z_1, \ldots, z_n)$ индексы $i_1, i_2, \ldots, i_n$ упорядочены в порядке возрастания надежности

$$|z_{i_1}| \leq |z_{i_2}| \leq \ldots \leq |z_{i_n}|.$$

Вектору z поставим в соответствие вектор $Z^{(k)}$, $k = 0, 1, \ldots, n$ из элементов 0, $+1$, -1, полученный (стиранием k самых ненадежных символов и заменой оставшихся символов на ± 1) в соответствии со следующей характеристикой ограничителя:

$$q_k(z_{i_j}) = \begin{cases} 0, & 1 \leq j \leq k,\ 0 \leq k \leq n, \\ 1, & z_{i_j} \geq 0,\ k+1 \leq j \leq n, \\ -1, & z_{i_j} < 0,\ k+1 \leq j \leq n. \end{cases}$$

Тогда вектор

$$z^{(k)} = (q_k(z_1), \ldots, q_k(z_k))$$

имеет k координат равных 0, а остальные координаты равны ± 1. Следующая теорема гарантирует, что если для некоторого слова x выполняется условие (1.1), то по крайней мере для одного вектора $z^{(k)}$ при $k < d$ декодер с исправлением ошибок и стираний найдет слово x.

Теорема 2 (Форни [1]). Если $(z, x) > n - d$, то для некоторого $k < d$ имеет место неравенство

$$(z^{(k)}, x) > n - d. \tag{1.2}$$

Как нетрудно видеть, условие (1.2) означает, что в векторе $z^{(k)}$ имеется k стираний и e_k ошибочных символов таких, что $2e_k + k < d$. Таким образом, необходимо не более d попыток декодирования векторов $z^{(k)}$, $k = 0, \ldots, d-1$ с исправлением ошибок и стираний, чтобы найти вектор x.

Как следует из результатов Форни [1] декодирование по МОР применимо к любому коду, для которого известен алгоритм исправления ошибок и стираний, гарантируемый его кодовым расстоянием. Тем самым этот алгоритм декодирования применим и к каскадному коду с каскадным декодированием, реализующим его кодовое расстояние [2]. Такое декодирование будем называть декодированием каскадного кода по МОР. При таком декодировании пропорционально кодовому расстоянию d увеличивается сложность декодирования, т. е. $d/2$ раз нужно провести декодирование кода длины $n = n_a n_b$ (где n_a — длина внешнего кода, n_b — длина внутреннего кода). Более целесообразным представляется при каскадном декодировании каскадного кода

декодировать по МОР только внутренние коды. Это декодирование будем называть каскадным декодированием по МОР, и наша задача — исследовать реализуемые корректирующие свойства при таком декодировании. Предварительные результаты по этому вопросу были опубликованы в [3].

2. Каскадное декодирование обобщенных каскадных кодов по МОР

Кратко опишем обобщенные каскадные (сокращенно ОК-коды) коды (подробнее см. [4], обозначениям которой будем следовать). Обозначим через $A(q, n, d, N)$ q-ный код длины n с минимальным расстоянием d и мощности N. Для m кодов $A_i = A_i(q_i, n_a, d_{a,i}, N_{a,i})$ одинаковой длины n_a (коды A_i называются внешними) введем множество матриц $\{M = M(A_1, \ldots, A_m)\}$, где матрица M размера $m \times n_a$ получена следующим образом. В i-ой строке ее находится произвольный кодовый вектор $a^{(i)} = (a_1^{(i)}, \ldots, a_{n_a}^{(i)})$ из кода A_i, $i = 1, \ldots, m$. Пусть теперь имеется код $B^{(1)} = B^{(1)}(q, n_b, d_{b,1}, N_{b,1})$ мощности $N_{b,1} = q_1 q_2 \ldots q_m$, который разбивается на q_1 кодов $B_{i_1}^{(2)}(q, n_b, d_{b,2}, N_{b,2})$, где $N_{b,2} = q_2 \ldots q_m$, а $i_1 = 0, 1, \ldots, q_1 - 1$. В свою очередь пусть каждый из кодов $B_{i_1}^{(2)}$ разбивается на q_2 кодов $B_{i_1, i_2}^{(3)} = B_{i_1, i_2}^{(3)}(q, n_b, d_{b,3}, N_{b,3})$, $N_{b,3} = q_3 \ldots q_m$, и т. д., всего m шагов подобных разбиений. На ϱ-ом шаге разбиения имеем

$$B_{i_1, \ldots, i_{\varrho-1}}^{(\varrho)}(q, n_b, d_{b,\varrho}, N_{b,\varrho}) = \bigcup_{i_\varrho = 0}^{q_\varrho - 1} B_{i_1, \ldots, i_\varrho}^{(\varrho+1)}(q, n_b, d_{b,\varrho+1}, N_{b,\varrho+1})$$

где

$$N_{b,\varrho} = q_\varrho \ldots q_m,\ N_{b,\varrho+1} = q_{\varrho+1} \ldots q_m,\ \varrho = 1, \ldots, m,\ q_{m+1} = N_{b,m+1} = 1,\ d_{b,m+1} = \infty.$$

Посредством такого разбиения можно устанавливать нумерацию кодовых векторов. Обозначим через $b(i_1, \ldots, i_m)$ ($(i_1, \ldots, i_m)$ — номер) кодовый вектор $B_{i_1, \ldots, i_m}^{(m+1)}$. Коды $B^{(i)}$ называются внутренними кодами.

На основе m внешних кодов A_i и m внутренних кодов $B^{(i)}$ строим ОК-код. В матрице $M(A_1, \ldots, A_m)$ каждый j-ый столбец $(a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(m)})$, $j = 1, \ldots, n_a$ заменяем на кодовый вектор $b(a_j^{(1)}, \ldots, a_j^{(m)})$. Полученную матрицу размера $n_b \times n_a$ из символов $0, 1, \ldots, q-1$ примем за кодовое слово ОК-кода $\{C\}$ порядка m. Ясно, что длина полученного кода $n = n_a n_b$, а мощность $N = N_{a,1} N_{a,2} \ldots N_{a,m}$. Пусть по определению

$$d_i = d_{a,i}\, d_{b,i}, \tag{2.1}$$

$$d^{(\text{н})} = \min_i \{d_i\}. \tag{2.2}$$

Как известно [4], кодовое расстояние d ОК-кода удовлетворяет неравенству: $d \geq d^{(\text{н})}$.

Перейдем теперь к декодированию по МОР двоичного ОК-кода $\{C\}$. Пусть в результате кодирования матрицы $M(A_1, \ldots, A_m)$ со строками $a^{(1)}, \ldots, a^{(m)}$ получено слово c, которое затем передано по каналу с непрерывным шумом. Пусть z — слово, принятое после ограничителя, $z_1, \ldots, z_{n_a}$ — его столбцы, $b_1, \ldots, b_{n_a}$ — столбцы переданного слова c. Согласно теореме 1, если

$$(c, z) > n - d^{(\text{н})}, \tag{2.3}$$

то, проведя $d^{(\text{н})}$ см. (2.1)—(2.2) попыток декодирования кодом $\{c\}$ длины n с исправлением ошибок и стираний, получим слово c. Ниже будет показано, что при выполнении (2.3) возможно каскадное декодирование по МОР. Оно осуществляется исправлением ошибок и стираний кодами $B^{(i)}$ длины n_b и кодами A_i длины n_a.

Алгоритм каскадного декодирования, который обозначим через $\psi(m)$, осуществляется последовательно за m шагов. На первом шаге определяется строка $a^{(1)} \in A_1$, т. е. первая строка матрицы M, на втором шаге определяется $a^{(2)} \in A_2$ и т. д. Наконец, на m-ом шаге определяется строка $a^{(m)} \in A_m$ и тем самым декодирование закончено.

На i-ом шаге декодирования известны строки $a^{(1)}, a^2, \ldots, a^{(i-1)}$ матрицы M. Столбец b_j матрицы c, $j = 1, \ldots, n_a$, принадлежит коду $B^{(i)}_{a^{(1)} \ldots a^{(i-1)}}$, который в дальнейшем будем обозначать для краткости $B^{(i)}_j$. Введем следующие определения. Пусть для некоторого натурального числа M_i задано множество

$$T_i = \{t_1^{(i)}, \ldots, t_{M_i}^{(i)}\}$$

из M_i действительных чисел таких, что $d_{b,i} = t_1^{(i)} > t_2^{(i)} > \ldots > t_{M_i}^{(i)} > 0$ (нам удобней обозначать d_{bi} через $t_1^{(i)}$). Пусть $B_j^{(i)}(d)$ означает следующее декодирование по МОР вектора z_j кодом $B_j^{(i)}$. Результатом его является вектор $\hat{b}_j \in B_j^{(i)}$ и число $r_j^{(i)} = n_b - (\hat{b}_j, z_j)$, если $(\hat{b}_j, z_j) > n_b - d_{b,i}$. В противном случае результатом декодирования является стирание; в этом случае положим, что $r_j^{(i)} = d_{b,i}$. Таким образом, параметр $r_j^{(i)}$ оценивает снизу расстояние, задаваемое скалярным произведением между принятым столбцом z_j и ближайшим словом $\hat{b}_j$ внутреннего кода $B_j^{(i)}$. В результате декодирования $B_j^{(i)}(d)$ всех n_a столбцов z_j, $j = 1, \ldots, n_a$, получаем строку $R^{(i)} = (r_1^{(i)}, \ldots, r_{n_a}^{(i)})$.

Реализация i-го шага декодирования осуществляется по следующей схеме:

$$z \to [\hat{a}^{(i)}, R^{(i)}] \to \{a^{(i)}(k), S^{(i)}(k)\} \to a^{(i)}.$$

При переходе $z \to [\hat{a}^{(i)}, R^{(i)}]$ проводим вначале декодирование $B_j^{(i)}(d)$, $j = 1, \ldots, n_a$, столбцов z_j матрицы z. В результате получаем столбцы $\hat{b}_j$ или стирания, а также строку $R^{(i)}$. Далее столбцы $\hat{b}_j$ для всех $j = 1, \ldots, n_a$ за-

меняем на их номера $(a_j^{(1)}, \ldots, a_j^{(i-1)}, \hat{a}_j^{(i)}, \ldots, \hat{a}_j^{(m)})$, а стертые столбцы заменяем на стертые номера. Отметим, что первые $i-1$ компонент номеров найдены верно, так как предполагается, что первые $i-1$ строк матрицы M декодированы правильно. В результате получаем строку $\hat{a}^{(i)} = (\hat{a}_1^{(i)}, \ldots, \hat{a}_{n_a}^{(i)})$, состоящую из i-ых компонент номеров и стираний.

При переходе $[\hat{a}^{(i)}, R^{(i)}] \to \{a^{(i)}(k), S^{(i)}(k)\}$ проводится декодирование ошибок и стираний в строке $\hat{a}^{(i)}$ с порогами t_k, $k = 1, \ldots, M_i$. Пусть $A^{(i)}(k)$, $k = 1, \ldots, M_i$ означает следующее декодирование строки $\hat{a}^{(i)}$. Вначале стираем в $\hat{a}^{(i)}$ все символы $\hat{a}_j^{(i)}$, для которых

$$r_j^{(i)} \geq t_k . \tag{2.4}$$

Полученную строку обозначим $\hat{a}^{(i)}(k)$ (отметим, что $\hat{a}^{(i)}(1) = \hat{a}^{(i)}$). Затем производим декодирование строки $\hat{a}^{(i)}(k)$ кодом A_i с исправлением ошибок и стираний. Если такое декодирование возможно, то в результате получаем некоторую строку $a^{(i)}(k) \in A_i$. Пусть $\{I^{(i)}(k)\}$ — множество тех позиций j, $j = 1, \ldots, n_a$, в которых строки $a^{(i)}(k)$ и $\hat{a}^{(i)}$ содержат различные нестертые символы (как отмечалось, в строке $\hat{a}^{(i)}$ могут находиться стирания). Иначе, если $\{V^{(i)}\}$ обозначает множество стертых позиций в $\hat{a}^{(i)}$, то

$$\{I^{(i)}(k)\} = \{j \notin V^{(i)} \mid a_j^{(i)}(k) \neq \hat{a}_j^{(i)}\} . \tag{2.5}$$

Для любого $j = 1, \ldots, n_a$ определим:

$$p_j^{(i)}(k) = \begin{cases} r_j^{(i)}, & \text{если} \quad j \notin \{I^{(i)}(k)\}, \\ 2d_{b,i} - r_j^{(i)}, & \text{если} \quad j \in \{I^{(i)}(k)\} . \end{cases} \tag{2.6}$$

Как будет показано далее, числа $p_j^{(i)}(k)$ оценивают снизу расстояния между столбцами z_j и b_j, $j = 1, \ldots, n_a$, если строка $a^{(i)}(k)$ является истинной строкой $a^{(i)}$ передаваемого каскадного слова c. Наконец, для каждой строки $a^{(i)}(k)$ определим число $S^{(i)}(k)$ следующим образом:

$$S^{(i)}(k) = \sum_{j=1}^{n_a} p_j^{(i)}(k).$$

Проведем декодирование $A^{(i)}(k)$ строки $\hat{a}^{(i)}$ для всех $k = 1, \ldots, M_i$. В результате получим множество $\{a^{(i)}(k), S^{(i)}(k)\}$, в котором содержится не более M_i строк $a^{(i)}(k)$ для различных k.

При переходе $\{a^{(i)}(k), S^{(i)}(k)\} \to a^{(i)}$ находим истинную строку $a^{(i)}$ переданного слова c. Для этого выберем из $\{a^{(i)}(k)\}$ строку $a^{(i)}(v)$, для которой

$$S^{(i)}(v) < d_i . \tag{2.7}$$

Ниже будет показано, что при выполнении условия (2.3) такая строка существует и единственна.

Лемма 2. Пусть скалярное произведение векторов c и z удовлетворяет условию:

$$(c, z) > n - d_i. \tag{2.8}$$

Если i-ая строка переданного слова c принадлежит множеству строк $\{a^{(i)}(k)\}$

$$a^{(i)} \in \{a^{(i)}(k)\}, \tag{2.9}$$

то именно эта строка удовлетворяет неравенству (2.7), и никакая другая строка из $\{a^{(i)}(k)\}$ этому неравенству не удовлетворяет.

Доказательство леммы 2. Пусть истинная строка $a^{(i)}$ принадлежит $\{a^{(i)}(k)\}$, т. е. $a^{(i)} = a^{(i)}(v)$ для некоторого v. Покажем вначале, что $S^{(i)}(v) < d_i$. Так как $a^{(i)}(v) = a^{(i)}$, то в множество $\{I^{(i)}(v)\}$ (см. (2.5)) входят те позиции, на которых в $\hat{a}^{(i)}$ стоят ошибочные символы. Следовательно, в результате декодирования $B_j^{(i)}(d)$ столбцов z_j для всех $j \in \{I^{(i)}(v)\}$ были получены ошибочные столбцы $\hat{b}_j \neq b_j$ (здесь b_j — столбцы истинного слова c). Так как $(z_j, \hat{b}_j) = n_b - r_j^{(i)}$, то, согласно лемме 1, $(z_j, b_j) \leq n_b - 2d_{b,i} + r_j^{(i)}$, или, учитывая выражение (2.6), $(z_j, b_j) \leq n_b - p_j^{(i)}(v)$. Если $j \notin \{I^{(i)}(v)\}$ то на j-ой позиции в строке $\hat{a}^{(i)}$ стоит или верный символ или стрение. В этом случае $(z_j, b_j) \leq n_b - p_j^{(i)}(v)$. Следовательно,

$$(c, z) = \sum_{j=1}^{n_a} (z_j, b_j) \leq \sum_{j=1}^{n_a} \left(n_b - p_j^{(i)}(v)\right) = n - S^{(i)}(v),$$

откуда, сравнивая с (2.8), получаем, что $S^{(i)}(v) < d_i$, т. е. неравенство (2.7).

Покажем теперь, что если $S^{(i)}(k) \neq S^{(i)}(v)$ для некоторого k, то $S^{(i)}(k) > d_i$. Строки $a^{(i)}(v)$ и $a^{(i)}(k)$ отличаются не менее, чем в $d_{a,i}$ позициях, множество которых обозначим через $J_{\mathfrak{d},k}$. Если для $j \in J_{\mathfrak{d},k}$ символ $\hat{a}_j^{(i)}$ стерт, то $p_j^{(i)}(v) + p_j^{(i)}(k) = 2d_{b,i}$. Если для $j \in J_{\mathfrak{d},k}$ символ $\hat{a}_j^{(i)}$ не стерт, то он отличен либо от $a^{(i)}(v)$, либо от $a_j^{(i)}(k)$, т. е. $p_j^{(i)}(v) + p_j^{(i)}(k) \geq 2d_{b,i}$. Следовательно,

$$S^{(i)}(v) + S^{(i)}(k) \geq \sum_{j \in J_{\mathfrak{d},k}} \left(p_j^{(i)}(k) + p_j^{(i)}(v)\right) \geq 2d_{a,i}\, d_{b,i}.$$

Отсюда $S^{(i)}(k) > d_i$, что доказывает лемму.

Как показано ниже в теореме 3, при определенном выборе системы порогов $\{T_i\}$ условие (2.9) всегда выполняется, если справедливо ограничение (2.8). Пусть принят вектор z и на i-ом шаге декодирования получена строка «расстояний» $R^{(i)} = (r_1^{(i)}, \ldots, r_{n_a}^{(i)})$. Обозначим через $J = \{1, 2, \ldots, n_a\}$ множество номеров всех n_a столбцов, а через J_0 множество номеров тех столбцов принятого слова z, которые стираются при декодировании столбцов по МОР,

$$J_0 = \{j \in J \mid r_j^{(i)} \geq d_{b,i}\},\ J' = J \setminus J_0; \tag{2.10}$$

здесь J' — множество номеров декодируемых столбцов. Пусть $R_1^{(i)} = \{r_1, r_2, \ldots, r_{M_i}\}$ представляет собой множество попарно различных «расстояний» $r^{(i)}$ для всех $j \in J'$, упорядоченное строго по убыванию, т. е.

$$d_{b,i} > r_1 > r_2 > \ldots > r_{M_i}.$$

Очевидно, $M_i = |J'|$, если «расстояния» $r^{(i)}$ для всех $j \in J'$ попарно различны и $M_i < |J'|$ в противном случае. Для каждого принятого вектора z возьмем на i-ом шаге декодирования систему порогов $T^{(i)}(z) = \{t_1, \ldots, t_{M_i}\}$, где

$$t_1 = d_{b,i},\ t_{k+1} = r_k \text{ для всех } k = 1, \ldots, M_i - 1. \tag{2.11}$$

Иными словами, при первой попытке декодирования внешним кодом A_i стираются только столбцы с номерами из множества J_0 (эти столбцы уже стерты при декодировании внутренних кодов по МОР). При второй попытке стираются все стольбцы $j \in J'$, для которых $r_j^{(i)} = r_1$, затем стираются все столбцы с «расстоянием» r_2 и т. д. После M_i попыток нестертыми остаются только столбцы с «расстоянием», которое условимся обозначать через t_{M_i+1}.

Теорема 3. (1) Если для принятого слова z выполняется условие (2.3), то в результате m шагов декодирования слова z системами порогов $T^{(i)}(z)$, $i = 1, \ldots, m$, будет найдено переданное слово c.

(2) Пусть строки $a^{(1)}, \ldots, a^{(i-1)}$ переданного слова c определены верно. Тогда, если выполняется условие (2.8), то система порогов $T^{(i)}(z)$ обеспечивает нахождение правильной строки $a^{(i)}$.

(3) На i-ом этапе декодирования системой порогов $T^{(i)}(z)$ достаточно не более $[(d_{a,i} + 1)/2]$ порогов для нахождения правильной строки $a^{(i)}$ для всех $i = 1, \ldots, m$.

Доказательство. Утверждение (1) является тривиальным следствием утверждения (2), к доказательству которого можно перейти. Предположим, что первые $i-1$ строк $a^{(1)}, \ldots, a^{(i-1)}$ определены верно. На i-ом шаге декодирования слова z, $1 \leq i \leq m$, получены строки $\hat{a}^{(i)}$ и $R^{(i)} = (r_1^{(i)}, \ldots, r_{n_a}^{(i)})$ и проводится декодирование строки $\hat{a}^{(i)}$ системой порогов (2.11). При этом последовательно получаются строки $\hat{a}^{(i)}(k)$, $k = 1, \ldots, M_i$, которые декодируются с исправлением ошибок и стираний.

Пусть ϱ_k и e_k обозначает число стертых и ошибочных символов соответственно в строке $\hat{a}^{(i)}(k)$. Как показано ниже, величина (c, z) оценивается сверху при помощи всех ϱ_k и e_k и если $\varrho_k + 2e_k \geq d_{a,i}$ для всех $k = 1, \ldots, M_i$, то $(c, z) < n - d_i$, что противоречит условию теоремы.

Нетрудно видеть, что при переходе от порога t_k к порогу t_{k+1}, $k = 1, \ldots, M_i - 1$, дополнительно стираются $\varrho_{k+1} - \varrho_k$ столбцов, из которых $e_k - e_{k+1}$ были ошибочными. Для системы порогов (2.11) при таком переходе в стирания попадают все столбцы с номерами $j \in J'$, для которых $r^{(i)} = t_{k+1}$. Обозначим множество этих столбцов через J_k, т. е.

$$(z_j, \hat{b}_j) = n_b - t_{k+1} \quad \text{для} \quad j \in J_k, \ k = 1, \ldots, M_i .$$

Согласно лемме 1 для всех $e_k - e_{k+1}$ ошибочных столбцов из J_k мы имеем оценку $(z_j, b_j) \leq n_b - 2d_{b,i} + t_{k+1}$, а для остальных $(\varrho_{k+1} - \varrho_k) - (e_k - e_{k+1})$ столбцов из $J_k - (z_j, b_j) \leq n_b - t_{k+1}$. Итак, для всех $k = 1, \ldots, M_i - 1$

$$\sum_{j \in J_k} (b_j, z_j) \leq (e_k - e_{k+1})(n_b - 2d_{b,i} + t_{k+1}) + ((\varrho_{k+1} - \varrho_k) - (e_k - e_{k+1}))(n_b - t_{k+1}) \quad (2.12)$$

Заметим, кроме того, что при декодировании порогом $t_1 = d_{b,i}$ в стирания попадают все ϱ_1 столбцов из J_0, а после декодирования порогом t_{M_i} остаются $n_a - \varrho_{M_i}$ столбцов, из которых e_{M_i} являются ошибочными. Поэтому

$$\sum_{j \in J_0} (b_j, z_j) \leq (n_b - d_{b,i})\, \varrho_1 , \quad (2.13)$$

$$\sum_{j \in J_{M_i}} (b_j, z_j) \leq e_{M_i}(n_b - 2d_{b,i} + t_{M_i+1}) + (n_a - \varrho_{M_i} - e_{M_i})(n_b - t_{M_i+1}) . \quad (2.14)$$

Так как

$$(c, z) = \sum_{j=1}^{n_a} (b_j, z_j) = \sum_{k=0}^{M_i} \sum_{j \in J_k} (b_j, z_j),$$

то из (2.13)–(2.14) после элементарных преобразований получаем, что

$$(c, z) \leq n - \sum_{k=1}^{M_i} (t_k - t_{k+1})(\varrho_k + 2e_k) - t_{M_i+1} \cdot n_a . \tag{2.15}$$

Предположим теперь, что для всех $k = 1, \ldots, M_i$

$$\varrho_k + 2e_k \geq d_{a,i} .$$

Тогда из (2.15) получаем

$$(c, z) \leq n - d_{b,i}\, d_{a,i} = n - d_i ,$$

что противоречит условию (2.8). Таким образом, найдется, по крайней мере, один порог t_v такой, что

$$\varrho_v + 2e_v < d_{a,i} . \tag{2.16}$$

Тогда при декодировании строки $\hat{a}^{(i)}(v)$ получим истинную строку $a^{(i)}$, что и требовалось доказать.

Покажем теперь, что из системы M_i порогов можно выбрать не более $[(d_{a,i} + 1)/2]$ порогов так, чтобы, по крайней мере для одного из них выполнялось условие (2.16). Нетрудно видеть, что при декодировании порогом t_{k_0}, где $k_0 = d_{a,i} + 1$, стерто не менее $d_{a,i}$ символов. Поэтому для всех порогов t_v при $k \geq k_0$ условие (2.16) не выполняется. Следовательно, «истинный» порог t_v (см. (2.16)) лежит среди $d_{a,i}$ порогов $t_1, t_2, \ldots, t_{d_{a,i}}$.

Пусть теперь P — подмножество номеров тех из $d_{a,i} - 1$ порогов $t_2, \ldots, t_{d_{a,i}}$, при декодировании которыми стирается ровно 1 столбец. Пусть при декодировании порогом t_p, $p \in P$, стирается столбец с номером j_s, $1 \leq s \leq n_a$. Заметим, что $\varrho_{p-1} = \varrho_p - 1$, а $e_{p-1} = e_p$, если столбец с номером j_s, не был ошибочным и $e_{p-1} = e_p + 1$ в противном случае. Поэтому

$$2e_{p-1} + \varrho_{p-1} = 2e_p + \varrho_p \pm 1.$$

Для каждого принятого вектора z выбросим из двух порогов t_{p-1} и t_p ($p \in P$) тот порог p' ($p' = p-1$ или $p' = p$), для которого $\varrho_{p'} \equiv d_{a,i} \pmod 2$. Если порог p' является «истинным» и удовлетворяет (2.16), то

$$2e_{p'} + \varrho_{p'} \leq d_{a,i} - 2.$$

Тогда оставшийся из двух порогов также удовлетворяет (2.16). При декодировании каждым из оставшихся после выбрасывания порогов стирается не менее двух столбцов, откуда следует, что достаточно $[(d_{a\,i} + 1)/2]$ порогов. Теорема доказана.

Замечание. В [1] также было показано, что принятый вектор z можно декодировать по МОР в слово c кода длины n с расстоянием d за $[(d + 1)/2]$ попыток исправления ошибок и стираний (если, конечно, $(c, z) > n - d$). При каскадном декодировании по МОР проводится декодирование по МОР внутренними кодами длины n_b, а затем (на i-ом шаге) осуществляем $[(d_{a,i} + 1)/2]\}$ попыток декодирования внешними кодами длины n_a.

Итак, при наличии $M_i^0 = [(d_{a,i} + 1)/2]$ порогов каскадное декодирование по МОР реализует «кодовое расстояние», задаваемое скалярным произведением ($(c, z) > n - d^{(\text{н})}$). Предположим теперь, что при декодировании внешних кодов имеется $M < M_i^0$ порогов. Определим, при каких условиях на скалярное произведение (c, z) переданного слова c и принятого слова z найдем слово c в результате каскадного декодирования по МОР системами из M порогов. В дальнейшем системы M порогов T_i, $i = 1, \ldots, m$, используемые на i-ом этапе декодирования, фиксированы для каждого из m этапов, т. е. не зависят от принятого вектора z.

Пусть на i-ом этапе декодирования, $1 \leq i \leq m$, выбрана равномерная система M порогов $P^{(i)} = \{p_1, \ldots, p_M\}$, где

$$p_k = d_{b,i}\left(\frac{2M - 2(k-1)}{2M+1}\right), \quad k = 1, \ldots, M. \tag{2.17}$$

Отметим, что при таком выборе $p_1 < d_{b,i}$. Следующая теорема показывает, что система порогов $P^{(i)}$, $i = 1, \ldots, m$, является наилучшей. Для обычных каскадных и итерированных кодов это было показано в [5].

Теорема 4. (1) Пусть строки $a^{(1)}, \ldots, a^{(i-1)}$ переданного слова c найдены правильно, $1 \leq i \leq m$. На i-ом этапе декодирования принятого слова z система M порогов $P^{(i)}$ (2.17) обеспечивает нахождение правильной строки $a^{(i)}$, если

$$(c, z) > n - d_i \cdot \frac{2M}{2M+1}. \tag{2.18}$$

(2) Если для принятого слова z выполняется неравенство

$$(c, z) > n - d^{(\text{н})} \cdot \frac{2M}{2M+1}, \tag{2.19}$$

то при декодировании системами порогов $P^{(i)}$ (2.17), $i = 1, \ldots, m$, будет найдено переданное слово c.

(3) Если в коде $\{c\}$ минимальное расстояние равно $d^{(\text{н})}$, то при каскадном декодировании любыми другими системами M фиксированных порогов T_i, $i = 1, \ldots, m$, найдется кодовое слово c и принятое слово z, удовлетворяющее (2.19), правильное декодирование которого каскадным алгоритмом невозможно.

Замечание. Фактически при доказательстве теоремы будет получен более сильный результат. Будет построена точная граница реализуемого расстояния для любой системы из M фиксированных порогов, которая зависит только от двух параметров δ и Δ, где $\delta = d_{b,i} - t_1$, а $\Delta = \max\limits_k (t_k - t_{k+1})$.

Доказательство. Пусть на i-ом шаге декодирования принятого вектора z используется система из M фиксированных порогов $T = (t_1, \ldots, t_M)$, где $d_{b,i} \geq t_1 > t_2 > \ldots > t_M > 0$. Для переданного слова c обозначим через $z^{(i)}(T, c)$ множество принятых векторов z, которые не могут быть декодированы в слово c на i-ом этапе системой порогов T. Это означает, что

$$2e_k + \varrho_k \geq d_{a,i} \quad \text{для всех} \quad k = 1, \ldots, M, \tag{2.20}$$

где, как и ранее, через ϱ_k и e_k обозначены число стертых и, соответственно, ошибочных символов в строке $\hat{a}^{(i)}(k)$ (полученной при использовании порога t_k). Пусть

$$D^{(i)}(T) = \sup (c, z) \text{ по всем } c \in \{c\}, \; z \in z^{(i)}(T, c).$$

Очевидно, что система порогов T заведомо обеспечивает нахождение строки $a^{(i)}$, если $(c, z) > D^{(i)}(T)$. Нужно доказать, что при использовании системы порогов $P^{(i)}$

$$D^{(i)}(P^{(i)}) = n - d_i \cdot \frac{2M}{2M+1},$$

а при использовании любой другой системы из M порогов T

$$D^{(i)}(T) \geq n - d_i \cdot \frac{2M}{2M+1},$$

т. е.

$$D^{(i)}(P^{(i)}) = \inf_{T} \{D^{(i)}(T)\}.$$

Оценим сверху величину $D^{(i)}(T)$ для заданной системы T из M порогов. Оценка будет проведена с помощью величин e_k и $\varrho_k (k = 1, \ldots, M)$ аналогично тому, как это было сделано в доказательстве теоремы 3.

Пусть принято слово Z. Разобьем все множество J номеров столбцов z_j, $j = 1, \ldots, n_a$, на $M + 1$ подмножеств следующим образом:

$$J_0 = \{j \mid (z_j, \hat{b}_j) \leq n_b - t_1\},$$

$$J_k = \{j \mid n_b - t_k < (z_j, \hat{b}_j) \leq n_b = t_{k+1}\}, \quad k = 1, \ldots, M,$$

где положим по определению $t_{M+1} = 0$. Очевидно, при переходе от порога t_k к t_{k+1}, $k = 1, \ldots, M - 1$, стирается $\varrho_{k+1} - \varrho_k$ столбцов z_j, $j \in J_k$, из которых $e_k - e_{k+1}$ являются ошибочными. Согласно лемме 1 для каждого из ошибочных столбцов z_j, $j \in J_k$, выполняется следующее неравенство:

$$(z_j, b_j) < n_b - 2d_{b,i} + t_k, \tag{2.22}$$

(так как $(z_j, \hat{b}_j) < n - t_k$, а $\hat{b}_j \neq b_j$). Отсюда для всех $k = 1, \ldots, M - 1$ имеем

$$\sum_{j \in J_k} (b_j, z_j) \leq (e_k - e_{k+1})(n_b - 2d_{b,i} + t_k) + \left((\varrho_{k+1} - \varrho_k) - (e_k - e_{k+1})\right)(n_b - t_{k+1}). \tag{2.23}$$

Аналогично,

$$\sum_{j \in J_k} (b_j, z_j) \leq (n_b - t_1)\,\varrho_1, \tag{2.24}$$

$$\sum_{j \in J_M}{}' (b_j, z_j) \leq e_M(n_b - 2d_{b,i} + t_M) + (n_a - \varrho_M - e_M)\,n_b. \tag{2.25}$$

Просуммировав (2.23)–(2.25) и проделав элементарные преобразования, получаем, что

$$(c, z) \leq n - \sum_{k=1}^{M} (\varrho_k + 2e_k)(t_k - t_{k+1}) + \sum_{k=1}^{M} (e_k - e_{k+1})(t_k - t_{k+1}) - 2e_1(d_{b,i} - t_1) \tag{2.26}$$

где $e_{M+1} = 0$.

Введем следующие определения. Пусть $E = (\varrho_1, \ldots, \varrho_M, e_1, \ldots, e_M)$ обозначает упорядоченный набор из $2M$ целых неотрицательных чисел, которые удовлетворяют условию (2.20), а также неравенствам:

$$0 \leq \varrho_1 \leq \varrho_2 \leq \ldots \leq \varrho_M, \tag{2.27}$$

$$e_1 \geq e_2 \geq \ldots \geq e_M \geq 0. \tag{2.28}$$

Пусть теперь $E^{(i)}$ обозначает множество всех наборов E, удовлетворяющих условиям (2.20), (2.27), (2.28). Определим величину $P(T, E)$ следующим образом:

$$P(T, E) = n - \sum_{k=1}^{M} (\varrho_k + 2e_k)(t_k - t_{k+1}) + \sum_{k=1}^{M} (e_k - e_{k+1})(t_k - t_{k+1}) - 2e_1(d_{b,i} - t_1). \tag{2.29}$$

Иными словами, $(c, z) \leq P(T, E)$, где E — набор чисел ϱ_k, e_k, $k = 1, \ldots, M$, полученный при декодировании слова Z системой порогов T. Отсюда следует, что

$$D^{(i)}(T) \leq \sup_{E \in E^{(i)}} \{P(T, E)\}. \tag{2.30}$$

Как будет показано ниже, в (2.30) имеет место равенство, поэтому вместо $D^{(i)}(T)$ достаточно найти $\sup_{E \in E^{(i)}} P(T, E)$, к чему сейчас и перейдем.

Пусть $E_0^{(i)}$ — множество таких наборов $E \in E^{(i)}$, для которых в (2.20) выполняется равенство для всех $k = 1, \ldots, M$, т. е.

$$\varrho_k + 2e_k = d_{a,i}, \quad k = 1, \ldots, M. \tag{2.31}$$

Отметим, что для таких наборов $\varrho_k = d_{a,i} \pmod 2$ для всех $k = 1, \ldots, M$.

Лемма 3

$$\sup_{E \in E^{(i)}} \{P(T, E)\} = \sup_{E \in E_0^{(i)}} \{P(T, E)\}.$$

Доказательство леммы дано в приложении.

Для всех наборов $E \in E_0^{(i)}$, как нетрудно видеть из (2.29) и (2.31), можно написать, что

$$P(T, E) = n - d_{a,i} t_1 + \sum_{k=1}^{M} (e_k - e_{k+1})(t_k - t_{k+1}) - 2e_1(d_{b,i} - t_1).$$

Пусть

$$\Delta = \max_{k=1, \ldots, M} \{t_k - t_{k+1}\}, \quad \delta = d_{b,i} - t_1,$$

и пусть v такое, что $t_v - t_{v+1} = \Delta$. Если таких v несколько, то выбираем любое из них. Очевидно, $\Delta \geq (d_{b,i} - \delta)/M$. Заметим, что

$$\sum_{k=1}^{M} (e_k - e_{k+1})(t_k - t_{k+1}) \leq \Delta \cdot \sum_{k=1}^{M} (e_k - e_{k+1}) = \Delta \cdot e_1,$$

причем равенство достигается, когда $e_v - e_{v+1} = e_1$, что эквивалентно условию

$$e_1 = \ldots = e_v, \; e_{v+1} = \ldots = e_M = 0.$$

Отсюда

$$P(T, E) \leq n - d_{a,i} \cdot t_1 + e_1(\Delta - 2\delta)$$

и

$$\sup_{E \in E_0^{(i)}} \{P(T, E)\} = \sup_{e_1} \{n - d_{a,i} \cdot t_1 + e_1(\Delta - 2\delta)\}, \tag{2.32}$$

причем sup $P(T, E)$ достигается на наборах $E \in E_0^{(i)}$, для которых

$$e_1 = \ldots = e_v, \; e_{v+1} = \ldots = e_M = 0,$$

$$\varrho_k = d_{a,i} - 2e_k, \quad k = 1, \ldots, M.$$

Рассмотрим теперь два случая.

Случай 1. Пусть T — система из M порогов, для которой $\Delta - 2\delta > 0$. Как видно из (2.32), в этом случае необходимо выбрать e_1 максимальным. Так как $2e_1 + \varrho_1 = d_{a,i}$, то $e_1 \leq [d_{a,i}/2]$. Взяв $e_1 = [d_{a,i}/2]$, получаем, что

$$\sup_{E \in E_0^{(i)}} P(T, E) = n - d_{a,i} \cdot t_1 + \left[\frac{d_{a,i}}{2}\right](\Delta - 2\delta),$$

причем равенство достигается при выборе E следующим образом

$$e_1 = \ldots = e_v = [d_{a,i}/2],\ e_{v+1} = \ldots = e_M = 0, \tag{2.33}$$

$$\varrho_1 = \ldots = \varrho_v = d_{a,i} - 2e_1,\ \varrho_{v+1} = \ldots = \varrho_M = d_{a,i}. \tag{2.34}$$

Докажем теперь, что $D^{(i)}(T) = \sup_{E \in E^{(i)}} P(T, E)$. Выберем два кодовых вектора c и c', отличающиеся в $d_{a,i}$ столбцах по $d_{b,i}$ символов в каждом. Это заведомо можно сделать, если в коде реализуется оценка $d_i = d_{a,i}\, d_{b,i}$ кодового расстояния i-го каскада. Пусть передано слово c. Построим слово z следующим образом. Выберем из этих $d_{a,i}$ столбцов слов c и c' произвольные $[d_{a,i}/2]$ столбцов. Пусть b_j представляет собой один из выбранных столбцов слова c, а b'_j — произвольный столбец слова c'. Пусть столбцы z_j декодируются по МОР в b'_j для всех $[d_{a,i}/2]$ выбранных столбцов. Очевидно, что выбор столбцов z_j можно осуществить при этом так, чтобы (z_j, b'_j) сколь угодно мало отличалось от $n_b - t_v$ и, более того, чтобы (z_j, b_j) сколь угодно мало отличалось от $n_b - 2d_{b,i} + t_v$. Кроме того, пусть для $d_{a,i} - 2[d_{a,i}/2]$ столбцов $(z_j, b_j) = n_b - t_1$, и для $[d_{a,i}/2]$ столбцов $(z_j, b_j) = n_b - t_{v+1}$. Пусть для всех остальных столбцов $z_j = b_j$. В соответствии с таким построением, скалярное произведение (c, z) сколь угодно мало отличается от $P(T, E)$ (см. неравенства (2.22)–(2.26). Кроме того, при таком построении вектора z удовлетворяются условия (2.33) и (2.34), т. е.

$$P(T, E) = n - d_{a,i}\, t_1 + [d_{a,i}/2]\,(\Delta - 2\delta).$$

Итак, для любой системы порогов T_i, для которой $\Delta \geq 2\delta$

$$D^{(i)}(T) = \sup_{\substack{'c \in \{c\} \\ z \in Z^{(i)}(T, C)}} \left\{n - d_{a,i} \cdot t_1 + \left[\frac{d_{a,i}}{2}\right](\Delta - 2\delta)\right\}. \tag{2.35}$$

Случай 2. Рассмотрим теперь все системы порогов T_i, для которых $\Delta - 2\,\delta \leq 0$. В этом случае в соответствии с (2.32) необходимо взять $e_1 = 0$ и

$$\sup_{E \in E_0^{(i)}} P(T, E) = n - d_{a,i} \cdot t_1,$$

причем равенство достигается при выборе E следующим образом

$$e_1 = e_2 = \ldots = e_M = 0, \tag{2.36}$$

$$\varrho_1 = \varrho_2 = \ldots = \varrho_M = d_{a,i}. \tag{2.37}$$

В этом случае для произвольного вектора c (со столбцами b_j) построим вектор z следующим образом. Выберем в c $d_{a,i}$ произвольных столбцов и для каждого из них построим столбцы z_i так, чтобы $(b_j, z_j) = n_b - t_1$. Пусть для остальных $n_a - d_{a,i}$ столбцов $z_j = b_j$. В этом случае $(c, z) = P(T, E)$, причем E удовлетворяет (2.36), (2.37).

Итак, для любой системы порогов T_i, для которой $\Delta \leq 2\delta$

$$D^{(i)}(T) = \sup_{\substack{c \in \{c\} \\ z \in Z^{(i)}(T, C)}} (c, z) = n - d_{a,i} \cdot t_1, \tag{2.38}$$

причем существуют c и z, для которых $(c, z) = n - d_{a,i} \cdot t_1$ такие, что правильное декодирование слова z невозможно.

Покажем теперь, что система порогов $P^{(i)}$ является наилучшей, т. е. $\inf_T D^{(i)}(T)$ достигается при выборе $P^{(i)}$ как в случае (2.35), так и (2.38). Пусть вначале выполняется (2.35), т. е. $\Delta \geq 2\delta$.

Заметим, что

$$\inf_{T:\, \Delta \geq 2\delta} D^{(i)}(T) = \inf_{\delta} \inf_{\Delta \geq 2\delta} D^{(i)}(T) = \inf_{\delta = \frac{\Delta}{2}} (n - d_{a,i} \cdot t_1) =$$

$$= \inf_{\delta = \frac{\Delta}{2}} (n - d_{a,i} \cdot d_{b,i} + d_{b,i} \cdot \delta).$$

Очевидно, необходимо выбрать наименьшее δ такое, что $\delta = \Delta/2$, но $\Delta \geq (d_{b,i} - \delta)/M$, откуда $\delta \geq (d_{b,i} - \delta)/2M$, т. е. $\delta \geq d_{b,i}/(2M + 1)$, причем равенство достигается, если $\Delta = (d_{b,i} - \delta)/M$, что возможно только, если $t_k - t_{k+1} = \Delta$ для всех $k = 1, \ldots, M$. В этом случае $\delta = d_{b,i}/(2M + 1)$, $t_k - t_{k+1} = \dfrac{2d_{b,i}}{2M + 1}$ для всех $k = 1, \ldots, M$, что соответствует системе порогов $P^{(i)}$. Итак,

$$\inf_{T:\, \Delta \geq 2\delta} D^{(i)}(T) = D^{(i)}(P^{(i)}) = n - d_{a,i}\, d_{b,i} \cdot \frac{2M}{2M + 1}.$$

Пусть теперь выполняется (2.38), т. е. рассматриваются системы порогов T_i, для которых $\Delta \leq 2\delta$. Аналогично

$$\inf_{T:\, \Delta \leq 2\delta} D^{(i)}(T) = \inf_{\Delta \leq 2\delta} (n - d_{a,i}\, d_{b,i} + d_{a,i} \cdot \delta).$$

В этом случае необходимо выбрать наименьшее δ, т. е. $\delta = \Delta/2$. Так как $\Delta \geq (d_{B,i} - \delta)/M$, то аналогично получаем, что $\delta_{\min} = d_{bi}/(2M+1)$ и $t_k - t_{k+1} = 2d_{b,i}/(2M+1)$ для всех $k = 1, \ldots, M$, что соответствует системе $p^{(i)}$.

Итак,

$$\inf_{T} \sup_{\substack{c \in \{c\} \\ z \in Z^{(i)}(T,C)}} (c, z) = n - d_{a,i}\, d_{b,l} \cdot \frac{2M}{2M+1},$$

причем наилучшей является система M порогов $P^{(i)}$. Теорема доказана.

Замечания. 1) Как видно из доказательства, равенство в (2.35) основывалось на выборе 2-х кодовых векторов c и c', находящихся на расстоянии $d^{(\text{н})}$ друг от друга. Если истинное кодовое расстояние $d > d^{(\text{н})}$, то (2.35) не должно обязательно выполняться для всех $i = 1, \ldots, M$. Возможно, в этом случае при использовании систем порогов (2.17) и (2.11) в каскадном алгоритме могут быть получены лучшие оценки, чем (2.19) и (2.3).

В случае ДСК столбцы z_j $(j = 1, \ldots, n_a)$ декодируются в $\hat{b}_j$ с исправлением $l_i = [(d_{b,i} - 1)/2]$ ошибок, а строка $\hat{a}^{(i)}$ — в множество строк $\{a^{(i)}(k)\}$ системой $l_i + 1$ порогов $L_i = \{l_i, l_i - 1, \ldots, 0\}$. Вместо скалярного произведения используется расстояние Хэмминга $d(z_j, \hat{b}_j)$, так что

$$p_j^{(i)}(k) = \begin{cases} l_i + 1 & \text{, если } \hat{b}_j \text{ стерт,} \\ d(z_j, \hat{b}_j) & \text{, если } \hat{b}_j \text{ не стерт и } \hat{a}_j^{(i)} = a_j^{(i)}(k), \\ d_{b,i} - d(z_j, \hat{b}_j), & \text{если } \hat{b}_j \text{ не стерт и } \hat{a}_j^{(i)} \neq a_j^{(i)}(k). \end{cases}$$

Алгоритм декодирования в ДСК предложен в [6]. Если произошло не более $[(d^{(H)} - 1)/2]$ ошибок, то алгоритм (2.7) обеспечивает выбор истинной строки $a^{(i)}$ из $\{a^{(i)}(k)\}$ для всех $i = 1, \ldots, m$.

В [7] исследована взаимосвязь каскадного декодирования в ДСК и декодирования по МОР.

Приложение

Доказательство леммы 3. Рассмотрим произвольный набор $E \in E^{(i)}$, $E = (\varrho_1, \ldots, \varrho_M, e_1, \ldots, e_M)$. Введем набор $E' = (\varrho_1, \ldots, \varrho'_M, e_1, \ldots, e'_M)$ следующим образом. Пусть для всех $k = 1, \ldots, M$

$$\varrho'_k = \begin{cases} d_{a,i}, & \text{если} \quad \varrho_k \geq d_{a,i}, \\ \varrho_k, & \text{если} \quad \varrho_k < d_{a,i} \text{ и } \varrho_k \equiv d_{a,i} \pmod 2, \\ \varrho_k - 1, & \text{если} \quad \varrho_k < d_{a,i} \text{ и } \varrho_k \not\equiv d_{a,i} \pmod 2 \end{cases}$$

и

$$e'_k = \begin{cases} 0, & \text{если} \quad \varrho_k \geq d_{a,i}, \\ (d_{a,i} - \varrho_k)/2, & \text{если} \quad \varrho_k < d_{a,i} \text{ и } \varrho_k \equiv d_{a,i} \pmod 2, \\ (d_{a,i} - (\varrho_k - 1))/2, & \text{если} \quad \varrho_k < d_{a,i} \text{ и } \varrho_k \not\equiv d_{a,i} \pmod 2. \end{cases}$$

Заметим, что $\varrho'_k + 2e'_k = d_{a,i}$ для всех $k = 1, \ldots, M$ и, если для набора E выполняются условия (2.27)–(2.28), то $0 \leq \varrho'_1 \leq \varrho'_2 \leq \ldots \leq \varrho'_M$ и $e'_1 \geq e'_2 \geq \ldots \geq e'_M \geq 0$. Таким образом, набор $E' \in E_0^{(i)}$, причем для всех $k = 1, \ldots, M$

$$\varrho'_k \leq \varrho_k, \qquad \text{(п. 1)}$$

$$e_k \leq e_k. \qquad \text{(п. 2)}$$

Неравенство (П. 1) очевидно. Неравенство (П. 2) следует из того, что если $\varrho_k \not\equiv \not\equiv d_{a,i} \pmod 2$, то $2e_k + \varrho_k \geq d_{a,i} + 1$.

Далее, из (2.29) получаем, что

$$P(T,\ E') - P(T, E) = \sum_{k=1}^{M} (\varrho_k - \varrho'_k)(t_k - t_{k+1}) + 2(e_1 - e'_1)(d_{b,i} - t_1) + \\ + \sum_{k=1}^{M} ((e_k - e'_k) + (e_{k+1} - e'_{k+1})(t_k - t_{k+1}).$$

С учетом (П.1) и (П.2) получаем, что

$$P(T, E') - P(T, E) \geq 0,$$

причем равенство возможно только, если $\varrho'_k = \varrho_k$ и $e'_k = e_k$ для всех $k = 1, \ldots, M$. Лемма доказана.

Литература

1. *Форни Д.* Каскадные коды. М., «Мир», 1970.
2. *Зиновьев В. А., Зяблов В. В.* Декодирование нелинейных обобщенных каскадных кодов. Пвоблемы передачи информации, 1978, **14,** *2*, 46–52.
3. *Зиновьев В. А., Зяблов В. В.* Каскадное декодирование по минимуму обобщенного расстояния. Труды V Международного симпозиума по теории информации, I. Тезисы докладов. Москва–Тбилиси, 1979, 161–162.
4. *Зиновьев В. А.* Обобщенные каскадные коды. Проблемы передачи информации, 1976, **12,** *1*, 5–15.
5. *Зяблов В. В.* Оптимизация алгоритмов каскадного декодирования. Проблемы передачи информации, 1973, **9,** *1*, 26–32.
6. *Думер И. И.* О декодировании обобщённых каскадных кодов. Труды V Всесоюзной школы-семинара по вычислительным сетям, часть 4, Москва–Владивосток, 1980, 92–95.
7. *Бояринов И. М.* Об исправлении ошибок прямыми суммами итеративных и каскадных кодов. ДАН СССР, 1980, **252,** *3*, 563–565.

Cascaded decoding according to minimal generalized distance

I. I. DUMER, V. A. ZINOV'EV, V. V. ZYABLOV

(Moscow)

The paper discusses the decoding of the generalized concatenated binary codes in the channel with continuous output. It is shown that the decoding according to the minimum of the generalized distance, suggested by Forney, may be realized for these codes in cascade way. Applying this method the internal and external codes are decoded separately and the complexity of decoding reduces. The correcting capability for various systems of decoding steps is investigated. The optimal decoding systems for arbitrary number of steps are defined. If the number of steps is sufficient, the suggested method realizes the code distance of the generalized concatenated codes. The results obtained may be applied for the discrete channels and provide more simple decoding of the generalized concatenated codes than the known before [2].

И. И. Думер, В. А. Зиновьев, В. В. Зяблов
Институт проблем передачи информации АН СССР
СССР, 101447 Москва, ГСП-4,
ул. Ермоловой, 19

Problems of Control and Information Theory, Vol. 10 (1), pp. 21—31

AN EXPERIMENTAL RECOGNITION SYSTEM USING PATTERN GRAMMARS

V. GALLÓ

(Budapest)

(Received 15 January, 1980)

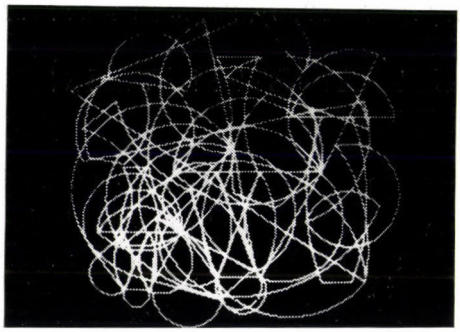
1. Introduction

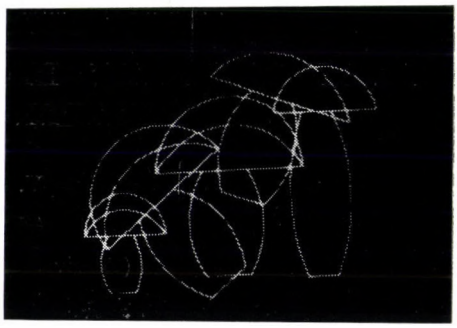
We all solved picture puzzles when we were children, e.g. found the hunter and hares hiding in a dense forest. Recognition of three-dimensional objects from contour images of real scenes is analogous to solve puzzles. Contour images of visible projections of the objects to be recognized are hidden in the "forest" of superfluous lines and contours resulting from the background, shadows, unhomogeneous illumination and so on.

Computers can be taught to solve puzzles like this. The following problems are to be solved:

1) How to pass or show a picture (a contour) to the computer;
2) How to define patterns to be recognized;
3) How to teach the computer to find the known patterns in an arbitrarily dense "forest", using pattern descriptions;
4) How to teach the computer to output the results.

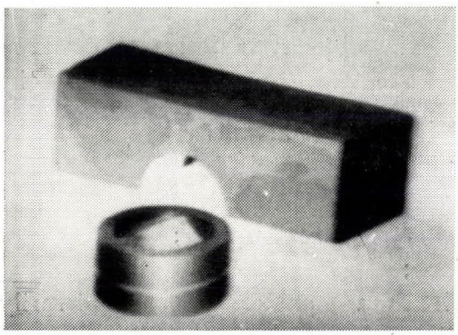
A solution to problems 2) and 3) in an experimental recognition system is reported. We discuss a language for the description of contour patterns and the algorithm applied in its interpreter program which solves picture puzzles via pattern descriptions. The program together with a preprocessing program (2) and a graphic package (3) makes up a recognition system with two alternative inputs. The contour picture comes either from a TV camera after preprocessing (Fig. 2c) or from a human operator using a graphic system (Fig. 1a). All the programs of picture processing and recognition use the graphic package to output their intermediate and final results on a graphic display (all the photos of the figures except 2a, 4 and 5a are taken from the display).

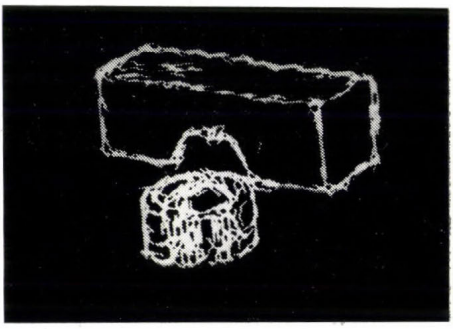




a *b*

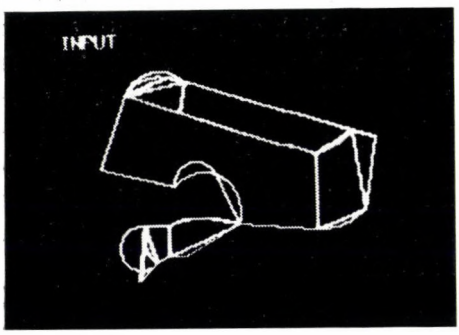
Fig. 1a. How many mushrooms are there in the picture?
Fig. 1b. All the mushrooms are found

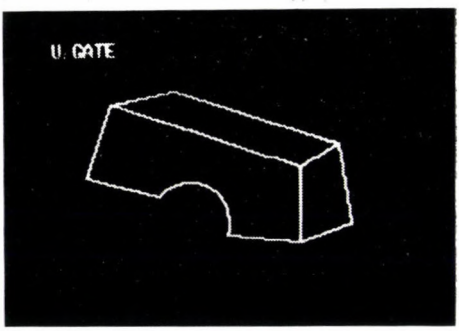




a *b*

Fig. 2a. A digital picture (negative)
Fig. 2b. The intermedial result of the preprocessing





c *d*

Fig. 2c. The result of preprocessing: all the possible interpretations of the paths between nodes
Fig. 2d. Recognition of the objects' contours by help of its structural description

2. The pattern description language and the interpreter

In defining the pattern description formalism and designing the interpreter we essentially used the ideas of (4) and (5). But our goal was a real-time recognition system that works within seconds on a minicomputer with 20 K words memory and 800 K words of disk storage. Limitations of memory and time and necessity of communication between several programs forced us to develop a new implementation. First we had to design and implement a fast, economical list processing system (6). This served as the basis for the implementation of the *C*ontour *P*attern *D*escription *L*anguage (CPCL). With this language recognition procedures for finding contour patterns in contour images can be described. In CPDL, contour patterns are described in a hierarchical way via a sequence of recognition procedures finding complicated picture elements as sets of more simple ones and setting in the process a hierarchical data structure for the actual input picture. The language includes instructions for choosing search strategies, for manipulating the data structure and ideal models and instructions for tracing, graphic output and so on.

The interpreter can received the CPDL program either in textual form or in form of a compact inner list structure. The latter is the result of a translation of a textual CPDL program into the list structure. This list structure may be either immediately intepreted on the actual input picture as a sequence of recognition procedures and/or it may be saved as a file in the system library and interpreted as a recognition program when needed.

CPDL is a generalization of the formalism of programmed context-free grammars. Constructions described by CPDL grammars are two-dimensional pictures. Picture elements are characterized not only by their syntactic types but also by their quantitative attributes, namely the two dimensional coordinates and their functions.

Relations between grammatical symbols (between picture elements) are not simple concatenations but various geometrical and topological relations that are indicated explicitly in production rules. Production rules do not determine syntax alone (i.e. syntactic types of the left-side and the right-side symbols) but the semantics as well. Here, semantics means geometrical and topological relations between symbols and definition of the left-side symbol attributes as functions of the right-side symbol attributes. The productions are generally invariant to rotation of the described patterns in the XY plane.

A production defines a nonterminal picture element as a set of the terminal and nonterminal picture elements for which certain geometrical and topological relations hold. A production rule is interpreted as a nonterminal picture-element recognition procedure that includes searching in the data

structure representing the picture, examining relations between the picture elements and modifying the picture's data structure. The sequence of recognition procedures is supplemented by a list of names of relation predicates and by a list of names of operations on the picture's data.

The running program can received the actual picture data either from the preprocessing program or from the graphic system but in both cases the list structure of the picture's data is the same. It contains lists of the nodes, lines and arcs linked by pointers according to the structure of the actual picture:

- NODES is a list of all nodes in the contour picture; coordinates and pointers to the lines emerging from nodes are indicated at every node;
- EDGES is a list of all lines in the contour picture; straight lines, arcs and small picture elements without types. Every element has pointers to the nodes, connected by this element.

If a contour is received from the preprocessing program, the picture's data structure is more complicated: the paths between nodes of degree ≥ 3 have several interpretations. (A path is a sequence of lines which does not contain nodes of degree ≥ 3.) Every line of the contour image is characterized not only by the pointers to its endpoints (nodes) but also by the number of the path and by the probabiilty of the interpretation containing this line. The lines which are parts of different interpretations of the same path are associated with the same path number but different probability values. The sum of the probabilities of all interpretations of the same path is equal to 1. The interpreter must not use different interpretations of the same path in the same pattern.

The interpreter supplements the initial picture's data by the list TYPES. The elements of the list TYPES have the same structure as the elements of the list NODES but instead of coordinates they are characterized by the names (i.e. types) of the picture elements and their pointers are directed to all the elements of this type occurring in the actual picture's data. Initially the list TYPES has three "nodes" which group the terminal elements of the actual picture in three lists according to their types:

- the "node" LINE has pointers to all straight lines
- the "node" ARC has pointers to all arcs
- the "node" INDF has pointers to all small indefinite lines.

This description of the actual input picture is the initial picture's data structure, on which a CPDL program defines a sequence of grammatical rules which will be interpreted as a sequence of recognition procedures for nonter-

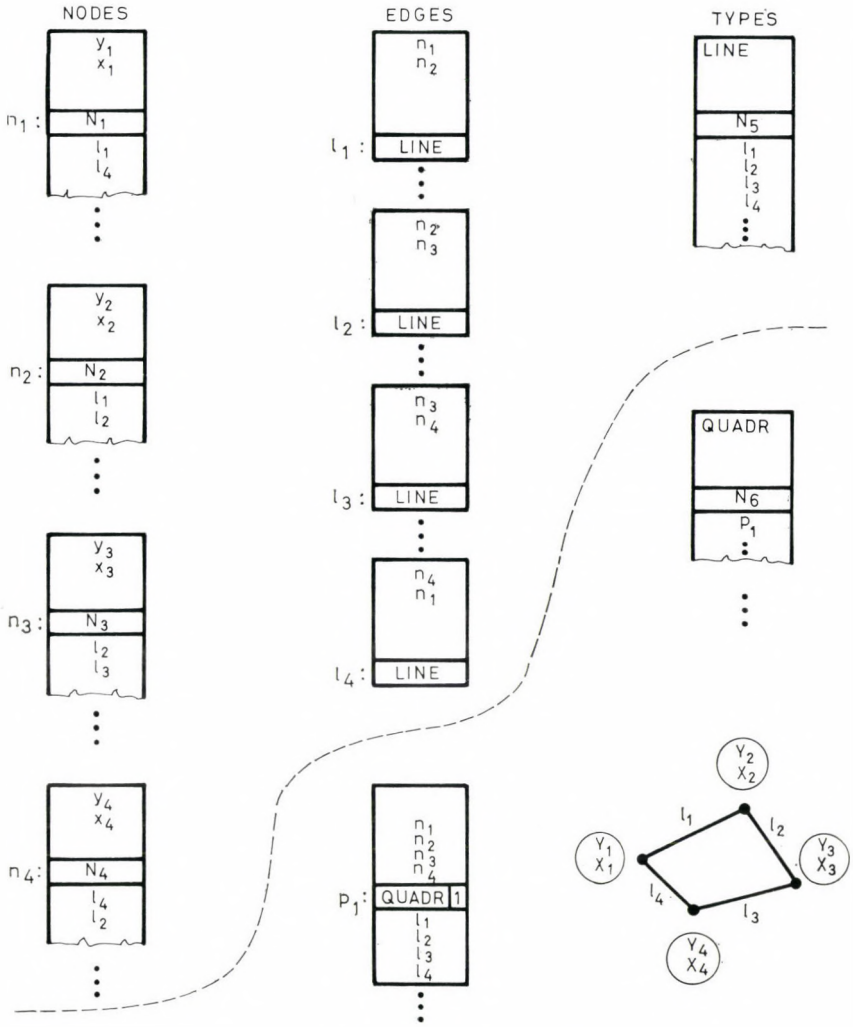


Fig. 3. Data structure for a quadrangle

$X_1Y_1, \ldots, X_4Y_4$	coordinates of nodes
n_1:, n_2:, . . . , n_4:	addresses in the data structure
l_1:, l_2:, . . . , l_4: p:	

minal picture elements. On this initial data structure a bottom-up grammatical analysis will start. During this process, over the initial data structure a hierarchical list structure is built, indicating the sequence of those nonterminal element recognition procedures which were interpreted with success on the actual picture (Fig. 3).

Every step of this process is a rather independent decision procedure searching for a nonterminal picture element as a combination of terminal and previously found nonterminal picture elements among which the relations indicated in the search procedure description hold. If a successful combination is found, a new element is added to the list EDGES, indicating the nonterminal picture element that has just been found. Its name is given by the value of the procedure name (i.e. the production's left-side symbol), as for example QUADR in Fig. 3. The new list element is added by the list of pointers to those EDGES elements that are its "parents", and by a list of attributes computed from the attributes of the "parents" according to the description of the procedure. So, topological and geometrical relations can be defined between nonterminal a well as between the terminals. The recognition procedures treat the terminal and the previously found nonterminal picture elements in the same way.

In Fig. 3 the data structure for a simple picture can be seen. The dotted line marks the initial state of the picture's data.

3. Syntax of the nonterminal picture element recognition procedure

The following is not the full and precise definition serving as documentation, but only an illustration of the main features:

A nonterminal element recognition procedure: =
(⟨Name⟩ (⟨Way of accessing EDGES⟩
(⟨Conditions for accepting a combination: the list of syntactic types and relations between them⟩),
⟨Way of modifying the picture's data⟩
(⟨List of data manipulations, ideal model manipulations and output operation calls⟩).
[⟨Repeat⟩],
[⟨Conditional goto⟩]))

⟨Way of accessing EDGES⟩: = STARTFN/STARTFT

⟨Way of modifying the picture's data⟩: = NODEND/SYNT

Here

STARTFN — *START From Nodes*: access to the EDGES via pointers from NODES (Fig. 3)

STARTFT — *START From Types*: access to the EDGES via pointers from TYPES (Fig. 3)

MODEND — modify the lists NODES, EDGES, TYPES

SYNT — modify only the lists EDGES and TYPES.

The rectangular brackets mark the optional parts. Lists are represented conventionally, i.e. its elements are enclosed in parentheses.

4. The algorithm of interpretation of the recognition procedure description

The simplified informal description of the interpretation algorithm is the following:

0. SUCCEES: = 0;
   if all the syntactic types indicated in ⟨Conditions for accepting a combination⟩ are presented in the list TYPES, then go to point 1; else go to 4.
1. Set up a combination of candidate parts for the current nonterminal element according to ⟨Way of accessing EDGES⟩; if ⟨Conditions for accepting a combination⟩ by substitution of the parameters of the candidate parts give value TRUE then SUCCESS: = SUCCESS +1 and go to 2; else: if there are more combinations then try a new one, else go to 4.
2. Add to the list EDGES a new element; assign to its identifier the procedure name; mark it with the flag as a nonterminal and set up the list of pointers to its parts; modify the lists NODES and EDGES according to the ⟨Way of modifying the picture's data⟩; execute the operations of ⟨List of data manipulations, ideal model manipulations and output operation calls⟩; go to 3.
3. If [⟨REPEAT⟩] is present in the procedure description then go to 1, also go to 4.
4. If [⟨conditional goto⟩] is present in the procedure description, go to the procedure indicated in the ⟨Conditional goto⟩, also go to the next procedure.

The description of step 1 is simplified; in fact the algorithm is more sophisticated in order to reduce the number of steps in the choice of combinations.

The order of execution of different nonterminal element recognition procedures depends on the order of their descriptions and on the conditional goto instructions. So the process of syntactic analysis can be enriched by some elements of heuristics by varying instructions STARTFN and STARTFT (accessing EDGES either from NODES or from TYPES), MODEND and SYNT (one of them modifies the list NODES while the other does not), and by using or not REPEAT and conditional goto.

5. Recognition of three-dimensional objects using descriptions of their projections

If the structure of a 3D object's edges and surfaces provides information sufficient for its recognition, the object can be identified by grammatical analysis of the contour image (7). If the objects are not too complex, every one has only a few projections differing from each other in the topology and geometry of their edges and nodes. Recognition experiments with objects similar to those shown in Fig. 4 were carried out with the described system, using structural descriptions of object projections. The grammatical analysis of the contour pictures shown in Fig. 2c and 5b takes less than 10 sec (including output of the results on the graphic display).

The sequence of the recognition procedures employed in the experiments was as follows: first in the contour image all those parts must be found which may be interpreted as contours of surfaces of the objects known to the system. Then all possible interpretations of contours as surfaces are used for obtaining all possible interpretations of the objects' projections.

The objects can be identified not only when its maximal projection is visible but also when the objects is partly occluded but its characteristic part is still visible (Fig. 5).

The programming of recognition procedures for objects in real scenes is a difficult task, because an input picture may contain contours of several objects and superfluous lines and the contours themselves of the object can be distorted.

Distortions of the object contour can be grouped in three categories:

1) Distortions in geometrical relations: in parallelity of lines, in relations of lengths of lines and so on;

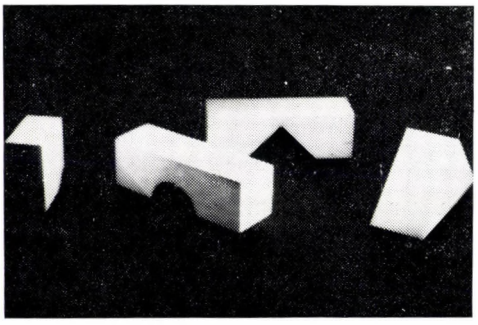


Fig. 4. Objects to be recognized

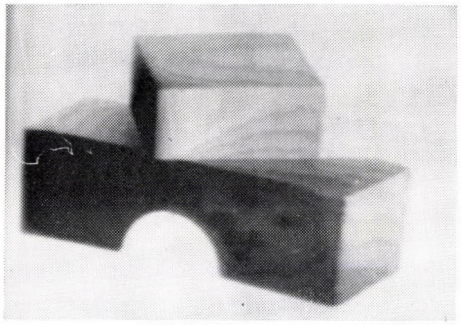


Fig. 5a. Digital image of a real scene (negative)

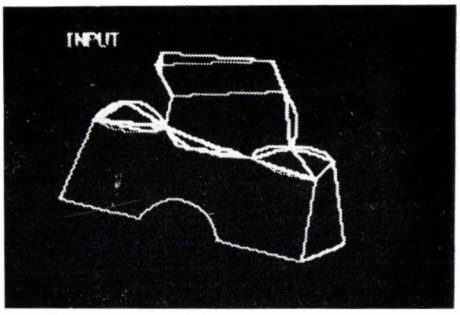


Fig. 5b

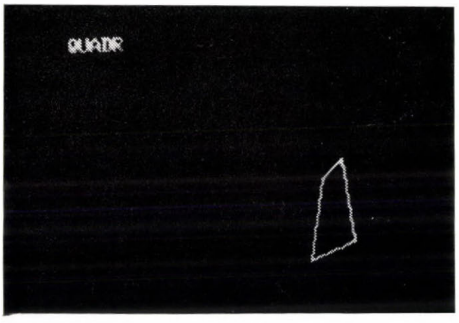


Fig. 5c

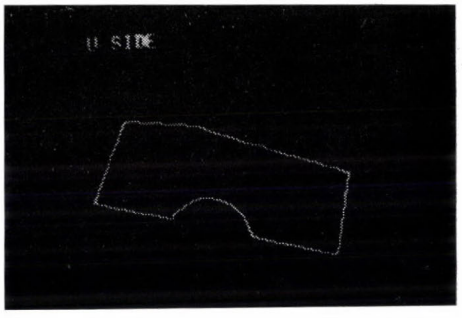


Fig. 5d

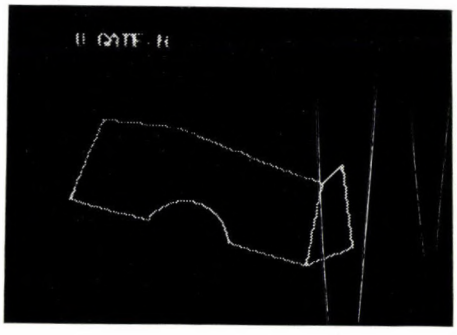


Fig. 5e

Figs. 5b, c, d, e show the sequence of the recognition procedures that were successful in recognizing the scene in Fig. 5a

2) distortions in contour topology: superfluous nodes, dividing the ideal lines of contour;
3) missing parts.

Recognition procedures must be programmed for the possible distortions of the input contours.

1) *Distortions in geometry*

In order to cope with distortions of this type, error parameters must be included in predicates of geometrical relations. These parameters are numerical constants and their values are given in the recognition procedures. For example, the predicate PARALL (x, y, k) gives value TRUE if the difference between the directions of the lines x and y is less than k. It is clear that by increasing k the number of examined combinations grows. By decreasing k the recognition algorithm becomes too sensitive to distortions.

2) *Superfluous nodes*

To cope with this type of distortion we include in the CPDL programs the recursive procedures that recognize nonterminal lines in the following manner: LINE → (LINE(x), LINE(y), FORWARD (x, y, k)), where FORWARD (x, y, k) is a predicate which is TRUE if the line y is a continuation of the line x within the angle error k. This procedure sets up all the possible combinations of the lines which approximately continue each other.

3) *Missing parts*

If no prior information about the scene is available, the recognition may become impossible because of missing parts. But if the characteristic parts of an object are visible, they can give information sufficient for identifying the object. It is the programmer of the CPDL program who determines the moment of identification and the moment of looking at the ideal contour of the corresponding objects' projection.

Procedures recognizing incomplete contours should be included in the set of recognition procedures. The following strategy of recognition can be applied: first try to find the contours of surfaces, then try to recognize the maximal complete projections, and — in case of failure — try to find the "characteristic" incomplete projections. Similar strategy was used in the recognition experiments, whose examples can be seen in Figs. 2a, b, c, d and in Figs. 5a, b, c, d, e.

In order to determine the difference between an ideal complete projection of an object and its incomplete projection recognized in input, the proce-

dure that recognized the object's projection in the input should be executed on the ideal projection's data structure. The procedure marks those parts of the ideal projection which correspond to the parts that are visible in the input picture.

References

1. *Vámos, T., Báthor, M., Mérő, L.*, A Knowledge-based Interactive Robot-vision System. 6th IJCAI, Tokyo, 1979
2. *Mérő, L.*, A Quasi-parallel Contour Following Algorithm. Proc. AISBGI Conf. on AI, Hamburg, 1978
3. *Báthor, M.*, Interactive Picture Manipulation. 2nd Hungarian Computer Science Conf., Budapest, 1977, pp. 168–177
4. *Evans, T. G.*, A Grammar-controlled Pattern Analyzer. IFIP Congress' 68, Appl. 3., North-Holland Publ. Co., Amsterdam, 1968
5. *Galló, V.*, A Program for Grammatical Pattern Recognition. 4th IJCAI (Tbilisi), 1975, pp. 628–634
6. *Galló, V.*, A List Processing System for An Intelligent Robot. 2nd Hungarian Computer Science Conf., Budapest, 1977, pp. 400–411 (in Russian)
7. *Vámos, T.*, Industrial Objects and Machine Parts Recognition in Applications of Syntactic Pattern Recognition (ed. K. S. Fu), Heidelberg, 1977, pp. 243–267

V. Galló
SzTAKI
H—1111 Budapest, Kende u. 13
Hungary

Problems of Control and Information Theory, Vol. 10 (1), pp. 33–42 (1981)

DYNAMIC CONTROL SYSTEM ESTIMATION UNDER UNCERTAINTY CONDITIONS, II

A. B. KURZHANSKII

(Sverdlovsk)

(Received 10 May, 1980)

1. Introduction

The given paper that continues the investigations of [1] is devoted to the problem of estimating the dynamics of a controlled plant under uncertain information on the system parameters. Specifying the inputs x^0, $u(t)$ of the system and being informed of the available output measurements as well as of the a priori restrictions on the parameters and on the measurement errors it is possible to obtain a description of certain informational domains in the phase space that are compatible with the current information on the system as well as with the restrictions given in advance. The evolution of these domains in time will characterize the dynamics of the estimation process. The description of this evolution will therefore give us the solution to the problem of *identification under uncertainty*. The knowledge of this description will allow us in particular to obtain minmax estimates for the unknown actual state of the system.

The construction of minmax estimates for the state space variables under uncertain information on the inputs of the system and of the measurement device, initiated by paper [2], as well as the description of the evolution of the informational sets that correspond to this problem are the object of the theory of filtering under uncertainty [3] (see also the bibliography for the first part of this paper, [1]).

In the present second part that deals with the problem of identification we consider systems with geometrical restrictions on the system parameters (these being given in advance). The proposed solution is achieved by being approximated with the solutions of certain appropriate identification problems with quadratic constraints on the system impulse transfer functions and on the disturbances. (The solutions of the latter problems are also given in the sequel.)

Special cases of solutions to the problem of identification of a dynamic plant under uncertainty were considered earlier in paper [4]. An example of an estimation of the parameters of a linear form under uncertain measurement errors with energy constraints is given in [5].

2. The problem of identification

Consider the system

$$\frac{dx}{dt} = A(t)\,x + B(t)\,u, \quad x(t_0) = x^0, \tag{2.1}$$

whose coefficients $A(t)$ are unknown in advance and restricted by an a priori constraint ($t \in T = [t_0, t_1]$)

$$A(t) \in \mathcal{A}, \ \mathcal{A} = \{A^*(t) + V(t)\}, \ V(t) \in \mathcal{V}, \tag{2.2}$$

where $A^*(t)$ is a given matrix function, $\mathcal{V}$ is a compact set in the space $\mathbf{M}$ of $n \times n$ matrices. Here $u(t)$ is a given input ($u \in R^p$), x^0 is a given initial vector. The information of the actual trajectory of the system (2.1) (for $x(t_0) = x^0$, $u = u(t)$ given) is determined by an available measurement vector

$$y = G(t)x + \xi, \ t \in T, \ \ y \in R^q, \tag{2.3}$$

where ξ is an uncertain disturbance bounded by the constraint

$$\xi(t) \in Q, \ t \in T \tag{2.4}$$

where Q is a given compact set in R^m.

On the basis of the available realization $y[t]$, $t_0 \leq t \leq \vartheta$ of the "signal" y, (2.3) we may specify in the phase space R^n a set $X(\vartheta, y[\cdot])$ of vectors x compatible with the realization $y[t]$ and with the a priori constraints on $A(t)$, $\xi(t)$.

Definition 2.1. The informational set $X(\vartheta, \cdot) = X(\vartheta, y[\cdot])$ of states $x(\vartheta)$ of system (2.1) compatible with the realization $y[\cdot]$ (the input x^0, $u(t)$ being given) is defined as the set of all vectors x for each of which there exists a pair $\overline{V}(t) \in \mathcal{V}$, $\overline{\xi}(t) \in Q$, $t \in T_\vartheta = [t_0, \vartheta]$ such that for the respective trajectory $x(t, (\cdot))$ of system (1.1) the following equalities will be true

$$x(\vartheta, \overline{V}(\cdot)) = x, \ G(t)\,x(t, \overline{V}(\cdot)) + \overline{\xi}(t) = y[t]$$

(for almost all $t \in T_\vartheta$).

The dynamics of the total system (2.1), (2.3) will now be described by the evolution of the domains $X(\vartheta, \cdot) = X[\vartheta]$. Note that under the restrictions of the above the set $X(\vartheta, \cdot)$ need not be convex. In the sequel we shall describe the evolution of the domains $coX[\vartheta]$ — the closed convex hulls of $X[\vartheta]$.

Denote with the symbol

$$x(t, V(\cdot)) = \int_{t_0}^{t} K(t, \tau, V(\cdot))\,B(\tau)\,u(\tau)\,d\tau + K(t, t_0, V(\cdot))\,x^0$$

the solution of system (2.1) with $V(t)$ given, i.e. $A(t) = A^*(t) + V(t)$, with $V(t) \in \mathcal{V}$ and

$$K(t, t_0, V(\cdot)) = \exp \int_{t_0}^{t} (A^*(\tau) + V(\tau))\, d\tau .$$

Due to the mapping

$$x(\vartheta, V(\cdot)) = \int_{t_0}^{t} K(\vartheta, \tau, V(\cdot))\, B(\tau)\, u(\tau)\, d\tau + K(\vartheta, t_0, V(\cdot))\, x^0 \tag{2.5}$$

one is to specify the image of those $V(\cdot)$ that generate for a given $y[t]$ the condition

$$y[t] \equiv G(t)\, x(t, V(\cdot)) + \xi(t), \quad t \in T_\vartheta \tag{2.6}$$

provided $\xi(t) \in Q$.

Let us now form a problem that is dual to the following:

Determine

$$\max \{(l, x)\} = \varepsilon^0$$

under the constraints (2.5), (2.6).

$$\gamma = \inf_{\lambda(\cdot)} \sup_{\omega(\cdot)} \mathfrak{L}(\omega(\cdot), \lambda(\cdot), \tag{2.7}$$

where

$$\mathfrak{L}(\omega(\cdot), \lambda(\cdot)) = (l, (x(\vartheta, V(\cdot)))) - \langle \lambda(\cdot), G(\cdot)\, x(\cdot, V(\cdot)) \rangle - \langle \lambda(\cdot), \xi(\cdot) \rangle + \\ + \langle \lambda(\cdot), y[\cdot] \rangle$$

and (l, x) and $\langle \lambda(\cdot), f(\cdot) \rangle$ denote respectively the scalar products in the space R^n and in the space $L_2^{(m)}$ of functions defined on T_ϑ and where the element $\omega(\cdot) = \{V(\cdot), \xi(\cdot)\}$ is restricted a priori by the inclusions (2.2), (2.4) and $\lambda(\cdot) \in L_2^m$. Under the above mentioned conditions the equality $\gamma^0 = \varepsilon^0$ is true [6]. Therefore the following proposition holds.

Lemma 2.1. The following equality is true

$$\gamma^0 = \varrho(l \mid X(\vartheta, \cdot)) = \max \{(l, x) \mid x \in X(\vartheta, \cdot)\}$$

Obviously $\varrho(l \mid co\, X(\vartheta, \cdot)) = \varrho(l \mid X(\vartheta, \cdot))$, [7]. The solution of the minmax problem (2.7) is a rather complicated procedure (especially in view of the fact that $V(t)$ is a matrix function). We shall therefore make use of an approach based on approximating the solution by means of appropriate problems with quadratic ("energy") constraints.

3. The augmented problem

Let $\omega(\cdot) = \{V(\cdot), \xi(\cdot)\}$ be an arbitrary fixed pair of elements $V(t)$, $(v_{ij}(\cdot) \in L_2)$, $\xi(t)$, $(\xi(\cdot) \in L_2^m)$. Now assume the function $y[t]$ to be generated not by system (2.6) but by the equation

$$y[t] \equiv G(t)\, x(t, V(\cdot)) + \xi(t) + \\ + G(t)\Big(\int_{t_0}^{t} H(t-\tau)\, B(\tau)\, u(\tau)\, d\tau + H(t-t_0)\, x^0\Big) + w(t), \tag{3.1}$$

where the functions $H(t)$, $\xi(t)$ are unknown in advance and restricted by the inequality

$$\int_{t_0}^{\vartheta} \{[H(\tau), N^{-1}(\tau)\, H(\tau)] + (w(\tau), M^{-1}(\tau)\, w(\tau))\}\, d\tau \leq \mu^2, \tag{3.2}$$

(where $M(t) > 0$ is an $n \times n$-matrix, $N(\tau)$ is a linear positive symmetrical operator defined on the space $\mathbf{M}$ of $(n \times n)$-matrix functions, $[H', H'']$ is the scalar product in $\mathbf{M}$).

Take $F^{(i)}$ to be an $n^2 \times n$-matrix, whose elements are $f_{k,s}^{(i)}$; $k = 1, \ldots, n^2$; $s = 1, \ldots, n$;

$$f_{k,s}^{(i)} = \delta_{k,s+n(i-1)}, \qquad (i = 1, \ldots, n)$$

and $e^{(i)}$ to be the orth of the i-th coordinate axis: $e_s^{(i)} = \delta_{i,s}$ (here $\delta_{k,s}$ is the Kronecker delta). Then the n^2-dimensional vector

$$\overline{H} = \sum_{i=1}^{n} F^{(i)}\, H e^{(i)}$$

and

$$\overline{(N(\tau)H)} = \overline{N}\,\overline{H}, \quad [H, NH] = (\overline{H}, \overline{N}\,\overline{H})$$

where $\overline{N}$ is a positive definite diagonal $n^2 \times n^2$-matrix with nonzero diagonal blocks N_i, $(i = 1, \ldots, n)$ of dimension $n \times n$. Note that the n^2-dimensional vector $\overline{H}$ is obtained from H by "stacking" the n-vector valued columns of H from up to down, beginning by the first.

Here $H(t)$ and $w(t)$ may be considered as fictitious perturbations of the pulse transfer function $K(\cdot)$ and of the disturbance $\xi(\cdot)$ respectively.

Denote the pair $\{N(\cdot), M(\cdot)\}$ with the symbol $\mathcal{A}$. The informational set $Z(\vartheta, y[\cdot]\,|\,\omega(\cdot), \mathcal{A})$ of states $z(\vartheta)$ of system

$$z(\vartheta) = z(\vartheta, V(\cdot), H(\cdot)) = \big(K(\vartheta, t_0) + H(\vartheta - t_0)\big)\, x^0 + \\ + \int_{t_0}^{t} \big(K(\vartheta, \tau, V(\cdot)) + H(\vartheta - \tau)\big)\, B(\tau)\, u(\tau)\, d\tau \tag{3.3}$$

that are compatible with the realization $y[t]$ (inputs x^0, $u(t)$ being given and $\omega(\cdot) = \{V(\cdot), \xi(\cdot)\}$ being fixed) is defined as the set of all vectors z for each of which there exists a pair $\eta(\cdot) = \{H(\cdot), w(\cdot)\}$ such that the respective values (3.3) and $y(t)$ (3.1) would satisfy the conditions $z(\vartheta) = z$, $y(t) \equiv y[t]$, under restriction (3.2).

It is useful to underline that for sufficiently large μ the sets $Z(\vartheta, y[\cdot]\,|\,|\omega(\cdot), \mathcal{A})$ are nonvoid. For computing the support function $\varrho(l\,|Z(\vartheta, y[\cdot]\,|\omega(\cdot), \mathcal{A})$ it is necesary to introduce the Lagrangian for the augmented problem. We have

$$\mathfrak{L}_0(\eta(\cdot), \omega(\cdot), \lambda(\cdot)) = \mathfrak{L}_1(\omega(\cdot), \lambda(\cdot)) + \mathfrak{L}_2(\eta(\cdot), \lambda(\cdot))$$

where

$$\mathfrak{L}_1(\omega(\cdot), \lambda(\cdot)) = (l, x(\vartheta, V(\cdot))) - \langle \lambda(\cdot), x(\cdot, V(\cdot))\rangle +$$

$$+ \langle \lambda(\cdot), y[\cdot]\rangle - \langle \lambda(\cdot), \xi(\cdot)\rangle,$$

$$\mathfrak{L}_2(\eta(\cdot), \lambda(\cdot)) = (l, H(\vartheta - t_0)\, x^0) - \langle \lambda(\cdot), w(\cdot)\rangle - \langle \lambda(\cdot), G(\cdot)\, z^*(\cdot)\rangle,$$

$$Z^*(t) = H(t - t_0)\, x^0 + \int_{t_0}^{t} H(t - \tau)\, B(\tau)\, u(\tau)\, d\tau.$$

Similarly to the previous point we arrive at the proposition

Lemma 3.1. The set $Z(\vartheta, \cdot\,|\,\omega(\cdot), \mathcal{A})$ is convex. The following equality is true

$$\varrho(l\,|Z(\vartheta, \cdot\,|\,\omega(\cdot), \mathcal{A})) = \min\{\mathfrak{L}_1(\omega(\cdot), \lambda(\cdot)) + \mathfrak{H}(\lambda(\cdot))\,|\,\lambda(\cdot)\}$$

$$\mathfrak{H}(\lambda(\cdot)) = \max\{\mathfrak{L}_2(\eta(\cdot), \lambda(\cdot))\,|\,\eta(\cdot)\}.$$

Here the maximum is to be taken over all the $\eta(\cdot)$ restricted according to (3.2). The minimum over $\lambda(\cdot)$ is to be taken over all elements of $L_2^{(m)}$ and is attained on an element of the following form

$$\lambda^0(\cdot) = (K + M)^{-1}(T_{tx}\bar{N}T^* + T_t\bar{N}T^*)\, l + \beta(\cdot), \tag{3.4}$$

where the mappings K, M, $\bar{N}$ are determined by the conditions

$$K\lambda(\cdot) = (T_{tx} + T_t)\,\bar{N}(T_{tx} + T_t)^*\,\lambda(\cdot)$$

$$M\lambda(\cdot) = M(\cdot)\,\lambda(\cdot)\ \ (M(\cdot)\,\lambda(\cdot) = M(t)\,\lambda(t),\ t \in T_\vartheta)$$

$$\bar{N}f(\cdot) = \bar{N}(\cdot)\,f(\cdot)\ \ (\bar{N}(\cdot)\,f(\cdot) = \bar{N}(\cdot)\,f(t),\ t \in T_\vartheta)$$

and where the following notations are introduced ($t_0 = 0$)

$$TH(\cdot) = \int_{t_0}^{\vartheta} H(\vartheta - t)\, B(t)\, u(t)\, dt = \int_{t_0}^{\vartheta} ((B(\vartheta - t)\, u(\vartheta - t)) \otimes E))'\, \bar{H}(t)\, dt,$$

$$T_x H(\vartheta) = H(\vartheta)\, x^0 = (x^{0\prime} \otimes E_n)\, \bar{H},$$

$$T_{tx} H(\cdot) = f(\cdot), \quad f(t) = G(t)\, (x^0 \otimes E_n)\, \bar{H}(t),$$

$$T_t H(\cdot) = f_1(\cdot), \; f_1(t) = G(t) \int_{t_0}^{t} \big((B(t-\tau)\, u(t-\tau)) \otimes E\big)\, \bar{H}(\tau)\, d\tau.$$

Here the prime stands for the transpose, the star for the conjugate mapping and the symbol $D \otimes F$ for the Kronecker product of matrices D, F [8].

Lemma 3.2. The domain $Z(\vartheta, \cdot \,|\, \omega(\cdot), \Lambda)$ is an ellipsoid with a support function

$$\varrho(l|Z(\vartheta, \cdot \,|\, \omega(\cdot), \Lambda)) = (l, x(\vartheta, V(\cdot))) + (l, z^0(\vartheta, \omega(\cdot), \Lambda))) + \\ + (\mu^2 - h^2(\vartheta, \omega(.), \Lambda))\, (l, P(\vartheta, \omega(\cdot), \Lambda)\, l)^{1/2} \tag{3.5}$$

where the vector

$$z^0(\vartheta, \omega(\cdot), \Lambda) = (T_{tx} \bar{N} T^* + T_t \bar{N} T^*)^* \, (K + M)^{-1} z(\cdot) \tag{3.6}$$

obtained from $z(\cdot) = z(t)$, $t \in T_\vartheta$,

$$z(t) \equiv y[t] - G(t)\, x(t, V(\cdot)) - \xi(t),$$

does not depend upon the value μ and

$$(l, P(\vartheta, \omega(\cdot), \Lambda)\, l) = (l, (T_x^* \bar{N} T_x + T^* \bar{N} T) - \\ - (T_{tx} \bar{N} T^* + T_t \bar{N} T^*)^* \, (K + M)^{-1} (T_{tx} \bar{N} T^* + T_t \bar{N} T^*)\, l.$$

We shall now turn to the description of the primary problem treating the obtained solutions as auxiliary relations.

4. Geometrical constraints

Consider the union $Z^0(\vartheta, \cdot \,|\, \Lambda)$ of the centres $z^0(\vartheta, \omega(\cdot), \Lambda) = x(\vartheta, V(\cdot) + z^0(\vartheta, \omega(\cdot), \Lambda)$ of the ellipsoids $Z(\vartheta, \,|\, \omega(\cdot), \Lambda)$ over all $\omega(\cdot)$

$$Z^0(\vartheta, \cdot \,|\, \Lambda) = \bigcup_{\omega(\cdot)} \{z^0(\vartheta, \omega(\cdot), \Lambda) \,|\, \omega(\cdot)\}. \tag{4.1}$$

Denoting $p(\cdot \,|\, l, \Lambda) = (K + M)^{-1} (T_{tx} \bar{N} T^* + T_t \bar{N} T^*)\, l$ we determine from (3.6)

Lemma 4.1. The following equality is true

$$\varrho(l \,|\, Z^0(\vartheta, \cdot \,|\, \Lambda)) = \Psi(l, \Lambda)$$

$$\Psi(l, \Lambda) = \sup \{(l, x(\vartheta, V(\cdot))) - \\ - (p(\cdot \,|\, l, \Lambda)\, G(\cdot)\, x(\cdot, V(\cdot))) + \langle p(\cdot, l, \Lambda), \; y[\cdot] - \xi(\cdot) \rangle \tag{4.2}$$

It is not difficult to observe that the right-hand side of expression (4.2) differs from the value to be minimized in (2.7) by the only substitution of the multiplier $\lambda(\cdot)$ by $p(\cdot \mid l, \mathcal{A})$. It then follows immediately

Theorem 4.1. Whatever is the pair $\{N(\cdot),\ M(\cdot)\}$, the inclusion is true

$$co X(\vartheta,\ \cdot) \subseteq Z^0(\vartheta,\ \cdot \mid \mathcal{A})\,. \tag{4.3}$$

Corollary 4.1. The inclusion

$$co\, X(\vartheta,\ \cdot) \subseteq \bigcap \{Z^0(\vartheta,\ \cdot \mid \mathcal{A}) \mid \mathcal{A}\} = Z^0(\vartheta,\ \cdot) \tag{4.4}$$

is true over all $\mathcal{A} = \{N(\cdot), M(\cdot) : \overline{N}(t) > 0,\ M(t) > 0\}$.

In analogy with theorem 5.1 of paper [1] one may establish the following assertion.

Theorem 4.2. The following equality is true

$$\varrho(l \mid Z^0(\vartheta,\ \cdot)) = f^{**}(l)$$

where

$$f(l) = \inf \{\varrho(l \mid Z^0(\vartheta,\ \cdot \mid \mathcal{A})) \mid \mathcal{A}\}$$

over all

$$\mathcal{A} = \{N(\cdot), M(\cdot) : \overline{N}(t) > 0,\ M(t) > 0\}$$

and $f^{**}(l)$ is a convex positively homogeneous function that turns to be the second conjugate relative to $f(l)$.

In other words $f^{**}(l)$ is the conjugate relative to the function

$$f^{*}(l) = \sup \{(x, l) - f(l) \mid l \in R^n\}$$

where $f^{*}(x)$ is the conjugate for $f(l)$, see [7]).

Hence the solution of the augmented problem generates a bundle $\{Z^0(\vartheta, \cdot \mid \mathcal{A})\}$ of sets $Z^0(\vartheta, \cdot \mid \mathcal{A})$ that are related to all possible $\mathcal{A} = \{N(\cdot),\ M(\cdot)\}$. The intersection of all the sets that belong to the bundle will already include the desired domain $X(\vartheta,\ \cdot)$. We shall obtain the equations that describe the evolution of the bundle $\{Z^0(\vartheta, \cdot \mid \mathcal{A})\}$.

5. Dynamic equations

For a completeness in the description of the solutions, we shall first state the equations that describe the dynamics of the ellipsoids $\{Z^0(\vartheta,\ \cdot \mid \omega(\cdot), \mathcal{A})\}$.

Theorem 5.1. The support function $\varrho(l \mid Z^0(\vartheta, \cdot \mid \omega(\cdot), \mathcal{A}))$ of the ellipsoid $Z(\vartheta,\ \cdot \mid \omega(\cdot), \mathcal{A})$ is described by formula (3.5) where $x(\vartheta,\ V(\cdot)) = x(\vartheta, \cdot)$ and

$z^0(\vartheta, \omega(\cdot), \Lambda) = z^0(\vartheta, \cdot)$, $P(\vartheta, \omega(\cdot), \Lambda) = P(\vartheta, \cdot)$, $h^2(\vartheta, \omega(\cdot), \Lambda) = h^2(\vartheta, \cdot)$ are the solutions of the following system of equations (provided $\dot{u} = Cu$, $C = \{c_{kj}\}$),

$$\frac{dx(\vartheta, \cdot)}{d\vartheta} = (A^*(\vartheta) + V(\vartheta))\, x(\vartheta, \cdot) + B(\vartheta)\, u(\vartheta), \quad x(t_0) = x^0, \tag{5.1}$$

$$z^0(\vartheta, \cdot) = \sum_{k=1}^{p} \sum_{j=1}^{p} z^{(k,j)}(\vartheta, \cdot)$$

$$\frac{dz^{(k,j)}(\vartheta, \cdot)}{d\vartheta} = C_{kj} E z^{(k,j)}(\vartheta, \cdot) + P^{(k,j)}(\vartheta, \cdot) G'(\vartheta) (M^{-1}(\vartheta) + S(\vartheta))^{-1} (y[\vartheta] -$$

$$- G(\vartheta)\, z(\vartheta, \cdot) - w(\vartheta)) - P^0(\vartheta, \cdot) G'(\vartheta) (M^{-1}(\vartheta) + S(\vartheta))^{-1} G(\vartheta)\, z^{(k,j)}(\vartheta, \cdot), \tag{5.2}$$

$$z^{(k,j)}(t_0) = 0, \quad (k, j = 1, \ldots, p),$$

$$S(\vartheta) = G(\vartheta)\, \bar{P}(\vartheta, \cdot) G'(\vartheta), \quad P^0(\vartheta_j) = P(\vartheta_j) - \bar{P}(\vartheta_j),$$

$$P(\vartheta, \cdot) = \sum_{j=1}^{p} \sum_{l=1}^{p} P^{(j,l)}(\vartheta, \cdot) + \bar{P}(\vartheta, \cdot),$$

$$\frac{dP^{(j,l)}(\vartheta, \cdot)}{d\vartheta} = c_{jl} P^{(j,l)}(\vartheta, \cdot) + c_{l,j} P^{(l,j)}(\vartheta, \cdot) -$$

$$- \frac{1}{2} \left(P^{(j,l)}(\vartheta, \cdot) G'(\vartheta) (M^{-1}(\vartheta) + S(\vartheta))\right)^{-1} G(\vartheta)\, P^0(\vartheta, \cdot) +$$

$$+ P^0(\vartheta, \cdot) G'(\vartheta) (M^{-1}(\vartheta) + S(\vartheta))^{-1} G(\vartheta)\, P^0(\vartheta, \cdot) - \tag{5.3}$$

$$- \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{p} \left(c_{jl}\, u_l(t_0)\, b_{ij}\, x_k^0 N_i(\vartheta) + c_{ij}(\vartheta)\, u_j(t_0)\, b_{il}\, x_k^0 N_i(\vartheta)\right) +$$

$$+ \delta_{jl} \sum_{i=1}^{p} \sum_{k=1}^{p} u_j(t_0)\, b_{ij} b_{ik}\, u_k(t_0)\, N_i(\vartheta),$$

$$P^{(j,l)}(t_0) = 0,$$

$$\bar{P}(\vartheta, \cdot) = (x^0 \otimes E_n)\, \bar{N}(\vartheta)\, (x^0 \otimes E_n), \tag{5.4}$$

$$\frac{dh^2(\vartheta, \cdot)}{d\vartheta} = (y[\vartheta] - G([\vartheta] z^0(\vartheta, \cdot) - G(\vartheta)\, x(\vartheta, \cdot)) \times$$

$$\times (M^{-1}(\vartheta) + S(\vartheta))^{-1} (y[\vartheta] - G(\vartheta)\, z^0(\vartheta, \cdot) - G(\vartheta)\, x^0(\vartheta, \cdot)), \tag{5.5}$$

$$h^2(t_0, \cdot) - 0,$$

where $x \otimes E$ stands for the Kronecker product of matrices x, E, [8] and δ_{jl} is the Kronecker delta.

In order to construct a set $Z^0(\vartheta, \cdot)$ that belongs to the bundle $\{Z^0(\vartheta, \cdot \mid \Lambda)\}$, one has to consider the variety of all the trajectories

$$\bar{z}^0(\vartheta, \cdot) = x(\vartheta, \cdot) + z^0(\vartheta, \cdot)$$

that correspond to all the possible values $V(t)$, $w(t)$ restricted by inclusions (2.2), (2.4).

Theorem 5.2. Each of the sets $Z^0(\vartheta, \cdot \mid \Lambda)$ may be determined as the variety of vectors $\bar{z}^0(\vartheta, \cdot)$ that are obtained for a given Λ as the sum of the solutions $x(\vartheta, \cdot) = x(\vartheta, V(\cdot))$ and $z^{(k,j)}(\vartheta, \cdot) = z^{(k,j)}(\vartheta, \omega(\cdot), \Lambda)$, $(k, j = 1, \ldots, p)$ of a system that consists of the inclusions (provided $\dot{u} = Cu$)

$$\frac{dx^0(\vartheta, \cdot)}{d\vartheta} \in (A^*(\vartheta) + \mathcal{P}) x^0(\vartheta, \cdot) + B(\vartheta) u(\vartheta), \tag{5.6}$$

$$x^0(t_0) = x^0,$$

$$\frac{dz^{(k,j)}(\vartheta, \cdot)}{d\vartheta} \in c_{kj} Ez^{(k,j)}(\vartheta, \cdot) + P^{(k,j)}(\vartheta, \cdot) G'(\vartheta) (M^{-1}(\vartheta) + S(\vartheta))^{-1} \times \tag{5.7}$$

$$\times (y[\vartheta] - G(\vartheta) z(\vartheta, \cdot) - Q) - P^0(\vartheta, \cdot) G'(\vartheta (M^{-1}(\vartheta) + \bar{P}(\vartheta, \cdot))^{-1} G(\vartheta) z^{(k,j)}(\vartheta, \cdot),$$

$$z^{(k,j)}(t_0) = 0$$

and of equations (5.3)–(5.5). The sets $Z^0(\vartheta, \cdot \mid \Lambda)$ are convex and compact.

The informational set $X(\vartheta, \cdot)$ may now be estimated by means of a procedure based on theorems 4.1 and 4.2. We also note that among the sets $Z^0(\vartheta, \cdot \mid \Lambda)$ it is possible to select a set optimal in a certain sense that includes $X(\vartheta, \cdot)$ — for example the one optimal with respect to the diameter or the volume or optimal with respect to a partial ordering generated by the operation of inclusion. The situation is similar to the one mentioned in pt. 8 of paper [1].

References

1. *Kurzanskii, A. B.*, Dynamic Control System Estimation Under Uncertainty Conditions. I. — Probl. Control & Inform. Theory, 1980.
2. *Krasovskii, N. N.*, On the Theory of Controllability and Observability of Linear Dynamic Systems. — Prikl. Mat. i. Mech., 1964, **28**, No. 1.
3. *Kurzanskii, A. B.*, Control and Observation Under Conditions of Uncertainty. Moscow, Nauka, 1977.
4. *Kurzanskii, A. B.*, On adaptive Control in Mechanical Systems. — Theoret. & Appl. Mech. Bulgar. Acad. Sci. (Sofia), 1978, vol. **9**, No. 2.
5. *Fogel, E.*, System Identification Via Membership Set Constraints with Energy Constrained Noise. — IEEE Trans. Automat. Control, 1979, vol. **AC–24**, No. 5.

6. *Ekeland, I., Temam, P.*, Convex Analysis and Variational Problems. North-Holland Publ. Co., 1976.
7. *Rockafellar, R. T.*, Convex Analysis. Princeton: Univ. Press, 1970.
8. *Lankaster, P.*, Theory of Matrices. Academic Press, 1969.

Об оценивании динамики управляемых систем в условиях неопределенности, II

А. Б. КУРЖАНСКИЙ

(Свердловск)

В настоящей работе рассматривается задача об оценивании по результатам измерения (наблюдения) динамики линейных управляемых систем в условиях неопределенной информации о параметрах системы. Располагая реализацией доступных измерению выходных переменных системы, а также зная априорные ограничения на параметры и на ошибки в системе измерения, оказывается возможным построить в фазовом пространстве некоторые информационные множества, совместимые с априорной и текущей информацией об исследуемом объекте. Эволюция указанных множеств во времени будет характеризовать динамику процесса оценивания, описание указанной эволюции доставит решение задачи об идентификации системы в условиях неопределенности.

В настоящей работе, посвященной задаче идентификации, рассматриваются системы с априорными геометрическими ограничениями на параметры. Предлагаемое решение опирается на аппроксимацию его при помощи решений подходящим образом подобранных задач идентификации с квадратичными ограничениями на весовые функции и возмущения в системе (решение последних задач также приводится в работе).

А. Б. Куржанский
Институт математики и механики УНЦ АН СССР
СССР, 620219 Свердловск ГСП-384,
ул. С. Ковалевской, 16

Problems of Control and Information Theory, Vol. 10 (1), pp. 43—52 (1981)

RECENT RESULTS ON NONPARAMETRIC REGRESSION ESTIMATE AND MULTIPLE CLASSIFICATION

L. GYÖRFI

(Budapest)

(Received 25 January, 1980)

In real-life information processing there is usually a partial or total lack of information about the underlying distribution, therefore a parametric model of the problem is not applicable. The main feature of nonparametric procedures is that one does not make use of any a priori information when devising the algorithm itself. However, consistency of these procedures is guaranteed under some assumption, which can be verified only under some a priori information. Universal consistency is therefore particularly important in using nonparametric techniques in real-life information processing, since by this consistency is guaranteed for all possible distributions. We prove universal consistency for nonparametric regression and nonparametric decision procedures of fundamental interest. More distinctly, strong and weak universal consistency of nonparametric regression estimate and multiple classification by the nearest neighbor and kernel estimates are investigated.

1. Nonparametric regression

Let (X, Y) be a random vector taking values in $\mathbf{R}^d \times \mathbf{R}$. If $\mathbf{E}|Y| < +\infty$, then the *regression function*

$$m(z) = \mathbf{E}(Y/X = z) \qquad z \in \mathbf{R}^d$$

can be defined.

The estimate of a regression function is based on a sample $Z^n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where (X_i, Y_i) $i = 1, 2, \ldots$ are independent and identically distributed random vectors having the same distribution as (X, Y). Assume that (X, Y) and the sample Z^n are independent. The estimate of the regression function is a measurable function of $X, (X_1, Y_1), \ldots, (X_n, Y_n)$. The estimate $m_n(X)$ is said to be *weakly consistent* if

$$\lim_{n \to \infty} \mathbf{E}\,|m_n(X) - m(X)|^p = 0$$

for each $p \geq 1$ for which $\mathbf{E}|Y|^p < \infty$. m_n is called *strongly consistent* if

$$\lim_{n \to \infty} m_n(X) = m(X) \qquad \text{a.s.}$$

We say that m_n is *strongly* (*weakly*) *universally consistent* if m_n is strongly (weakly) consistent for all distributions of (X, Y). Two types of conditions on

the distribution of (X, Y) are applied. On the one hand the existence of the density of X is assumed, on the other hand some continuity conditions on the regression function are used.

Stone (1977) investigated a general class of regression estimates of the form

$$m_n(X) = \sum_{i=1}^{n} W_{n,i}(X, X_1, \ldots, X_n)\, Y_i \stackrel{\triangle}{=} \sum_{i=1}^{n} W_{n,i}(X)\, Y_i$$

and proved the following useful tool for verifying the weak universal consistency for several choices of weights $\{W_{n,i}\}$:

Theorem 1 (Stone, 1977). Assume that there exist $C \geq 1$, $D \geq 0$ such that for every nonnegative Borel function f

$$\mathbf{E} \sum_{i=1}^{n} |W_{n,i}(X) f(X_i)| \leq C \mathbf{E} f(X) \quad \text{for all} \quad n \geq 1 \tag{1}$$

and

$$\mathbf{P}\left(\sum_{i=1}^{n} |W_{n,i}(X)| \leq D \right) = 1 \qquad \text{for all} \quad n \geq 1; \tag{2}$$

$$\sum_{i=1}^{n} |W_{n,i}(X)| \chi_{\{\|X_i - X\| \geq a\}} \to 0 \tag{3}$$

in probability for all $a > 0$;

$$\sum_{i=1}^{n} W_{n,i}(X) \to 1 \quad \text{in probability;} \tag{4}$$

$$\max_{i \leq n} |W_{n,i}(X)| \to 0 \quad \text{in probability;} \tag{5}$$

Then m_n is weakly consistent. ■

Unfortunately, there is no general result on strong consistency which makes it possible to show the strong universal consistency of particular estimates.

2. Nonparametric multiple classification

For a multiple classification problem the observation X is a random vector taking values in $\mathbf{R}^d$ and the random variable (label) Y takes values in $\{1, 2, \ldots, M\}$. The classification problem is to decide on Y given X, minimizing the probability of misdecision. If the distribution of (X, Y) is known then the Bayesian decision solves this problem. Let P_i denote the a posteriori probability of $\{Y = i\}$ given X, i.e.

$$P_i(z) \stackrel{\triangle}{=} \mathbf{P}(Y = i / X = z) \qquad i = 1, 2, \ldots, M.$$

The Bayesian decision is defined as follows:

$$g_B(X) \stackrel{\triangle}{=} i \quad \text{if} \quad P_i(X) > P_j(X) \qquad j < i$$
$$\text{and} \quad P_i(X) \geq P_j(X) \qquad j > i,$$

then

$$L^* \stackrel{\triangle}{=} \mathbf{P}(g_B(X) \neq Y) = 1 - \mathbf{E}(\max_i P_i(X)) = \inf_{g: \mathbf{R}^d \to \{1,2,\ldots,M\}} \mathbf{P}(g(X) \neq Y).$$

If the a posteriori probabilities are unknown, some estimates can be applied. Observe that P_i is the regression function for the indicator of the event $\{Y = i\}$ $(i = ,1\ 2, \ldots, M)$. Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, $P_{i,n}$ denotes the estimate of P_i $(i = 1, 2, \ldots, M)$, and the decision will be

$$g_n(X) \stackrel{\triangle}{=} i \quad \text{if} \quad P_{i,n}(X) > P_{j,n}(X) \qquad j < i$$
$$\text{and} \quad P_{i,n}(X) \geq P_{j,n}(X) \qquad j > i.$$

This decision is qualified by the conditional error probability

$$L_n \stackrel{\triangle}{=} \mathbf{P}(g_n(X) \neq Y/(X_1\, Y_1), \ldots, (X_n, Y_n)).$$

The decision g_n is called *strongly* (*weakly*) *Bayes risk consistent* if $L_n \xrightarrow{n} L^*$ a.s. (in probability) and is called *strongly* (*weakly*) *universally Bayes risk consistent* if it is strongly (weakly) Bayes risk consistent for all finite M and all distributions of (X, Y).

By the bound of Györfi (1978, Lemma 4)

$$0 \leq L_n - L^* \leq \mathbf{E}\left(\sum_{i=1}^{n} |P_{i,n}(X) - P_i(X)|/Z^n\right)$$

therefore applying the Dominated Convergence Theorem we get

Theorem 2. If $P_{i,n}$ is a weakly (strongly) universally consistent estimate of P_i $(i = 1, 2, \ldots, M)$ then the decision rule g_n is weakly (strongly) universally Bayes risk consistent. ■

Theorem 2 gives a relation between the estimate and the classification: it is sufficient to deal with the nonparametric estimates of the a posteriori probabilities.

3. Nearest neighbor estimate

The idea of the nearest neighbor (NN) regression estimate came from the classification problems (Cover, Hart 1967, Cover 1968) and from the density estimate (Loftsgaarden, Quesenberry 1965, Fukunaga, Hostetler 1973, Mack, Rosenblatt 1979). Györfi, Györfi (1975), Stone (1977), Beck (1979) investigated the properties of the NN estimates.

For a fixed $z \in \mathbf{R}^d$ let $(X^z_{1,n}, Y^z_{1,n}), \ldots, (X^z_{n,n}, Y^z_{n,n})$ be the nearest neighbor ordering of the sample Z^n according to increasing values of $||z - X_i||$. In case of tie, i.e. $||X_i - z|| = ||X_j - z||$, $i < j$, X_i preceeds X_j in the ordering. Let $v_{n,1} \geq v_{n,2} \geq \ldots \geq v_{n,n} \geq 0$ $(n = 1, 2, \ldots)$ be a double array of weights for which

$$\sum_{i=1}^{n} v_{n,i} = 1 \qquad (n = 1, 2, \ldots)$$

then the NN estimate is defined by

$$m_n(z) = \sum_{i=1}^{n} v_{n,i} Y^z_{i,n}. \tag{6}$$

Stone (1977) proved the weak universal consistency of a slightly complicated version of (6):

Theorem 3 (Stone, 1977). If

$$\lim_{n \to \infty} v_{n,1} = 0$$

and

$$\lim_{n \to \infty} \sum_{i=k_n}^{n} v_{n,i} = 0$$

for a sequence k_n for which

$$\lim_{n \to \infty} k_n/n = 0,$$

then the NN estimate is weakly universally consistent. ■

Concerning the strong consistency, first Devroye, Wise (1979) proved a universal result for an iterative modification of the NN rule, where the estimate (and therefore the proof) is based on an independent segmentation of the sample. For (6) Devroye (1980) gave the conditions of the strong universal consistency:

Theorem 4 (Devroye, 1980). Assume that there exists a sequence k_n, $n = 1, 2, \ldots$ for which

$$\lim_{n\to\infty} k_n/n = 0 \quad \lim_{n\to\infty} k_n/\log n = \infty \tag{7}$$

$$\sup_n k_n v_{n,1} < +\infty$$

$$\lim_{n\to\infty} \sum_{i=k_n}^{n} v_{n,i} = 0$$

then the NN estimate is a strongly universally consistent estimate provided that $|Y|$ is bounded a.s. ■

The most popular choice of $\{v_{n,i}\}$ is the uniform NN weights, where

$$v_{n,i} = \begin{cases} \dfrac{1}{k_n} & \text{if} \quad i \leq k_n \\ 0 & \text{if} \quad i > k_n, \end{cases}$$

In this case (7) is the only condition on k_n. The concept on the NN decision rule was introduced for more general sample space, where X takes values from an arbitrary separable metric space with the metric ϱ, and the Euclidean metric of $\mathbf{R}^d$ is replaced by the actual metric ϱ (Cover, Hart 1967). The most interesting feature of the NN procedures was that the asymptotic error probability was independent of the metric ϱ. Theorem 4 can also be formulated in metric space:

Theorem 5. Let X take values in a separable metric space $\mathbf{Z}$, then under the condition (7) the uniform k_n-NN estimate is strongly universally consistent provided that Y is bounded a.s. ■

Proof. We prove that $\lim_{n\to\infty} m_n(z) = m(z)$ a.s. for each $z \in \mathbf{Z}$ mod μ, where μ stands for the probability measure of X. Let $S_{z,r}$ denote the sphere in $\mathbf{Z}$ with radius r centered at z. For an arbitrary $r > 0$ and $z \in \mathbf{Z}$ for which $\mu(S_{z,r}) > 0$ the joint conditional distribution of $(X^z_{1,n}, \ldots, X^z_{k_n,n})$ given $\varrho(z, X^z_{k_n+1,n}) = r$ is the same as the distribution of $(\tilde{X}^z_{1,k_n}, \ldots, \tilde{X}^z_{k_n,k_n})$, which is the nearest neighbor ordering of the independent and identically distributed random elements $(\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_{k_n})$, where

$$\mathbf{P}(\tilde{X}_i \in A) = \frac{\mu(A \cap S_{z,r})}{\mu(S_{z,r})},$$

therefore for each $\varepsilon > 0$

$$\mathbf{P}\left\{\left|\frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}^z) - \frac{1}{\mu(S_{z,r})}\int_{S_{z,r}} m(u)\,\mu(du)\right| \geq \frac{\varepsilon}{2} \Big/ \varrho(z, X_{k_n+1,n}^z) = r\right\} =$$

$$= \mathbf{P}\left\{\left|\frac{1}{k_n}\sum_{i=1}^{k_n} m(\tilde{X}_{i,k_n}^z) - \frac{1}{\mu(S_{z,r})}\int_{S_{z,r}} m(u)\,\mu(du)\right| \geq \frac{\varepsilon}{2}\right\} = \tag{8}$$

$$= \mathbf{P}\left\{\left|\frac{1}{k_n}\sum_{i=1}^{k_n} m(\tilde{X}_i) - \frac{1}{\mu(S_{z,r})}\int_{S_{z,r}} m(u)\,\mu(du)\right| \geq \frac{\varepsilon}{2}\right\} \leq 2\,e^{-\frac{\varepsilon^2}{4c^2}k_n}$$

independently of r, where the Berstein–Chernoff–Hoeffding inequality was used and C stands for the a.s. bound of $|Y|$. Put

$$A_1 = \{z;\ \mu(S_{z,r}) > 0,\ \forall\, r > 0\}.$$

$\mathbf{Z}$ is a separable metric space, therefore A_1 is a Borel set of μ measure 1 (Cover, Hart 1967). Denote

$$A_2 = \left\{z \in A_1;\ \lim_{r\to 0}\frac{1}{\mu(S_{z,r})}\int_{S_{z,r}} m(u)\,\mu(du) = m(z)\right\}$$

then, by the Pointwise Density Theorem $\mu(A_2) = 1$ (Federer 1969, Theorem 2.9.8), provided $\int |m(u)|\,\mu(du) < \infty$. Put

$$A_3 = \left\{z;\ \sum_{n=1}^{\infty}\mathbf{P}(\varrho(z, X_{k_n+1,n}^z) > r) < +\infty,\ \forall\, r > 0\right\}.$$

If $\mathbf{Z}$ is separable and $\lim_{n\to\infty} k_n/n = 0$ then $A_1 = A_3$. We show that for each $z \in A_1 \cap A_2 \cap A_3 = A_2$

$$\lim_{n\to\infty}\frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}^z) = m(z) \quad \text{a.s.}$$

For an arbitrary fixed $\varepsilon > 0$ and $z \in A_2$ choose $r_0 > 0$ such that for each $0 < r < r_0$

$$\left|\frac{1}{\mu(S_{z,r})}\int_{S_{z,r}} m(u)\,\mu(du) - m(z)\right| < \frac{\varepsilon}{2}$$

then by (8)

$$\mathbf{P}\left\{\left|\frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}^z) - m(z)\right| > \varepsilon\right\} \leq \mathbf{P}(\varrho(z, X_{k_n+1,n}^z) > r_0) +$$

$$+ \mathbf{P}\left\{\varrho(z, X_{k_n+1,n}^z) \leq r_0,\ \left|\frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}^z) - \frac{1}{\mu(S_{z,r})}\int_{S_{z,r}} m(u)\,\mu(du)\right| > \frac{\varepsilon}{2}\right\} \leq$$

$$\leq \mathbf{P}(\varrho(z, X_{k_n+1}^z) > r_0) + \chi_{\{r\leq r_0\}}\,2\cdot e^{-\frac{\varepsilon^2}{4c^2}k_n}. \tag{9}$$

Because of $z \in A_3$ and $\lim_{n\to\infty} \frac{k_n}{\log n} = \infty$, the Borel–Cantelli Lemma implies that

$$\lim_{n\to\infty} \frac{1}{k_n} \sum_{i=1}^{k_n} m(X_{i,n}^z) = m(z) \quad \text{a.s.}$$

Györfi, Györfi (1975, Lemma 3) proved that

$$\lim_{n\to\infty} \left(\frac{1}{k_n} \sum_{i=1}^{k_n} Y_{i,n}^z - \frac{1}{k_n} \sum_{i=1}^{k_n} m(X_{i,n}^z) \right) = 0 \quad \text{a.s.}$$

thus the proof is complete. ∎

4. Kernel estimates

Let K be an arbitrary density defined on $\mathbf{R}^d$, and h_n $n = 1, 2, \ldots$ a sequence of positive numbers, then the most popular kernel estimate is as follows

$$m_n(X) = \sum_{i=1}^{n} \frac{K\left(\frac{X - X_i}{h_n}\right)}{\sum_{j=1}^{n} K\left(\frac{X - X_j}{h_n}\right)} Y_i \tag{10}$$

which is derived from the Rosenblatt–Parzen density estimate. This estimate was introduced by Nadaraya (1964). An iterative version of (10) was first motivated by classification problems (Wolwerten, Wagner 1969):

$$m_n(X) = \sum_{i=1}^{n} \frac{\frac{1}{h_i^d} K\left(\frac{X - X_i}{h_i}\right)}{\sum_{j=1}^{n} \frac{1}{h_j^d} K\left(\frac{X - X_j}{h_j}\right)} Y_i . \tag{11}$$

(See Greblicki 1974). The most simple recursive estimate is due to Révész (1973):

$$m_{n+1}(X) = m_n(X) + \frac{1}{(n+1)\, h_{n+1}^d} \left(Y_{n+1} - m_n(X)\right) K\left(\frac{X - X_{n+1}}{h_{n+1}}\right). \tag{12}$$

(See also Györfi, Kóczy (1979).)

Stone (1977) did not verify the weak universal consistency of kernel estimates. Devroye, Wagner (1980) proved the following:

Theorem 6 (Devroye, Wagner 1980). Suppose the constants $0 < r_1 \leq r_2$, $0 < c_1 \leq c_2$ such that

$$c_1 \chi_{\{x \in S_{0,r_1}\}} \leq K(x) \leq c_2 \chi_{\{x \in S_{0,r_2}\}} \tag{13}$$

and

$$\lim_{n \to \infty} h_n = 0, \quad \lim_{n \to \infty} n h_n^d = \infty. \tag{14}$$

Then the estimate (10) is weakly universally consistent. ■

Devroye (1980) proved the strong universal consistency of (10):

Theorem 7 (Devroye, 1980). Under the conditions of Theorem 6 (10) is strongly universally consistent if, in addition, $|Y|$ is bounded a.s. and $\lim_{n \to \infty} \frac{n h_n^d}{\log n} = \infty$. ■

For the estimate (11) neither weak nor strong universal result exists, for (12) there is a weak universal result:

Theorem 8. Under the conditions of Theorem 6 assume that

$$K(z) = K(-z) \quad \text{for each} \quad z \in \mathbf{R}^d \tag{15}$$

and

$$\frac{1}{n h_n^d} c_2 \leq 1 \qquad n = 1, 2, \ldots$$

Then (12) is weakly universally consistent. ■

Proof. The proof is the verification of the conditions of Theorem 1. In case of (12), $W_{n,i}$ might be written in the form

$$W_{n,i}(z) = \begin{cases} \prod_{l=i+1}^{n} \left(1 - \frac{1}{l h_l^d} K\left(\frac{z - X_l}{h_l}\right)\right) \frac{1}{i h_i^d} K\left(\frac{z - X_i}{h_i}\right) & i \neq n \\ \frac{1}{n h_n^d} K\left(\frac{z - X_n}{h_n}\right). & i = n. \end{cases} \tag{16}$$

It is easy to check by induction that

$$\sum_{i=1}^{n} W_{n,i}(z) = 1 \quad \text{for each} \quad z \in \mathbf{R}^d, \quad n \geq 1 \tag{17}$$

therefore (2) and (4) are met. (5) is proved if

$$\lim_{n \to \infty} \sum_{i=1}^{n} \mathbf{E} W_{n,i}^2(z) = 0 \quad \text{for each} \quad z \in \mathbf{R}^d \bmod \mu. \tag{18}$$

By (16)

$$\sum_{i=1}^{n} \mathbf{E}\, W_{n,i}^2(z) \leq \sum_{i=1}^{n} \mathbf{E}\, W_{n,i}(z)\, \frac{\left(\frac{1}{ih_i^d}\right)^2 \mathbf{E}\, K^2\left(\frac{z - X_i}{h_i}\right)}{\frac{1}{ih_i^d}\, \mathbf{E}\, K\left(\frac{z - X_i}{h_i}\right)} \leq \sum_{i=1}^{n} \mathbf{E}\, W_{n,i}(z)\, \frac{c_2}{ih_i^d}.$$

Because of $\lim_{n\to\infty} nh_n^d = \infty$ and (17) the relation (18) is met applying the Toeplitz Theorem (Ash 1972, Lemma 7.1.1) if for each fixed i

$$\lim_{n\to\infty} \mathbf{E}\, W_{n,i}(z) = 0 \quad \text{for each} \quad z \in \mathbf{R}^d \bmod \mu.$$

The sample is i.i.d. therefore

$$\mathbf{E}\, W_{n,i}(z) \leq e^{-c_1 \sum_{l=i+1}^{n} \frac{1}{lh_l^d} \mu(S_{z,r_1 h_l})} \cdot \frac{c_2}{ih_i^d}.$$

By a nice measure theoretic result of Devroye (1980, Lemma 2.2)

$$\lim_{n\to\infty} \frac{h_n^d}{\mu(S_{z,r_1 h_n})}$$

exists for each $z \in \mathbf{R}^d \bmod \mu$, therefore

$$\sum_{l=1}^{\infty} \frac{1}{lh_l^d}\, \mu(S_{z,r_1 h_l}) = +\infty \quad \text{for each} \quad z \in \mathbf{R}^d \bmod \mu.$$

(3) can be proved also by Toeplitz argument, since $\lim_{n\to\infty} h_n = 0$ and K has finite support. For the remaining condition (1) let $b_{n+1}(X)$ be the expression of $m_{n+1}(X)$ in (12) if Y_{n+1} is replaced by $f(X_{n+1})$ where f is an arbitrary nonnegative Borel function, then one has to prove that there exists C such that

$$\mathbf{E}\,|b_n(X)| = \mathbf{E}\, b_n(X) = C\, \mathbf{E}\, f(X).$$

By the recursive form (12)

$$\mathbf{E}(b_{n+1}(X)/Z^n) = \mathbf{E}(b_n(X)/Z^n) +$$

$$+ \frac{1}{(n+1)\, h_{n+1}^d} \int\int \big(f(x) - b_n(y)\big)\, K\left(\frac{x - y}{h_n}\right) Q(dx)\, Q(dy) = \mathbf{E}\big(b_n(X)/Z^n\big) +$$

$$+ \frac{1}{(n+1)\, h_{n+1}^d} \int\int \big(f(x) - b_n(x)\big)\, K\left(\frac{x - y}{h_n}\right) Q(dx)\, Q(dy) =$$

$$= \mathbf{E}\big(b_n(x)/Z^n\big) + \frac{1}{(n+1)\, h_{n+1}^d}\, \mathbf{E}\left(\big(f(X) - b_n(X)\big)\, K\left(\frac{X - X_{n+1}}{h_{n+1}}\right) \Big/ Z^n\right) \quad (19)$$

where the condition (15) was applied. From (17) and (19)

$$\mathbf{E}\, b_n(X) = E \sum_{i=1}^{n} W_{n,i}(X)\, f(X) = E f(X). \quad \blacksquare$$

References

Ash, R. B., Real analysis and Probability, Academic Press, 1972.

Cover, T. M., Estimation by nearest neighbor rule, IEEE Trans. on Information Theory, vol. **IT–14**, pp. 50–55, 1968.

Cover, T. M., Hart, P. E., Nearest neighbor pattern classification, IEEE Trans. on Information Theory, vol. **IT–13**, pp. 21–27, 1967.

Devroye, L., On the almost everywhere convergence of nonparametric regression function estimates, Submitted to the Annals of Statistics, 1980.

Devroye, L. P., Wagner, T. J., Disribution-free consistency results in nonparametric discrimination and regression function estimation, Annals of Statistics, 1980 (to appear).

Devroye, L. P., Wise, G. L., Consistency of a recursive nearest neighbor regression function estimate, J. Multivariate Analysis 1980 (to appear).

Fukunaga, K., Hostetler, L. D., Optimization of k-nearest neighbor density estimates, IEEE Trans. on Information Theory, vol. **IT–19**, pp. 320–326, 1973.

Greblicki, W., Asymptotically optimal probabilistic algorithms for pattern recognition and identification. Monografie No. *3*, Prace Naukowe Instytutu Cybernetyki Technicznej Politechniki Wroclawsjiej No. *18*, Wroclaw, Poland, 1974.

Györfi, L., On the rate of convergence of nearest neighbor rules. IEEE Trans. on Information Theory, vol. **IT–24**, pp. 509–512, 1978.

Györfi, L., Györfi, Z., On the nonparametric estimate of aposteriori probabilities of simple statistical hypotheses, in Topics in Information Theory, ed. I. Csiszár, P. Elias, North-Holland, 1975.

Györfi, L., Kóczy, L. T., An algorithm for nonparametric decision motivated by fuzzy approach, Problems of Control and Information Theory, vol. **8**, pp. 229–238, 1979.

Loftsgaarden, D. O., Quesenberry, C. P., A nonparametric estimate of a multivariate density function, Annals of Mathematical Statistics, vol. **36**, pp. 1049–1051, 1965.

Mack, Y. P., Rosenblatt, M., Multivariate k-nearest neighbor density estimates, J. Multivariate Analysis, 1979.

Nadaraya, E. A., On estimating regression, Theory of Probability and its Applications, vol. **9**, pp. 141–142, 1964.

Stone, C. J., Consistent nonparametric regression estimate, Annals of Statistics, vol. **5**, pp. 595–645, 1977.

Beck, J., The exponential rate of convergence of error for k_n-NN nonparametric regression and decision, Problems of Control and Information Theory, vol. **8**, pp. 303–312, 1979.

Federer, H., Geometric Measure Theory. Springer Verlag, 1969.

Wolverton, C. T., Wagner, T. J., Asymptotically optimal discriminant functions for pattern classification. IEEE Trans. on Information Theory, vol. **IT–15**, pp. 258–265, 1969.

Révész, P., Robbins—Monroe procedures in a Hilbert space and its application in the theory of learning processes, Studia Sci. Math. Hung., vol. **8**, pp. 391–398, 1973.

Последные результаты о непараметрических регрессии и классификации

Л. ДЁРФИ

(Будапешт)

В стати исследована слабая и сильная универсальная состоятельность непараметрических оценок, с особенным интересом по ближайщему саседу и ядерным функциям.

L. Györfi
Technical University of Budapest
H—1111 Budapest, Stoczek u. 2.
Hungary

Problems of Control and Information Theory, Vol. 10 (1), pp. 53—65 (1981)

О СТАТИСТИЧЕСКИХ МОДЕЛЯХ СЛОЖНЫХ МНОГОЭКСТРЕМАЛЬНЫХ ФУНКЦИЙ И ИХ ПРИМЕНЕНИИ ДЛЯ ПОСТРОЕНИЯ АЛГОРИТМОВ ОПТИМИЗАЦИИ

А. Г. ЖИЛИНСКАС

(Вильнюс)

(Поступила в редакцию 20 августа 1979 г.)

1. Введение

Алгоритмы нелинейной оптимизации обычно строятся на основании тех или иных моделей оптимизируемых функций, в рамках которых интерпретируются результаты предыдущих шагов и планируется проведение последующих. Например, для построения локальных алгоритмов часто используется квадратическая модель [1]. Однако в случае оптимизации сложных многоэкстремальных функций модель должна быть адекватной значительно большей неопределенности в поведении функции. В данной работе статистическая модель строится на основании довольно общего предположения о сравнимости интервалов возможных значений функций по их правдоподобию. Показано, что семейство случайных величин является довольно естественной статистической моделью, причем в классе таких моделей могут быть построены более простые с вычислительной точки зрения чем гауссовские случайные поля (обычно используемые в качестве моделей сложных функций в условиях неопределенности). Алгоритм оптимизации комбинирует глобальную стратегию поиска, которая определяется на основании идей рационального выбора, с обычными локальными алгоритмами.

2. Представление информации о многоэкстремальной целевой функции семейством случайных величин

Пусть единственной объективной информацией о минимизируемой функции $f(x)$, $x \in A \subset R^n$ являются значения $f(\cdot)$, вычисленные в точках $x_i \in A$: $y_i = f(x_i)$, $i = \overline{1, k}$. Кроме того имеется субъективная информация (например, опыт решения близких задач минимизации в прошлом) о сложности и многоэкстремальности $f(\cdot)$. Наиболее слабым представляющим интерес предположением относительно имеющейся информации является, повидимому, предположение о сравнимости интервалов возможных значений $f(x)$, $x \neq x_i$, $i = \overline{1, k}$ по их правдоподобию. Предположим заданное от-

носительное правдоподобие (ОП), обозначаемое символом $\geq_x$, где $(a, a') \geq \geq_x (b, b')$ означает, что событие $f(x) \in (a, a')$ не менее правдоподобно, чем $f(x) \in (b, b'$; индекс x может быть опущен, если из контекста ясно, о какой точке идет речь. Если $(a, a') \geq_x (b, b')$ и $(b, b') \geq_x (a, a')$, то это обозначается как $(a, a') \sim_x (b, b')$; $(a, a') >_x (b, b')$ означает, что $(a, a') \geq_x (b, b')$, но несправедливо, что $(a, a') \sim_x (b, b')$. Пусть фиксирована некоторая точка $x \neq x_i$, $i = \overline{1, k}$. Имеющаяся информация о $f(\cdot)$ часто не противоречит следующим предположениям рациональности ОП:

A1. 1) для любых двух интервалов (a, a'), (b, b') или $(a, a') \geq (b, b')$ или $(b, b') \geq (a, a')$,

A1. 2) если $(a, a') \geq (b, b')$ и $(b, b') \geq (c, c')$ то $(a, a') \geq (c, c')$,

A1. 3) $[a, a'] > \emptyset$ тогда и только тогда, когда $\mu[a, a'] > 0$, где $\mu(\cdot)$ обозначаем меру Лебега, $\emptyset$ — пустое множество;

A1. 4) пусть $(a, a') \cap (b, b') = \emptyset$, $(a, a') \cap (c, c') = \emptyset$, $a = b'$ или $a' = b$ и $a = c'$ или $a' = c$. $(b, b') \geq (c, c')$ тогда и только тогда, когда $[a, a'] \cup [b, b'] \geq [a, a'] \cup [c, c']$,

A1. 5) если $(a, a') \geq (b, b') > \emptyset$, существуют такие a_1, a_2 $a \leq a_i \leq a'$, $i = 1, 2$ что $(a, a_1) \sim (a_2, a') \sim (b, b')$.

Аксиомами A1.1–A1.5 ОП определено лишь для простейших множеств, т. е. интервалов. Основание продолжению ОП на более сложные множества дает следующее утверждение [2]: для любого $X = \bigcup_{j=1}^{n} I(j)$, где $I(j) = (a_j, a_j')$, $I(i) \cap I(j) = \emptyset$, $i \neq j$, существует возрастающая последовательность $b_1, b_2, \ldots, b_n$, такая, что $(-\infty, b_1) \sim I(1)$, $(b_{j-1}, b_j) \sim I(j)$, $j = \overline{2, n}$, причем b_n не зависит от нумерации интервалов $I(m)$, $j = \overline{1, m}$ и $b_m < \infty$, если $\mu((-\infty, \infty)/X) > 0$.

A1. 6) интервал $(-\infty, b_m)$ одинаково правдоподобен с X, т. е. $(-\infty, b_m) \sim X$.

Аксиома A1.6 позволяет продолжить ОП на алгебру конечных объединений интервалов, сохраняя при этом основные свойства ОП. Имеет место следующая теорема [2].

Теорема 1.1. Пусть ОП аксиомами A1.1–A1.5 задано для интервалов и посредством A1.6 продолжено на алгебру конечных объединений интервалов. Существует единственная плотность вероятности $p_x(\cdot)$, согласованная с ОП, т. е. $X_1 \geq_x X_2$ тогда и только тогда, когда

$$\int_{X_1} p_x(z)\, dz \geq \int_{X_2} p_x(z)\, dz,$$

где $X_j = \bigcup_{i=1}^{m_j} (a_{ji}, a_{ji}')$.

Утверждение теоремы позволяет интерпретировать значение $f(x)$ как случайную величину ξ_x с плотностью распределения $p_x(\cdot)$ и, следовательно, семейство ξ_x, $x \in A$, $x \neq x_i$, $i = \overline{1, k}$ как статистическую модель исследуемой функции.

Так как А1.1 требует сравнимости довольно простых множеств (интервалов), то она является слабее обычного для работ близкой тематики [3, 4, 5,6] предположения о сравнимости элементов алгебры. Транзитивность ОП, выражаемая А1.2, является одним из наиболее фундаментальных предположений рациональности отношения правдоподобия и обсуждалась многими авторами (см., например, [7]). Интуитивная приемлемость аксиом А1.1, А1.2 для специалистов, работающих в различных областях кибернетики и сталкивающихся в своей работе со сложными оптимизационными задачами, подтверждается результатами психологического эксперимента, описанного в [2]. А1.4 выражает аддитивность ОП и является одним из обычных предположений рациональности [3, 4, 5, 6]. Аксиомы А1.3 и А1.5 специфичны для данной работы. А1.3, выражая сложность исследуемой функции, утверждает, что невозможно ни точно предсказать значение $f(x)$, ни выбрать интервал (a, a') так, чтобы $\mu[a, a'] > 0$ и гарантированно $f(x) \notin (a, a')$. Непрерывность изменения ОП при изменении интервала возможных значений $f(x)$ представляется весьма естественной; аксиома А1.5 выражает свойство непрерывности ОП, повидимому, наиболее наглядным образом. Аксиоматическая характеризация ОП, заданного лишь для простейших множеств (интервалов) и последующее его продолжение посредством А1.6 представляется интуитивно более приемлемым, чем характеризация ОП сразу для сложной системы множеств.

Из принятой системы аксиом вытекает существование и единственность $p_x(\cdot)$, однако для построения алгоритмов оптимизации необходимо конструктивное построение плотности и ее параметров как функций от x. В [2] показано, что целесообразно в качестве $p_x(\cdot)$ выбирать гауссовскую плотность. Вопросы определения математического ожидания и дисперсии этой плотности рассматриваются ниже.

Отметим, что случайная функция может рассматриваться как семейство случайных величин и, таким образом, является частным случаем рассматриваемых здесь статистических моделей. В работе [8] аксиоматическая характеризация ОП распространена на вектора значений функции в нескольких точках и показано, что в таком случае существует случайная функция, согласованная с ОП, но аксиомы работы [8] более сложные и не являются столь естественными как А1.1–А1.6. Для построения алгоритмов оптимизации не менее важно и то, что в классе моделей, соответствующих А1.1–А1.6 можно построить более простые с вычислительной точки зрения чем гауссовские случайные поля.

3. Аксиоматическое определение характеристик статистической модели

Как показано выше, довольно естественные предположения относительно информации о $f(\cdot)$ открывают возможность принятия семейства гауссовских случайных величин ξ_x, $x \in A$, $x \neq x_i$, $i = \overline{1, k}$, в качестве статистической модели $f(\cdot)$. Математическое ожидание $m_k(x, (x_i, y_i), i = \overline{1, k})$ и дисперсию $s_k(x, (x_i, y_i), i = \overline{1; k})$ случайной величины ξ_x определим аксиоматически, следуя, как и выше, идеям рациональности. Не формально $m(x, (x_i, y_i), i = \overline{1, k})$ можно описать как ожидаемое, наиболее правдоподобное, репрезентативное значение исследуемой функции в точке x, а $s_k(x, (x_i, y_i), i = \overline{1, k})$ как оценку возможного отклонения $f(x)$ от $m_k(x, (x_i, y_i), i = \overline{1, k})$ или как условную (при $f(x_i) = y_i$, $i = \overline{1, k}$) неопределенность значения $f(x)$. Отметим, что построение $m_k(x, (x_i, y_i), i = \overline{1, k})$, $s_k(x, (x_i, y_i), i = \overline{1, k})$ представляет и самостоятельный интерес как задача рационального построения экстраполятора в условиях неопределенности [9] и оценки его погрешности; поэтому характеризация $m_k(\cdot)$, $s_k(\cdot)$ рассматривается как самостоятельная задача, непосредственно не связанная с А1.1–А1.6.

А2. 1) $m_k(x, (x_i, cy_i), i = \overline{1, k}) = cm_k(x, (x_i, y_i), i = \overline{1, k})$,

А2. 2) $m_k(x, (x_i, y_i + z), i = \overline{1, k}) = m_k(x, (x_i, y_i), i = \overline{1, k}) + z$,

А2. 3) для любой перестановки индексов $\{j(i), i = \overline{1, k}\} = \{1, 2, \ldots, k\}$, $m_k(x, (x_{j(i)}, y_{j(i)}), i = \overline{1, k}) = m_k(x, (x_i, y_i), i = \overline{1, k})$,

А2. 4) $m_k(x_j, (x_i, y_i), i = \overline{1, k}) = y_j$, $j = \overline{1, k}$,

А2. 5) существует функция $v_k(\cdot) : A \times (A \times R)^{k-1} \to B$, такая что $m_k(x, (x_i, y_i), i = \overline{1, k}) = n_k(x, (x_i, z_i), i = \overline{1, k})$, где $z_i = z$, $i = \overline{1, k-1}$, $z = v_k(x, x_i, y_i, i = \overline{1, k-1})$, $z_k = y_k$.

Имеет место следующее утверждение [9].

Теорема 2.1. Единственной функцией, удовлетворяющей А2.1–А2.5, является

$$m_k(x, (x_i, y_i), i = \overline{1, k}) = \sum_{i=1}^{k} y_i w_i^k(x, x_j, j = \overline{1, k}), \tag{2.1}$$

причем веса $w_i^k(\cdot)$ имеют следующие свойства: а) $\sum_{i=1}^{k} w_i^k(x, x_j, j = \overline{1, k}) = 1$, б) $w_i^k(x, x_1, \ldots, x_k) = w_p^k(x, x_1, \ldots, x_{j-1}, x_l, x_{j+1}, \ldots, x_{l-1}, x_j, x_{l+1}, \ldots, x_k)$, где $p = i$ при $i \neq j$, $i \neq l$; $p = j$ при $i = l$; $p = l$ при $i = j$, в) $w_j^k(x_l, x_i, i = \overline{1, k})$ равно единице при $j = l$ и нулю при $j \neq l$.

Аксиомы А2.1–А2.4 кажутся естественными; А2.5, постулируя возможность агрегирования данных, ограничивает сложность функции $m_k(\cdot)$. Из аксиом А2.1–А2.5 вытекает, что $m_k(x, (x_i, y_i), i = \overline{1,k})$ является взвешенным средним известных значений исследуемой функции y_i; сравнение А2.1–А2.5 с аксиоматическими характеризациями экстраполятора взвешенного среднего из [10, 11] приведено в [12]. В [11] отмечена возможность аксиоматической характеризации весов $w_i^k(\cdot)$, но в этом направлении сделано еще очень мало. Поэтому веса выбираются на основании эвристических соображений; некоторые из них приведены в [13]. Выражение (2.1) с весами

$$w_i^k(x, x_1, \ldots, x_k) = d(x, x_i) \Big/ \sum_{j=1}^{k} d(x, x_j), \tag{2.2}$$

где $d(x, x_i) = \exp(-c\,||x - x_i||^2)/||x - x_i||$, $c > 0$, $||\cdot||$ — эвклидова норма в R^n, исследовано в [12] как аппроксимирующая формула (экстраполятор) и показано, что она является достаточно точной при весьма малых затратах машинного времени, необходимого для ее вычисления; на этом основании такие же веса использованы и в данной работе.

Рассматривая аксиоматическую характеризацию $s_k(\cdot)$, для краткости обозначим $w_i^k = w_i^k(x, x_j, j = \overline{1,k})$

А3. 1) существует функция $\sigma(x, z)$: $A^2 \to R$, такая что $\sigma(x, z) = \sigma(z, x)$, $\sigma(x, x) > \sigma(x, z)$, $x \neq z$,

А3. (2) $s_k(x, (x_i, y_i), i = \overline{1,k}) = \gamma_k S_k(\sigma, (\sigma_i, w_i^k), i = \overline{1,k})$, где $\gamma_k = \gamma_k(y_1, \ldots, y_k)$, $\sigma = \sigma(x, x)$, $\sigma_i = \sigma(x, x_i)$,

А3. 2) $S_k(\sigma + \Delta, \sigma_i, w_i^k), i = \overline{1,k}) = S_k(\sigma, (\sigma_i, w_i^k), i = \overline{1,k}) + \Delta$,

А3. 3) $S_k(\sigma + \Delta, (\sigma_i + \Delta, w_i^k), i = 1, k) = S_k(\sigma, (\sigma_i, w_i^k), i = \overline{1,k})$,

А3. 4) $S_k(c\sigma, (c\sigma_i, w_i^k), i = \overline{1,k}) = cS_k(\sigma, (\sigma_i, w_i^k), i = \overline{1,k})$,

А3. 5) для любой перестановки индексов $\{j(i), i = 1, k\} = \{1, 2, \ldots, k\}$ $S_k(\sigma, (\sigma_{j(i)}, w_{i(i)}^k), i = \overline{1,k}) = S_k(\sigma, (\sigma_i, w_i^k), i = \overline{1,k})$,

А3. 6) существует функция $u_k(\cdot): R \times (R^2)^{k-1} \to R$, такая что $S_k(\sigma, (\sigma_i', w_i^k), i = \overline{1,k}) = S_k(\sigma, (\sigma_i, w_i^k), i = 1, k)$, где $\sigma_i' = u$, $i = 1, k-1$, $\sigma_k' = \sigma_k$, $u = u_k(\sigma, \sigma_i, w_i^k, i = \overline{1,k-1})$,

А3. 7) пусть $\tilde{w}_i^k = w_i^k$, $i \neq j$, $i \neq k$, $\tilde{\sigma}_i = \sigma_i$, $i = j$, $\sigma_k = 0$ и $w_j^k \sigma_j = \tilde{w}_j^k \tilde{\sigma}_j$ тогда $S_k(\sigma, (\sigma_i, w_i^k), i = \overline{1,k}) = S_k(\sigma, (\tilde{\sigma}_i, \tilde{w}_i^k), i = \overline{1,k})$.

Аналогично предыдущей может быть доказана следующая теорема [14].

Теорема 2.2. Единственной функцией, удовлетворяющей А3.1–А3.7, является

$$S_k(x, (x_i, y_i), i = \overline{1,k}) = \gamma_k \sum_{i=1}^{k} (\sigma(x, x) - \sigma(x, x_i))\, W_i^k(x, x_j, j = \overline{1,k}), \tag{2.3}$$

А3.1 определяет структуру $S_k(x, (x_i, y_i), i = \overline{1,k})$, постулируя зависимость условной неопределенности значения $f(x)$ (при известных $f(x_i) = y_i$, $i = \overline{1,k}$) от весов $w_k^i(x, x_j, j = \overline{1,k})$ и $\sigma(x, x_i)$, $i = \overline{1,k}$ где $\sigma(x, z)$ — функция выражающая информацию о $f(x)$, получаемую при измерении $f(x)$. Смысл такого неформального описания раскрывают аксиомы А3.2–А3.5. Так как информация о $f(x)$, получаемая при его же измерении может быть интерпретирована как безусловная неопределенность $f(x)$, то А3.2 утверждает, что возрастание безусловной неопределенности влечет такое же увеличение условной неопределенности. Согласно А3.3 возрастание безусловной неопределенности может быть компенсировано возрастанием получаемой информации. Смысл аксиом пропорциональности А3.4 и симметрии А3.5 очевиден, а А3.6 постулирует возможность агрегирования данных. А3.7 утверждает, что увеличение получаемой информации от измерения функции в точке x эквивалентно пропорциональному увеличению веса этого измерения.

Как утверждает теорема, условная неопределенность значения $f(x)$ при известных $f(x_i)$, $i = \overline{1,k}$ равна (с точностью до множителя) взвешенному среднему разностей $\sigma(x, x) - \sigma(x, x_i)$, которые выражают условную неопределенность значения $f(x)$ при единственном известном $f(x_i)$. Довольно естественно предположить, что неопределенность $f(x)$ при известном $f(x_i)$ пропорциональна расстоянию между x и x_i, т. е.

$$\sigma(x, x) - \sigma(x, x_i) = a\,||x - x_i||, \quad a > 0. \tag{2.4}$$

Возвращаясь к статистической модели исследуемой функции и интерпретации $n_k(\cdot)$ и $s_k(\cdot)$ как математического ожидания и дисперсии ξ_x, множитель γ_k можно определить методом наименьших квадратов: $\gamma_k = \gamma_{k-1}^*$, где γ_{k-1}^* минимизирует сумму

$$\sum_{i=1}^{k} [(y_i - n_k(x_i, (x_j, y_j), j = \overline{1,k}, j \neq i)^2 - s_{k-1}(x_i, (x_j, y_j), j = \overline{1,k}), j \neq i)]^2.$$

Как было отмечено выше, случайные функции являются частным случаем статистических моделей ξ_x. Поэтому интерес представляет вопрос: при каких условиях (2.1) и (2.3) соответствуют выражениям условного математического ожидания и условной дисперсии гауссовской случайной функции.

А2.6) пусть $z_i = y_i$, $i = \overline{1,k-1}$, $z_k = n_{k-1}(x_k, (x_i, y_i), i = \overline{1,k-1})$, тогда $n_k(x, (x_i, z_i), i = \overline{1,k}) = n_{k-1}(x, (x_i, y_i), i = \overline{1,k-1})$,

А2.7) $\sigma(x, \tilde{x})$ положительно определена и $n_2(x, (x_i, \tilde{z}_i), i = 1, 2) = n_2(\tilde{x}, (x_i, z_i), i = 1, 2)$, где $z_i = \sigma(x, x_i)$, $z_i = \sigma(\tilde{x}, x_i)$, $i = 1, 2$.

Аксиома А2.6 выражает свойство согласованности $n_{k-1}(\cdot)$ и $n_k(\cdot)$. Так как $\tilde{z}_i$ выражает информацию о $f(\tilde{x})$, получаемую при измерении $f(x_i)$, то

$\sigma_2(x, \tilde{x}) = n_2(x, (x_i, z_i),\ i = 1, 2)$ соответствует ожидаемой (по известным $\tilde{z}_i = \sigma(\tilde{x}, x_i)$) информации о $f(\tilde{x})$ при измерении $f(x)$; аксиома А2.7 постулирует симметричность $\sigma_2(\cdot)$ относительно ее аргументов. В [9] показано, что присоединение к А2.1–А2.5 аксиом А2.6, А2.7 влечет следующее выражение для весов: $(w_1^k, \ldots, w_k^k)^t = \Sigma_k^{-1}(\sigma_1, \ldots, \sigma_k)^t$, где Σ_k — матрица, составленная из элементов $\sigma(x_i, x_j)$, $i, j = \overline{1, k}$, t — знак транспонирования. В этом случае (2.1) и (2.3) соответствуют выражениям условного математического ожидания и условной дисперсии гауссовской случайной функции с математическим ожиданием μ и корреляционной функцией $\sigma(\cdot)$, т. е. могут быть записаны следующим образом:

$$m_k(x, (x_i, y_i),\ i = \overline{1, k}) = (y_1, \ldots, y_k)\,\Sigma_k^{-1}(\sigma_1, \ldots, \sigma_k)^t, \tag{2.5}$$

$$s_k(x, (x_i, y_i),\ i = \overline{1, k}) = \sigma - (\sigma_1, \ldots, \sigma_k)\,\Sigma_k^{-1}(\sigma_1, \ldots, \sigma_k)^t, \tag{2.6}$$

где μ сокращается вследствие пункта а) теоремы 2.1.

Формулы (2.5), (2.6), являясь частным случаем (2.1), (2.3), включают довольно сложную в вычислительном отношении операцию обращения матрицы. Как показано в [12], на вычисление (2.1) с весовой функцией (2.2) требуется в десятки раз меньше машинного времени, чем на вычисление (2.5). Кроме того, для вычисления по формулам (2.5), (2.6) необходим дополнительный объем оперативной памяти ЦВМ для хранения матрицы Σ_k^{-1}. Отметим, что при специальной структуре матрицы Σ_k ее обращение может и не вызывать никаких вычислительных затруднений; например, если $n = 1$ и $\sigma(\cdot)$ — корреляционная функция марковского случайного процесса. С другой стороны, выбор весов иногда легче обосновать чем выбор соответствующей случайной функции, например в одномерном случае для $x_i \le x \le x_{i+1}$ весьма естественны веса

$$w_i(x, x_j\ j = \overline{1, k}) = \frac{x_{i+1} - x}{x_{i+1} - x_i},\ w_{i+1}(x, x_j, j = \overline{1, k}) = 1 - w_i(x, x_j, j = \overline{1, k}), \tag{2.7}$$

и

$$\sigma(x, x) - \sigma(x, x_i) = a|x - x_i|\,.$$

Тогда $n_k(x, (x_i, y_i),\ i = \overline{1, k})$ и $s_k(x, (x_i, y_i),\ i = \overline{1, k})$ соответствуют выражениям условного математического ожидания и условной дисперсии винеровского процесса, т. е. процесса, выборочные функции которого недифференцируемы почти всюду. Вопрос о приемлемости винеровского процесса в качестве модели многоэкстремальных функций более детально обсуждается в [2].

Самые общие предположения относительно рациональности информации об исследуемой функции определяют лишь структуру формул (2.1), (2.3), оставляя свободу выбора функций $w_i^k(x, x_j, j = \overline{1, k})$ и $\sigma(x, x) - \sigma(x, x_i)$, выра-

жающих веса и неопределенность $f(x)$ при известном $f(x_i)$ соответственно. Так как эти функции имеют очевидный смысл, вполне естественно их выбирать исходя из эвристических соображений. Тем не менее теоретический интерес представляет также и аксиоматическая их характеризация. Например, аксиомами А2.6, А2.7 выделяется класс статистических моделей, соответствующих гауссовским случайным функциям. В некоторых случаях (если, например, статистическая модель предназначается для исследования многошаговых процедур принятия решения) А2.6 вполне естественна. Что касается А2.7, то она с интуитивной точки зрения наименее оправдана. Значение аксиом А2.6, А2.7 заключается в том, что они устанавливают связь между предложенными на основании принципов рациональности статистическими моделями и гауссовскими случайными функциями.

4. Построение алгоритма оптимизации

Пусть минимизируется функция $f(x)$, $x \in A \subset R^n$, где A — конечная прямоугольная область. Выбор алгоритмом очередной точки вычисления можно интерпретировать на основании статистической модели $f(\cdot)$ как выбор между плотностями распределения вероятностей $p_x(\cdot)$ (или, пользуясь более выразительной терминологией из [3, 4, 15], как выбор между лотереями, определяемыми $p_x(\cdot)$. В [3, 4, 15] показано, что из довольно общих предположений рациональности выбора вытекает существование функции полезности (прибыли) $u(\cdot)$ такой, что выбор плотности $p(\cdot)$ предпочтительнее выбора $g(\cdot)$ тогда и только тогда, когда $\int_{-\infty}^{\infty} u(t)\, p(t)\, dt > \int_{-\infty}^{\infty} u(t)\, g(t)\, dt$. Рассмотрим вопрос построения функции полезности $u(\cdot)$, соответствующей представлениям о рациональном поиске минимума. Предположим, что проведено k вычислений минимизируемой функции: $y_i = f(x_i)$, $i = \overline{1, k}$, и планируется проведение $k+1$-го. Так как ξ_x — семейство гауссовских случайных величин, рассматриваются лишь предпочтения между гауссовскими плотностями, понимаемые как предпочтения между векторами их параметров (m, s), в том числе при $s = 0$.

А4. 1) для любых $m_1 < m_2$, $s_1 > 0$ существует s такое, что $(m_1, s_1) > (m_2, s_2)$ при $s_2 \leq s$,

А4. 2) для любых m_1, $s_1 > 0$, $m_2 \geq y_{0k}$, где $y_{0k} = \min\limits_{1 \leq i \leq k} y_i$, $(m_1, s_1) > (m_2, 0)$,

А4. 3) $(m_1, s_1) > (m_2, s_2)$ тогда и только тогда, когда $(m_1, ks_1) > (m_2, ks_2)$,

А4. 4) $u(\cdot)$ — неотрицательная ступенчатая функция вида: $u(t) = a \sum_{i=1}^{M} I(t_i - t)\,\alpha_i$, $I(t) = 0$ при $t < 0$, $I(t) = 1$ при $t \geq 0$, $a > 0$, $t_i \geq t_{i+1}$, $|\alpha_i| = 1$.

Аксиома А4.1 утверждает, что выбирать точку, в которой ожидается сравнительно большое значение функции может быть рациональным только при достаточно большом уровне неопределенности. Согласно А4.2 выбирать точку, в которой гарантировано получение значения минимизируемой функции не меньше чем y_{0k}, не рационально. А4.3 предполагает инвариантность отношения предпочтения относительно выбора масштаба измерения неопределенности.

Требования к функции полезности, постулируемые аксиомой А4.4 довольно сильные. Однако с прикладной точки зрения они не слишком стеснительны, так как любую неотрицательную интегрируемую с гауссовской плотностью функцию $u(t)$ можно аппроксимировать ступенчатой функцией $u^*(t) = \varepsilon \sum_{i=1}^{M} I(t_i - t)\,\alpha_i$, причем $\int_{-\infty}^{\infty} |u(t) - u^*(t)|\, n(t|\, m,\, s)\, dt < \varepsilon$ для любых $\varepsilon > 0$, $m^- \leq m \leq m^+$, $0 \leq s \leq s^+$, где $n(\cdot|m, s)$ — гауссовская плотность с математическим ожиданием m и дисперсией s; m^-, m^+, s^+ — фиксированы и $M = M(\varepsilon)$.

Теорема 3.1. Единственной функцией, удовлетворяющей А4.1–А4.4 является (с точностью до постоянного множителя)

$$u(t) = I(z_{0k} - t), \text{ где } z_{0k} < y_{0k}. \tag{3.1}$$

Доказательство теоремы приведено в приложении, а из ее утверждения вытекает, что полезность выбора x в качестве точки $k+1$-го вычисления пропорциональна вероятности

$$P(\xi_x < z_{0k}) = \Pi\{(z_{0k} - n_k(x, (x_i, y_i), i = \overline{1, k}))/(s_k(x, (x_i, y_i), i = \overline{1, k}))^{1/2}\},$$

где $\Pi(\cdot)$ — функция нормального распределения. Следовательно, алгоритм минимизации, соответствующий приведенным выше предположениям, $k+1$-ое вычисление $j(\cdot)$ проводит в точке максимума этой вероятности, которая ($\Pi(\cdot)$ — монотонно возрастающая) совпадает с точкой максимума функции $v_{k+1}(x) = (z_{0k} - n_k(x, (x_i, y_i), i = \overline{1, k})/(s_k(x, (x_i, y_i), i = \overline{1, k}))^{1/2}$.

Алгоритм одномерной многоэкстремальной минимизации, соответствующий приведенным выше результатам (т. е. планирующий вычисление $j(\cdot)$ в точках максимума вероятности превышения некоторого уровня винеровским процессом, выбранным в качестве модели $j(\cdot)$) был предложен на

основании эвристических соображений в [17]. В [18, 19] использовалась другая функция полезности ($\tilde{u}(t) = 0$ при $t \geq y_{0k}$, $\tilde{u}(t) = y_{0k} - t$ при $t < y_{0k}$), которая соответствует предположению об одношаговой байесовской оптимальности алгоритма [20, 21]; $\tilde{u}(\cdot)$ удовлетворяет аксиомам А4.1, А4.2, но не удовлетворяет А4.3. Результаты исследования одномерного алгоритма, построенного, используя в качестве статистической модели винеровский процесс и функцию полезности $\tilde{u}(\cdot)$, показали, что при соответствующем (не имеющем, правда, теоретического обоснования) выборе параметров модели алгоритм является весьма эффективным [18]. Байесовские алгоритмы (т. е. оптимальные в смысле средней погрешности относительно случайной функции, используемой в качестве статистической модели) исследованы в [20, 21]. Статистические модели многоэкстремальных функций используются также при построении информационных алгоритмов минимизации [22].

Так как при построении статистической модели имеется довольно большой произвол при выборе весов (2.2) и выражения (2.4), а при построении функции полезности — уровня z_{0k}, то этот произвол отражается на точку максимума функции $\nu_{k+1}(\cdot)$. Поэтому при реализации алгоритма, повидимому, достаточно функцию $\nu_{k+1}(x)$ максимизировать весьма грубо, рассчитывая, в основном, на исключение вычислений $f(\cdot)$ в точках с низкой полезностью. Следовательно, предлагаемый алгоритм целесообразно использовать лишь для приближенной (с малой точностью) оценки основных локальных минимумов, которые впоследствии могут быть вычислены с высокой точностью при помощи хорошо известных локальных методов. Зоны локальных минимумов $f(\cdot)$ можно оценить предполагая, что при определенных условиях окрестность точки локального минимума $n_k(\cdot, (x_i, y_i), i = \overline{1,k})$ содержит также и локальный минимум $f(\cdot)$ [24, 25]. Основание такому предположению дает тот факт, что предлагаемый алгоритм выбирает точки вычисления $f(\cdot)$ в окрестностях малых значений $f(\cdot)$ с большей плотностью и поэтому в этих окрестностях $m_k(\cdot, (x_i, y_i), i = \overline{1,k})$ является достаточно хорошей аппроксимацией $f(\cdot)$ (в случае весов (2.2) — сплайн — аппроксимация [16], в одномерном случае (2.7) — кусочно-линейная аппроксимация).

Подробное описание одномерного алгоритма, его более ранней версии, результаты тестирования и ФОРТРАН программа приведены в [23, 24, 25, 26] соответственно. Многомерный алгоритм описан в [27]. Результаты тестирования [23, 26, 27] показали сравнительно высокую эффективность алгоритмов в смысле количества вычислений минимизируемой функции. На основании аналогичных идей могут быть построены и алгоритмы минимизации в присутствии помех; описание одномерного алгоритма дано в [23], а его АЛГОЛ программа в [28].

5. Выводы

Задача поиска глобального минимума многоэкстремальной функции рассматривается как задача принятия решения в условиях неопределенности. Аксиоматически формулируются предположения относительно неопределенности поведения минимизируемой функции и рациональности выбора точки очередного вычисления. Показано, что, исходя из этих аксиом, может быть построена статистическая модель минимизируемой функции, а на ее основании и алгоритм минимизации.

6. Приложение

Доказательство теоремы 3.1. Из А4.1 вытекает, что $u(\cdot)$ — невозрастающая функция. Допустим противное. Пусть для $m_1 < m_2$ $u(m_2) - u(m_1) = \varepsilon > 0$. Так как $u(\cdot)$ — кусочно-непрерывна, можно выбрать s_0, такое, что $\left| \int_{-\infty}^{\infty} u(t)\, n(t|\, m_i,\, s_i)\, dt - u(m_i) \right| < \frac{\varepsilon}{2}$ для всех $s_i < s_0$, $i = 1,\ 2$. Но тогда $\int_{-\infty}^{\infty} (u(t|\, n(t)\, m_2,\, s_2)\, dt > \int_{-\infty}^{\infty} (u(t|\, n(t)\, m_1,\, s_1)\, dt$, что противоречит А4.1.

Из А4.2 вытекает, что $u(t) = c$ при $t \geq y_{0k}$. Допустим противное. Пусть (так как $u(\cdot)$ — невозрастающая) для $y_{0k} \leq m_2 < m_1$ и $u(m_2) - u(m_1) = \varepsilon > 0$. Но тогда, выбрав достаточно малое s_1, такое что $\left| \int_{-\infty}^{\infty} u(t)\, n(t|\, m_1,\, s_1)\, dt - u(m_1) \right| < \varepsilon$ получаем $u(m_2) > \int_{-\infty}^{\infty} u(t_2)\, n(t|\, m_1,\, s_1)\, dt$, что противоречит А4.2. Не ограничивая общность, можно положить $c = 0$, $a = \frac{1}{M}$.

Следовательно, в выражении $u(t)$, $\alpha_i = 1$, $i = \overline{1, M}$ и $\max t_i < y_{0k}$. Если при любом сколь угодно малом ε, $t_i = t_1$, $i = \overline{2, M}$, то $u(t) = I(t_1 - t)$ и легко показать, что имеет место А4.3. Эта функция является единственной удовлетворяющей требованию А4.3, так как в случае $t_1 - t_M > 0$ (предположим, что точки t_i упорядочены: $t_M \leq t_{M-1} \leq \ldots \leq t_1$) можно выбрать m_1, s_1, m_2, s_2, k так, чтобы

$$\sum_{i=1}^{M} \Pi\left(\frac{t_i - m_1}{s_1}\right) > \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{s_2}\right), \quad \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_1}{k s_1}\right) < \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{k s_2}\right).$$

Пусть T_-, T_0, T_+ определены из уравнений $\Pi(T_-) = \varepsilon/M$, $\Pi(T_0) = 0.25$, $\Pi(T_+) = 0.5 - \varepsilon/M$ и

$$k = \frac{t_M - t_1 + T_-}{T_+},\ m_2 = t_1 - T_-,\ s_2 = 1,\ s_1 = \frac{t_M - t_1}{T_0(k+1)},\ m_1 = t_1 + \frac{t_M - t_1}{k+1},$$

Тогда при $\varepsilon < 0.25$

$$\sum_{i=1}^{M} \Pi\left(\frac{t_i - m_1}{s_1}\right) > 0.25 > \varepsilon > \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{s_2}\right),$$

$$\sum_{i=1}^{M} \Pi\left(\frac{t_i - m_1}{ks_1}\right) < 0.25 + \frac{M-1}{2} < M\left(\frac{1}{2} - \frac{\varepsilon}{M}\right) < \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{ks_2}\right)$$

что и требовалось доказать.

Литература

1. Численные методы условной оптимизации. Под редакцией Ф. Гилла и У. Мюррэй, Мир (1977).
2. *Žilinskas, A.*, On Statistical Models for Multimodal Optimization. Math. Operat. Stat., ser. Statistics, vol. **9**, No. *2*, 255–266 (1978).
3. *Savage, L. J.*, Fundations of Statistics, Wiley New York (1954).
4. *De Groot, M. H.*, Optimal Statistical Decisions. Mc Graw-Hill, New York (1970).
5. *Fine, T. L.*, Theories of Probability. Academic Press, New York (1973).
6. *Gottinger, H. W.*, Konstruktion der subjektiven Wahrscheinlichkeiten. Math. Operat. Stat., vol. **5**, No. *6*, 509–539 (1974).
7. *Mac Crimmon*, Descriptive and Normative Implications of the Decision Theory Postulates. In Risk and Uncertainty (eds. K. Broch and J. Mossin), St. Martin, New York (1968).
8. *Жилинскас А., Каткаускайте А.* Построение статистических моделей сложных функций, включающих элементы неопределенности. VII Всесоюзная конференция по теории кодирования и передачи информации, доклады, часть I, Москва–Вильнюс (1978).
9. *Жилинскас А. Г.* Аксиоматический подход к проблеме экстраполяции в условиях неопределенности. Автоматика и телемеханика, № *12*, 66–70 (1979).
10. *Fine, T.*, Extrapolation when Very Little is Known about the Source. Inform. Contr., vol. **16**, 331–359 (1970).
11. *Goldman, J.*, An Approach to Estimation and Extrapolation with Possible Applications in an Incomplitely Specified Environment. Inform. Contr., vol. **30**, 203–223 (1976).
12. *Жилинскас А. Г.* Исследование задач многомерной экстраполяции в условиях неопределенности. Теория оптимальных решений, вып. 4, Вильнюс, 27–53 (1978).
13. *Shepard, D.*, A Two-Dimensional Interpolation Function for Irregularly-Spaced Data. Proc. 23rd Nat. Conf. ACM, 517–524 (1965).
14. *Жилинскас А. Г.* Об аксиоматической характеризации статистических моделей многоэкстремальных функций. Применения случайного поиска, ВИНИТИ, 106–110 (1980).
15. *Фишберн П.* Теория полезности для принятия решений, Наука (1978).
16. *Gordon, W., Wixom, A.*, Shepard's Method of "Metric Interpolation" to Bivariate and Multivariate Interpolation. Mathematics of Computation, vol. **32**, No. *141*, 253–264 (1978).
17. *Kushner, H.*, A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. Trans ASME, ser. D. J. Basic Eng., vol. **86**, No. 1, 97–105 (1964).
18. *Жилинскас А. Г.* Одношаговый байесовский метод поиска экстремума функций одной переменной. Кибернетика, № *1*, 139–144 (1975).
19. *Шалтянис В. Р.* Об одном методе многоэкстремальной оптимизации. Автоматика и вычислительная техника, № *3*, 33–38 (1971).
20. *Моцкус И. Б.* О байесовых методах поиска экстремума. Автоматика и вычислительная техника, № *3*, 53–62 (1972).
21. *Mockus, J.*, On Bayesian Methods of Seeking the Extremum and their Applications. In Information Processing 77 (ed. B. Gilchrist), North-Holland (1977).

22. *Стронгин Р. Г.* Численные методы многоэкстремальной минимизации, Наука, (1978).
23. *Žilinskas, A.*, Two Algorithms for One-Dimensional Multimodal minimization. Math. Operat. Stat., ser. Optimization (in print).
24. *Жилинскас А. Г.* Метод одномерной многоэкстремальной минимизации. Изв. АН СССР, Техническая кибернетика, № *4*, 71–74 (1976).
25. *Žilinskas, A.*, Optimization of One-Dimensional Multimodal Functions, Algorithm AS 133. Applied Statistics. vol. **27**, 367–375 (1978).
26. *Žilinskas, A.*, On One-Dimensional Multimodal Minimization. Trans. of Eight Prague Conf. on Inform. Theory, Stat. Dec. Func., Rand. Proc., vol. **B**, 393–402 (1978).
27. *Žilinskas, A.*, On the use of statistical models of multimodal functions for the construction of the optimization algorithms. In Proceedings of 9th IFIP Conference on Optimization Techniques (1980).
28. *Žilinskas, A.*, MIMUM-optimization algorithm of one-dimensional multimodal functions in the presence of noise. Aplikace Matematiky, vol. **25**, No. *1*, 234–240 (1980).

On statistical models of multimodal functions and their application for the construction of the minimization algorithms

A. ŽILINSKAS

(Vilnius)

The information on a multimodal objective function is formalized as a binary relation of comparative probability (CP) between intervals of possible values of the function. The intuitively acceptable assumptions about CP imply that the family of random variables may be interpreted as a statistical model of the function. The problem of further characterization of the statistical model is considered as a problem of extrapolation under uncertainty. The algorithm of minimization is characterized by axioms of rational choice and it is shown that such an algorithm may be reduced to the optimal procedure with respect to the statistical model.

А. Г. Жилинскас
Институт математики и кибернетики АН Литовской ССР
СССР, Вильнюс 232600,
К. Пожелос, 54

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

ENGLISH TRANSLATION OF THE PAPERS IN RUSSIAN
Problems of Control and Information Theory, Vol. **10**, No. *1*

CASCADED DECODING WITH RESPECT TO MINIMAL GENERALIZED DISTANCE

I. I. DUMER, V. A. ZINOV'EV, V. V. ZYABLOV

Decoding of binary generalized cascaded codes in continuous-output channels is considered, and an algorithm is proposed for cascaded decoding of such codes with respect to minimal generalized distance. For various systems of decoding thresholds, realizable correcting properties are studied. In the presence of sufficient number of thresholds this method realises the guaranteed code distance of generalized cascaded codes. The results may be extended to discrete channels and provide a new decoding algorithm for generalized cascaded codes.

1. Introduction

Consider decoding with respect to minimal generalized distance (**MGD**) [1] for binary symbols transmitted through a channel with additive continuous noise. At the receiving end, symbols of the demodulated word $y = (y_1, \ldots, y_n)$ enter the likelihood detector generating for each symbol y_i a logarithm of likelihood ratio

$$f_i = \ln \frac{P(y_i \mid 1)}{P(y_i \mid 0)}$$

(where $P(y_i|a)$, $a = 0, 1$ is the probability of symbol y_i for transmitted a) rather than a pair of conditional probabilities $P(y_i|1)$ and $P(y_i|0)$. Let (a, b) denote the scalar product of the vectors a and b in the field of real numbers (in what follows, it will always be assumed that binary code vectors consist of the symbols -1 and $+1$ substituted for 1 and 0, respectively). For the transmission of the code word c through a memoryless channel, the probability $P(y/c)$ is known to be maximal if and only if the scalar product (f, c) is maximal, where $f = (f_1, \ldots, f_n)$ [1]. The detector is followed by a limiter with the following characteristic

$$q(x) = \begin{cases} 1, & x \geq T \\ \dfrac{x}{T}, & -T \leq x \leq T \\ -1 & x \leq -T \end{cases}$$

where T is a threshold. This results in the vector z equal to $(q(f_1), \ldots, q(f_n))$. Constraint $-1 \leq q(x) \leq 1$ enables the proof of a theorem underlying decoding with respect to the MGD criterion.

Theorem 1. (Forney [1]). In a code of length n with minimal distance d there is at most one code vector x for which

$$(z, x) > n - d \tag{1}$$

where z is the word received after limitation.

Proof of Theorem 1 is based on the following fact which we shall need further.

Lemma 1. Let for some vector x_1 from the code C with the distance d

$$(z, x_1) > n - m$$

where z is an arbitrary vector received after the limiter, and $m \leq d$. Then for any vector $x_2 \in C$, $x_2 \neq x_1$,

$$(z, x_2) < n - 2d + m.$$

Proof. Although this fact is proved in Forney's book [1, p. 70] we repeat the proof because its reasoning will be useful below. Let x_2 be a code word differing from x_1 in d positions, i.e.

$$x_{1,i} \neq x_{2,i}, \quad i \in S, \quad |S| = d.$$

Then

$$(z, x_1) = \sum_{i \notin S} z_i x_{1,i} + \sum_{i \in S} z_i x_{1,i} = m_1 + m_2,$$

$$(z, x_2) = \sum_{i \notin S} z_i x_{2,i} + \sum_{i \in S} z_i, x_{2,i} = m_1 - m_2$$

but by virtue of the condition,

$$m_1 + m_2 > n - m \quad \text{and} \quad m_1 \leq n - d;$$

hence the lemma follows.

Theorem 1 holds for any characteristic of the limiter $-1 \leq q(x) \leq 1$. In particular, it holds for the case of errors and erasures to which the following limiter characteristic corresponds

$$q(x) = \begin{cases} +1, & x \geq T, \\ 0, & -T < x < T, \\ -1, & x \leq -T. \end{cases}$$

Decoding with correction of errors and erasures has turned out to be useful for decoding with respect to MGD if condition (1) is met. Let for any $z = (z_1, \ldots, z_n)$ the indices $i_1, i_2, \ldots, i_n$ be ordered in the order of increasing reliability,

$$|z_{i_1}| \leq |z_{i_2}| \leq \ldots \leq |z_{i_n}|.$$

To the vector z assign the vector $z^{(k)}$, $k = 0, 1, \ldots, n$ consisting of elements $0, +1, -1$ and obtained (by means of the erasure of the k most unreliable symbols and substitution of ± 1 for the rest of them) in accordance with the following limiter characteristic

$$q_k(z_{i_j}) = \begin{cases} 0, & 1 \leq j \leq k, \quad 0 \leq k \leq n, \\ 1, & z_{i_j} \geq 0, \qquad k+1 \leq j \leq n, \\ -1, & z_{i_j} < 0, \qquad k+1 \leq j \leq n. \end{cases}$$

Then the vector

$$z^{(k)} = (q_k(z_1), \ldots, q_k(z_n))$$

has k coordinates equal to 0, the balance being equal to ± 1. The following theorem guarantees that if condition (1) is met for some word x, the error-and-erasure correcting decoder will find the word x for $k < d$ at least for one vector $z^{(k)}$.

Theorem 2 (Forney [1]). If $(z, x) > n - d$, the following inequality holds for some $k < d$

$$(z^{(k)}, x) > n - d. \tag{2}$$

As it may be easily seen, condition (2) means that there are k erasures and l_k erroneous symbols in the vector $z^{(k)}$ such that $2l_k + k < d$. Thus, in order to determine the vector x, one needs at most d attempts to decode the vectors $z^{(k)}$, $k = 1, \ldots, d$, with the correction of errors and erasures.

It follows from the results obtained by Forney [1] that MGD decoding is applicable to any code for which an error-and-erasure correcting algorithm is known guaranteed by its code distance. This decoding algorithm is, thus, applicable to cascaded code with cascaded decoding realizing its code distance [2]. In such a decoding referred to as MGD decoding of cascaded code, complexity grows in proportion to the code distance d, i.e. code of length $n = n_a n_b$ should be decoded $d/2$ times (where n_a and n_b are, respectively, lengths of external and internal codes). For cascaded decoding, it seems more reasonable to MGD decode only internal codes of cascaded code. This decoding will be referred to as cascaded MGD decoding. We are going to study realizable correcting properties of this decoding. Preliminary results on this matter were published in [3].

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

2. MGD cascaded decoding of generalized cascaded codes

We now describe briefly the generalized cascaded codes (GC-codes). For more details see [4], whose notations will be used here. Denote by $A(q, n, d, N)$ a q-ary code of length n with minimal code distance d and power N. For m codes $A_i = A_i(q_i, n_a, d_{a,i}, N_{a,i})$ of identical length n_a (the codes A_i are called external) introduce a set of matrices $\{M = M(A_1, \ldots, A_m)\}$ where the matrix M of size $m \times n_a$ is obtained as follows. Its i-th row contains an arbitrary code vector $a^{(i)} = (a_1^{(i)}, \ldots, a_{n_a}^{(i)})$ from code A_i, $i = 1, \ldots, m$. Let the code $B^{(1)} = B^1(q, n_b, d_{b,1}, N_{b,1})$ of power $N_{b,1} = q_1, q_2, \ldots, q_m$ be decomposable into q_1 codes $B_{i_1}^{(2)}(q, n_b, d_{b,2}, N_{b,2})$, where $N_{b,2} = q_2, \ldots, q_m$ and $i_1 = 0, 1, \ldots, q_1 - 1$. Let, in turn, each code $B_{i_1}^{(2)}$ be decomposable into q_2 codes $B_{i_1,i_2}^{(3)} = B_{i_1,i_2}^{(3)}(q, n_b, d_{b,3}, N_{b,3})$, $N_{b,3} = q_3, \ldots, q_m$, etc., the decomposition having m steps. At the s-th decomposition step one has

$$B_{i_1,\ldots,i_{s-1}}^{(s)}(q, n_b, d_{b,s}, N_{b,s}) = \bigcup_{i_s=0}^{q_s-1} B_{i_1,\ldots,i_s}^{(s+1)}(q, n_b, d_{b,s+1}, N_{b,s+1})$$

where $N_{b,s} = q_s, \ldots, q_m$, $N_{b,s+1} = q_{s+1}, \ldots, q_m$, $\quad s = 1, \ldots, m$

$$q_{m+1} = N_{b,m+1} = 1, \quad d_{b,m+1} = \infty.$$

This decomposition enables the enumeration of code vectors. Denote by $b(i_1, \ldots, i_m)$ ($b(i_1, \ldots, i_m)$ are numbers) the code vector $B_{i_1,\ldots,i_m}^{(m+1)}$. Codes $B^{(i)}$ are called internal codes.

GC-code is constructed on the basis of m external codes A_i, and m internal ones, $B^{(i)}$. In the matrix $M(A_1, \ldots, A_m)$ the code vector $b(a_j^{(1)}, \ldots, a_j^{(m)})$ is substituted for each j-th column $(a_j^{(1)}, a_j^{(2)}, \ldots, a_j^{(m)})$, $j = 1, \ldots, n_a$. The resulting matrix of size $n_b \times n_a$ consisting of symbols $0, 1, \ldots, q-1$ will be regarded as a code word of GC-code $\{c\}$ of order m. Understandably, the length of the resulting code is $n = n_a n_b$ and its power is $N = N_{a,1}, N_{a,2}, \ldots, N_{a,m}$. Let by definition

$$d_i = d_{a,i} d_{bi} \tag{3}$$

$$d^{(H)} = \min_i \{d_i\}. \tag{4}$$

The code distance d of a GC-code is known [4] to satisfy inequality $d \geq d^{(H)}$.

Consider now MGD decoding of a binary GC-code $\{c\}$. Let the coding of the matrix $M(A_1, \ldots, A_m)$ with rows $a^{(1)}, \ldots, a^{(m)}$ result in the word c transmitted through a continuous-noise channel. Let z be a word received after the limiter, $z_1, \ldots, z_{n_a}$ be its columns, and $b_1, \ldots, b_{n_a}$ be columns of the transmitted word c. According to Theorem 1, if

$$(c, z) > n - d^{(H)} \tag{5}$$

after $d^{(H)}$ (see (3) and (4)) attempts of decoding by the code $\{c\}$ of length n with correction of errors and erasures, we obtain the word C. Below, cascaded MGD decoding will be shown to be feasible provided that (5) is met. It is performed by correction of errors and erasures by means of the codes $B^{(i)}$ of length n_b and the codes A_i of lenght n_a.

The cascaded decoding algorithm denoted by $\psi(m)$ is done successively in m steps. At the first step, the row $a^{(1)} \in A_1$ is determined, i.e. the first row of the matrix M, at the second step $a^{(2)} \in A_2$ is determined, etc. Finally, at the m-th step the row $a^{(m)} \in A_m$ is determined, thus completing the decoding.

The rows $a^{(1)}, a^{(2)}, \ldots, a^{(i-1)}$ of the matrix M are known at the i-th step of decoding. Column b_j of the matrix c, $j = 1, \ldots, n_a$ belongs to the code $B^{(i}_{a_j^{(1)}, \ldots, a_j^{(i-1)}}$, denoted below for the sake of brevity as $B_j^{(i)}$. Introduce the following definitions. Let the set

$$T_i = \{t_1^{(i)}, \ldots, t_{M_i}^{(i)}\}$$

be defined for some natural number M_i. The set consists of M_i real numbers such that $d_{b,i} = t_1^{(i)} > t_2^{(i)} > \ldots > t_{M_i}^{(i)} > 0$ (it is more convenient to denote $d_{b,i}$ as $t_1^{(i)}$). Let $B_j^{(i)}(d)$ denote the current MGD decoding of the vector z_j by code $B_j^{(i)}$ resulting in the vector $\hat{b}_j \in B_j^{(i)}$ and the number $r_j^{(i)} = n_b - (\hat{b}_j, z_j)$ if $(\hat{b}_j, z_j) > n_b - d_{b,i}$. Otherwise, decoding results in erasure; assume that in this case $r_j^{(i)} = d_{b,i}$. Thus, the parameter $r_j^{(i)}$ lower-bounds the distance defined by the scalar product between the received column z_j and the nearest word $\hat{b}_j$ of the internal code $B_j^{(i)}$. As the result of $B_j^{(i)}(d)$ decoding of all the n_a columns z_j, $j = 1, \ldots, n_a$, we obtain row $R^{(i)} = (r_1^{(i)}, \ldots, r_{n_a}^{(i)})$.

The i-th decoding step is realized by the following scheme:

$$z \to [\hat{a}^{(i)}, R^{(i)}] \to \{a^{(i)}(k), S^{(i)}(k)\} \to a^{(i)}.$$

At the passage $z \to [\hat{a}^{(i)}, R^{(i)}]$, the decoding $B_j^{(i)}(d)$, $j = 1, \ldots, n_a$ of the columns z_j of the matrix z is first performed. As a result, we obtain the columns $\hat{b}_j$ or erasures, and also the row $R^{(i)}$. Next, for all $j = 1, \ldots, n_a$ columns, $\hat{b}_j$ are substituted by their numbers $(a_j^{(k)}, \ldots, a_j^{(i-1)}, \hat{a}_j^{(i)}, \ldots, \hat{a}_j^{(m)})$, and the erased columns are substituted by erased numbers. Note that the first $i - 1$ components of the numbers are correct because the first $i - 1$ rows of the matrix M are assumed to be decoded correctly. This results in the row $\hat{a}^{(i)} = (\hat{a}_1^{(i)}, \ldots, \hat{a}_{n_a}^{(i)})$ consisting of the i-th components of numbers and erasures.

At the passage $[\hat{a}^{(i)}, R^{(i)}] \to \{a^{(i)}(k), S^{(i)}(k)\}$, errors and erasures are decoded in the row $\hat{a}^{(i)}$ with the threshold t_k, $k = 1, \ldots, M_i$. Let $A^{(i)}(k)$, $k = 1, \ldots, M_i$, denote the following decoding of the row $\hat{a}^{(i)}$. At first, erase in $\hat{a}^{(i)}$ all symbols $\hat{a}_j^{(i)}$ for which

$$r_j^{(i)} \geq t_k. \tag{6}$$

Let the resulting row be denoted by $\hat{a}^{(i)}(k)$ (note that $\hat{a}^{(i)}(1) = \hat{a}^{(i)}$). Next, the row $\hat{a}^{(i)}(k)$ is decoded by the code A_i with correction of errors and erasures. If possible, this decoding generates some row $a^{(i)}(k) \in A_i$. Let $\{I^{(i)}(k)\}$ be a set of those positions j, $j = 1, \ldots, n_a$, where the rows $a^{(i)}(k)$ and $\hat{a}^{(i)}$ contain different non-erased symbols (as was noted above, the row $\hat{a}^{(i)}$ may involve erasures). In other words, if $\{V^{(i)}\}$ denotes the set of erased positions in $\hat{a}^{(i)}$,

$$\{I^{(i)}(k)\} = \{j \notin V^{(i)} \mid a_j^{(i)}(k) \neq \hat{a}_j^{(i)}\}. \tag{7}$$

For any $j = 1, \ldots, n_a$, define

$$p_j^{(i)}(k) = \begin{cases} r_j^{(i)} & \text{if } j \notin \{I^{(i)}(k)\} \\ 2d_{b,i} - r_j^{(i)} & \text{if } j \in \{I^{(i)}(k)\}. \end{cases} \tag{8}$$

The numbers $p_j^{(i)}(k)$ will be demonstrated below to lower-bound the distance between the columns z_j and b_j, $j = 1, \ldots, n_a$, if the row $a^{(i)}(k)$ is a true row $a^{(i)}$ of the transmitted cascaded word c. Finally define the number $S^{(i)}(k)$ for each row $a^{(i)}(k)$ in the following manner:

$$S^{(i)}(k) = \sum_{j=1}^{n_a} p_j^{(i)}(k).$$

For all $k = 1, \ldots, M_i$, perform the decoding $A^{(i)}(k)$ of row $\hat{a}^{(i)}$ resulting in the set $\{a^{(i)}(k), S^{(i)}(k)\}$ where for different k at most M_i rows $a^{(i)}(k)$ are contained.

At the passage $\{a^{(i)}(k), S^{(i)}(k)\} \to a^{(i)}$, determine the true row $a^{(i)}$ of the transmitted word c. To this end, take the row $a^{(i)}(v)$ from $\{a^{(i)}(k)\}$, for which

$$S^{(i)}(v) < d_i. \tag{9}$$

It will be shown below that such a row exists and is unique provided that condition (5) is observed.

Lemma 2. Let the scalar product of the vectors c and z meet the condition

$$(c, z) > n - d_i. \tag{10}$$

If the i-th row of the transmitted word c belongs to the row set $\{a^{(i)}(k)\}$,

$$a^{(i)} \in \{a^{(i)}(k)\}, \tag{11}$$

namely this row satisfies inequality (9) and no other row from $\{a^{(i)}(k)\}$ satisfies it.

Proof. Let the true row $a^{(i)}$ belong to $\{a^{(i)}(k)\}$, i.e. $a^{(i)} = a^{(i)}(v)$ for some v. Demonstrate first that $S^{(i)}(v) \leq d_i$. Since $a^{(i)}(v) = a^{(i)}$, the set $\{I^{(i)}(v)\}$ in-

volves those positions where erroneous symbols stand in $\hat{a}^{(i)}$ (see (7)). Hence, as the result of the decoding $B_j^{(i)}(d)$ of the rows z_j for all $j \in \{I^{(i)}(v)\}$ erroneous columns $\hat{b}_j \neq b_j$ have been obtained, where b_j are columns of the true word c. Since $(z_j, \hat{b}_j) = n_b - r_j^{(i)}$, according to Lemma 1 $(z_j, b_j) \leq n_b - 2d_{b,i} + r_j^{(i)}$ or, with regard to (8), $(z_j, b_j) \leq n_b - p_j^{(i)}(v)$. If $j \notin \{I^{(i)}(v)\}$, the j-th position of the row $\hat{a}^{(i)}$ contains either true symbol or erasure. In this case, $(z_j, b_j) \leq$ $\leq n_b - p_j^{(i)}(v)$. Consequently,

$$(c, z) = \sum_{j=1}^{n_a} (z_j, b_j) \leq \sum_{j=1}^{n_a} \left(n_b - p_j^{(i)}(v)\right) = n - S^{(i)}(v),$$

whence obtain by comparing with (10) that $S^{(i)}(v) < d_i$, i.e. inequality (9) follows.

Demonstrate now that if $S^{(i)}(k) \neq S^{(i)}(v)$ for some k, then $S^{(i)}(k) > d_i$. The rows $a^{(i)}(v)$ and $a^{(i)}(k)$ differ at least in $d_{a,i}$ positions whose set is denoted by $J_{v,k}$. If for $j \in J_{v,k}$, the symbol $\hat{a}_j^{(i)}$ is erased, $p_j^{(i)}(v) + p_j^{(i)}(k) =$ $= 2d_{b,i}$. If for $j \in J_{v,k}$, the symbol $\hat{a}_j^{(i)}$ is not erased, it differs either from $a_j^{(i)}(v)$, or from $a_j^{(i)}(k)$, i.e. $p_j^{(i)}(v) + p_j^{(i)}(k) \geq 2d_{b,i}$. Consequently,

$$S^{(i)}(v) + S^{(i)}(k) \geq \sum_{j \in J_{v,k}} \left(p_j^{(i)}(k) + p_j^{(i)}(v)\right) \geq 2d_{a,i}\, d_{b,i}.$$

Hence $S^{(i)}(k) > d_i$.

Q.E.D.

As Theorem 3 shows below, under definite selection of the threshold system $\{T_i\}$, condition (11) is always observed if constraint (10) holds. Let the vector z be received, and at the i-th step of decoding the "distance" row $R^{(i)} = (r_1^{(i)}, \ldots, r_{n_a}^{(i)})$ be determined. Denote by $J = \{1, 2, \ldots, n_a\}$ the set of numbers of all n_a columns, and by J the set of numbers of those columns of the received word z which are erased at MGD decoding of columns,

$$J_0 = \{j \in J \mid r_j^{(i)} \geq d_{b,i}\},\ J' = J \setminus J_0; \tag{12}$$

where J' is the set of the numbers of decoded columns. Let $R_1^{(i)} = \{r_1, r_2, \ldots, r_{M_i}\}$ be the set of pairwise different "distances" $r_j^{(i)}$ for all $j \in J'$ strictly ordered in decreasing order, i.e.

$$d_{b,i} > r_1 > r_2 > \ldots > r_{M_i}.$$

Obviously, $M_i = |J'|$ if the "distances" $r_j^{(i)}$ for all $j \in J'$ are pairwise different. Otherwise $M_i < |J'|$. For each received vector z, take at the i-th decoding step a threshold system $T^{(i)}(z) = \{t_1, \ldots, t_{M_i}\}$, where

$$t_1 = d_{b,i},\ t_{k+1} = r_k \text{ for all } k = 1, \ldots, M_i - 1. \tag{13}$$

In other words, at the first attempt of decoding by the external code A_i, only columns with numbers from J_0 are erased (these columns were already erased at MGD decoding of internal codes). At the second attempt all those columns $j \in J'$ are erased for which $r_j^{(i)} = r_1$ then all columns with the "distance" r_2 are erased, etc. After M_i attempts, only columns with "distance" r_{M_i} are left conventionally denoted by t_{M_i+1}.

Theorem 3. (1) If for received word z condition (5) is met the transmitted word c will be restored as the result of m steps of decoding of the word z by the threshold system $T^{(i)}(z)$, $i = 1, \ldots, m$.

(2) Let the rows $a^{(1)}, \ldots, a^{(i-1)}$ of the transmitted word c be determined correctly. Then the threshold system $T^{(i)}(z)$ provides correct determination of the row $a^{(i)}$ if condition (10) is met.

(3) At the i-th step of decoding by threshold system $T^{(i)}(z)$, at most $[(d_{a,i} + 1)/2]$ thresholds suffice for determination of the correct row $a^{(i)}$ for all $i = 1, \ldots, m$.

Proof. Statement (1) is a trivial corollary to Statement (2) whose proof will now be considered. Assume that the first $i - 1$ rows $a^{(1)}, \ldots, a^{(i-1)}$ are correctly determined. At the i-th step of decoding of the word z, $1 \leq i \leq m$, the rows $\hat{a}^{(i)}$ and $R^{(i)} = (r_1^{(i)}, \ldots, r_{n_a}^{(i)})$ were obtained and the row $\hat{a}^{(i)}$ is decoded by threshold system (13). In doing so, the rows $\hat{a}^{(i)}(k)$, $k = 1. \ldots, M_i$ are successively defined which are decoded with correction of errors and erasures.

Let ϱ_k and e_k denote the numbers of erased and erroneous symbols in the row $\hat{a}^{(i)}(k)$ respectively. As shown below, (c, z) is upper-bounded by means of all ϱ_k and e_k and if $\varrho_k + 2e_k \geq d_{a,i}$ for all $k = 1, \ldots, M_i$, $(c, z) < n - d_i$ which contradicts the condition of the theorem.

One may readily see that at passage from the threshold t_k to t_{k+1}, $k = 1, \ldots$ $\ldots, M_{i-1}$ additionally $\varrho_{k+1} - \varrho_k$ rows are erased of which $e_k - e_{k+1}$ were erroneous. For threshold system [13] at such a passage all the columns with numbers $j \in J'$ for which $r_j^{(i)} = t_{k+1}$ are erased. Denote the set of these columns by J_k, i.e.

$$(z_j, \hat{b}_j) = n_b - t_{k+1} \quad \text{for} \quad j \in J_k, \; k = 1, \ldots, M.$$

According to Lemma 1 for all $e_k - e_{k+1}$ erroneous columns from J_k we have the estimate $(z_j, b_j) \leq n_b - 2d_{b,i} + t_{k+1}$, and for the remaining $(\varrho_{k+1} - \varrho_k) -$ $- (e_k - e_{k+1})$ columns from J_k we have the estimate $(z_j, b_j) \leq n_b - t_{k+1}$. Thus, for all $k = 1, \ldots, M_i - 1$

$$\sum_{j \in J_k} (b_j, z_j) \leq (e_k - e_{k+1})(n_b - 2d_{b,i} + t_{k+1}) + \\ + ((\varrho_{k+1} - \varrho_k) - (e_k - e_{k+1}))(n_b - t_{k+1}). \tag{14}$$

Moreover, note that at decoding by the threshold $t_1 = d_{b,i}$, all ϱ_1 columns from J_0 are erased, and after decoding by the threshold t_{M_i}, $n_a - \varrho_{M_i}$ columns are left, of which e_{M_i} are erroneous. Therefore

$$\sum_{j \in J_0} (b_j, z_j) \leq (n_b - d_{b,i})\, \varrho_1, \tag{15}$$

$$\sum_{j \in J_{M_i}} (b_j, z_j) \leq e_{M_i}(n_b - 2d_{b,i} + t_{M_i+1}) + (n_a - \varrho_{M_i} - e_{M_i})\,(n_b - t_{M_i+1}). \tag{16}$$

Since

$$(c, z) = \sum_{j=1}^{n_a} (b_j, z_j) = \sum_{K=0}^{M_i} \sum_{j \in J_k} (b_j, z_j),$$

we obtain after elementary transformations from (14)–(16)

$$(c, z) \leq n - \sum_{k=1}^{M_i} (t_k - t_{k+1})\,(\varrho_k + 2e_k) - t_{M_i+1} \cdot n_a. \tag{17}$$

Assume now that for all $k = 1, \ldots, M_i$

$$\varrho_k + 2e_k \geq d_{a,i}.$$

Then obtain from (17)

$$(c, z) \leq n - d_{a,i}\, d_{b,i} = n - d_i$$

which contradicts condition (10). Thus, there will be at least one threshold t_v such that

$$\varrho_v + 2e_v \leq d_{a,i}. \tag{18}$$

Then decoding of the row $\hat{a}^{(i)}(v)$ will result in the true row $a^{(i)}$.

Q.E.D.

Now, demonstrate that from a system of M_i thresholds one may select at most $[(d_{a,i} + 1)/2]$ thresholds so that condition (18) is satisfied at least for one of them. One may easily see that at least $d_{a\,i}$ symbols are erased at decoding by the threshold t_{k_0} where $k_0 = d_{a,i} + 1$. Therefore, for all the thresholds t_k at $k \geq k_0$ condition (18) is not met. Consequently, the "true" threshold t_v (see (18)) is among $d_{a,i}$ thresholds $t_1, t_2, \ldots, t_{d_{a,i}}$.

Now let P be a subset of numbers of those of $d_{a,i} - 1$ thresholds $t_2, \ldots, t_{d_{a,i}}$ which erase at decoding exactly one column. Let the column number j_s, $1 \leq s \leq n_a$ be erased at decoding by the threshold t_p, $p \in P$. Note that $\varrho_{p-1} = \varrho_p - 1$ and $e_{p-1} = e_p$ if the column with number j_s was not erroneous, or $e_{p-1} = e_p + 1$, otherwise. Therefore,

$$2e_{p-1} + \varrho_{p-1} = 2e_p + \varrho_p \pm 1.$$

For each received vector z, reject that the threshold p' ($p' = p - 1$ or $p' = p$) of two thresholds t_{p-1} and t_p ($p \in P$) for which $\varrho_{p'} = d_{a,i} \pmod 2$. If the threshold p' is "true" and satisfies (18),

$$2e_{p'} + \varrho_{p'} \leq d_{a,i} - 2.$$

Then another threshold also satisfies (18). Decoding by any of the thresholds left after rejection erases at least two columns; whence the sufficiency of $[(d_{a,i} + 1)/2]$ thresholds follows. This completes the proof.

Remark. It was also demonstrated in [1] that the received vector Z may be MGD decoded into the word c of code of length n with distance d at $[(d + 1)/2]$ attempts to correct errors and erasures (of course, if $(c, z) > n - d$). At cascaded MGD decoding, one performs MGD decoding by internal codes of length n_b followed at the i-th step by $[(d_{a\,i} + 1)/2]$ attempts to decode by external codes of length n_a.

In the presence of $M_i^0 = [(d_{a,i} + 1)/2]$ thresholds, cascaded MGD decoding realizes "code distance" defined by the scalar product $((c, z) > n - d^{(H)})$. Assume now that decoding of external codes makes use of $M < M_i^0$ thresholds. Let us determine under which conditions imposed on the scalar product of the transmitted and received words c and z, respectively, the word c will be found as a result of cascaded MGD decoding by systems of M thresholds. In what follows, the systems of M thresholds T_i, $i = 1, \ldots, m$ used at the i-th decoding step are fixed for each of the m steps, i.e. are independent of the received vector z.

Let at the i-th decoding step, $1 \leq i \leq m$, a uniform system of M thresholds $P^{(i)} = \{p_1, \ldots, p_m\}$ be used where

$$p_k = d_{b,i}\left(\frac{2M - 2(k-1)}{2M+1}\right), \quad k = 1, \ldots, M. \tag{19}$$

Note that in this case $p_1 < d_{b,i}$. The following theorem shows that the threshold system $P^{(i)}$, $i = 1, \ldots, m$, is the best. For conventional cascaded and iterated codes this fact was demonstrated in [5].

Theorem 4. (1) Let the rows $a^{(1)}, \ldots, a^{(i-1)}$ of the transmitted word c be determined correctly, $1 \leq i \leq m$. At the i-th step of decoding of the received word z, the system of M thresholds $P^{(i)}$ (19) ensures correct determination of the row $a^{(i)}$ if

$$(c, z) > n - d_i \cdot \frac{2M}{2M+1}. \tag{20}$$

(2) If the inequality

$$(c, z) > n - d^{(H)} \cdot \frac{2M}{2M+1} \tag{21}$$

holds for the received word z, the transmitted word c will be determined through decoding by the threshold systems $P^{(i)}$ (19), $i = 1, \ldots, m$.

(3) If in the code $\{c\}$ minimal distance is $d^{(H)}$, for cascaded decoding by any other systems of M fixed thresholds T_i, $i = 1, \ldots, m$, there will be a code word c and received word z satisfying (21), whose correct decoding by cascaded algorithm is not possible.

Remark. In point of fact, a stronger result will be obtained in the course of proving the theorem. Exact bound will be constructed for realizable distance for any system of M fixed thresholds which depends only on two parameters δ and Δ where $\delta = d_{b,i} - t_1$ and $\Delta = \max\limits_k (t_k - t_{k+1})$.

Proof. Let at the i-th of decoding of the received vector z the system of M fixed thresholds $T = (t_1, \ldots, t_M)$ be used where $d_{b,i} \geq t_1 > t_2 > \ldots > t_M > 0$. For the transmitted word c denote by $z^{(i)}(T, c)$ the set of received vectors z which cannot be decoded into the word c at the i-th step by the threshold system T. This means that

$$2e_k + \varrho_k \geq d_{a,i} \quad \text{for all} \quad k = 1, \ldots, M \tag{22}$$

where, as above, ϱ_k and e_k denote the numbers of erased and erroneous symbols in the row $\hat{a}^{(i)}(k)$, respectively (the row is the result of application of the threshold t_k). Let

$$D^{(i)}(T) = \sup (c, z) \qquad \text{over all} \qquad c \in \{c\}, z \in Z^{(i)}(T, c).$$

Obviously, the threshold system T provides determination of the row $a^{(i)}$ if $(c, z) > D^{(i)}(T)$. We need to prove that under the threshold system $P^{(i)}$

$$D^{(i)}(P^{(i)}) = n - d_i \cdot \frac{2M}{2M+1}$$

and under any other system of M thresholds T

$$D^{(i)}(T) > n - d_i \cdot \frac{2M}{2M+1},$$

i.e.

$$D^{(i)}(P^{(i)}) = \inf_T \{D^{(i)}(T)\}.$$

For a given system T of M thresholds, let us find the upper bound of $D^{(i)}(T)$. Similarly to the proof of Theorem 3, the estimation will be carried out by means of e_k and ϱ_k $(k = 1, \ldots, M)$.

Let the word z be received. Decompose the set J of the column numbers of z_j, $j = 1, \ldots, n_a$ into $M + 1$ subsets as follows

$$J_0 = \{j | (z_j, \hat{b}_j) \leq n_b - t_1\},$$

$$J_k = \{j | n_b - t_k < (z_j, \hat{b}_j) < n_b - t_{k+1}\}, \quad k = 1, \ldots, M,$$

where $t_{M+1} = 0$ by definition. Obviously, at passage from threshold t_k to t_{k+1}, $k = 1, \ldots, M-1$, $\varrho_{k+1} - \varrho_k$ columns z_j, $j \in J_k$, are erased of which $e_k - e_{k+1}$ are erroneous. According to Lemma 1, for each of the erroneous columns z_j, $j \in J_k$, the following inequality holds

$$(z_j, b_j) < n_b - 2d_{b,i} + t_k \tag{24}$$

because $(z_j, \hat{b}_j) > n - t_k$ and $\hat{b}_j \neq b_j$. Hence, for all $k = 1, \ldots, M-1$ one has

$$\sum_{j \in J_k} (b_j, z_j) \leq (e_k - e_{k+1})(n_b - 2d_{b,i} + t_k) + ((\varrho_{k+1} - \varrho_k) - (e_k - e_{k+1})(n_b - t_{k+1})). \tag{25}$$

Similarly

$$\sum_{j \in J_0} (b_j, z_j) \leq (n_b - t_1)\varrho_1, \tag{26}$$

$$\sum_{j \in J_M} (b_j, z_j) \leq e_M(n_b - 2d_{b,i} + t_M) + (n_a - \varrho_M - e_M) \cdot n_b. \tag{27}$$

After summation of (25)–(27) and some elementary transformations, we obtain

$$(c, z) \leq n - \sum_{k=1}^{M} (\varrho_k + 2e_k)(t_k - t_{k+1}) + \sum_{k=1}^{M} (e_k - e_{k+1})(t_k - t_{k+1}) - 2e_1(d_{b,i} - t_1), \tag{28}$$

where $e_{M+1} = 0$.

Introduce the following definitions. Let $E = (\varrho_1, \ldots, \varrho_M, e_1, \ldots, e_M)$ denote the ordered set of $2M$ non-negative integers satisfying condition (22) and inequalities

$$0 \leq \varrho_1 \leq \varrho_2 \leq \ldots \leq \varrho_M$$

$$e_1 \geq e_2 \geq \ldots \geq e_M \geq 0.$$

Let now $E^{(i)}$ denote the set of all ensembles E satisfying conditions (22), (29), (30). Define $P(T, E)$ as follows:

$$P(T, E) = n - \sum_{k=1}^{M} (\varrho_k + 2e_k)(t_k - t_{k+1}) + \sum_{k=1}^{M} (e_k - e_{k+1}) - 2e_1(d_{b,i} - t_1). \tag{31}$$

Stated differently, $(c, z) \leq P(T, E)$, where E is an ensemble of numbers ϱ_k, e_k, $k = 1, \ldots, M$, obtained at decoding of the word z by the system of thresholds T. Hence,

$$D^{(i)}T \leq \sup_{E \in E^{(i)}} \{P(T, E)\}. \tag{32}$$

As will be shown below, in (32) equality takes place; therefore, it is sufficient to determine $\sup_{E \in E^{(i)}} P(T, E)$ instead of $D^{(i)}T$.

Let $E_0^{(i)}$ be a set of ensembles $E \in E^{(i)}$ such that in (22) equality is observed for all $k = 1, \ldots, M$, i.e.

$$\varrho_k + 2e_k = d_{a,i}, \quad k = 1, \ldots, M. \tag{33}$$

Note that for such ensembles $\varrho_k \equiv d_{a,i} \pmod 2$ for all $k = 1, \ldots, M$.

Lemma 2.

$$\sup_{E \in E^{(i)}} \{P(T, E)\} = \sup_{E \in E_0^{(i)}} \{P(T, E)\}.$$

See the proof in the appendix.

It may be readily seen from (31) and (33) that for all ensembles $E \in E_0^{(i)}$

$$P(T, E) = n - d_{a,i}\, t_1 + \sum_{k=1}^{M} (e_k - e_{k+1})(t_k - t_{k+1}) - 2e_1(d_{b,i} - t_1).$$

Let

$$\Delta = \max_{k=1,\ldots,M} \{t_k - t_{k+1}\}, \qquad \delta = d_{b,i} - t_1$$

and let v be such that $t_v - t_{v+1} = \Delta$. If there are several v, any of them can be selected. Obviously, $\Delta \geq (d_{b,i} - \delta)/M$. Note that

$$\sum_{k=1}^{M} (e_k - e_{k+1})(t_k - t_{k+1}) \leq \Delta \cdot \sum_{k=1}^{M} (e_k - e_{k+1}) = \Delta \cdot e_1,$$

equality being attained when $e_v - e_{v+1} = e_1$ which is equivalent to the condition $e_1 = \ldots = e_v$, $e_{v+1} = \ldots = e_M = 0$. Whence

$$P(T, E) \leq n - d_{a,i} \cdot t_1 + e_1(\Delta - 2\delta)$$

and

$$\sup_{E \in E_0^{(i)}} \{P(T, E)\} = \sup_{e_1} \{n - d_{a,i} \cdot t_1 + e_1(\Delta - 2\delta)\} \tag{34}$$

$\sup P(T, E)$ being attained at ensembles $E \in E_0^{(i)}$ for which

$$e_1 = \ldots = e_v, \; e_{v+1} = \ldots = e_M = 0,$$

$$\varrho_k = d_{a,i} - 2e_k, \quad k = 1, \ldots, M.$$

Consider now two cases.

Case 1. Let T be a system of M thresholds for which $\Delta - 2\delta > 0$. As it may be seen from (34), e_1 in this case should be taken as maximal. Since $2e_1 + \varrho_1 = d_{a,i}$, $e_1 \leq [d_{a,i}/2]$. By taking $e_1 = [d_{a,i}/2]$, we obtain that

$$\sup_{E \in E_0^{(i)}} P(T, E) = n - d_{a,i} \cdot t_1 + \left[\frac{d_{a,i}}{2}\right](\Delta - 2\delta),$$

equality being attained if E is taken as follows

$$e_1 = \ldots = e_v = [d_{a,i}/2], \qquad e_{v+1} = \ldots = e_M = 0, \tag{35}$$

$$\varrho_1 = \ldots = \varrho_v = d_{a,i} - 2e_1, \quad \varrho_{v+1} = \ldots = \varrho_M = d_{a,i}. \tag{36}$$

Let us now prove that $D^{(i)}(T) = \sup_{E \in E^{(i)}} P(T, E)$. Take two code vectors c and c' differing in $d_{a,i}$ columns having $d_{b,i}$ symbols each. Knowing, this may be done if the code realizes the estimate $d_i = d_{a,i}, d_{b,i}$ of code distance of the i-th cascade. Let the word c be transmitted. Construct the word z in the following manner. Select arbitrary $[d_{a,i}/2]$ columns from these $d_{a,i}$ of the words c and c'. Let b_j be one of the selected columns of the word c, and b'_j be an arbitrary column of the word c'. Let columns of z_j be MGD decoded into b'_j for all $[d_{a,i}/2]$ selected columns. Obviously, the columns z_j may be selected so as to make the difference between (z_j, b'_j) and $n_b - t_v$ as small as desired and, moreover, to make the difference between (z_j, b_j) and $n_b - 2d_{b,i} + t_v$ also as small as desired. In addition, let for $d_{a,i} - 2[d_{a,i}/2]$ columns $(z_j, b_j) = n_b - t_1$, and for $[d_{a,i}/2]$ columns $(z_j, b_j) = n_b - t_{v+1}$. Let for the rest of the columns $z_j = b_j$. According to this construction, the scalar product (c, z) differs as little as desired from $P(T, E)$ (see inequalities (24)–(28)). Moreover, under such a construction of z, conditions (35) and (36) are satisfied, i.e.

$$P(T, E) = n - d_{a,i} t_1 + [d_{a,i}/2](\Delta - 2\delta).$$

And so, for any system of thresholds T_i for which $\Delta \geq 2\delta$,

$$D^{(i)} T = \sup_{c \in \{C\}} \left\{ n - d_{a,i} \cdot t_1 + \left[\frac{d_{a,i}}{2}\right](\Delta - 2\delta) \right\}, \quad z \in Z^{(i)}(T, C). \tag{37}$$

Case 2. Consider now all systems of thresholds T_i for which $\Delta - 2\delta \leq 0$. According to (34), one should take in this case $e_1 = 0$ and

$$\sup_{E \in E_0^{(i)}} P(T, E) = n - d_{a,i} \cdot t_1$$

equality being attained if E is selected as follows

$$e_1 = e_2 = \ldots = e_M = 0, \tag{38}$$

$$\varrho_1 = \varrho_2 = \ldots = \varrho_M = d_{a,i}. \tag{39}$$

In this case, for an arbitrary vector c (with columns b_j) construct the vector z as follows. Select $d_{a,i}$ arbitrary columns in c and for each of them construct the columns z_j so as to provide $(b_j, z_j) = n_b - t_1$. In this case $(c, z) = P(T, E)$, and satisfies (38) and (39).

Thus, for any system of thresholds T_i for which $\Delta \leq 2\delta$

$$D^{(i)}(T) = \sup_{c \in \{C\}} (c, z) = n - d_{a,i} \cdot t_1, \quad z \in Z^{(i)}(T, C) \qquad (40)$$

and there exist c and z for which $(c, z) = n - d_{a,i} \cdot t_1$ are such that correct decoding of the word z is not possible.

Let us demonstrate now that the system of thresholds $P^{(i)}$ is the best, i.e. $\inf_T D^{(i)}(T)$ is attained in the case of (37) and (40). At first, let (37) be met, i.e. $\Delta \geq 2\delta$. Note that

$$\inf_{T:\Delta > 2\delta} D^{(i)}(T) = \inf_{\delta} \inf_{\Delta \geq 2\delta} D^{(i)}(T) =$$

$$= \inf_{\delta = \frac{\Delta}{2}} (n - d_{a,i} \cdot t_1) = \inf_{\delta = \frac{\Delta}{2}} (n - d_{a,i} \cdot d_{b,i} + d_{a,i} \cdot \delta).$$

Obviously, one has to select the smallest δ such that $\delta = \Delta/2$ but $\Delta \geq$ $\geq (d_{b,i} - \delta)/M$, whence $\delta \geq (d_{b,i} - \delta)/2M$, i.e. $\delta \geq d_{b,i}/(2M + 1)$, equality being attained if $\Delta = (d_{b,i} - \delta)/M$ which is possible only if $t_k - t_{k+1} = \Delta$ for all $k = 1, \ldots, M$. In this case $\delta = d_{b,i}/(2M + 1)$, $t_k - t_{k+1} = \dfrac{2d_{b,i}}{2M + 1}$ for all $k = 1, \ldots, M$ which corresponds to the threshold system $P^{(i)}$. Thus,

$$\inf_{T:\Delta \geq 2\delta} D^{(i)}(T) = D^{(i)}(P^{(i)}) = n - d_{a,i} \cdot d_{b,i} \cdot \frac{2M}{2M + 1}.$$

Let now (40) be met, i.e. the system of thresholds $T^{(i)}$ are considered for which $\Delta \leq 2\delta$. Similarly,

$$\inf_{T:\Delta \leq 2\delta} D^{(i)}(T) = \inf_{\Delta \leq 2\delta} (n - d_{a,i} \cdot d_{b,i} + d_{a,i} \cdot \delta).$$

In this case it is required to select the smallest δ, i.e. $\delta = \Delta/2$. Since $\Delta \geq$ $\geq (d_{b,i} - \delta)/M$, we obtain similarly that $\delta_{\min} = d_{b,i}/(2M + 1)$ and $t_k - t_{k+1} =$ $= 2d_{b,i}/(2M + 1)$ for all $k = 1, \ldots, M$ which corresponds to the system $P^{(i)}$. Thus,

$$\inf \sup_{c \in \{C\}} (c, z) = n - d_{a,i} \cdot d_{b,i} \cdot \frac{2M}{2M + 1}, \quad z \in Z^{(i)}(T, C)$$

the system M of thresholds $P^{(i)}$ being the best.

Q.E.D.

Note that algorithm (8–9) for the selection of the correct row was suggested in [6] for discrete channel. In this case the columns $Z_j, j = 1, \ldots, n_a$ are decoded in $\hat{b}_j$ with correction not more than $l_i = [(d_{b,i} - 1)/2]$ errors. The row $\hat{a}^{(i)}$ is decoded into $\{a^{(i)}(k)\}$ with threshold system $L_i = \{l_i\, l_i - 1, \ldots, 0\}$. Let $d(x, y)$ denote the Hamming distance between x and y. Then

$$P_j^{(i)}(k) = \begin{cases} l_i + 1, & \text{if } \hat{a}_j^{(i} \text{ is erased} \\ d(z_j, \hat{b}_j), & \text{if } \hat{a}_j^{(i)} \text{ is not erased and } \hat{a}_j^{(i)} = a_j^{(i)}(k) \\ d_{b,i} - d(z_j \hat{b}_j), & \text{if } \hat{a}_j^{(i)} \text{ is not erased and } \hat{a}_j^{(i)} \neq a_j^{(i)}(k). \end{cases}$$

The decoding is true if the number of errors is not more than $[(d^{(H)} - 1)/2]$. The connection between discrete cascaded decoding and MGD decoding is investigated in [7].

Appendix

Proof of Lemma 2. Consider the arbitrary ensemble $E \in E^{(i)}$, $E = (\varrho_1, \ldots, \varrho_M, e_1, \ldots, e_M)$. Introduce the ensemble $E = (\varrho', \ldots, \varrho'_M, e'_1, \ldots, e'_M)$ in the following manner. Let for all $k = 1, \ldots, M'$

$$\varrho'_k = \begin{cases} d_{a,i} & \text{if } \varrho_k \geq d_{a,i} \\ \varrho_k & \text{if } \varrho_k < d_{a,i} \text{ and } \varrho_k \equiv d_{a,i} \pmod 2 \\ \varrho_{k-1} & \text{if } \varrho_k < d_{a,i} \text{ and } \varrho_k \not\equiv d_{a,i} \pmod 2 \end{cases}$$

and

$$l'_k = \begin{cases} 0 & \text{if } \varrho_k \geq d_{a,i} \\ (d_{a,i} - \varrho_k)/2 & \text{if } \varrho_k < d_{a,i} \text{ and } \varrho_k \equiv d_{a,i} \pmod 2 \\ (d_{a,i} - (\varrho_k - 1))/2 & \text{if } \varrho_k < d_{a,i} \text{ and } \varrho_k \not\equiv d_{a,i} \pmod 2. \end{cases}$$

Note that $\varrho'_k + 2e' = d_{a,i}$ for all $k = 1, \ldots, M$ and if conditions (29) and (30) are met for the ensemble E, then $0 \leq \varrho'_1 \leq \varrho'_2 \leq \ldots \leq \varrho'_M$ and $e'_1 \geq e'_2 \geq \ldots \geq e'_M \geq 0$. Thus, the ensemble $E' \in E_0^{(i)}$, and for all $k = 1, \ldots, M$

$$\varrho'_k \leq \varrho_k, \tag{A.1}$$

$$l'_k \leq e_k. \tag{A.2}$$

Inequality (A.1) is obvious. Inequality (A.2) follows from the fact that if $\varrho_k \not\equiv d_a$, (mod 2), then $2e_k + \varrho_k \geq d_{a,i} + 1$.

Further, we obtain from (31) that

$$P(T, E') - P(T, E) = \sum_{k=1}^{M} (\varrho_k - \varrho'_k)(t_k - t_{k+1}) +$$

$$+ 2(e_1 - e'_1)(d_{b,i} - t_1) + \sum_{k=1}^{M} \left((e_k - e'_k) + (e_{k+1} - e'_{k+1})\right)(t_k - t_{k+1}).$$

Allowing for (A.1) and (A.2), we obtain

$$P(T, E') - P(T, E) \geq 0,$$

equality being possible only if $\varrho'_k = \varrho_k$ and $e'_k = e_k$ for all $k = 1, \ldots, M$.

Q.E.D.

References

1. *Forni, D.*, Kaskadnye kody. Moscow. Mir, 1970.
2. *Zinov'ev, V. A., Zyablov, V. V.*, Decoding of non-linear generalized cascaded codes. Problemy peredachi informatsii, 1978, **14**, *2*, 46–52 (in Russian).
3. *Zinov'ev, V. A., Zyablov, V. V.*, Cascaded decoding with respect to minimum of generalized distance. Proc. of 5-th Int. symp. on information theory, I, Abstracts, Moscow—Tbilisi, 1979, 161–162 (in Russian).
4. *Zinov'ev, V. A.*, Generalized cascaded codes. Problemi peredachi informatsii, 1976, **12**, *1*, 5–15 (in Russian).
5. *Zyablov, V. V.*, Optimization of cascaded decoding algorithms, Problemi peredachi informatsii, 1973, **9**, *1*, 26–32 (in Russian).
6. *Dumer I. I.*, On decoding of generalized concatenated codes. Proc. of the fifth Soviet conference on computer networks, p. 4, Moscow—Vladivostok, 1980, 92–95 (in Russian).
7. *Boiarinov I. M.*, On correcting errors by direct sums of iterative and concatenated codes, Doklady of the Soviet Academy of Sciences, **252**, *3*, pp. 563–566 (in Russian).

ON STATISTICAL MODELS OF COMPLEX MULTIMODAL FUNCTIONS AND THEIR APPLICATION TO THE DESIGN OF OPTIMIZATION ALGORITHMS

A. G. ŽILINSKAS

1. Introduction

Nonlinear optimization algorithms usually rely on some models of the functions to be optimized; in the framework of these models the results of the preceding steps are interpreted and subsequent steps are planned. Thus, a quadratic model is often used in design of local algorithms [1]. However, in the case of optimizing multimodal functions the model would recognize significantly larger uncertainty in the behaviour of the function. In this article a statistical model proceeds from a rather general assumption that the intervals of possible function values are comparable likelihood-wise. A family of random values is shown to be a rather natural statistical model and that in the class of such models random fields can be obtained which are computationally simpler than Gaussian (usually employed as models of complex functions under uncertainty). The optimization algorithm combines a global search strategy which is obtained from ideas of rational choice with common local algorithms.

2. Representing the information on a multimodal function as a family of random values

Assume that there is no objective information on the function to be minimized $f(x)$, $x \in A \subset R^n$ other than values of $f(\cdot)$ computed in the points $x_i \in A : y_i = f(x_i)$, $i = \overline{1, k}$. Also, there is subjective information (such as solutions of kindred minimization problems) on the complexity of $f(\cdot)$ which has many extrema. The weakest constructive assumption on the available information is probably comparability of intervals of the possible values of $f(x)$, $x \neq x_i$, $i = \overline{1, k}$ in terms of likelihood. Assume that the relative likelihood (RL) is specified and denoted as $\geq x$ where $(a, a') \geq_x (b, b')$ implies that the event $f(x) \in (a, a')$ is at least equiprobable with $f(x) \in (b, b')$; x can be omitted if it is clear what point is meant. If $(a, a') \geq_x (b, b')$ and $(b, b') \geq_x (a, a')$, then this is denoted as $(a, a') \sim_x (b, b')$; $(a, a') >_x (b, b')$ implies that $(a, a') \geq (b, b')$ but it not true that $(a, a') \sim_x (b, b')$. Let a certain point, $x \neq x_i$,

$i = \overline{1,k}$ be fixed. The available information on $f(\cdot)$ is often not in conflict with the following assumptions on the rationality of RL:

A1. 1) for any two intervals (a, a') and (b, b') either $(a, a') \geq (b, b')$ or $(b, b') \geq (a, a')$;

A1. 2) if $(a, a') \geq (b, b')$ and $(b, b') \geq (c, c')$, then $(a, a') \geq (c, c')$;

A1. 3) $[a, a'] > \emptyset$ if and only if $\mu[a, a'] > 0$ where $u(\cdot)$ denotes the Lebesgue measure and $\emptyset$ is an empty set;

A1. 4) let $(a, a') \cap (b, b') = \emptyset\,(a, a') \cap (c, c') = \emptyset$, $a = b'$ either $a' = b$ and $a = f'$ or $a' = c(b, b') \geq (c, c')$ if and only if $[a, a'] \cup [b, b'] \geq$ $\geq [a, a'] \cup [c, c']$;

A1. 5) if $(a, a') \geq (b, b') > \emptyset$, then there exist a_1, a_2, $a \leq a_i < a'$, $i = 1, 2$ such that $(a, a_1) \sim (a_2, a') \sim (b, b')$.

Axioms A.1.1–A.1.5 define RL only for the simplest possible sets which are intervals. Extension of RL to more involved sets is substantiated by the following contention [2]: for any $x = \bigcup_{i=1}^{m} I(j)$, where $I(j) = (a_j, a'_j)$, $I(i) \cap I(j) =$ $= \emptyset$, $i \neq j$, there is an increasing sequence $b_1, b_2, \ldots, b_m$ such that $(-\infty, b_1) \sim$ $\sim I(1)$ $(b_{j-1}, b_j) \sim I(j)$, $j = \overline{2,m}$ where b_m is independent of the enumeration of the intervals $I(j)_j = \overline{1,m}$ and $b_m < \infty$ if $\mu((-\infty, \infty) \backslash x) > 0$.

A1.6) the interval $(-\infty, b_m)$ is equiprobable with x or

$$(-\infty, b_m) \sim x.$$

A1.6 permits extending RL to the algebra of finite unions of intervals while retaining all basic properties of RL. The following Theorem holds [2].

Theorem 1.1. Assume that axioms A1.1–A1.5 specify RL for intervals and A1.6 extends this to the algebra of finite unions of intervals. There is a unique probability density $p_x(\cdot)$ coordinated with RL, or $X_1 \geq X_2$ if and only if

$$\int_{x_1} p(z)\,dz \geq \int_{x_2} p_x(z)\,dz$$

where

$$X_j = \bigcup_{i=1}^{m_j} (a_{ji}, a'_{ji}).$$

The contention of the Theorem permits interpreting the value of $f(x)$ as a random quantity ξ_x with a distribution density $p_x(\cdot)$ and, consequently, the family ξ_x, $x \in A$, $x \neq x_i$, $i = \overline{1,k}$ as a statistical model of the function.

Since A1.1 requires the comparability of rather simple sets (interval), it is weaker than the comparability of the algebra elements, an assumption common in related papers [3–6]. RL transitivity expressed by A1.2 is one of the most fundamental assumptions on rationality of the likelihood relation and has been discussed by many writers (see e.g. [7]). The intuitive acceptability of axioms A1.1 and A1.2 for control engineers who have to solve complex optimization problems is confirmed in a psychological experiment [2]. A.1.4 expresses RL additivity and is a common rationality assumption [3–6]. Axioms A1.3 and A1.5 are specific of this research. A1.3 while expressing complexity of the function, asserts that one neither accurately forecast the value of $f(x)$ nor choose the interval (a, a') so that $\mu[a, a') > 0$ and $f(x) \in (a, a')$ be insured. Continuity of RL variation with change of the possible range of $f(x)$ seems very natural; axiom A1.5 expresses RL continuity in an obviously more illustrative way. The axiomatic characterization of RL specified only for simplest possible sets (intervals) and its subsequent extension with the aid of A1.6 intuitively seems more acceptable than characterizing RL directly for a complex system of sets.

From this system of axioms follow the existence and uniqueness of $p_x(\cdot)$; however, to obtain optimization algorithms, the density and its parameters should be constructed as functions of x. Gaussian density has been shown [2] to work as $p_x(\cdot)$. The mean and variance of this density are discussed below.

Note that a random function can be regarded as a family of random quantities and, consequently, is a particular case of the statistical models considered here. Reference [8] has extended the axiomatic characterization of RL to vectors of function values in several points and has shown that in this case there exists a random function coordinated with RL but the axioms of [8] are more complicated and are not as natural as A1.1–A1.6. To obtain optimization algorithms it is equally important that in the class of models associated with A1.1–A1.6 one can obtain fields which would be computationally simpler than random Gaussian fields.

3. Axiomatic definition of statistic model characteristics

As has been shown above, the rather natural assumptions on the information of $f(\cdot)$ make it possible to accept the family of random Gaussian quantities ξ_x, $x \in A$, $x \neq x_i$, $i = 1, k$ as a statistical model of $f(\cdot)$. The mean $m_k(x,(x_i, y_i), i = 1, k)$ and variance $s_k(x_i, y_i)$, $i = \overline{1, k}$) of the random quantity ξ_x will be defined axiomatically following, as above, the rationality ideas.

Informally $m_k(x, (x_i, y_i), i = \overline{1,k})$ is an estimate of the possible deflection of $f(x)$ from $m_k(x, (x_i, y_i)), i = \overline{1,k}$ or as conditional (with $f(x_i) = y_i, i = \overline{1,k}$) uncertainty of the value of $f(x)$. Note that the design of $m_k(x, (x_i, y_i)), i = \overline{1,k}$, $s_k(x, (x_i, y_i), i = \overline{1,k})$ is of interest per se as a problem of rational design of an extrapolating function under uncertainty [9] and of estimating its error; therefore characterization of $n_k(\cdot)$ and $s_k(\cdot)$ is regarded as an independent problem which is not directly related to A1.1–A1.6.

A2. 1) $m_k(x, (x_i, cy_i), i = \overline{1,k}) = cm_k(x, (x_i, y_i), i = \overline{1,k})$;

A2. 2) $m_k(x_i, y_i + z), i = \overline{1,k} = m_k(x, (x_i, y_i), i = \overline{1,k}) + z$;

A2. 3) for any permutation of indices $\{j(i), i = \overline{1,k}\} = \{1, 2, \ldots, k\}$ $m_k(x, (x_{j(i)}, y_{j(i)}), i = \overline{1,k} = m_k(x_i, y_i), i = \overline{1,k})$;

A2. 4) $m_k(x_j, (x_i, y_i), i = \overline{1,k}), j = \overline{1,k}$;

A2. 5) there is a function $v_k(\cdot) : (A \times R)^{k-1} \to R$ such that $m_k(x, (x_i, y_i), i = \overline{1,k}) = m_k(x, (x_i, z_i), i = \overline{1,k})$ where $z_i = z, \quad i = \overline{1,k-1}$, $z = v_k(x_i, y_i, i = \overline{1,k-1}, \zeta_k = y_k)$.

The following contention holds [9]:

Theorem 2.1. The only function that satisfies A2.1–A2.5 is

$$m_k(x, (x_i, y_i), i = \overline{1,k} = \sum_{i=1}^{k} y_i w_i^k(x, x_j, j = \overline{1,k}) \tag{2.1}$$

where the weights $w_i^k(\cdot)$ have certain properties: a) $\sum_{i=1}^{k} w_i^k(x, x_i) = \overline{1,k} = 1$; b) $w_i^k(x, x_1, \ldots, x_k) = w_p^k(x, x_1, \ldots, x_{j-1}, x_l, x_{j+1}, \ldots, x_j, x_{l-1}, x_j, x_{l+1}, \ldots, x_k)$ where $p = i$ with $i \neq j$; $i \neq l$ with $i = l$; $p = l$ with $i = j$; c) $w_j^k(x_l, x_i, i = \overline{1,k})$ is equal to one with $j = l$ and zero with $j \neq l$.

Axioms A2.1–A2.4 are natural beyond refutation. A2.5, which postulates the possibility of data aggregation, limits the function $m_k(\cdot)$. From axioms A2.1–A2.5 it follows that $m_k(x, (x_i, y_i), i = \overline{1,k})$ is the mean weighted value of the values of the function y_i; comparison of A2.1–A2.5 with axiomatic characterizations of the extrapolating function of the weighted mean from [10, 11] is made in [12]. In [11] is noted the possibility of axiomatically characterizing the weights $w_i^k(\cdot)$ but very little has been done for this thus far. Therefore the weights are chosen from heuristic reasoning some of which is given in [13]. Expression (2.1) with weights

$$w_i^k(x, x_1, \ldots, x_k) = d(x, x_i) \Big/ \sum_{j=1}^{k} d(x, x_j) \tag{2.2}$$

where $d(x, x_i) = \exp(-c||x - x_i||^2)/||x - x_i||$, $c > 0$; $||\cdot||$ is a Euclidean norm in R^n, has been investigated in [12] as an approximating function (extrapolator) and has been found accurate enough with very little computer load; the same weights are used in this article.

Considering the axiomatic characterization of $s_k(\cdot)$, for brevity denote $w_k^i = w_k^i(x, x_j,\ j = \overline{1,k})$.

A3.1) there is a function $\sigma(x, z)$: $A^2 \to R$ such that $\sigma(x, z) = \sigma(z, x)$, $\sigma(x, x) > \sigma(x, z)$, $x \neq z$;

A3. 2) $s_k(x(x_i, y_i),\ i = \overline{1,k}) = \gamma_k S_k(\sigma(\sigma_i, w_i^k),\ i = \overline{1,k})$ where $\gamma_k = \gamma_k(y_1, \ldots, y_k)$, $\sigma = \sigma(x, x)$, $\sigma_i = \sigma(x, x_i)$;

A3. 3) $S_k(\sigma + \Delta(\sigma_i, w_i^k),\ i = \overline{1,k} = S_k(\sigma(\sigma_i, w_i^k),\ i = \overline{1,k}) + \Delta$;

A3. 4) $S_k(c\sigma, (c\sigma_i, w_i^k),\ i = \overline{1,k}) = cS_k(\sigma(\sigma_i, w_i^k),\ i = \overline{1,k})$;

A3. 5) for any permutation of indices $\{j(i), i = \overline{1,k}\} = \{1, 2, \ldots, k\}$ $S_k(\sigma(\sigma_{j(i)}, w_{j(i)}^k),\ i = \overline{1,k}) = S_k(\sigma(\sigma_i, w_i^k),\ i = \overline{1,k})$;

A3. 6) there is a function $u_k(\cdot) : (R^2)^{k-1} \to R$ such that $S_k(\sigma, (\sigma_i, w_i^k),\ i = \overline{1,k}) = S_k(\sigma, (\sigma, w_i^k),\ i = \overline{1,k})$ where $\sigma_i' = u$, $i = \overline{1,k-1}$, $\sigma_k' = \sigma_k$, $u = u_k(\sigma_i, w_i^k, i = \overline{1,k-1})$;

A3. 7) let $\tilde{w}_i^k = w_i^k$, $i \neq j$, $i \neq k$, $\hat{\sigma}_i = \sigma_i$, $i \neq j$, $\sigma_k = 0$ and $w_j^k \sigma_j = \tilde{w}_j^k \tilde{\sigma}_j$ then $S_k(\sigma, (\sigma_i, w_i^k),\ i = \overline{1,k}) = S_k(\sigma(\tilde{\sigma}_i, \tilde{w}_i^k),\ i = \overline{1,k})$.

In the same way one can prove the following [14].

Theorem 2.2. The only function that satisfies A3.1–A3.7 is

$$S_k(x(x, (x_i, y_i),\ i = \overline{1,k}) = \gamma_k \Sigma(\sigma(x, x) - \sigma(x, x_i)) w_i^k(x, x_j),\ j = \overline{1,k}. \quad (2.3)$$

A3.1 defines the structure of $S_k(x, (x_i, y_i,\ i = \overline{1,k}))$ by postulating the dependence of conditional uncertainty of the values of $f(x)$ (with $f(x_i) = y_i$, $i = \overline{1,k}$ known) on the weights $w_k^i(x, x_j, j = \overline{1,k})$ and $\sigma(x, x_i)\ i = \overline{1,k}$ where $\sigma(x, z)$ is a function which expresses the information on $f(x)$ obtained in measuring $f(z)$. The significance of this informal description is revealed by axioms A3.2–A3.5. Since the information on $f(x)$ obtained by measuring it can be interpreted as unconditional uncertainty of $f(x)$, A3.2 asserts that the increase of unconditional uncertainty entails the same increase in conditional uncertainty. In compliance with A3.3 increased unconditional uncertainty may be compensated by an increased amount of information. The meaning of the proportionality axiom A3.3 and the symmetry axiom A3.5 is obvious and A3.6 postulates the possibility of data aggregation. A3.7 asserts that increase in

the information obtained from measuring the function in the point x is equivalent to proportional increase in the weight of that measurement.

As the Theorem asserts, the conditional uncertainty of the value of $f(x)$ with $f(x_i)$, $i = \overline{1, k}$ known is equal (with the accuracy of a multiplier) to the weighted mean of the differences $\sigma(x, x) - \sigma(x, x_i)$ which express the conditional uncertainty in the value of $f(x)$ with a unique available $f(x_i)$. It is rather natural to assume that uncertainty of $f(x)$ with $f(x_i)$ known is proportional to the distance between x and x_i, or

$$\sigma(x, x) - \sigma(x, x_i) = a\,||x - x_i||, a > 0. \tag{2.4}$$

Returning to the statistical model of the function and to the interpretation of $m_k(\cdot)$ and $s_k(\cdot)$ as the mean and variance respectively of ξ_x, the multiplier γ_k can be determined by the method of least squares, $\gamma_k = \gamma^*_{k-1}$ where γ^*_{k-1} minimizes the sum

$$\sum_{i=1}^{k} [y_i - m_k(x_i, (x_j, y_j),\ j = \overline{1, k}),\ j = i)^2 - s_{k-1}(x_i, (x_j, y_j))],\ j = \overline{1, k},\ j \neq i.$$

As noted above, random functions are a particular case of the statistical models of ξ_x. Therefore it would be worthwhile to answer the question, under which conditions (2.1) and (2.2) are expressions of conditional mean and conditional variance of the random Gaussian function.

A2. 6) let $z_i = y_i$, $i = \overline{1, k-1}$, $z_k = m_{k-1}(x_i, y_i)$, $i = \overline{1, k-1}$ then $m_k(x, (x_i, z_i),\ i = \overline{1, k} = m_{k-1}(x, (x_i, y_i),\ i = \overline{1, k-1}$;

A2. 7) $\sigma(x, \tilde{x})$ is positive definite and $m_2(x, (x_i, \tilde{z}_i)\ i = 1, 2) = m_2(\tilde{x}_i, (x_i, z_i)$ $i = 1, 2$, where $z_i = \sigma(x, x_i)$, $\tilde{z}_i = \sigma(\tilde{x}, x_i)$, $i = 1, 2$.

Axiom A2.6 expresses coordination of $m_{k-1}(\cdot)$ and $m_k(\cdot)$. Since $\tilde{z}_i$ expresses information on $f(\tilde{x})$ obtained in measuring $f(x_i)$, then $\sigma_2(x, \tilde{x}) = m_2(x, (x_i, z_i)$ $i = 1, 2)$ represents the expected (from available $\tilde{z}_i = \sigma(\tilde{x}, x_i)$) informtaion of $f(\tilde{x})$ in measuring $f(x)$; axiom A2.7 postulates the symmetry of $\sigma_2(\cdot)$ with respect to its arguments. Supplementing A2.1–A2.5 with A2.6 and A2.7 has been shown in [9] to entail the following expression for the weights $(w_1^k, \ldots, w_k^k)^t = \Sigma_k^{-1}(\sigma_1, \ldots, \sigma_k)^t$ where Σ_k is a matrix made of elements $\sigma(x_i, x_j)$, $i, j = \overline{1, k}$ and t denotes transposition. In this case (2.1) and (2.3) express conditional mean and conditional variance of the random Gaussian function with mean μ and a correlation function $\sigma(.)$, or can be represented as

$$m_k(x, (x_c, y_i)\ i = \overline{1, k}) = (y_1, \ldots, y_k)\,\Sigma_k^{-1}(\sigma_1, \ldots, \sigma_k)^t \tag{2.5}$$

$$S_k(x(x_i, y_i),\ i = \overline{1, k}) = \sigma - (\sigma_1, \ldots, \sigma_k)\,\Sigma_k^{-1}(\sigma_n, \ldots, \sigma_k)^t \tag{2.6}$$

where u is canceled by virtue of item a) of Theorem 2.1.

Formulae (2.5) and (2.6) which are a particular case of (2.1) and (2.3), involve computationally a rather complex operation of matrix inversion. Computing (2.1) with weighting function (2.2) has been shown [12] to require dozens of times less computer time than (2.5) does. Also, computation by formulae (2.5) and (2.6) requires an additional amount of computer core memory for storing the matrix Σ_k^{-1}. Note that with a special structure of the matrix Σ_k its inversion may cause no computing difficulties, for instance if $n = 1$ and $\sigma(\cdot)$ is a correlational function of the random Markov process. On the other hand, the choice of weights is sometimes easier to explain than the choice of the associated random function; thus in a one-dimensional case for $x_i \leq x \leq x_{i+1}$ the very natural weights are

$$w_i(x, x_j, j = \overline{1, k}) = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \quad w_{i+1}(x, x_j, j = \overline{1, k}) = 1 - w_1(x_j), j = \overline{1, k} \tag{2.7}$$

and

$$\sigma(x, x) - \sigma(x, x_i) = a(x - x_i).$$

Then $m_k(x, (x_i, y_i)), i = \overline{1, k}$ and $S_k(x(x_i, x_i)), i = \overline{1, k}$ are expressions of the conditional mean and conditional variance of the Wiener process, or a process whose choice function are nondifferentiable almost everywhere. Acceptability of the Wiener process as a model of multimodal functions is described in more detail elsewhere [2].

Most general assumptions on rationality of information on the function determine only the structure of formulae (2.1) and (2.3) leaving free the choice of the functions $w_i^k(x, x_j), j = \overline{1, k}$ and $\sigma(x, x) - \sigma(x, x_i)$ which express the weight and uncertainty of $f(x)$, respectively, with $f(x_i)$ given. Since these express that the weight and uncertainty of meaning are obvious, these functions are naturally chosen heuristically. Nevertheless, their axiomatic characterization is also of theoretical interest. Thus axioms A2.6 and A2.7 locate a class of statistical models which are associated with random Gaussian functions. In some cases (if, for instance, the statistical model is intended for investigation of multi-step decision making procedures) A2.6 is quite natural. As for A2.7, it is the least convincing intuitively. The value of axioms A2.6 and A2.7 is that they link statistical models which follow from rationality principles and random Gaussian functions.

4. Designing an optimization algorithm

Assume that a function $f(x), x \in A \subset R^n$ is to be minimized where A is a finite rectangular region. The choice of a subsequent computation point by the algorithm can, be virtue of the statistial model $f(\cdot)$, be interpreted as

choice between the probability distribution densities $p_x(\cdot)$ (or, using a more expressive terminology of [3, 4, 15], as choice between lotteries defined by $p_x(\cdot)$). In these References have been shown that from rather general assumptions on rationality of choice follows existence of a utility function (profit) $u(\cdot)$ such that choice of the density $p(\cdot)$ is preferable over choice of $g(\cdot)$ if and only if $\int_{-\infty}^{\infty} u(t)\, p(t)\, dt > \int_{-\infty}^{\infty} u(t)\, g(t)\, dt$. Let us consider design of a utility function $u(\cdot)$ which is consistent with ideas on rational search for a minimum. Assume that computations of the function to be minimized have been completed, $y_i = f(x_i)$, $i = \overline{1, k}$, and a $k+1$-st is planned. Since ξ_k is a family of random Gaussian quantities, only those preferences between Gaussian densities are considered which are understood as preferences between vectors of their parameters (m, s), including the case of $s = 0$.

A4. 1) for any $m_1 < m_2$, $s_1 > 0$ there is an s such that $(m_1, s_1) > (m_2, s_2)$ with $s_2 \leq s$;

A4. 2) for any $m_1, s_1 > 0$, $m_2 \geq y_{0k}$, where $y_{0k} = \min y_i$, $1 \leq i \leq k$ $(m_1, s_1) > (m_2, 0)$;

A4. 3) $(m_1, s_1) > (m_2, s_2)$ if and only if $(m_1, ks_1) > (m_2, ks_2)$;

A4. 4) $u(\cdot)$ is a nonnegative stepwise function of the form $a \sum_{i=1}^{M} I(t_i - t)\, a_i$, $I(t) = 0$, with $t < 0$, $I(t) = 1$ with $t \geq 0$, $a > 0$, $t_i \geq t_{i+1}$, $|\alpha_i| = 1$.

Axiom A4.1 asserts that the choice of a point where a comparatively large value fo the function is expected cannot be reasonable unless the uncertainty is rather high. By virtue of A4.2, the choice of a point where the value of the function to be minimized is at least y_{0k} is not reasonable. A4.3 assumes invariance of the preference with respect to the choice of the uncertainty measuring scale.

The requirements on the utility function as postulated by Axiom A4.4 are rather strong. However, they are not unduly restrictive as far as applications are concerned because any nonnegative function $u(t)$ integrable with a Gaussian density can be approximated as a stepwise function $u^*(t) = \varepsilon \sum_{i=1}^{M} I(t_i - t)\, \alpha_i$ with $\int_{-\infty}^{\infty} |u(t) - u^*(t)|\, n(t)\, m, s)\, dt < \varepsilon$ for any $\varepsilon > 0$, $m^- \leq m \leq m^+$, $0 \leq s \leq s^+$, where $n(\cdot)\, m, s)$ is a Gaussian density with mean m and variance s; m^-, m, m^+ are fixed and $M = M(\varepsilon)$.

Theorem 3.1. The only function which satisfies A4.1–A4.4 is (with an accuracy of a constant factor)

$$u(t) = I(z_{0k} - t) \text{ where } z_{0k} < y_{0k}. \tag{3.1}$$

The proof is given in the Appendix and from the assertion it follows that the utility of choosing x as the point of the $k+1$-st computation is proportional with the probability

$$P(\xi_k < z_{0k}) = \Pi\{(z_{0k} - m_k(x, (x_i, y_i), i = \overline{1,k}))/(s_k(x(x_i, y_i), i = \overline{1,k}))^{1/2}\}$$

where $\Pi(\cdot)$ is a function of normal distribution. Consequently, the minimization algorithm which is consistent with the above assumptions performs the $k+1$-st computation of $f(\cdot)$ in the maximum point of that probability which ($\Pi(\cdot)$ being monotonously increasing) coincides with the point of maximum of the function $v_{k+1}(x) = z_{0k} - m_k(x, (x_i, y_i)), i = \overline{1,k})/(s_k(x, (x_i, y_i)\ i = \overline{1,k})^{1/2}$.

The algorithm of a one-dimensional multimodal minimization which is consistent with the above results (or which plans computing $f(\cdot)$ in points where the probability of exceeding $f(\cdot)$ a certain level by the Wiener process chosen as a model is at its maximal) was proposed from heuristic reasoning [17]. Elsewhere [18, 19] another utility function was used ($\tilde{u}(t) = 0$ with $t \geq y_{0k}$, $\tilde{u}(t) = y_{0k} - t$ with $t < y_{0k}$) which is associated with the assumption of Bayesian optimality of the algorithm in a single step [20, 21]; $\tilde{u}(t)$ satisfies A4.1 and A4.2 but not A4.3. Investigation of a one-dimensional algorithm obtained with the Wiener process as a statistical model and with a utility function $u(\cdot)$ have revealed that with a proper (though theoretically unsubstantiated) choice of the model parameters the algorithm is very effective [18]. Bayesian algorithms (optimal in terms of mean error of the random function used as the statistical model) have been studied in [20, 21]. Statistical models of multimodal functions are also used in designing informational minimization algorithms [22].

Since in designing a statistical model, the choice of weights (2.2) and the expression (2.4) is rather arbitrary, which is also the case of choosing the level of y_{0k} in designing the utility function, this arbitrariness influences the choice of the maximum point for the function $v_{k+1}(\cdot)$. Therefore on implementing the algorithm it is sufficient to maximize the function $v_{k+1}(x)$ very crudely, expecting that computation of $f(\cdot)$ in points of low utility will be eliminated. Consequently, the algorithm is useful only for the approximate estimate of main local minima which can be subsequently computed very accurately with the aid of conventional local methods. Zones of local minima of $f(\cdot)$ can be estimated by assuming that under certain conditions the vicinity of the local minima $m_k(\cdot, (x_i, y_i), i = \overline{1,k})$ contains also the local minimum of $f(\cdot)$ [24, 25]. This

is confirmed by the fact that the algorithm chooses the points for computing $f(\cdot)$ in the vicinities of small values of $f(\cdot)$ with large density and so in these vicinities $m_k(\cdot, (x_i, y_i), i = \overline{1, k})$ is a sufficiently good approximation of $f(\cdot)$ (in the case of weights (2.2) this is spline-approximation [16], and in the one-dimensional case (2.7) this is piecewise-linear approximation).

A detailed description of the one-dimensional algorithm, its earlier version, testing results, and FORTRAN programs are given in [23, 24, 25, 26], respectively. The multi-dimensional algorithm has been described in [27]. The testing results [23, 26, 27] have shown rather high effectiveness of the algorithms in terms of computing the functions to be minimized. The same approach can lead to algorithms of minimization in the presence of noise; a one-dimensional algorithm is described in [23] and its ALGOL program in [28].

5. Conclusions

Search for a global minimum of a multimodal function is regarded as a decision making problem under uncertainty. Assumptions on uncertain behaviour of the function to be minimized and of the rationality of choice of a point for subsequent computation are formulated axiomatically. These axioms are shown to lead to a statistical model of the function to be minimized and subsequently to an optimization algorithm.

6. Appendix

Proof of Theorem 3.1. From A4.1 it follows that $u(\cdot)$ is a non-increasing function. Assume that the opposite is true; let for $m_1 < m_2$ $u(m_2) - u(m_1) : \varepsilon > 0$. Since $u(\cdot)$ is piecewise-continuous, an s_0 can be chosen such that $\left| \int_{-\infty}^{\infty} u(t)\, n(t)\, m_i, s_i)\, dt - u(m_i) \right| < \frac{\varepsilon}{2}$ for all $s_i < s_0$, $i = 1, 2$. But then $\int_{-\infty}^{\infty} u(t)\, n(t)\, m_2, s_2)\, dt > \int_{-\infty}^{\infty} u(t)\, n(t)\, m_1 s_1\, dt$, is in conflict with A4.1.

From A4.2 it follows that $u(t) = 0$ with $t \geq y_{0k}$. Assume the opposite; let (since $u(\cdot)$ is non-increasing) for $y_{0k} \leq m_2 < m_1$ and $u(m_2) - u(m_1) = \varepsilon > 0$. But then with the choice of sufficiently small s_1 such that $\left| \int_{-\infty}^{\infty} u(t)\, n(t)\, m_1, s_1)\, dt - u(m_1) \right| < \varepsilon$ we have $u(m_2) > \int_{-\infty}^{\infty} u(t)\, n(t)\, m_1 s_1)\, dt$, which is in conflict with A4.2. Without limiting the generality, one can assume that $c = 0$ and $a = \frac{1}{M}$.

Consequently, in the expression for $u(t)$, $\alpha_i = 1$, $i = \overline{1, M}$, and $\max t_i < y_{0k}$ If with ε as small as desired $t_i = t_1$, $i = \overline{2, M}$, then $u(t) = I(t_1 - t)$ and it is easy to show that A4.3 holds. This function is the only one which satisfies the requirement A4.3 because in the case of $t_1 - t_M > 0$ (assuming that the points t_i are ordered $t_M \leq t_{m-1} \leq \ldots \leq t_1$) one can choose m_1, s_1, m_2 and k so that

$$\sum_{i=1}^{M} \Pi\left(\frac{t_i - m_1}{s_1}\right) > \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{s_2}\right), \quad \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_1}{k s_1}\right) < \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{k s_2}\right).$$

Let T_{-1}, T_0, and T_+ be determined from the equations $\Pi(T_-) = \varepsilon/M$, $\Pi(T_0) = 0.25$, $\Pi(T_+) = 0.5 - \varepsilon/M$ and

$$k = \frac{t_M - t_1 + T_-}{T_+}, \quad m_2 = t_1 - T_-, \quad s_2 = 1, \quad s_1 = \frac{t_M - t_1}{T_0(k+1)}, \quad m_1 = t_1 + \frac{t_M - t_1}{k+1}.$$

Then with $\varepsilon < 0.25$

$$\sum_{i=1}^{M} \Pi\left(\frac{t_i - m_i}{s_1}\right) > 0.25 > \varepsilon > \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{s_2}\right),$$

$$\sum_{i=1}^{M} \Pi\left(\frac{t_i - m_1}{ks_1}\right) < 0.25 + \frac{M-1}{2} < M\left(\frac{1}{2} - \frac{\varepsilon}{M}\right) < \sum_{i=1}^{M} \Pi\left(\frac{t_i - m_2}{ks_2}\right).$$

Q.E.D.

References

1. Chislennye metody uslovnoj optimizatsii (eds. F. Gill and U. Murrey), Mir, 1977.
2. *Žilinskas, A.*, On Statistical Models for Multimodal Optimization. Math. Operat. Stat., ser. Statistics, vol. **9**, No. *2*, 255–266 (1978).
3. *Savage, L. J.*, Foundations of Statistics, Wiley, New York (1954).
4. *De Groot, M. H.*, Optimal Statistical Decisions. McGraw-Hill, New York (1970).
5. *Fine, T. L.*, Theories of Probability, Academic Press, New York (1973).
6. *Gottinger, H. W.*, Konstruktion der subjektiven Wahrscheinlichkeiten. Math. Operat. Stat., vol. **5**, No. *6*, 509–539 (1974).
7. *Mac Crimmon*, Descriptive and Normative Implications of the Decision Theory Postulates. In Risk and Uncertainty (eds. K. Borch and J. Mossin), st. Martin, New York (1968).
8. *Žilinskas, A., Katkauskajte, A.*, Design of Statistical Models of Complex Functions Which Include Elements of Uncertainty. Sixth All-Union Conference on Encoding and Data Transmission Theory. Proceedings. Part I. Moscow—Vilnius (1978) (in Russian).
9. *Žilinskas, A. G.*, An Axiomatic Approach to Extrapolation under Uncertainty. Avtomatika i telemekhanika (translated as Automation and Remote Control), Vol. **40**, No. *12*, pp. 66–70.
10. *Fine, T.*, Extrapolation when Very Little is Known about the Source. Inform. Contr., vol. **16**, 331–359 (1970).
11. *Goldman, J.*, An Aproach to Estimation and Extrapolation with Possible Applications in an Incompletely Specified Environment. Inform. Contr., vol. **30**, 203–223 (1976).
12. *Žilinskas, A. G.*, Investigating Problems of Multidimensional Extrapolation under Uncertainty. Theory of Optimal Solutions Issue 4, Vilnius, pp. 27–53 (1978) (in Russian).
16. *Shepard, D.*, A Two-Dimensional Interpolation Function for Irregulary-Spaced Data. Proc. 23rd Nat. Conf. ACM, 517–524 (1965).
14. *Žilinskas, A. G.*, On Axiomatic Characterization of Statistical Models of Multimodal Functions. Application of Random Search. Kemerovo (in print) (in Russian).
15. *Tishbern, P.*, Teoria poleznosty dlia priniatia reshenii. Nauka, 1978.
16. *Gordon, W., Wixom, A.*, Shepard's Method of "Metric Interpolation" to Bivariate and Multivariate Interpolation. Mathematics of Computation, vol. **42**, No. *141*, 253–264 (1978).
17. *Kushner, H.*, A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. Trans. ASME, ser. D. J. Basic Eng. vol. **86**, No. *1*, 97–105 (1964).
18. *Žilinskas, A. G.*, One-Step Bayesian Method of Search for the Extremum of Functions of One Variable. Kibernetika, No. *1*, pp. 139–144 (1975) (in Russian).
19. *Šaltianis, V. R.*, On One Method of Multimodal Optimization. Avtomatika i vychislitel'naya tekhnika. No. *3*, 33–38 (1971).
20. *Mockus, I. B.*, On Bayesian Methods of Search for the Extremum. Avtomatika i vychislitel'naya tekhnika, No. *3*, 53–62 (1972).

21. *Mockus, J.*, On Baysian Methods of Seeking the Extremum and their Applications. In Information Processing *77*, ed. B. Gilchist, North-Holland (1977).
22. *Strongin, R. G.*, Numerical Methods of Multimodal Minimization. Nauka, Moscow (1978) (in Russian).
23. *Žilinskas, A.*, Two Algorithms for One-Dimensional Multimodal Minimization. Math. Operat. Stat., ser. Optimization (in print).
24. *Žilinskas, A. G.*, A Method of One-Dimensional Multimodal Minimization. Izv. AN SSSR, Tekhnicheskaya kibernetika, No. *4*, pp. 71–74 (1976).
25. *Žilinskas, A.*, Optimization of One-Dimensional Multimodal Functions, Algorithms AS 133. Applied Statistics, vol. **27**, 367–375.
26. *Zilinskas, A.*, On One-Dimensional Multimodal Minimization. Trans. of Eighth Prague Conf. on Inform. Theory, Stat. Dec. Func., Rand. Proc., vol. **B**, 393–402 (1978).
27. *Žilinskas, A.*, On the use of statistical models of multimodal functions for the construction of the optimization algorithms. In Proceedings of 9th IFIP Conference on Optimization Techniques (in print).
28. *Žilinskas, A.*, MIMUN-optimization algorithm of one-dimensional multimodal functions in the presence of noise. Aplikace Matematiky (in print).

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

L. Györfi

Technical University of Budapest

H–1111 Budapest XI., Stoczek u. 2, 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 (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 th paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) 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 статей в четырех экземплярах на русском языке или в двух на английском и в двух на русском языке следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50–100 слов и приложено резюме – реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в двух экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Думер И. И., Зиновьев В. А., Зяблов В. В.: Каскадное декодирование по минимуму обобщенного расстояния (*Dumer, I. I., Zinov'ev, V. A., Zyablov, V. V.*: Cascaded decoding according to minimal generalized distance) 3

Galló, V.: An experimental recognition system using pattern grammars (*Галло, В.*: Система для экспериментов по распознаванию, использующая грамматику образов) 21

Kurzhanskii, A. B.: Dynamic control system estimation under uncertainty conditions, II (*Куржанский А. Б.*: Об оценивании динамики управляемых систем в условиях неопределенности, II) 33

Györfi, L.: Recent results on nonparametric regression estimate and multiple classification (*Дёрфи Л.*: Последные результаты о непараметрических регрессии и классификации) 43

Жилинскас А. Г.: О статистических моделях сложных многоэкстремальных функций и их применении для построения алгоритмов оптимизации (*Žilinskas, A. G.*: On statistical models of multimodal functions and their application for the construction of the minimization algorithms) 53

HU ISSN 0370—2529

Index: 26 660

316.920

VOL. 10 * NUMBER 2
TOM НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1981

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

Is an international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Roumania, Viet-Nam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H–1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1981 Subscription Rate US Dollars 110.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)

G. BOGNÁR (Budapest)

COORDINATING EDITORS

S. V. EMELYANOV (Moscow)
N. N. KRASOVSKII (Sverdlovsk)
E. P. POPOV (Moscow)
V. S. PUGACHEV (Moscow
V. I. SIFOROV (Moscow)
E. D. TERYAEV (Moscow)
T. VÁMOS (Budapest)
L. VARGA (Budapest)
A. PRÉKOPA (Budapest)
S. CSIBI (Budapest)
I. CSISZÁR (Budapest)
J. KOCSIS (Budapest)
J. BENEŠ (Prague)
V. STREJC (Prague)

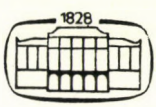
РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)

Г. БОГНАР (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

С. В. ЕМЕЛЬЯНОВ (Москва)
Н. Н. КРАСОВСКИЙ (Свердловск)
Е. П. ПОПОВ (Москва)
В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Е. Д. ТЕРЯЕВ (Москва)
Т. ВАМОШ (Будапешт)
Л. ВАРГА (Будапешт)
А. ПРЕКОПА (Будапешт)
Ш. ЧИБИ (Будапешт)
И. ЧИСАР (Будапешт)
Я. КОЧИШ (Будапешт)
И. БЕНЕШ (Прага)
В. СТРЕЙЦ (Прага)



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 10 (2), pp. 67—81 (1981)

AN AUGMENTED LAGRANGIAN METHOD FOR A CLASS OF CONVEX CONTINUOUS OPTIMAL CONTROL PROBLEMS

J. V. OUTRATA, Z. SCHINDLER

(Prague)

(Received April 10, 1980)

For the solution of convex continuous optimal control problems with certain type of inequality state-space constraints a numerical method is proposed, based on the generalized quadratic penalty idea. A Sobolev space was chosen as perturbation space so that for the construction of the appropriate augmented Lagrangian, a special projection theorem must be applied. The proposed conceptual algorithm can easily be implemented.

0. Introduction and notations

The application of duality theory for the construction of numerical methods for the solution of infinite-dimensional extremal problems and in particular for continuous optimal control problems is by far not so numerous as for the finite-dimensional problems of mathematical programming. One of the probably most important contributions is paper [10] where almost all the important results concerning finite-dimensional problems were extended to the case in which the constraint space is Hilbert. In [5] and [6] the results of [10] were applied to so-called "abstract convex optimal control problem" and, particularly, to a class of optimization problems with systems described by linear ordinary differential equations. In the present paper we treat, similarly as in [5] and [6], the state-space constraints by means of an appropriate generalized penalty term, but we choose a different perturbation space, namely the Sobolev space W_2^1. This is well justified e.g. in the following case: We want to control a nonlinear plant by the convex feedback method described in [2]. Hence, we have to solve many times convex subproblems where the dynamics and possibly also control and state-space constraints are linearized. In this case, the range space of state-space constraints can really be chosen as W_2^1.

The paper presumes a certain knowledge of basic definitions and propositions of convex analysis. Therefore, we refer to [1] and [9] and we shall extensively exploit the relevant results.

1. Preliminaries

Let U be a reflexive Banach space (control space), Y be a linear normed space (state-space), and H be a Hilbert space (constraint space). The extremal problem

$$I(u, y) \to \inf$$

subj. to

$$\begin{aligned} &y = \Pi u + y_0, \\ &y \in \Omega \subset U, \\ &-q(y) \in D \subset H, \end{aligned} \tag{1.1}$$

where

(i) $I[U \times Y \to \bar{R}]$ is a proper convex l.s.c. functional with $\Omega \times \{y \in Y | -q(y) \in D\} \subset \operatorname{dom} I$,
(ii) $\Pi \in \mathcal{L}[U, Y]$, $y_0 \in Y$,
(iii) Ω is a closed convex subset of U (set of admissible controls),
(iv) D is a closed convex cone in H with vertex at the origin,
(v) $q[Y \to H]$ is a D-convex continuous mapping,

will be termed *abstract convex optimal control problem* with state-space inequality constraints (ACOCP).

We involve now the functional $f[U \to \bar{R}]$

$$f(u) = I(u, \Pi u + y_0), \tag{1.2}$$

and the set

$$G = \Omega \cap \{u | -q(\Pi u + y_0) \in D\}, \tag{1.3}$$

and impose the following assumptions on (1.1):

AS1: G is nonempty.
AS2: Either Ω is bounded or f is coercive on Ω.

In [5] and [6] it is shown that using the perturbed essential objective

$$\begin{aligned} F_r(u, p) = I(u, \Pi u + y_0) + \delta(u | \Omega) + \delta(\Pi u + y_0 | \{y | p - q(y) \in D\}) + \\ + r||p||_H^2, \quad r \in R^+, \end{aligned} \tag{1.4}$$

we obtain a family of dual problems to (1.1)

$$\gamma_r(p^*) \to \sup$$

subj. to (1.5)

$$p^* \in H,$$

parametrized by the nonnegative scalar r. In (1.5) the dual objective

$$\gamma_r(p^*) = \inf_{u \in W} L_r(u, p^*), \tag{1.6}$$

where the functional L_r defined on $U \times H$ is called the *augmented Lagrangian*, and

$$L_r(u, p^*) = I(u, \Pi u + y_0) - \left\langle p^*, \left(\frac{p^*}{2r} - q\right)^D + q \right\rangle_H +$$

$$+ r \left\| \left(\frac{p^*}{2r} - q\right)^D + q \right\|_H^2 \text{ if } u \in \Omega,$$

$$L_r(u, p^*) = +\infty \quad \text{if} \quad u \notin \Omega. \tag{1.7}$$

(Arguments at q were dropped due to notational simplicity.)

The assumptions imposed imply that

(i) there exists at least one solution of (1.1), and the corresponding value of the objective is finite,

(ii) problem (1.1) is with respect to perturbations (1.4) normal, which implies in particular that

$$\sup_{p^* \in H} \gamma_r(p^*) = \min_{\substack{u \in \Omega \\ -q(\Pi u + y_0) \in D}} I(u, \Pi u + y_0). \tag{1.8}$$

For proofs of these assertions see [1] and [6]. In [5] and [6] a numerical method is proposed for solving (1.5) and simultaneously (1.1) based on the well-known primal-dual method of Rockafellar for finite-dimensional convex programming problems (cf. [8]). This method reads as follows:

Given $r > 0$, $p_0^* \in H$ and a sequence $\{\alpha_k\}$ with $0 \leq \alpha_k \to 0$. k-th step:

1) Given $p_k^* \in H$, determine $u_k \in \Omega$ such that

$$L_r(u_k, p_k^*) \leq \inf_{u \in \Omega} L_r(u, p_k^*) + \alpha_k. \tag{1.9}$$

2) Set

$$p_{k+1}^* = p_k^* - 2r\left[q(\Pi u_k + y_0) + \left(\frac{p_k^*}{2r} - q(\Pi u_k + y_0)\right)^D\right]. \tag{1.10}$$

Concerning the convergence properties of this numerical scheme, the following proposition is proved in [6].

Theorem 1.1. Let the sequence of multipliers $\{p_k^*\}$ generated by (1.9), (1.10) be bounded. Then the corresponding sequence of controls $\{u_k\}$ is asymptotically minimizing for (1.1). In particular, $\{u_k\}$ possesses weak cluster points and every weak cluster point is an actual solution of (1.1).

In this sequel the numerical scheme (1.9)–(1.10) will be applied to the solution of an optimization problem with a control system described by linear ordinary differential equations and a convex objective, control and state-space inequality constraints. This was originally recommended in [10], and in [5] and [6] this application was investigated in detail. However, only for the case $H = L_2[0, T, R^m] \times R^k$. Here, we restrict ourselves to those cases where the state-space constraints map the trajectories into $W_2^1[0, T, R^m] \times R^k$. The different form of the inner product in W_2^1 reflects in a different form of the appropriate augmented Lagrangian so that the intermediate minimization, in (1.9) as well as the multiplier iteration (1.10) attain in this case a different structure. This will be investigated in the following sections.

2. Main results

Henceforth we shall be concerned with the following basic problem (BP):

$$\varphi(x(T)) + \int_0^T \psi(u(t))\, dt \to \inf$$

subj. to

$$\begin{aligned} &\dot{x}(t) = Ax(t) + Bu(t) \text{ a.e. on } [0, T], \\ &x(0) = x_0, \\ &u(t) \in \Omega \subset L_2[0, T, R^l], \\ &Fx(t) + d_0 \leq \theta \text{ on } [0, T], \\ &a(x(T)) \leq \theta, \end{aligned} \tag{2.1}$$

where A, B, F are constant $[n \times n]$, $[n \times l]$, $[m \times n]$ matrices, respectively, $u \in L_2[0, T, R^l]$, $x \in W_2^1[0, T, R^n]$, $\varphi\,[R^n \to \bar{R}]$, $\psi\,[R^l \to \bar{R}]$, $a\,[R^n \to R^k]$, $d_0 \in R^m$, T is the fixed terminal time, and the inequality sign is valid for all coordinates.

We denote

$$D_0 = \{v \in W_2^1[0, T, R^1] \,|\, v(t) \geq 0 \,\forall t \in [0, T]\}, \tag{2.2}$$

and impose the following assumptions on (1.1):

(i) φ is a proper convex l.s.c. function with

$$\operatorname{dom} \varphi \supset \{z \in R^n \,|\, a(z) \leq \theta\},$$

(ii) ψ is a proper convex l.s.c. function with

$$\operatorname{dom}\psi \supset \{z \in R^l \mid z \in \bigcup_{\substack{u \in \Omega \\ t \in [0,T]}} u(t)\},$$

(iii) Ω is convex and closed,

(iv) the functions a^i, $i = 1, 2, \ldots, k$ are convex and continuous,

(v) there exists a control $u_0 \in w$ such that the absolutely continuous trajectory $\bar{x}$ corresponding to u_0 satisfies the inequalities

$$F\bar{x}(t) + d_0 < \theta, \quad t \in [0, T],$$

$$a^i(\bar{x}(T)) < 0, \quad i = 1, 2, \ldots, k,$$

(vi) either Ω is bounded or the objective in (2.1) is on Ω coercive,

(vii) the system matrix A is nonsingular.

Remark. The first part of requirement (ii) ensures that the integral in the objective of (2.1) is a proper convex l.s.c. functional of u due to the results of [7]. The Slater condition (v) implies that the sequence $\{p_k^*\}$ generated by (1.9)–(1.10) is bounded.

If we now set $U = L_2[0, T, R^l]$, $Y = W_2^1[0, T, R^n]$, $H = W_2^1[0, T, R^m] \times \times R^k$, $y = x(t)$, $u = u(t)$,

$$q(y) = \big(Fx(t) + d_0,\ a(x(T))\big),$$

$$\Pi u = \int_0^t e^{A(t-\tau)} Bu(\tau)\, d\tau,\ y_0 = e^{At} x_0, \tag{2.3}$$

$$D = \{(w, z) \in H \mid w^i \in D_0,\ i = 1, 2, \ldots, m,\ z^j \geq 0,\ j = 1, 2, \ldots, k\},$$

and I to be equal to the objective of (2.1), our BP has immediately the form of the abstract convex optimal control problem (1.1) specified in the previous section. Moreover, it satisfies all assumptions stated there.

Our constraint space H is clearly Hilbert with the inner product

$$\langle (v_1, \mu_1), (v_2, \mu_2) \rangle_H = \langle v_1(0), v_2(0) \rangle_{R^m} + \int_0^T \sum_{i=1}^m \dot{v}_1^i(t)\, \dot{v}_2^i(t)\, dt + \langle \mu_1, \mu_2 \rangle_{R^k}. \tag{2.4}$$

Therefore, the multiplier p^* will be henceforth represented by a couple $(v, \mu) \in H$, $v \in W_2^1[0, T, R^m]$, $\mu \in R^k$. In order to be able to compute the projections in (1.7) in this case, we shall first investigate the projection of the functions $p(t) \in W_2^1[0, T, R^1]$ onto the positive cone (2.2). As it is shown in [10], every

element $p \in W_2^1[0, T, R^1]$ can be uniquely decomposed into two orthogonal components

$$p = p^{D_0} + p^{D_0^*} \tag{2.5}$$

where D_0^* is a negative dual cone to D_0, i.e.

$$D_0^* = \{p \in W_2^1[_0 0, T, R^1] \,|\, p \text{ is nondecreasing, } p(0) \leq \dot{p}(t) \leq 0 \text{ a.e.}\} \tag{2.6}$$

cf. [10].

Theorem 2.1. Let $p \in W_2^1[0, T, R^1]$. Then $p^{D_0^*}$ is the restriction of co $\hat{p}$ on the interval $[0, T]$ where $\hat{p}[[-1, T] \to R^1]$ is given by

$$\hat{p}(t) = \begin{cases} 0 & \text{for } t \in [-1, 0), \\ p(t) & \text{for } t \in [0, T), \\ \min\{0, \min\limits_{t \in [0,T]} p(t)\} & \text{for } t = T. \end{cases} \tag{2.7}$$

Proof. As p is absolutely continuous, its minimum on $[0, T]$ is achieved; if it is nonnegative, the case is trivial because then clearly $p \in D_0$. Hence, let us investigate the case $\min\limits_{t \in [0,T]} p(t) < 0$. First we note that all elements of D_0^* are convex functions. Indeed, for a nonconvex absolutely continuous function p, it can easily be shown that its derivative $\dot{p}$ cannot be nondecreasing a.e. Next we utilize the following equivalence of [10]. For any $\overline{p} \in D_0^*$ the equality $\overline{p} = p^{D_0^*}$ holds iff $p - \overline{p} \in D_0$, and $\langle \overline{p}, p - \overline{p} \rangle_{W_2^1} = 0$.

Henceforth we denote the above-mentioned restriction of co $\hat{p}$ to the interval $[0, T]$ as b. Because $b \in D_0^*$ due to (2.7), and $p - b \in D_0$, it remains to prove that $\langle b, p - b \rangle_{W_2^1} = 0$, i.e. that

$$b(0)\,(p(0) - b(0)) + \int_0^T \dot{b}(\dot{p} - \dot{b})\, dt = 0. \tag{2.8}$$

Because $p - b$ is continuous, it attains its minimum on $[0, T]$ and this minimum equals zero. Let

$$\alpha = \min \sigma, \quad \beta = \max \sigma$$

where

$$\sigma = \arg \inf_{t \in [0,T]} (p - b),$$

because the minima (maxima) of σ really exist. The values α, β divide the interval $[0, T]$ into three (possibly degenerate) subintervals. Then (2.8) has the form

$$b(0)\,(p(0) - b(0)+) \int_0^{\alpha} \dot{b}(\dot{p} - \dot{b})\,dt + \int_{\alpha}^{\beta} \dot{b}(\dot{p} - \dot{b})\,dt + \int_{\beta}^{T} \dot{b}(\dot{p} - \dot{b})\,dt. \quad (2.9)$$

Clearly, the convex hull b is on $[0, \alpha]$ an affine function. Hence, $\dot{b}(t)$ is on $[0, \alpha)$ constant and due to (2.7) equal to $b(0)$. That is why

$$b(0)\,(p(0) - b(0)) + \int_0^{\alpha} \dot{b}(\dot{p} - \dot{b})\,dt = b(0)\,(p(0) - b(0)) + b(0) \int_0^{\alpha} (\dot{p} - \dot{b})\,dt =$$

$$= b(0)\,(p(0) - b(0)) + b(0)\,(p(\alpha) - b(\alpha) - p(0) + b(0)) = 0.$$

As b is on $[\beta, T]$ constant, the third integral equals zero as well. The interval $[\alpha, \beta]$ can be decomposed into two subsets $\mathcal{S}$ and $\mathcal{Q}$. $\mathcal{S}$ is the set where $p - b \equiv 0$ which is closed due to the continuity of p, b. Therefore, $\dot{p} - \dot{b} = 0$ a.e. on $\mathcal{S}$. $\mathcal{Q}$ is the open set $[\alpha, \beta] \setminus \mathcal{S}$ which is the countable union of open subintervals $\mathcal{Q}_i$. Evidently, on each such subinterval $\dot{b}$ is constant. Hence,

$$\int_{\alpha}^{\beta} \dot{b}(\dot{p} - \dot{b})\,dt = \int_{\mathcal{S}} \dot{b}(\dot{p} - \dot{b})\,dt + \int_{\mathcal{Q}} \dot{b}(\dot{p} - \dot{b})\,dt = \sum_i \dot{b}_i \int_{\mathcal{Q}_i} (\dot{p} - \dot{b})\,dt.$$

As $p - b$ at boundary points of $cl\,\mathcal{Q}_i$ equals zero, the whole above sum equals zero as well. This completes the proof.

Thus, for a particular multiplier $p^* = (\nu, \mu)$ and a particular trajectory x

$$\left(\frac{p^*}{2r} - q\right)^D = (c, d) \in H \quad (2.9)$$

where

$$c^i(t) = \left(\frac{\nu^i(t)}{2r} - (Fx(t))^i - d_0^i\right)^{D_0}, \qquad i = 1, 2, \ldots, m, \quad (2.10)$$

$$d^i = \max\left\{a^i(x(T)), \frac{\mu^i}{2r}\right\} - a^i(x(T)), \; i = 1, 2, \ldots, k, \quad (2.11)$$

and we are able to compute the functions $c^i(t)$ according to the above proposition.

$$L_r(u, p^*) = L_r'(u(t), x(t), \nu(t), \mu) = \varphi(x(T)) + \int_0^T \psi(u(t))\,dt -$$

$$- \sum_{i=1}^{k} \mu^i \max\left\{a^i(x(T)), \frac{\mu^i}{2r}\right\} + r \sum_{i=1}^{k} \max^2\left\{a^i(x(T)), \frac{\mu^i}{2r}\right\} -$$

$$- \langle \nu(0), c(0) + Fx(0) + d_0 \rangle_{R^m} - \int_0^T \langle \dot{\nu}, \dot{c} + F\dot{x} \rangle_{R^m} dt + r\,||c(0) + Fx(0) + d_0||_{R^m}^2 +$$

$$+ r \int_0^T ||\dot{c} + F\dot{x}||_{R^m}^2\,dt, \quad (2.12)$$

where the trajectory $x(t)$ corresponds to the control $u(t)$ according to the system dynamics equation.

If we compute $\gamma_r(p^*) = \gamma_r(v(t), \mu)$ with the specified accuracy in the intermediate problem (1.9), we have to solve the optimal control problem

$$\begin{aligned} \Lambda(x, u) = \varphi(x(T)) - \sum_{i=1}^{k} \mu^i \max\left\{a^i(x(T)), \frac{\mu^i}{2r}\right\} + \\ + r \sum_{i=1}^{k} \max^2\left\{a^i(x(T)), \frac{\mu^i}{2r}\right\} - \langle v(0), c(0) + Fx(0) + d_0\rangle_{R^m} + \\ + r\,||c(0) + Fx(0) + d_0||^2_{R^m} + \int_0^T [\psi(u(t))\,t - \langle \dot{v}(t), \dot{c}(t) + F\dot{x}(t)\rangle_{R^m} + \\ + r\,||\dot{c}(t) + F\dot{x}(t)||^2_{R^m}]\,dt \to \inf \end{aligned} \tag{2.13}$$

subj. to $\dot{x}(t) = Ax(t) + Bu(t), x(0) = x_0,$

$u \in \Omega,$

(as the value of $-\langle v(0), c(0) + Fx(0) + d_0\rangle_{R^m} + r\,||c(0) + Fx(0) + d_0||^2_{R^m}$ cannot be omitted). The extremal problem (2.13) may seem to be very difficult even if the state-space contraints are taken off. However, Theorem 2.2 and its corollaries below can substantially help us under only slightly more restrictive assumptions. Henceforth, the argument t will be mostly dropped to simplify the notation.

Theorem 2.2. Let the functions φ and a be continuously differentiable. Suppose, moreover, that the trajectory x corresponds to an admissible control u and a function $\varkappa$ is the solution of the costate equation

$$-\dot{\varkappa} = A^T \varkappa - 2rF^T\left(F\dot{x} - \frac{v}{2r} + \dot{c}\right) \tag{2.14}$$

backwards on $[0, T]$ from the terminal condition

$$\varkappa(T) = (A^T)^{-1}\left[-\frac{\partial\varphi(x(T))}{\partial x} - \sum_{i=1}^{k} \max\{-\mu^i + 2ra^i(x(T)), 0\}\,\frac{da^i(x(T))}{dx}\right]. \tag{2.15}$$

Then for $v \in \Omega$

$$\begin{aligned} \Lambda(x(v), v) - \Lambda(x(u), u) = \\ = \int_0^T [\langle B^T\dot{\varkappa}, v - u\rangle_{R^l} + \psi(v) - \psi(u)]\,dt + o(||v - u||_{L_2}). \end{aligned} \tag{2.16}$$

The proof depends heavily on Proposition 1 (Section 9.6 of [4]), which we state here, but without proof. It concerns the extremal problem

$$g(x, u) \to \inf$$

subj. to

$$\mathcal{H}(x, u) = \theta,$$

$$u \in \Omega \subset U,$$

where $x \in X$, X, U are normed linear spaces, Ω is a prescribed subset of U, $g[X \times U \to R^1]$, and $\mathcal{H}[X \times U \to X]$, the equality constraint defines a unique implicit function $x(u)$. Futhermore, we assume that $\mathcal{H}$ and g are Fréchet differentiable with respect to x and that the derivatives $\mathcal{H}_x(x, u)$ and $g_x(x, u)$ are continuous on $X \times U$. Finally, we assume that the implicit function $x(u)$ satisfies the Lipschitz condition of the form

$$||x(u) - x(v)||_X \leq K||u - v||_U.$$

Theorem 2.3. For any $u \in \Omega$ let λ be a solution of the equation

$$\mathcal{H}_x^*(x(u), u)\,\lambda + g_x(x(u), u) = \theta. \tag{2.17}$$

Then for $v \in \Omega$

$$g(x(v), v) - g(x(u), u) = \langle \lambda, \mathcal{H}(x(u), v)\rangle_X + g(x(u), v)$$
$$- \langle \lambda, \mathcal{H}(x(u), u)\rangle_X - g(x(u), u) + o(||v - u||_U).$$

Proof of Theorem 2.2. In our case

$$U = L_2[0, T, R^l], X = W_2^1[0, T, R^n],$$

$$g(x, u) = L_r'(u, x, \nu, \mu),$$

$$\mathcal{H}(x, u) = x_0 + \int_0^t [Ax(\tau) + Bu(\tau)]\,d\tau - x(t).$$

Now we recall that generally

$$L_r(u, p^*) = \Psi_{2r}\left(u, \frac{p^*}{2r}\right) - \frac{1}{4r}\|p^*\|_H^2,$$

where Ψ is the shifted penalty functional of [10]. In the case of ACOCP

$$\Psi_{2r}\left(u, \frac{p^*}{2r}\right) = I(u, y) + r\left\|\left(q(y) - \frac{p^*}{2r}\right)^{+D^*}\right\|_H^2 \tag{2.18}$$

where y corresponds to u with respect to the system dynamics. The penalty part of (2.18) is Fréchet differentiable and its derivative equals

$$2rq_y^*(y)\left(q(y) - \frac{p^*}{2r}\right)^{+D^*},$$

if only $q(y)$ is Fréchet differentiable (cf. [10]). Thus, for problem (2.1) and increments $h \in W_2^1[0, T, R^n]$, $h(0) = \theta$

$$\langle g_x g(x, u), h\rangle_X = \left\langle \frac{\partial \varphi(x(T))}{\partial x}, h(T)\right\rangle_{R^n} + \tag{2.19}$$

$$+ \left\langle \sum_{i=1}^{k} \max\{-\mu^i + 2ra^i(x(T)), 0\} \frac{\partial a(x(T))}{\partial x}, h(T)\right\rangle_{R^n} +$$

$$+ 2r \int_0^T \left\langle F^T \left(F\dot{x} - \frac{\dot{\nu}}{2r} + \dot{c}\right), \dot{h}\right\rangle_{R^n} dt$$

because

$$\left(q(y) - \frac{p^*}{2r}\right)^{+D^*} = q(y) - \frac{p^*}{2r} + \left(\frac{p^*}{2r} - q(y)\right)^D = \left(Fx + d_0 - \frac{\nu}{2r} + c, z\right)$$

with

$$z = \max\left\{a^i(x(T)) - \frac{\mu^i}{2r}, 0\right\}.$$

Clearly,

$$\mathcal{H}_x(x, u)\, h = \int_0^t Ah(\tau)\, d\tau - h(t). \tag{2.20}$$

All assumptions of Theorem 2.3 will be satisfied if $\varkappa$, solving our costate equation (2.14) from the terminal conditions (2.15), corresponds to λ in the general costate equation (2.17) or, more suitably, in the modified equation

$$\langle \lambda, \mathcal{H}_x(x(u), u)\, h\rangle_X + \langle g_x(x(u), u), h\rangle_X = 0 \text{ for all } h \in W_2^1[0, T, R^n],\ h(0) = \theta.$$

Substituting from (2.19), (2.20), we obtain

$$\int_0^T \langle \dot{\lambda}, Ah(t) - \dot{h}(t)\rangle_{R^n} dt + 2r \int_0^T \left\langle F^T \left(F\dot{x} - \frac{\dot{\nu}}{2r} + \dot{c}\right), \dot{h}\right\rangle_{R^n} dt +$$

$$+ \left\langle \frac{\partial \varphi(x(T))}{\partial x}, h(T)\right\rangle_{R^n} + \left\langle \sum_{i=1}^{k} \max\{-\mu^i + 2ra^i(x(T)), 0\} \frac{\partial a^i(x(T))}{\partial x}, h(T)\right\rangle_{R^n} = 0$$

$$\text{for all } h \in W_2^1[0, T, R^n],\ h(0) = \theta$$

which implies after some computation that indeed $\lambda = \varkappa$. Therefore, Theorem 2.3 is really applicable and gives us

$$\Lambda(x(v), v) - \Lambda(x(u), u) = \int_0^T [\langle B^T \dot{\varkappa}, v - u\rangle_{R^l} + \psi(v) - \psi(u)]\, dt + o(\|v - u\|_{L_2}),$$

and we are done. ∎

Corollary 2.2.1. Let there exists a constant $[l \times l]$ matrix Q and a constant scalar M such that

$$\psi(v) \leq \langle v, Qv \rangle_{R^l} + M \quad \forall\, v \in R^l.$$

Then $\Theta(u) = \Lambda(x(u), u)$ is subdifferentiable for all $u \in \Omega$ (in fact for all $u \in L_2[0, T, R^l]$), and

$$B^T \dot{\varkappa} + s \in \partial \Theta(u)$$

whenever $\varkappa$ satisfies (2.14), (2.15), and

$$s(t) \in \partial \psi(u(t)) \text{ a.e. on } [0, T]. \tag{2.21}$$

Proof. Under the above assumptions, Θ is finite over $L_2[0, T, R^l]$, hence it is continuous (cf. [1]) and subdifferentiable. Thus, for any control vector $u \in L_2[0, T, R^l]$ there exists a function $w \in L_2[0, T, R^l]$ such that

$$\Theta(v) - \Theta(u) \geq \int_0^T \langle w, v - u \rangle_{R^l} dt \quad \forall\, v \in L_2(0, T, R^l].$$

Combining (2.16) and (2.21) we obtain that

$$\Theta(v) - \Theta(u) \geq \int_0^T \langle B^T \dot{\varkappa} + s, v - u \rangle_{R^l} dt + o(||v - u||_{L_2}),$$

and convexity of Θ implies that $o(||v - u||_{L_2}) \geq 0$. Thus, $B^T \dot{\varkappa} + s \in \partial \Theta(u)$ which was to be proved. ∎

Corollary 2.2.2. Let the assumptions of Corollary 2.2.1 be satisfied, and let the function ψ be continuously differentiable. Then, the functional Θ is Fréchet differentiable for all $u \in \Omega$ (in fact for all $u \in L_2[0, T, R^l]$), and

$$\Theta_u(u) = B^T \dot{\varkappa} + \frac{\partial \psi(u)}{\partial u}.$$

Proof. It is obvious that $\Theta(u)$ is Gâteau differentiable over $L_2[0, T, R^l]$ with $B^T \dot{\varkappa} + \dfrac{\partial \psi(u)}{\partial u}$ being its Gâteau derivative. To prove that $B^T \dot{\varkappa} + \dfrac{\partial \psi(u)}{\partial u}$ is in fact the Fréchet derivative of $\Theta(u)$, it suffices to show the $(L_2[0, T, R^l], L_2[0, T, R^l])$ continuity of the map

$$u \to B^T \dot{\varkappa} + \frac{\partial \psi(u)}{\partial u}$$

which reduces furthermore to the proof of $(L_2[0, T, R^l), L_2[0, T, R^l])$ continuity of the map $u \to \dot{\varkappa}$ given by

$$\dot{\varkappa}(t) = -A^T\left[e^{A^T(T-t)}(A^T)^{-1}\left(-\frac{\partial\varphi(x(T))}{\partial x} - \sum_{i=1}^{k}\max\{-\mu^i + 2ra^i(x(T)), 0\}\cdot\right.\right.$$

$$\left.\left.\cdot\frac{\partial a^i(x(T))}{\partial x}\right) - 2r\int_t^T e^{A^T(\tau-t)}F^T\left(F\dot{x} - \frac{\dot{\nu}}{2r} + \dot{c}\right)dt\right] + 2rF^T\left(F\dot{x} - \frac{\dot{\nu}}{2r} + \dot{c}\right).$$

The exact proof is not difficult but a little bit lengthy and therefore we omit it here. ∎

Concerning the multiplier iteration (1.10), it attains the form

$$\nu_{s+1} = \nu_s - 2r[Fx_s + d_0 + c_s], \tag{2.22}$$

and

$$\mu^i_{s+1} = \mu^i_s - 2r\max\left\{a^i(x_s(T)), \frac{\mu^i_s}{2r}\right\}, \quad i = 1, 2, \ldots, s \tag{2.23}$$

where $x_s = x(u_s)$ and $c_s = c(x_s)$.

3. Some remarks on numerical properties

If we wish to solve BP numerically we can either discretize it immediately at the beginning and for the resulting finite-dimensional problem apply some mathematical programming technique, or construct some conceptual infinite-dimensional algorithm and find some finite-dimensional implementable version of it. The algorithm described in the previous section is a conceptual one and our task is to find some suitable version of it. Let us first represent the control functions $u^i(t) \in L_2[0, T, R^1]$ by e_1 values $u^i(t_j)$, $0 = t_1 < t_2 < \ldots t_{e_1} = T$, $i = 1, 2, \ldots, l$. The numerical experiments show that in most cases it is sufficient to take controls as either affine on each subinterval $[t_j, t_{j+1}]$, or even constant (cf. [2]). Similarly, let the trajectories $x^i(t) \in W_2^1[0, T, R^1]$ be represented by e_2 values $x^i(t_j)$, $0 = t_1 < t_2 < \ldots t_{e_2} = T$, $i = 1, 2, \ldots, l$. The type of the trajectory approximation by values $x^i(t_j)$ decides how exactly the given state-space constraints will be satisfied and influences, of course, the convergence of the applied method. If the method described in [5] and [6] is used, the state-space constraints affect the objective in the intermediate problems by means of the generalized penalty term

$$K_1(x) = \int_0^T\left[-\sum_{i=1}^{m}\nu^i(t)\max\left\{(Fx(t))^i + d_0^i, \frac{\nu^i(t)}{2r}\right\} + \right.$$

$$\left. + r\sum_{i=1}^{m}\max^2\left\{(Fx(t))^i + d_0^i, \frac{\nu^i(t)}{2r}\right\}\right]dt, \tag{3.1}$$

where $\nu^i(t) \in L_2[0, T, R^1]$ are the corresponding multipliers. Thus, if the last integral is evaluated by means of the rectangular rule, it corresponds in fact to the approximation of trajectories $x^i(t)$ by piecewise constant functions. Hence, it suffices to set the multipliers $\nu^i(t)$ to be also piecewise constant. The projections $\max\left\{(Fx(t))^i + d_0^i, \frac{\nu^i(t)}{2r}\right\}$ remain in such case constant on each subinterval $[t_j, t_{j+1})$. If, however, the integral (3.1) is evaluated by means of the trapezoidal rule, it corresponds to the approximation of the trajectories $x^i(t)$ and the corresponding multipliers $\nu^i(t)$ by piecewise affine functions. But then the projection $\max\left\{(Fx(t))^i + d_0^i, \frac{\nu^i(t)}{2r}\right\}$ need not be necessarily affine on each subinterval $[t_j, t_{j+1}]$ so that the quadrature in (3.1) is to be somehow adapted. If still more advanced quadrature formulae are applied, the situation becomes still more complicated.

The generalized penalty term proposed here, i.e.

$$K_2(x) = \int_0^T [-\langle \dot{\nu}(t), \dot{c}(t) + F\dot{x}(t)\rangle_{R^m} + r\,||\dot{c}(t) + F\dot{x}(t)||^2_{R^m}]\,dt \qquad (3.2)$$

possesses an important property with respect to its numerical quadrature. Namely, for piecewise affine $x^i(t)$ and $\nu^i(t)$ the projections $c^i(t)$ are affine on each subinterval of the appropriate division. Hence, the rectangular quadrature formula suffices completely to the evaluation of (3.2). Besides, we believe that the application of the penalty term (3.2) instead of (3.1) for problems (2.1) will result in the faster convergence of the dual iterational scheme similarly as it was found in [3] for problems with equality constraints. Still one aspect seems to be worth mentioning. The increase of the penalty parameter r influences strongly the behaviour of the costate equation (2.14) so that the integration becomes more difficult. The errors of this integration then cause the difficulties in the minimization technique applied to (2.13), because the gradients or subgradients are not sufficiently precise.

In the illustrative example below we have used a variant of the Polak–Ribière conjugate gradient method for the minimization in (1.9). Definite integrals were evaluated using the trapezoidal and rectangular rule and the integrations were performed by the 4-th order Runge–Kutta method.

Numerical example. Suppose that we have to solve the extremal problem (2.1) with the minimum energy objective $\frac{1}{2}\,||u||^2_{L_2}$,

$$A = \begin{bmatrix} 0 & 1 \\ 0.1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad T = 4$$

$$\Omega = \{u \mid -1 \leq u(t) \leq 1 \text{ a.e. on } [0, 4]\}, \; x^2(t) \geq -0.7,$$

$$x(T) \leq \theta.$$

We have chosen $r = 5$, $v_0(t) = 0$ on $[0, T]$, $\mu_0 = \theta$. The iteration process was stopped after the 10-th dual iteration where the augmented constraints were satisfied with the maximum error 10^{-5}. The objective reached 0.9056 and $\mu_{10} = \begin{bmatrix} -0.569 \\ 0.0 \end{bmatrix}$. The trajectories and the control are plotted in Fig. 1.

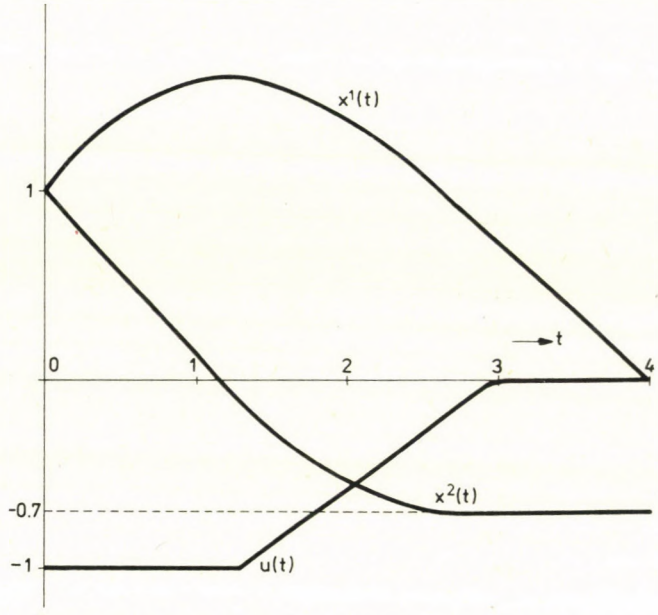


Fig. 1. Resulting control $u(t)$ and trajectories $x^1(t)$, $x^2(t)$

List of symbols

$\bar{R}$	extended real line
$L_2[0, T, R^n]$	space of (equivalence classes of) square integrable vector-valued functions on $[0, T]$ with values in R^n
$W_2^1[0, T, R^n]$	space of absolutely continuous vector-valued functions on $[0, T]$ with values in R^n and derivatives in $L_2[0, T, R^n]$
$\mathfrak{L}[X, Y]$	space of all continuous linear operators mapping X into Y
$\langle \cdot, \cdot \rangle_X$	inner product in a Hilbert space X
$(.)^D$	projection on a set D
$^+D^*$	positive dual cone to a cone D
$\delta(\cdot \mid A)$	indicatory functional of a set A
dom f	effective domain of a functional f
cl A	closure of a set A

co f	convex hull of a function f
$\partial f(x)$	subdifferential of a functional f at x
$\mathcal{A}^*$	conjugate operator to a linear operator $\mathcal{A}$
x^j	j-th coordinate of a vector x
l.s.c.	lower semi-continuous

References

1. *Ekeland, I., Temam, R.*, Analyse Convexe et Problémes Variationnels. Dunod, Paris, 1974.
2. *Eldin, N.*, A report on the convex feedback method with applications. Proceedings of the 2-nd IFAC/IFORS symposium "Optimization Methods/Applied Aspects", Varna, 1979.
3. *Kurcyusz, S.*, Some Remarks on Generalized Lagrangians. Proc. 7th IFIP Conference on Optimization Techniques, Nice, 1975.
4. *Luenberger, D. E.*, Optimization by Vector Space Methods. J. Wiley and Sons, N.Y., 1968.
5. *Outrata, J. V.*, A multiplier method for convex optimal control problems. Proceedings of the 2-nd IFAC/IFORS symposium "Optimization Methods/Applied Aspects", Varna, 1979.
6. *Outrata, J. V., Kříž, O. F.*, An application of conjugate duality for numerical solution of continuous convex optimal control problems. Kybernetika, to appear.
7. *Rockafellar, R. T.*, Integrals which are convex functionals. Pac. J. of Math., Vol. **24**, 1968, pp. 525–539.
8. *Rockafellar, R. T.*, A dual approach to solving nonlinear programming problems by unconstrained optimization. Math. Programming, Vol. **3**, 1973, pp. 354–373.
9. *Rockafellar, R. T.*, Conjugate Duality and Optimization. Regional Conference Series in Applied Mathematics No. *16*, SIAM, Philadelphia, 1974.
10. *Wierzbicki, A. P., Kurcyusz, S.*, Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space. SIAM J. Contr. and Optimiz., Vol. **15**, 1977, pp. 25–56.

Метод обобщенных функций Лагранжа для решения одного класса выпуклых непрерывных задач оптимального управления

И. В. ОУТРАТА, З. ШИНДЛЕР

(Прага)

Для решения выпуклых задач оптимального управления непрерывными системами при наличии ограничений типа неравенств на фазовые координаты предлагается численный подход, основанный на идее обобщенной квадратической штрафной функции. В качестве пространства пертурбаций избрано подходящее пространство Соболева. Поэтому для конструкции соответствующей обобщенной функции Лагранжа надо пользоваться специальной теоремой о проекции. Для предлагаемого принципиального алгоритма легко найти подходящий конечноразмерный вариант.

J. V. Outrata, Z. Schindler
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4
182 08 Praha 8
Czechoslovakia

Problems of Control and Information Theory, Vol. 10 (2), pp. 83–94 (1981)

OPTIMIZATION OF THE DISCRETE DESCRIPTION FOR CONTINUOUS SYSTEMS

A. TUZAR, Z. BERAN

(Prague)

(Received April 10, 1980)

When using digital computers the natural question is to find a suitable discrete description of the studied continuously varying system. Practically such requirement implies the replacement of either differential or integral operators by combinations of function values in discrete times. The contribution deals with an optimization problem for such approximation of linear operators on the class of function exhibiting certain smoothness properties.

More exactly, a general method determining approximate derivatives is suggested which exploits only function values in several given points. The method is based on the theory of splines and its application is illustrated by concrete examples.

Introduction

The main purpose of this paper is a simple approach for optimal determination of finite dimensional approximation for differential operators (in reasonable sense). Such a problem appears in control theory, when we search for a discrete model of some continuous physical process.

§ 1. Theoretical background

The problem which we will to solve, is a special case of approximation of bounded linear functionals on real Hilbert space H. Every such functional can be expressed by means of scalar product $\langle h, .\rangle_H$, $h \in H$. Let $\Phi(.)$ be a bounded linear mapping of H onto Euclidean space $\mathbf{R}^n$. The general problem of optimal approximation is as follows: For arbitrary given $h \in H$ find $c^0 \in \mathbf{R}^n$ for which the minimum of the expression

$$\min_{c \in \mathbf{R}^n} \max_{\substack{x \in H \\ ||x|| \leq 1}} \left| \langle h, x\rangle_H - \langle c, \Phi(x)\rangle_n \right| \tag{1.1}$$

is attained.

Here $\langle .,. \rangle_n$ is the scalar product in $\mathbf{R}^n$. The solution of this problem will be called optimal vector and its coordinates $c_1^0, \ldots, c_n^0$ optimal coefficients. Let $e_1, \ldots, e_n$ be an orthonormal basis in $\mathbf{R}^n$. Evidently, there exist elements $\varphi_i \in H$, $i = 1, \ldots, n$, such that

$$\forall x \in H\colon \langle \varphi_i, x\rangle_H = \langle e_i, \Phi(x)\rangle_n, \quad i = 1, \ldots, n. \tag{1.2}$$

Since $\Phi(.)$ is an onto mapping, the elements φ_i are linearly independent. Let S be the linear subspace in H, spanned over the elements $\varphi_1, \ldots, \varphi_n$. Besides the *natural* basis $\{\varphi_1, \ldots, \varphi_n\}$ there evidently exists in S the unique *cardinal* basis $\{\sigma_1, \ldots, \sigma_n\}$ such, that

$$\Phi(\sigma_i) = e_i, \quad i = 1, \ldots, n,$$

or equivalently, using Kronecker's delta symbol

$$(\varphi_i, \sigma_j) = \delta_{ij}, \quad i, j = 1, \ldots, n. \tag{1.3}$$

The problem of an optimal approximation of bounded linear functional on H is equivalent to the approximation of an element $h \in H$ by an element of S, so that the norm

$$\left\| h - \sum_{i=1}^{n} c_i^0 \varphi_i \right\|_H = \min_{c \in \mathbf{R}^n} \left\| h - \sum_{i=1}^{n} c_i \varphi_i \right\|_H \tag{1.4}$$

is minimized. For purposes of practical evaluation we formulate

Theorem 1.1. The real numbers $c_1^0, \ldots, c_n^0$ are optimal coefficients for the approximation of a bounded linear functional $\langle h, . \rangle_H$ by $\left\langle \sum_{i=1}^{n} c_i \varphi_i, . \right\rangle_H$, iff (i.e. if and only if) one of the following (equivalent) conditions holds:

$$\forall\, s \in S: \langle \Sigma c_i^0 \varphi_i, s \rangle_H = \langle h, s \rangle_H, \tag{1.5}$$

$$\sum_{i=1}^{n} c_i^0 \langle \varphi_i, \varphi_j \rangle_H = \langle h, \varphi_j \rangle_H, \quad j = 1, \ldots, n, \tag{1.6}$$

$$c_i^0 = \langle h, \sigma_i \rangle, \qquad i = 1, \ldots, n. \tag{1.7}$$

Proof. Condition (1.5) follows immediately from the well-known fact that the nearest (in the sense of the norm in H) element of a linear subspace S to an element $h \in H$ is the orthogonal projection of h into S and therefore $h - \sum_{i=1}^{n} c_i^0 \varphi_i$ is orthogonal to S. Condition (1.6) expresses this fact for the elements of natural basis, condition (1.7) for the elements of cardinal basis.

§ 2. Approximation on Sobolev spaces

Now let us solve the problem of the optimal approximation for a special case $H = W_2^{(q)}(a, b)$, where $q > 1$ is an integer. The Sobolev space $W_2^{(q)}(a, b)$ is the Hilbert space whose elements are continuous functions on $[a, b]$ with quadratically integrable generalized derivatives (in the sense of theory of

distributions). The scalar product on $W_2^{(q)}(a, b)$ is introduced by

$$\langle f, g\rangle_w = \int_a^b f^{(q)}(x)\, g^{(q)}(x)\, dx + \langle \Pi_q(f), \Pi_q(g)\rangle_P, \tag{2.1}$$

and the corresponding norm by

$$||f||_w^2 = \langle f, f\rangle_w, \tag{2.2}$$

where the symbol $\Pi_q(.)$ is a projection operator onto the linear space $P_{q-1}(a, b)$ of polynomials of degree $d \leq q - 1$. We introduce the scalar product on $P_{q-1}(a, b)$ in the following way: we assume that there are given fixed points

$$-\infty < a < x_1 < \ldots < x_q < b < +\infty \tag{2.3}$$

and put for $p_1, p_2 \in P_{q-1}(a, b)$

$$\langle p_1, p_2\rangle_P = \sum_{i=1}^{q} p_1(x_i)\, p_2(x_i), \quad ||p||_P^2 = \langle p, p\rangle_P. \tag{2.4}$$

Setting $\Pi_q(f)$ equal to the (unique) polynomial of degree $d \leq q - 1$ which coincides with $f(.)$ in $x_1, \ldots, x_q$, we can consider $P_{q-1}(a, b)$ as a closed linear subspace of $W_2^{(q)}(a, b)$. Let us remark that in the literature we can find different definitions of norms in Sobolev spaces, which are, as it is not difficult to verify, equivalent to the norm introduced here.

Now, let $1 < q < n$ and let the mapping $\Phi(.)$ be given as

$$\Phi(f) = [f(x_1), \ldots, f(x_n)] \in \mathbf{R}^n, f \in W_2^{(q)}(a, b), \tag{2.5}$$

where

$$-\infty < a < x_1 < \ldots < x_q < \ldots < x_n < b < +\infty. \tag{2.6}$$

In order to determine the subspace S from §1, we shall start with its natural basis $\{\varphi_1, \ldots, \varphi_n\}$. By (1.2), (2.1) and (2.5), for arbitrary $f \in W_2^{(q)}(a, b)$

$$\int_a^b \varphi_i^{(q)}(x)\, f^{(q)}(x)\, dx + \sum_{j=1}^{q} \varphi_i(x_j)\, f(x_j) = f(x_i), \quad i = 1, \ldots, n. \tag{2.7}$$

Theorem 2.1. Let the points (2.6) be given and let $q > 1$. Then there exists for every $i = 1, \ldots, n$ just one function $\varphi_i(.) \in W_2^{(q)}(a, b)$ such that (2.7) holds for all $f(.) \in W_2^{(q)}(a, b)$. This function $\varphi_i(.)$ has the following properties:

(S1) φ_i is equal to a polynomial of degree $d \leq 2q - 1$ on each

$$]x_j, x_{j+1}[, \quad j = 1, \ldots, n - 1.$$

(S2) φ_i is equal to a polynomial of degree $d \leq q-1$ on $]a_1, x_1[$ and on $]x_n, b]$.
(S3) $\varphi_i^{(2q-2)}$ is continuous on $[a, b]$.

Proof. The functional $\langle \varphi_i, f \rangle_w = f(x_i)$ is linear and, according to Sobolev's theorem of embedding, also bounded, as (for $q > 1$)

$$|\langle \varphi_i, f \rangle_w| = |f(x_i)| \leq ||f||_{C(a,b)} \leq \text{const.}\ ||f||_w.$$

Therefore by theorem of F. Riesz there exists in $W_2^{(q)}(a, b)$ exactly one φ_i, for which (2.7) holds.

Let us suppose for a moment that φ_i has continuous derivatives of order $2q-2$ on $[a, b]$ and of order $2q-1$ on the intervals $[a_1 x_1[\ , \ ]x_1, x_2[\ , \ldots,]\ x_n, b]$. To simplify the notation, let us denote $a = x_0$ and $b = x_{n+1}$. Integrating by parts we obtain

$$\int_a^b \varphi_i^{(q)}(x)\, f^{(q)}(x)\, dx = \varphi_i^{(q)} f^{(q-1)} \Big|_a^b - \varphi_i^{(q+1)} f^{(q-2)} \Big|_a^b + \ldots + (-1)^{q-2} \varphi_i^{(2q-2)} f' \Big|_a^b +$$

$$+ \sum_{j=0}^{n} \left\{ (-1)^{q-1} \varphi_i^{(2q-1)} f \Big|_{x_j}^{x_{j+1}} + (-1)^q \int_{x_j}^{x_{j+1}} \varphi_i^{(2q)}(x)\ f(x)\ dx \right\}. \tag{2.8}$$

After inserting this expression into (2.7), which must be true for arbitrary $f \in W_2^{(q)}(a, b)$, we obtain the following necessary conditions for φ_i:

$$\varphi_i^{(2q)}(x) = 0 \quad \text{for} \quad x \in]x_j, x_{j+1}[, \quad j = 0, \ldots, n, \tag{2.9}$$

$$(-1)^q [\varphi_i^{(2q-1)}(x_j +) - \varphi_i^{(2q-1)}(x_j -)] + u(q-j)\, \varphi_i(x_j) = \delta_{ij}, \quad j = 1, \ldots, n, \tag{2.10}$$

where $u(t) = 1$ for $t \geq 0$, $u(t) = 0$ for $t < 0$,

$$\varphi_i^{(q)}(a) = \varphi_i^{(q)}(b) = \varphi_i^{(q+1)}(a) = \varphi_i^{(q+1)}(b) = \ldots = \varphi_i^{(2q-1)}(a) = \varphi_i^{(2q-1)}(b) = 0, \tag{2.11}$$

$$\varphi_i(.), \ldots, \varphi_i^{(2q-2)}(.) \text{ are continuous on } [a, b]. \tag{2.12}$$

On the other hand, it is easy to see that every solution of system (2.9)–(2.12) fulfils all requirements of the theorem concerning φ_i and also conditions (S1)–(S3). If we prove that (2.9)–(2.12) has exactly one solution (which is evidently as smooth as we have supposed above in order to derive (2.8)), then this solution will be identical with the function, the existence and uniqueness of which follows from the Riesz's theorem. In fact we will construct the solution of the system (2.9)–(2.12).

Let us start with the interval $]a, x_1[$. $\varphi_i(.)$ is according to (2.9) a polynomial $p_i(.)$ of degree $d \leq 2q-1$, which has, according to (2.11) for $x = a$ all derivatives of order $q, \ldots, 2q-1$ equal to zero. Therefore its degree is at most $q-1$. So $\varphi_i(x)$ is on $]a, x_1[$ a polynomial of degree $d \leq q-1$ with

q unknown coefficients. Let us denote the function $\varphi_i(.)$ on $]x_j, x_{j+1}[$ as $\Phi_{ij}(.)$. Using the symbol

$$(y)_+ = \frac{y + |y|}{2}, \tag{2.13}$$

it is easy to verify that the following function is a solution of the system (2.9)–(2.12):

$$\Phi_{ij}(x) = \Phi_{i,j-1}(x) + (-1)^q \frac{\delta_{ij} - u(q-j)\,\Phi_{i,j-1}(x_j)}{(2q-1)!}\,(x - x_j)_+^{2q-1}, \quad j = 1, \ldots, n,$$

where $\Phi_{ij}(x_k) = \lim_{x \to x_k^-} \Phi_{ij}(x) = \lim_{x \to x_k^+} \Phi_{ij}(x)$, $k = 1, \ldots, n$.

Therefore for $k = 1, \ldots, n$ and $x \in [x_k, x_{k+1}]$ there is

$$\Phi_{ik}(x) = p_i(x) + \frac{(-1)^q}{(2q-1)!} \sum_{j=1}^{k} [\delta_{ij} - u(q-j)]\,\Phi_{i,j-1}(x_j)\,(x - x_j)_+^{2q-1} \tag{2.14}$$

and for $x \in [x_n, b]$

$$\Phi_{in}(x) = p_i(x) + \frac{(-1)^q}{(2q-1)!}\left[(x - x_i)_+^{2q-1} - \sum_{j=1}^{q} \Phi_{i,j-1}(x_j)\,(x - x_j)_+^{2q-1}\right]. \tag{2.15}$$

The unknown q coefficients of $p_i(x)$ and q values $\Phi_{i0}(x_1), \Phi_{i1}(x_1), \ldots, \Phi_{i,q-1}(x_q)$ can be determined in the following way: Applying conditions (2.11) to the function (2.15) for $x = b$, we obtain evidently a system of linear equations with a regular matrix for unknown $\Phi_{i,j-1}(x_i)$ not depending on $p_i(.)$. Using then $\Phi_{i0}(x_1) = p_i(x_1)$ and putting $x = x_{k+1}$ for $k = 1, \ldots, q-1$ into (2.14), we obtain q equations; each of them determines exactly one value of $p_i(x_j)$, $j = 1, \ldots, q$. By those values is the polynomial $p_i(.)$ uniquely determined. ∎

The linear subspace S, which is spanned on functions from Theorem 2.1, is usually called the space of polynomial splines of order q. Besides the *natural* basis $\{\varphi_1, \ldots, \varphi_n\}$ there exists by Theorem 1.1 another basis which consists of the *cardinal* splines σ_i, $i = 1, \ldots, n$.

§ 3. Approximation of derivatives

Theorem 2.1 implies that the cardinal splines in the space $W_2^{(q)}(a, b)$, $q \geq 2$, can be written in the form

$$\sigma_k^{(q,n)}(x) = \begin{cases} p_k(x) = \sum_{i=0}^{q-1} a_i x_i + \sum_{i=1}^{n-1} \lambda_i (x - x_i)_+^{2q-1}, & x \in [a, a + nh], \\ r_k(x) = \sum_{i=0}^{q-1} b_i x_i, & x \in [a + nh, b], \end{cases}$$

$$\sigma_k^{(q,n)}(x_j) = \delta_{kj}; \quad k, j = 1, \ldots, n,$$

where

$$a < a + h < \ldots < a + nh < b$$

is an equidistant partition of the interval $[a, b]$ with the step

$$h = \frac{b - a}{n + 1} > 0.$$

From this expression it is easy to see that $\sigma_k^{(q,n)}(.) \in C^{(2q-2)}([a, a + nh])$ and for the determination of the $\sigma_k^{(q,n)}$ it is necessary to find the $n + 2q - 1$ unknown coefficients $a_0, a_1, \ldots, a_{q-1}, \lambda_1, \ldots, \lambda_{n-1}, b_0, \ldots, b_{q-1}$. The requirement $\sigma_k^{(q,n)}(.) \in C^{(2q-2)}([a, b])$ implies that

$$p_k^{(i)}(a + nh) = r_k^{(i)}(a + nh), \quad i = 0, 1, \ldots, 2q - 2.$$

For the determination of the cardinal splines it is necessary to solve n systems of linear algebraic equations. Using the explicit form of the cardinal splines we can obtain (via Theorem 1.1) those expressions for the optimal coefficients of difference formulae for the l-th numerical derivative of a function from $W_2^{(q)}(a, b)$ in the mesh-point z:

$$c_i^0 = D^l \sigma_i^{(q,n)}(x)\big|_{x=z}. \tag{3.1}$$

For the spaces $W_2^{(2)}(0, 4h)$, $W_2^{(2)}(0, 5h)$, $W_2^{(2)}(0, 6h)$, $W_2^{(3)}(0, 5h)$ we obtain these difference formulae:

1) the space $W_2^{(2)}(0, 4h)$: 3-points formulae

$$\begin{aligned} f'(x)\big|_{x=h} &= \frac{1}{4h}[-5f(h) + 6f(2h) - f(3h)], \\ f'(x)\big|_{x=2h} &= \frac{1}{2h}[f(3h) - f(h)], \\ f'(x)\big|_{x=3h} &= \frac{1}{4h}[f(h) - 6f(2h) + 5f(3h)]. \end{aligned} \tag{3.2}$$

2) the space $W_2^{(2)}(0, 5h)$: 4-points formulae:

$$f'(x)\big|_{x=h} = \frac{1}{15h}[-19f(h) + 24f(2h) - 6f(3h) + f(4h)],$$

$$f'(x)\big|_{x=2h} = \frac{1}{15h}[-7f(h) - 3f(2h) + 12f(3h) - 2f(4h)], \tag{3.3}$$

$$f'(x)\big|_{x=3h} = \frac{1}{15h}[2f(h) - 12f(2h) + 3f(3h) + 7f(4h)], \tag{3.4}$$

$$f'(x)\big|_{x=4h} = \frac{1}{15h}[-f(h) + 6f(2h) - 24f(3h) + 19f(4h)].$$

3) the space $W_2^{(2)}(0, 6h)$: 5-points formulae:

$$f'(x)|_{x=h} = \frac{1}{56h}[-71f(h) + 90f(2h) - 24f(3h) + 6f(4h) - f(5h)],$$

$$f'(x)|_{x=2h} = \frac{1}{28h}[-13f(h) - 6f(2h) + 24f(3h) - 6f(4h) + f(5h)],$$

$$f'(x)|_{x=3h} = \frac{1}{8h}[f(h) - 6f(2h) + 6f(4h) - f(5h)],$$

$$f'(x)|_{x=4h} = \frac{1}{28h}[-f(h) + 6f(2h) - 24f(3h) + 6f(4h) + 13f(5h)], \qquad (3.5)$$

$$f'(x)|_{x=5h} = \frac{1}{56h}[f(h) - 6f(2h) + 24f(3h) - 90f(4h) + 71f(5h)].$$

4) the space $W_2^{(3)}(0,5h)$: 4-points formulae:

$$f'(x)|_{x=h} = \frac{1}{132h}[-223f(h) + 339f(2h) - 141f(3h) + 25f(4h)],$$

$$f'(x)|_{x=2h} = \frac{1}{132h}[-47f(h) - 57f(2h) + 123f(3h) - 19f(4h)],$$

$$f'(x)|_{x=3h} = \frac{1}{132h}[19f(h) - 123f(2h) + 57f(3h) + 47f(4h)], \qquad (3.6)$$

$$f'(x)|_{x=4h} = \frac{1}{132h}[-25f(h) + 141f(2h) - 339f(3h) + 223f(4h)]. \qquad (3.7)$$

In the space $W_2^{(3)}(0,5h)$ it is possible to approximate also the second derivative.

Using (3.1) we obtain:

$$f''(x)|_{x=h} = \frac{1}{22h^2}[31f(h) - 71f(2h) + 49f(3h) - 9f(4h)],$$

$$f''(x)|_{x=2h} = \frac{1}{132h^2}[146f(h) - 306f(2h) + 174f(3h) - 14f(4h)],$$

$$f''(x)|_{x=3h} = \frac{1}{132h^2}[-14f(h) + 174f(2h) - 306f(3h) + 146f(4h)],$$

$$f''(x)|_{x=4h} = \frac{1}{22h^2}[-9f(h) + 49f(2h) - 71f(3h) + 31f(4h)].$$

§ 4. General form of an error functional in the spaces $L_2^{(q)}(a, b)$

The space $L_2^{(q)}(a, b)$, $q \geq 2$ is defined as a factor space

$$L_2^{(q)}(a, b) = W_2^{(q)}(a, b)/P_{q-1}(a, b);$$

the space $L_2^{(q)}(a, b)$ is a Hilbert space, if we define a scalar product in the form

$$\langle f, g\rangle_L = \int_a^b f^{(q)}(x)\, g^{(q)}(x)\, dx$$

for any $f, g \in L_2^{(q)}(a, b)$.

We say that $g(x)$ is an *extremal function* of a continuous linear functional F in the space $L_2^{(q)}(a, b)$, if

$$F(g) = \langle f, g\rangle_L = ||F||_{L^*}\, ||g||_L,$$

where $[L_2^{(q)}(a, b)]^*$ is a dual space of $L_2^{(q)}(a, b)$, for some $f \in L_2^{(q)}(a, b)$ (as follows from the Riesz representation theorem).

Let $q \geq 2$, then according to Sobolev's imbedding theorem for all $x \in]a, b[$ it holds

$$||f^{(p)}(x)||_{C(a,b)} \leq \text{const}\ ||f||_w, \quad p = 1, \ldots, q-1$$

and therefore the p-th derivative, $p = 1, \ldots, q-1$ in an arbitrary fixed point is a bounded linear functional. We approximate it by a suitable linear combination of function values in some given points, i.e.

$$f^{(p)}(x)\big|_{x=x_r} \sim \sum_{i=1}^{n} c_i f(x_i), \quad p = 1, \ldots, q-1,\ r = 1, \ldots, n, \tag{4.1}$$

where $a = x_0 < x_1 < \ldots < x_{n+1} = b$ is a fixed partition of $[a, b]$. We solve the approximation problem (1.9) and for this problem we define *error functionals* $F_r(.)$, $r = 1, \ldots, n$ as

$$F_r(f) = f^{(p)}(x)\big|_{x=x_r} - \sum_{i=1}^{n} c_i f(x_i).$$

It can be shown that

$$F_r(f) = \int_a^b \left[\sum_{i=1}^{n} c_i\, u'(x - x_i) + (-1)^{p+1} u^{(p+1)}(x - x_r)\right] f(x)\, dx,$$

where $u(t)$ is the distribution introduced in (2.10).

Now we can prove

Theorem 4.1. Let $F_r(.)$ be the error functional for problem (4.1).

Then

$$||F_r||^2_{L^*} = (-1)^q \left\{ \frac{1}{(2q-1)!} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} c_i c_j (x_i - x_j)^{2q-1} + \right.$$

$$\left. + \frac{1}{(2q-p-1)!} \left[(-1)^{p+1} \sum_{i=r+1}^{n} c_i (x_i - x_r)^{2q-p-1} - \sum_{i=1}^{r-1} c_i (x_r - x_i)^{2q-p-1} \right] \right\}.$$

Proof. From the definition of the scalar product in the space $L_2^{(q)}(a, b)$ and Riesz's representation theorem it follows that

$$F_r(f) = \langle f_r, f \rangle_L = \int_a^b f_r^{(q)}(x) f^{(q)}(x)\, dx,$$

where $f_r(.) \in L_2^{(q)}(a, b)$. Further, using the integration by part, it can be shown that

$$\langle f_r, f \rangle_L = (-1)^q \int_a^b f_r^{(2q)}(x) f(x)\, dx + \sum_{i=0}^{q-1} (-1)^i f_r^{(q+i)}(x) f^{(q-i-1)}(x) \Big|_a^b .$$

From the integral representation of the error functional f_r, it is easy to see that the function $f_r(.)$ can be obtained as the solution of the boundary value problem

$$\begin{cases} f_r^{(2q)}(x) = (-1)^q F_r(f)(x) = (-1)^q \left[\sum_{i=1}^{n} c_i u'(x - x_i) + (-1)^{p+1} u^{(p+1)}(x - x_r) \right] \\ f_r^{(q+i)}(a+) = 0, \quad f_r^{(q+i)}(b-) = 0, \quad i = 0, 1, \ldots, q-1. \end{cases} \tag{4.2}$$

Now, the function $F_r \in [L_2^{(q)}(a, b)]^*$ iff

$$\langle f_r, x^\alpha \rangle_L = 0, \qquad \alpha = 0, 1, \ldots, q-1. \tag{4.3}$$

Then all solutions of (4.2) are given in the form

$$f_r(x) = (-1)^q \left[\frac{1}{(2q-1)!} \sum_{i=1}^{n} c_i (x - x_i)_+^{2q-1} + \frac{(-1)^{p+1}}{(2q-p-1)!} (x - x_r)_+^{2q-p-1} \right] + \omega(x),$$

where $\omega(x) \in P_{q-1}(a, b)$ and $(y)_+$ is defined by (2.13). This solution is unique in the sense of space $L_2^{(q)}(a, b)$. Thus $f_r(.)$ is an extremal function of the problem (4.1) in the space $L_2^{(q)}(a, b)$. Further, it can be shown that

$$||F_r||^2_{L^*} = F_r(f_r) \tag{4.4}$$

and the end of the proof of the theorem follows now using (4.4) from easy calculations of integrals. ∎

If the partition of $[a, b]$ is equidistant with the step $h = \dfrac{b-a}{n+1} > 0$, i.e.

$$a < a + h < \ldots < a + nh < b$$

then it holds:

Corollary 4.2. Let F_r be the error functional for problem (4.1). Then

$$||F_r||^2_{L^*} = Qh^{2q-2p-1}$$

with

$$Q = (-1)^q \left\{ \frac{1}{(2q-1)!} \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} d_i d_j (i-j)^{2q-1} + \right.$$

$$\left. + \frac{1}{(2q-p-1)!} \left[(-1)^{p+1} \sum_{i=r+1}^{n} d_i (i-r)^{2q-p-1} - \sum_{i=1}^{r-1} d_i (r-i)^{2q-p-1} \right] \right\},$$

where $d_i = h^p c_i$, $i = 1, \ldots, n$ and c_i are the coefficients from (4.1).

Remark 1. From (4.3) follows that the above difference formulas are precise for polynomials of degree $d \leq q - 1$.

Remark 2. Using Corollary 4.2, it can be shown that the norms of the error functionals for the difference formulas in §3 are in all cases better then the norms of the well-known difference formulae that were obtained using a Lagrange interpolation (for those formulae see [2]).

§ 5. Some numerical results

1. Calculate the response of the following simple problem

$$\dot{y}(t) = y(t), \; y(0) = 1, \; t \in [0, 1]$$

with the discretization step $h = 0,1$. We denote the error $E = \tilde{y}(t) - y(t)$, where $y(t)$ is an exact solution in the point t and $\tilde{y}(t)$ is the corresponding approximate solution. This problem was solved using formulae (3.2), (3.4), (3.5), (3.7). For the sake of comparison, this problem was solved using alternative difference formulae which were derived by the Lagrange interpolation schema (see [2]) and classical Euler's method. Let us remark that the k-point difference formula, which is derived by the Lagrange interpolation schema, is precise on $P_{k-1}(a, b)$, while optimal formulae in $W_2^{(2)}(a, b)$ and $W_2^{(3)}(a, b)$ are precise on $P_1(a, b)$ and $P_2(a, b)$, respectively. In Table 1 are given the errors of $y(1)$, which were obtained using all the above-mentioned difference formulae.

Here E_{opt} is the error of an optimal formula, E_{cl} is the error of a corresponding formula according to [2] and $||F_r||_{L^*}$ is the norm the error functional

due to Corollary 4.2. Euler's method gives the error $E = -0.1245\ldots$ From the results in Table 1, it can be concluded that in all cases the error E_{opt} is much less than E_{cl}. It is even less than the norm of the error functional F_r.

Table 1

Formula	E_{opt}	E_{cl}	$\|\|F_r\|\|_{L^*}$
(3.2)	$-64.1 \cdot 10^{-3}$	$-284.6 \cdot 10^{-3}$	0.1708
(3.4)	$-4.5 \cdot 10^{-3}$	$38.9 \cdot 10^{-3}$	0.1247
(3.5)	$13.3 \cdot 10^{-3}$	$502.3 \cdot 10^{-3}$	0.1244
(3.7)	$3.8 \cdot 10^{-3}$	—	$8.3299 \cdot 10^{-3}$

2. The second problem deals with a control prediction problem when the motion of a satellite is stabilized in the prescribed direction under the presence of small initial deviations. The problem can be described by a system of differential equations (see [8]):

$$\dot{x}_1 = a_1 x_2, \quad \dot{x}_2 = -a_1 x_1 + a_2 x_4 + u_1,$$

$$\dot{x}_3 = a_3 x_4, \quad \dot{x}_4 = -a_2 x_2 - a_3 x_3 + u_2,$$

where a_1, a_2, a_3 are constant parameters and u_1, u_2 are controls. It can be shown that the solution of this stabilization problem can be expressed in the form

$$u_1^* = \frac{1}{2a_1}(2a_1^2 - \alpha^2)\, x_1 - \alpha x_2 - a_2 x_4,$$

$$u_2^* = a_1 x_2 + \frac{1}{2a_3}(2a_3^2 - \alpha^2)\, x_3 - \alpha x_4,$$

where $\alpha > 0$ is a given coefficient of exponential stability.

The problem was solved on the interval $[0, 1]$ for $a_1 = 1, a_2 = 2, a_3 = -1$, $\alpha = 2$, with the discretization step $h = 0.2$, using formula (3.2). From Corollary 4.2 it follows that the norm of error functional for (3.2) and $h = 0.2$ is $||F_r||_{L^*} = 0.2415$. Table 2 summarizes the obtained numerical results for controls u_1, u_2 in the mesh-points of $[0, 1]$ and simultaneously the deviations $E_{u_i} = u_i - u_i^*(t)$ $i = 1, 2$, from exact solutions $u_1(.), u_2(.)$ in the mesh-points.

Table 2

t	u_1	E_{u_1}	u_2	E_{u_2}
0.2	−2.265	$-2.0 \cdot 10^{-2}$	−1.728	$5.0 \cdot 10^{-2}$
0.4	−1.575	$1.5 \cdot 10^{-2}$	−2.099	$8.4 \cdot 10^{-2}$
0.6	−1.078	$-2.9 \cdot 10^{-2}$	−2.188	$5.8 \cdot 10^{-2}$
0.8	−0.692	$-7.6 \cdot 10^{-2}$	−2.109	0.138
1	−0.393	−0.107	−1.927	0.128

From these results it can be seen that the errors E_{u_1}, E_{u_2} are in all cases less than the norm of error functional F_r.

Remark. The optimal difference formulae have an additional interesting property. The roots of the corresponding characteristic polynomials have less absolute values than the roots of characteristic polynomials of those formulae which are derived by Lagrange interpolation schema. This result represents relatively better numerical stability of optimal formulae.

References

1. *Laurent, P. J.*, Approximation et optimisation. Hermann, Paris (1972) 496 pp.
2. *Положий, Т. Н. и др.* Математический практикум. Физматгиз, Москва (1960) 512 стр.
3. *Соболев С. Л.* Введение в теорию кубатурных формул. Наука, Москва (1974) 808 стр.
4. *Schoenberg, J. J.*, Spline interpolation and higher derivatives, Proc. Nat. Acad. Sci., **51** (1964), *1*, pp. 24–28.
5. *Tuzar, A.*, Bemerkung zur Optimisation der numerischen Lösung für die Randwertaufgaben, Acta technica ČSAV, **3** (1967), pp. 304–313.
6. *Tuzar, A.*, Optimal approximation of linear functionals in Hilbert space with applications to numerical methods, Research report of ÚTIA ČSAV No. *988* (1979), 39 pp.
7. *Varga, R.*, Functional analysis and approximation theory in numerical analysis, Soc. for Industrial and Applied Math., Philadelphia, Pensylvania (1971) 126 pp.
8. *Фурасов, В. Д.* Устойчивость движения, оценки и стабилизация. Наука, Москва (1977), 247 стр.

Оптимизация дискретного представления непрерывных систем

А. ТУЗАР, З. БЕРАН

(Прага)

При использовании цифровой вычислительной техники очевидно возникает проблема нахождения подходящего дискретного описания заданной непрерывной системы. На практике это представляет замену дифференциальных или интегральных операторов на линейные комбинации значений функции в дискретные моменты времени. В работе решена проблема оптимизации такой замены линейных ограниченных операторов на некоторых классах гладких функций. В частности, разработан метод вывода формул приближенного вычисления производных, использующий значения функции в нескольких заданных точках. Метод основан на теории сплайн-функций и проиллюстрирован на конкретных примерах.

A. Tuzar, Z. Beran
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4, 182 08 Praha 8, ČSSR

Problems of Control and Information Theory, Vol. 10 (2), pp. 95–114 (1981)

THE FINITE-DIMENSIONAL DISTRIBUTIONS OF A RANDOM PROCESS DETERMINED BY A STOCHASTIC DIFFERENTIAL EQUATION AND THEIR APPLICATION TO CONTROL PROBLEMS

V. S. PUGACHEV

(Moscow)

(Received June 23, 1980)

The equations determining the finite-dimensional characteristic functions of a random process determined by a stochastic differential equation are given. These equations are applied to the problems of the statistical analysis of non-linear systems. For linear systems the explicit formulae for all the finite-dimensional characteristic functions are obtained. The equations determining the finite-dimensional characteristic functions are also used to solve the problems of non-linear filtering and prediction. Besides the methods of conditionally optimal estimation of the state vector and unknown parameters of a system described by a stochastic differential equation recently developed by the author a new method for conditionally optimal extrapolation of the state vector of such a system is given. Some approximate methods for the solution of the problems considered are discussed.

1. Introduction

Consider the Ito stochastic differential equation

$$dX = a(X, t)\,dt + b(X, t)\,dW \tag{1}$$

where a and b are non-random functions and W is a process with independent increments.* The processes $X = X(t)$ and $W = W(t)$ are assumed to have their values in some finite-dimensional real spaces, say in R^p and R^q respectively. Then the function $a(x, t)$ is a vector-valued function mapping $R^p \times R$ into R^p, and the function $b(x, t)$ is a $p \times q$ matrix-valued function mapping $R^p \times R$ into R^{pq}. We consider the random process $X(t)$ representing the strong solution of equation (1) satisfying the initial condition $X(t_0) = X_0$, X_0 being a random variable independent of $W(s) - W(t)$ for any $s > t \geq t_0$.

The problem is to find the finite-dimensional distributions of the process $X(t)$ given the distribution of X_0 and the one-dimensional distribution of $W(t)$.

The problem stated arises in various areas of applications of mathematical methods. In particular, it is encountered in control theory where the distribution of the output of a system described by differential equations is needed

* Unlike in classical cases of Ito stochastic differential equations where W represents a Wiener process we consider more general equations where W may be any process with independent increments. The author seems to be first who studied such equations [1].

given the distribution of the random disturbances. If the latter may be considered as white noises then the equations describing the system represent themselves stochastic differential equations. In other problems the random disturbances may be considered as random processes determined by stochastic differential equations as for instance stationary random processes with rational spectral densities. In such cases the differential equations describing the behaviour of the system together with the equations determining the random inputs respesent a set of stochastic differential equations.

2. Equations for finite-dimensional characteristic functions

There are various approaches to solving the problem stated. If in particular W represents a Wiener process then the one-dimensional and transient densities (if they exist) may be found by integrating the respective Fokker–Plank–Kolmogorov equation, after which all the finite-dimensional densities may be expressed in terms of the one-dimensional and transient densities since the process $X(t)$ determined by equation (1) is a Markov process. Another way to find the one-dimensional and transient distributions of the process $X(t)$ is solution of the author's equation determining the one-dimensional and transient characteristic functions [1–4]. The method of characteristic functions is preferable in many aspects from the viewpoint of applications. We shall now show that all the finite-dimensional characteristic functions of a process determined by a stochastic differential equation may by found in a unified way by solving similar equations [5]. Namely we shall prove the following theorem.

Theorem 1. If W in equation (1) *represents a second-order process with independent increments whose expectation $m(t)$ and covariance function $k(t)$ are given by*

$$m(t) = m(t_0) + \int_{t_0}^{t} m'(\tau)\, d\tau, \quad k(t) = k(t_0) + \int_{t_0}^{t} v(\tau)\, d\tau \tag{2}$$

and whose one-dimensional characteristic function $h_1(\mu; t)$ is differentiable with respect to t and the functions $a(x, t)$ and $b(x, t)$ satisfy the conditions[1]

$$M\,|a(Z, t)|^2 < k_1 + k_2\, M\,|Z|^2, \tag{3}$$

$$M\ \mathrm{tr}\ b(Z, t)\, v(t)\, b(Z, t)^{\mathrm{T}} < k_3 + k_4\, M\,|Z|^2, \tag{4}$$

$$M\,|a(Z', t) - a(Z, t)|^2 < c_1^2\, M\,|Z' - Z|^2, \tag{5}$$

$$M\ \mathrm{tr}\,[b(Z', t) - b(Z, t)]\, v(t)\, [b(Z', t)^{\mathrm{T}} - b(Z, t)^{\mathrm{T}}] < c_2^2\, M\,|Z' - Z|^2 \tag{6}$$

[1] We denote by MX the expectation of the random variable X and by $|x|$ the Euclidean norm of the vector x.

for some $k_1, k_2, k_3, k_4, c_1, c_2 > 0$ *and for any random variables* Z, Z' *with finite second moments, then the finite-dimensional characteristic functions* $g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n)$ $(n = 1, 2, \ldots)$ *of the process* X *determined by equation* (1) *and the initial condition* $X(t_0) = X_0$, X_0 *being a random variable with finite second moment, satisfy the following equations and initial conditions*:

$$\frac{\partial}{\partial t_n} g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n) =$$

$$= M\{i\lambda_n^{\mathrm{T}} a(X_{t_n}, t_n) + \chi(b(X_{t_n}, t_n)^{\mathrm{T}} \lambda_n; t_n)\} \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\} \quad (n = 1, 2, \ldots), \tag{7}$$

$$g_1(\lambda_1, t_0) = g_0(\lambda_1), \tag{8}$$

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_{n-1}) = g_{n-1}(\lambda_1, \ldots, \lambda_{n-1} + \lambda_n; t_1, \ldots, t_{n-1}) \tag{9}$$

where $\chi(\mu; t)$ *represents the logarithmic derivative of the one-dimensional characteristic function* $h_1(\mu; t)$ *of the process* W *with respect to* t:

$$\chi(\mu; t) = \frac{\partial}{\partial t} \ln h_1(\mu, t) \tag{10}$$

and $g_0(\lambda)$ *the characteristic function of the random variable* X_0.

Proof. First of all we observe that under the conditions of the theorem, equation (1) has a unique strong solution $X(t)$ satisfying $X(t_0) = X_0$ which represents a second-order random process [6]. We have then for any $t_1 < \ldots < t_n < s$

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, s) - g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n) =$$

$$= M\left[\exp\left\{i \sum_{k=1}^{n-1} \lambda_k^{\mathrm{T}} X_{t_k} + i\lambda_n^{\mathrm{T}} X_s\right\} - \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\}\right] =$$

$$= M\{\exp\{i\lambda_n^{\mathrm{T}}(X_s - X_{t_n}) - 1\}\} \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\}. \tag{11}$$

Now we notice that by (1)

$$X_s - X_{t_n} = a(X_{t_n}, t_n)(s - t_n) + b(X_{t_n}, t_n)(W_s - W_{t_n}) +$$

$$+ \int_{t_n}^{s} [b(X_\tau, \tau) - b(X_{t_n}, t_n)]\, dW_\tau + o_p(s - t_n).$$

Substituting into (11) gives

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, s) - g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_n) =$$
$$= M\left[\exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\}\{i\lambda_n^{\mathrm{T}} a(X_{t_n}, t_n)(s - t_n) +\right.$$
$$+ i\lambda_n^{\mathrm{T}} \int_{t_n}^{s} [b(X_\tau, \tau) - b(X_{t_n}, t_n)]\, dW_\tau + o_p(s - t_n)\} +$$
$$\left. + \exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\} - 1\right] \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\}. \tag{12}$$

But

$$\left| M\left[\exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\} - 1\right] \int_{t_n}^{s} [b(X_\tau, \tau) - \right.$$
$$\left. - b(X_{t_n}, t_n)]\, dW_\tau \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\}\right|^2 \leq$$
$$\leq M\left|\exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\} - 1\right|^2 M\left|\int_{t_n}^{s} [b(X_\tau, \tau) - \right.$$
$$\left. - b(X_{t_n}, t_n)]\, dW_\tau\right|^2 = [2 - M\exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\} -$$
$$- M\exp\{-i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\}] \int_{t_n}^{s} \mathrm{tr}\, M[b(X_\tau, \tau) -$$
$$- b(X_{t_n}, t_n)]\nu_\tau [b(X_\tau, \tau)^{\mathrm{T}} - b(X_{t_n}, t_n)^{\mathrm{T}}]\, d\tau. \tag{13}$$

Averaging at first with respect to $W_s - W_{t_n}$ for fixed X_{t_n} and taking into account that $W_s - W_{t_n}$ is independent of X_{t_n} we get

$$M\exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\} = Mh(b(X_{t_n}, t_n)^{\mathrm{T}} \lambda_n; t_n, s)$$

where $h(\mu; t, s)$ is the characteristic function of the random variable $W_s - W_t$. Remembering that

$$h(\mu; t, s) = \frac{h_1(\mu; s)}{h_1(\mu; t)} = 1 + \frac{1}{h_1(\mu; t)} \frac{\partial h_1(\mu; t)}{\partial t}(s - t) + o(s - t) =$$
$$= 1 + \chi(\mu; t)(s - t) + o(s - t)$$

we obtain

$$M\exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)(W_s - W_{t_n})\} =$$
$$= 1 + \chi(b(X_{t_n}, t_n)^{\mathrm{T}} \lambda_n; t_n)(s - t_n) + o(s - t_n). \tag{14}$$

Now we consider the integral in the last member of (13). By (6)

$$\int_{t_n}^{s} \operatorname{tr} M\,[b(X_\tau, \tau) - b(X_{t_n}, t_n)]\,\nu_\tau\,[b(X_\tau, \tau)^{\mathrm{T}} - b(X_{t_n}, t_n)^{\mathrm{T}}]\,d\tau <$$

$$< c_2^2 \int_{t_n}^{s} M\,|X_\tau - X_{t_n}|^2\,d\tau = o(s - t_n) \tag{15}$$

since $M\,|X_\tau - X_{t_n}|^2$ is a continuous function of τ and is equal to zero at $\tau = t_n$. Using (14), (15) and (2) we obtain from (13)

$$M \exp\left\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)\,(W_s - W_t) + i\sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\} \int_{t_n}^{s} [b(X_\tau, \tau) -$$

$$- b(X_{t_n}, t_n)]\,dW_\tau = M \exp\left\{i\sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\} \int_{t_n}^{s} [b(X_\tau, \tau) - b(X_{t_n}, t_n)]\,dW_\tau +$$

$$+ o(s - t_n) = \int_{t_n}^{s} M \exp\left\{i\sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\} [b(X_\tau, \tau) - b(X_{t_n}, t_n)]\,m'_\tau\,d\tau +$$

$$+ o(s - t_n) = o(s - t_n) \tag{16}$$

since $b(X_\tau, \tau) \to b(X_{t_n}, t_n)$ in probability in virtue of (6). Using (16), equation (12) becomes

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, s) - g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_n) =$$

$$= M\left[\exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)\,(W_s - W_{t_n})\}\, i\lambda_n^{\mathrm{T}} a(X_{t_n}, t_n)\,(s - t_n)\right.$$

$$\left. + \exp\{i\lambda_n^{\mathrm{T}} b(X_{t_n}, t_n)\,(W_s - W_{t_n})\} - 1\right] \exp\left\{i\sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\} + o(s - t_n).$$

Averaging with respect to $W_s - W_{t_n}$ for fixed $X_{t_1}, \ldots, X_{t_n}$ and taking into account that $W_s - W_{t_n}$ is independent of $X_{t_1}, \ldots, X_{t_n}$ we obtain

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, s) - g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_n) =$$

$$= M\left[i\lambda_n^{\mathrm{T}} a(X_{t_n}, t_n)\, h(b(X_{t_n}, t_n)^{\mathrm{T}} \lambda_n, t_n, s) +\right.$$

$$\left. + h(b(X_{t_n}, t_n)^{\mathrm{T}} \lambda_n; t_n, s) - 1\right] \exp\left\{i\sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\} + o(s - t_n). \tag{17}$$

Observing that in virtue of (10)

$$h(\mu; t, s) - 1 = \frac{h_1(\mu; s) - h_1(\mu; t)}{h_1(\mu; t)} = \frac{1}{h_1(\mu; t)} \frac{\partial h_1(\mu; t)}{\partial t}\,(s - t) + o(s - t) =$$

$$= \chi(\mu; t)\,(s - t) + o(s - t)$$

we get

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, s) - g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_n) =$$

$$= M\{i\lambda_n^{\mathrm{T}} a(X_{t_n}, t_n)\, h(b(X_{t_n}, t_n)^{\mathrm{T}} \lambda_n; t_n, s) +$$

$$+ \chi(b(X_{t_n}, t_n)^{\mathrm{T}} \lambda_n; t_n)\} \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\} (s - t_n) + o(s - t_n). \qquad (18)$$

The expectation here exists according to (3) and (4) since

$$|\chi(\mu; t)| < c\,|\mu|^2$$

for some $c > 0$. Dividing (18) by $(s - t_n)$ and putting then $s \downarrow t_n$ yields (7). The initial condition (8) for equation (7) with $n = 1$ is obvious. The initial condition (9) for $g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n)$, $n \geq 2$, follows from the obvious relation

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, t_{n-1}) = M \exp\left\{i \sum_{k=1}^{n-2} \lambda_k^{\mathrm{T}} X_{t_k} + i(\lambda_{n-1}^{\mathrm{T}} + \lambda_n^{\mathrm{T}}) X_{t_{n-1}}\right\} =$$

$$= g_{n-1}(\lambda_1, \ldots, \lambda_{n-1} + \lambda_n; t_1, \ldots, t_{n-1})$$

proving the theorem.

We emphasize that our proof is valid only if (1) represents a stochastic differential equation in the sense of Ito since (17) is true only in this case.

Now we shall extend Theorem 1 to the case of a discontinuous process with independent increments W. Let $W = W(t)$ in (1) be a process with independent increments with random jumps W_k (certainly independent of one another) at fixed points τ_k $(k = 1, 2, \ldots)$. The one-dimensional characteristic function $h_1(\mu; t)$ of the process $W(t)$ has in this case the jumps $[g_k(\mu) - 1]\, h_1(\mu; \tau_k)$ at the points τ_k, $g_k(\mu)$ being the characteristic function of $W_k (k = 1, 2, \ldots)$. Hence the derivative of $h_1(\mu; t)$ with respect to t can exist only as a generalized function containing the linear combination of δ-functions $\delta(t - \tau_k)$ with the coefficient $[g_k(\mu) - 1]\, h_1(\mu; \tau_k)$, and respectively the function $\chi(\mu; t)$ determined by (10) can exist only as a generalized function containing the linear combination of δ-functions $\delta(t - \tau_k)$ with the coefficients $g_k(\mu) - 1$. Now we observe that the solution $X(t)$ of (1) has in this case the random jumps $b(X_{\tau_k}, \tau_k)\, W_k$ at the points τ_k. Hence the n-dimensional characteristic function $g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n)$ of the process $X(t)$ considered as a function of t_n for fixed $t_1, \ldots, t_{n-1}$ has the jumps $[g_k(b(X_{\tau_k}, \tau_k)^{\mathrm{T}} \lambda_n) - 1]\, g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_{n-1}, \tau_k)$ at the points τ_k $(k = 1, 2, \ldots)$. These jumps are exactly the same jumps which are determined by equation (7) with $\chi(\mu; t)$ containing the linear combination of δ-functions $\delta(t - \tau_k)$ with the coefficients $g_k(\mu) - 1$.

Thus the following theorem is true.

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

Theorem 2. If W in (1) *is a second-order process with independent increments with random jumps at fixed sequence of points* $\{\tau_k\}$ *whose expectation* $m(t)$ *and covariance function* $k(t)$ *are representable in the form* (2) *where* $\nu(\tau)$ *contains the linear combination of* δ*-functions* $\delta(t - \tau_k)$ *and whose one-dimensional characteristic function* $h_1(\mu; t)$ *is differentiable with respect to* t *at all* t *different from* τ_k *and the functions* $a(x, t)$ *and* $b(x, t)$ *in* (1) *satisfy conditions* (3)–(6) *for all* t *different from* τ_k, *then the finite-dimensional characteristic functions of the process* $X(t)$ *determined by equation* (1) *with the initial condition* $X(t_0) = X_0$ *satisfy* (7)–(9) *with* $\chi(\mu; t)$ *containing the respective linear combination of* δ*-functions* $\delta(t - \tau_k)$ $(k = 1, 2, \ldots)$.

Observe that the function $\chi(\mu; t)$ determined by (10) may also be determined in two other ways.

Differentiating

$$h(\mu; t, s) = \frac{h_1(\mu; s)}{h_1(\mu; t)}$$

with respect to s and then putting $s = t$ yields

$$\left[\frac{\partial}{\partial s} h(\mu; t, s)\right]_{s=t} = \frac{1}{h_1(\mu; t)} \frac{\partial h_1(\mu; t)}{\partial t}.$$

Consequently

$$\chi(\mu; t) = \frac{\partial}{\partial t} \ln h_1(\mu; t) = \left[\frac{\partial}{\partial s} h(\mu; t, s)\right]_{s=t}. \tag{19}$$

To obtain the third expression for $\chi(\mu; t)$ we notice that for any function $x(t)$ the characteristic function of the random variable

$$U = \int_t^s b(x_\tau, \tau) dW(\tau)$$

is equal to

$$h_u(\lambda; x_t^s, t, s) = \exp\left\{\int_t^s \chi(b(x_\tau, \tau)^T \lambda; \tau) d\tau\right\}$$

where x_t^s denotes the set of values of the function $x(\tau)$ corresponding to all τ in the interval $[t, s]$. Differentiating with respect to s and putting afterwards $s = t$ give

$$\chi(b(x_t, t)^T \lambda; t) = \left[\frac{\partial}{\partial s} h_u(\lambda; x_t^s, t, s)\right]_{s=t}. \tag{20}$$

We shall not consider here the questions of existence and uniqueness of the solutions of equations (7)–(9), but we confine ourselves to the observation that equation (7) represents ordinary homogeneous linear differential equation with unbounded operator in Banach space $C(R^{pn})$ of bounded continuous functions. Therefore the existence and uniqueness of solutions of (7)–(9) is established by the general theorems of the theory of differential equations in Banach spaces [7].

3. Case of linear equation

In the case of a linear equation (1):

$$dX = (a_0 + a_1 X)\,dt + b\,dW \tag{21}$$

a_0, a_1 and b being functions of t only equation (7) becomes

$$\frac{\partial g_n}{\partial t} = M\{i\,\lambda_n^{\mathrm{T}}(a_{0t_n} + a_{1t_n} X_{t_n}) + \chi(b_{t_n}^{\mathrm{T}}\lambda_n; t_n)\} \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} X_{t_k}\right\}$$

or

$$\frac{\partial g_n}{\partial t_n} = \lambda_n^{\mathrm{T}} a_{1t_n} \frac{\partial g_n}{\partial \lambda_n} + [i\,\lambda_n^{\mathrm{T}} a_{0t_n} + \chi(b_{t_n}^{\mathrm{T}}\lambda_n; t_n)]\,g_n \quad (n = 1, 2, \ldots) \tag{22}$$

$\partial/\partial\lambda_n$ being the matrix-column $[\partial/\partial\lambda_{n1} \ldots \partial/\partial\lambda_{np}]^{\mathrm{T}}$.

Hence all the finite-dimensional characteristic functions of the random process determined by a linear stochastic differential equation, each considered as a function of its last arguments λ_n, t_n, satisfy the same linear equation in partial derivatives of the first order

$$\frac{\partial g}{\partial t} = \lambda^{\mathrm{T}} a_1 \frac{\partial g}{\partial \lambda} + [i\,\lambda^{T} a_0 + \chi(b^{T}\lambda; t)]\,g. \tag{23}$$

To integrate (23) we write the corresponding set of ordinary differential equations

$$\frac{d\lambda}{dt} = -\,a_1^{\mathrm{T}}\lambda, \quad \frac{dg}{dt} = [i\,\lambda^{\mathrm{T}} a_0 + \chi(b^{\mathrm{T}}\lambda; t)]\,g. \tag{24}$$

Let $u(t, \tau)$ as a function of t for fixed τ be the integral of the homogeneous equation $du/dt = a_1 u$ satisfying the condition $u(\tau, \tau) = I$. Then the general integral of equations (24) is given by

$$\lambda = u(t, \alpha)^{-1\mathrm{T}} c_1,$$

$$g = c_2 \exp\left\{\int_{\alpha}^{t} [ic_1^{\mathrm{T}} u(\tau, \alpha)^{-1} a_0(\tau) + \chi(b(\tau)^{\mathrm{T}} u(\tau, \alpha)^{-1\mathrm{T}} c_1; \tau)]\,d\tau\right\}$$

where α is an arbitrary value of t. Solving with respect to c_1 and c_2 yields

$$c_1 = u(t, \alpha)^{\mathrm{T}} \lambda,$$

$$c_2 = g \exp\left\{-\int_{\alpha}^{t} \left[i \lambda^{\mathrm{T}} u(t, \alpha)\, u(\tau, \alpha)^{-1} a_0(\tau) + \chi(b(\tau)\, u(\tau, \alpha)^{-1^{\mathrm{T}}} u(t, \alpha)^{\mathrm{T}} \lambda; \tau)\right]\right\}.$$

Then taking into account that

$$u(t, \tau)\, u(\tau, \alpha) = u(t, \alpha)$$

the complete integral of equation (23) is determined by solving the equation

$$g \exp\left\{-\int_{\alpha}^{t} \left[i \lambda^{\mathrm{T}} u(t, \tau)\, a_0(\tau) + \chi(b(\tau)^{\mathrm{T}} u(t, \tau)^{\mathrm{T}} \lambda; \tau)\right] d\tau\right\} = \varphi(u(t, \alpha)^{\mathrm{T}} \lambda)$$

with respect to g, $\varphi(\mu)$ being an arbitrary function. As a result we get for $t \geq \alpha$

$$g(\lambda; t) = \varphi(u(t, \alpha)^{\mathrm{T}} \lambda) \exp\left\{\int_{\alpha}^{t} \left[i \lambda^{\mathrm{T}} u(t, \tau)\, a_0(\tau) + \chi(b(\tau)^{\mathrm{T}} u(t, \tau)^{\mathrm{T}} \lambda; \tau)\right] d\tau\right\}. \quad (25)$$

Putting $\lambda = \lambda_1$, $t = t_1$, $\alpha = t_0$ and using the initial condition (8) we obtain from (25) for $t_1 \geq t_0$

$$g_1(\lambda_1; t_1) = g_0(u(t_1, t_0)^{\mathrm{T}} \lambda_1) \exp\left\{\int_{t_0}^{t_1} \left[i \lambda_1^{\mathrm{T}} u(t_1, \tau)\, a(\tau) + \chi(b(\tau)^{\mathrm{T}} u(t_1, \tau)^{\mathrm{T}} \lambda_1; \tau)\right] d\tau\right\}. \quad (26)$$

Putting $\lambda = \lambda_n$, $t = t_n$, $\alpha = t_{n-1}$ and using the initial condition (9) we obtain from (25) for $t_n \geq t_{n-1}$

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n) = g_{n-1}(\lambda_1, \ldots, \lambda_{n-1} + u(t_n, t_{n-1})^{\mathrm{T}} \lambda_n; t_1, \ldots, t_{n-1}) \times$$

$$\times \exp\left\{\int_{t_{n-1}}^{t_n} \left[i \lambda_n^{\mathrm{T}} u(t_n, \tau)\, a(\tau) + \chi(b(\tau)^{\mathrm{T}} u(t_n, \tau)^{\mathrm{T}} \lambda_n; \tau)\right] d\tau\right\} \quad (n = 2, 3, \ldots). \quad (27)$$

Finally we find from (26) and (27) by induction for any $t_0 \leq t_1 \leq \ldots \leq t_n$

$$g_n(\lambda_1, \ldots, \lambda_n; t_1, \ldots, t_n) = g_0\left(\sum_{k=1}^{n} u(t_k, t_0)\, \lambda_k^{\mathrm{T}}\right) \exp\left\{i \sum_{k=1}^{n} \lambda_k^{\mathrm{T}} \int_{t_0}^{t_k} u(t_k, \tau)\, a(\tau)\, d\tau + \right.$$

$$\left. + \sum_{k=1}^{n} \int_{t_{k-1}}^{t_k} \chi\left(b(\tau)^{\mathrm{T}} \sum_{l=k}^{n} u(t_l, \tau)^{\mathrm{T}} \lambda_l; \tau\right) d\tau\right\} \quad (n = 1, 2, \ldots). \quad (28)$$

This formula gives explicit expressions for all the finite-dimensional characteristic functions of the process $X(t)$ determined by equation (21) and satisfying $X(t_0) = X_0$. We have thus proved the following theorem.

Theorem 3. If the process W *in* (21) *satisfies the conditions of Theorem 2, then the finite-dimensional characteristic functions of the process* X *determined by equation* (21) *are given by* (28).

4. Application to problems of statistical analysis of systems

Formula (28) gives a complete and exact statistical description of the output of any linear system described by ordinary linear differential equations since the random disturbances in such systems may practically always be considered as random processes determined by linear stochastic differential equations.

In the case of non-linear systems described by differential equations the complete and exact statistical description of the output may be obtained in principle by solving equations (7) with the initial conditions (8) and (9). But the solution of equations of the type (7) in this case is a very difficult problem. We do not know a method giving exact solutions of equations (7)–(9) in the general case. Only in a trivial case of a linear function $\chi(\mu; t) = \chi_1(t)i\mu$ equations (7)–(9) can be solved in a closed form for any functions $a(x, t)$, $b(x, t)$ [1]. But this case represents in fact the case where only the initial value X_0 of the process $X(t)$ is random and there are no random disturbances acting on the system (i.e. the covariance function of the white noise $V(t)$ is zero, $v = 0$, and only its expectation $m'(t) = \chi_1(t)$ may differ from zero).

Thus the problem of approximate solution of equations (7)–(9) arises. The first method for approximate solution of equation (7) with the initial condition (8) for $g_1(\lambda; t)$ was proposed by the author in the early forties [1]. This method is based on the approximation of the unknown characteristic function $g_1(\lambda; t)$ by the normal characteristic function with indefinite expectation and covariance matrix. Substituting this normal characteristic function into (7) with $n = 1$ and (8) yields ordinary differential equations with the respective initial conditions for the expectation and covariance matrix of the random variable $X(t)$. In the case of $b(x, t)$ independent of x this method gives the same equations for $MX(t)$ and MX^0X^{0T} as the statistical linearization of the function $a(x, t)$ with respect to x by the second method of Kazakov [8]. This method was generalized by Demukh, who proposed to approximate $g_1(\lambda; t)$ by any known characteristic function depending on a finite number of indefinite parameters. Substituting this characteristic function into (7) with $n = 1$ and (8) yields ordinary differential equations with the respective initial conditions for all indefinite parameters [9]. Dashevsky developed this method approximating $\ln g_1(\lambda; t)$ by a truncated power series in λ and elabor-

ated artificial intelligence programs for automatic derivation and solution of ordinary differential equations for the respective moments of $X(t)$ for the case of polynomial functions $a(x, t)$ and $b(x, t)$ (with respect to x) [10].

The approximate methods for solution of equation (7) for $g_1(\lambda; t)$ developed by Demukh and Dashevsky may obviously be extended to equation (7) for $n \geq 2$. These methods permit in principle to find solutions of equations of the form (7)–(9) with any degree of accuracy. In fact any characteristic function satisfying some fairly general conditions may be approximated with any degree of accuracy by some truncated orthogonal expansion, e.g. by a truncated series of Gram–Charlier type. The more terms of such an expansion are taken, the higher will be the accuracy. Hence taking a sufficiently large number of terms of such an expansion we may obtain the solution of equation (7) with any degree of accuracy.

Thus we have approximate methods for the statistical analysis of any non-linear systems described by ordinary differential equations. These methods enable us to find statistical characteristics of the outputs of such non-linear systems with any degree of accuracy.

5. Application to problems of non-linear filtering

Many problems of non-linear filtering may be reduced to the problem of estimating a set of components of a vector random process determined by a stochastic differential equation using observations of some other components of this process. Such is the problem of estimating the state vector Y of a system described by a stochastic differential equation of the form

$$dY = \varphi(Y, t)\, dt + \psi(Y, t)\, dW \tag{29}$$

using continuous noisy measurements of some vector function of Y and t. The result of these measurements represents a random process $X(t)$ of the form

$$X = \varphi_1(Y, t) + \psi_1(Y, t)U \tag{30}$$

where the noise $U = U(t)$ is a random function generally independent of $W = W(t)$. If U represents a white noise in the strict sense (i.e. the weak mean square derivative of a process with independent increments), then introducing the random processes

$$Z(t) = \int_{t_0}^{t} X(\tau)\, d\tau, \quad W_1(t) = \int_{t_0}^{t} U(\tau)\, d\tau$$

and denoting the compound vector process $[W^{\mathrm{T}} W^{\mathrm{T}}]^{\mathrm{T}}$ again by W, equations

(29) and (30) become a special case of the set of stochastic differential equations

$$\begin{aligned} dY &= \varphi(Y, Z, t)\,dt + \psi(Y, Z, t)\,dW, \\ dZ &= \varphi_1(Y, Z, t)\,dt + \psi_1(Y, Z, t)\,dW. \end{aligned} \tag{31}$$

The problem is to find an estimate $\hat{Y} = \hat{Y}(t)$ of the process $Y = Y(t)$ for any $t > t_0$ using continuous observations of the process Z in the interval $[t_0, t]$.

The problem of finding an absolutely optimal estimate $\hat{Y}$ of the process Y using some statistical criterion, e.g. mean square error criterion, is extremely difficult. All known methods for the solution of this problem lead to tremendous and often unrealizable computations using observations. This stimulated the author to develop a new theory of conditionally optimal estimation of the process Y constrained by the requirement that the estimate $\hat{Y}$ satisfy some differential equations [11, 12]. This constraint provides simple computation of estimates using observations.

Thus the problem arises to find an estimate $\hat{Y}$ of the process Y yielding the conditional minimum of the mean square error $M\,|\hat{Y} - Y|^2$ for any $t > t_0$ under the condition that $\hat{Y}$ satisfy a differential equation of the form

$$d\hat{Y} = \alpha\xi(Z, \hat{Y}, t)\,dt + \beta\eta(Z, \hat{Y}, t)\,dZ + \gamma\,dt \tag{32}$$

where the functions ξ and η are given and the functions $\alpha = \alpha(t)$, $\beta = \beta(t)$, $\gamma = \gamma(t)$ may be varied.

We confine ourselves here to equation (32) of the first order. In [11, 12] conditionally optimal estimate $\hat{Y}$ satisfying a differential equation of any given order p:

$$d\hat{Y} = \hat{Y}'\,dt,\ d\hat{Y}' = \hat{Y}''\,dt, \ldots, d\hat{Y}^{(p-2)} = \hat{Y}^{(p-1)}\,dt,$$

$$d\hat{Y}^{(p-1)} = \alpha\xi(Z, \hat{Y}, \ldots, \hat{Y}^{(p-1)}, t)\,dt + \beta\eta(Z, \hat{Y}, \ldots, \hat{Y}^{(p-1)}, t)\,dZ + \gamma\,dt$$

is determined. Some conditions on smoothness of the functions φ, ψ, φ_1, ψ_1 depending on the distribution of W are necessary in the cases of $p > 1$.

We assume that all the random processes considered are finite-dimensional real vector-valued random functions and respectively all non-random functions represent finite-dimensional real vector or matrix-valued functions. If for instance $Y(t), Z(t)$ and $W(t)$ represent respectively a R^n-valued, R^m-valued and R^q-valued random processes, then $\varphi(y, z, t)$ and $\varphi_1(y, z, t)$ are known non-random vector-valued functions mapping $R^n \times R^m \times R$ into R^n and R^m respectively, $\psi(y, z, t)$ and $\psi_1(y, z, t)$ known non-random matrix-valued func-

tions mapping $R^n \times R^m \times R$ into R^{nq} and R^{mq} respectively. In accordance with this the random-process $\hat{Y}(t)$ is R^n-valued, $\xi(z, \hat{y}, t)$ is a given non-random vector-valued function mapping $R^m \times R^n \times R$ into R^r and $\eta(z, \hat{y}, t)$ a given non-random matrix-valued function mapping $R^m \times R^n \times R$ into R^{rm}. The number r of rows of the matrices ξ and η may be taken arbitrarily. The more is r the more will be the accuracy of the estimate $\hat{Y}$. Increasing r we may in principle obtain an approximation to optimal functions ξ and η with any degree of accuracy.

The problem is to find optimal functions α, β, γ providing minimum of $M|\hat{Y}_t - Y_t|^2$. For this purpose we write from (31) and (32)

$$\hat{Y}_s - \hat{Y}_t = (\alpha\xi_t + \gamma)(s-t) + \beta\left[\eta_t \varphi_{1t}(s-t) + \int_t^s \eta_\tau \psi_{1\tau} dW_\tau\right] + o_p(s-t) \quad (33)$$

where $f_t = f(t)$, $\xi_t = \xi(Z_t, \hat{Y}_t, t)$, $\eta_t = \eta(Z_t, \hat{Y}_t, t)$, $\varphi_{1t} = \varphi_1(Y_t, Z_t, t)$, $\psi_{1t} = \psi_1(Y_t, Z_t, t)$. Taking into account that $\hat{Y}_s$ is an optimal estimate of Y_s for infinitesimal $s - t$ if and only if the right member of (33) represents a linear mean square regression of $Y_s - \hat{Y}_t$ on the random variables ξ_t and $\eta_t \varphi_{1t}(s-t) + \int_t^s \eta_\tau \psi_{1\tau} dW_\tau$ for infinitesimal $s - t$ and using the standard technique of linear regression theory we find [12, 13]

$$\alpha = (\varkappa_{01} - \beta\varkappa_{21})\varkappa_{11}^{-1}, \quad \beta = \varkappa_{02}\varkappa_{22}^{-1}, \quad (34)$$

$$\gamma = m_0 - \alpha m_1 - \beta m_2 \quad (35)$$

where

$$\begin{aligned} &\varkappa_{01} = M(\varphi_t - M\varphi_t)\xi_t^{\mathrm{T}}, \ \varkappa_{02} = M(Y_t - \hat{Y}_t)\varphi_{1t}^{\mathrm{T}}\eta_t^{\mathrm{T}} + M\psi_t\nu_t\psi_{1t}^{\mathrm{T}}\eta_t^{\mathrm{T}}, \\ &\varkappa_{11} = M(\xi_t - M\xi_t)\xi_t^{\mathrm{T}}, \ \varkappa_{12} = \varkappa_{21}^{\mathrm{T}} = M(\xi_t - M\xi_t)\varphi_{1t}^{\mathrm{T}}\eta_t^{\mathrm{T}}, \\ &\varkappa_{22} = M\eta_t\psi_{1t}\nu_t\psi_{1t}^{\mathrm{T}}\eta_t^{\mathrm{T}}, \ m_0 = M\varphi_t, \ m_1 = M\xi_t, \ m_2 = M\eta_t\varphi_{1t}. \end{aligned} \quad (36)$$

In deriving these formulae we assume that $W = W(t)$ represents a second-order process with independent increments with zero expectation and the covariance function given by

$$K_w(t_1, t_2) = k_0 + \int_{t_0}^{\min(t_1, t_2)} \nu(\tau)\, d\tau, \quad (37)$$

and that the functions $\varphi, \varphi_1, \psi, \psi_1, \xi$ and η satisfy the conditions of the form (3)–(6) sufficient for the existence and uniqueness of the solution of the set of stochastic differential equations (31) and (32).

The expectations in (36) are completely determined by the one-dimensional distribution of the random process $[Y^{\mathrm{T}} Z^{\mathrm{T}} \hat{Y}^{\mathrm{T}}]^{\mathrm{T}}$. To find this distribution equation (7) corresponding to the set of equations (31), (32) may be used. This equation is

$$\frac{\partial}{\partial t} g_1(\lambda_1, \lambda_2, \mu; t) = M\{i\lambda_1^{\mathrm{T}} \varphi(Y, Z, t) + i\lambda_2^{\mathrm{T}} \varphi_1(Y, Z, t) +$$

$$+ i\mu^{\mathrm{T}} [\alpha\xi(Z, \hat{Y}, t) + \beta\eta(Z, \hat{Y}, t)\varphi_1(Y, Z, t) + \gamma] +$$

$$+ \chi(\psi(Y, Z, t)^{\mathrm{T}} \lambda_1 + \psi_1(Y, Z, t)^{\mathrm{T}} \lambda_2 +$$

$$+ \psi_1(Y, Z, t)^{\mathrm{T}} \eta(Z, \hat{Y}, t)^{\mathrm{T}} \beta^{\mathrm{T}} \mu; t)\} \exp\{i\lambda_1^{\mathrm{T}} Y + i\lambda_2^{\mathrm{T}} Z + i\mu^{\mathrm{T}} \hat{Y}\}. \qquad (38)$$

The initial condition (8) has in this case the form

$$g_1(\lambda_1, \lambda_2, \mu; t_0) = g_0(\lambda_1, \lambda_2, \mu) \qquad (39)$$

$g_0(\lambda_1, \lambda_2, \mu)$ being the characteristic function of the random variables $Y_0 = Y(t_0)$, $Z_0 = Z(t_0)$, $Y_0 = Y(t_0)$ which must be given (or taken to some extent arbitrarily if no prior information about $Y_0, Z_0, \hat{Y}_0$ can be obtained).

Equation (32) with optimal α, β, γ yields unbiased estimate $\hat{Y}_t$ of Y_t for any $t > t_0$ if $M\hat{Y}_0 = MY_0$.

Confidence regions for Y_t may be found knowing $g_1(\lambda_1, \lambda_2, \mu; t)$ since the joint distribution of Y_t and $\hat{Y}_t$ is completely determined by $g_1(\lambda_1, \lambda_2, \mu; t)$.

Equations (34), (35) and (38) represent the set of four equations determining $g_1, \alpha, \beta, \gamma$ and thus solving the problem.

Thus using equation (7) determining the one-dimensional characteristic function of the process determined by a stochastic differential equation we may solve a lot of problems of nonlinear filtering under fairly general conditions.

In the special case where

$$\varphi(y, z, t) = a(t)\, y + a_1(t)\, z + a_0(t),$$

$$\varphi_1(y, z, t) = b(t)\, y + b_1(t)\, z + b_0(t),$$

$$\psi(y, z, t) = \psi(t), \quad \psi_1(y, z, t) = \psi_1(t),$$

$$\xi(z, \hat{y}, t) = [z^{\mathrm{T}} \hat{y}^{\mathrm{T}}]^{\mathrm{T}}, \quad \eta(z, \hat{y}, t) = I,$$

the method outlined gives a generalization of the well-known linear filtering of Kalman and Bucy obtained by Liptser and Shiryayev [14, 15].

The extension of the above theory to the case where the observed process is given by

$$X = \chi(Y, U, t)$$

where $U = U(t)$ is determined by a stochastic differential equation was given by I. D. Siluyanova [16].

The case in which the functions $\varphi, \psi, \varphi_1, \psi_1$ in (31) depend on a finite-dimensional unknown parameter, θ, may also be reduced to equations of the form (31) with completely determined functions $\varphi, \psi, \varphi_1, \psi_1$ by introducing the random process $\Theta(t) = \theta$ satisfying the equation $\dot{\Theta} = 0$ and replacing Y by $[Y^{\mathrm{T}}\, \Theta^{\mathrm{T}}]^{\mathrm{T}}$. Then the above method will give simultaneous estimates of the process Y and parameter θ.

6. Application to problems of prediction

Now we shall consider the problem of extrapolation of the process Y using observations of the process Z.

The problem is to find optimal functions α, β, γ in (32) in such a way that the mean square error $M|\widehat{Y}_t - Y_{t+T}|^2$ be minimal for a given $T > 0$. We shall solve this problem assuming that the functions φ and ψ in (31) are independent of Z, and W, ψ and ψ_1 are representable in the form $W = [W_1^{\mathrm{T}}\, W_2^{\mathrm{T}}]^{\mathrm{T}}$, $\psi(y, t) = [\psi'(y, t)\, 0]$, $\psi_1(y, z, t) = [0\, \psi_1'(y, z, t)]$ where $W_1 = W_1(t)$, $W_2 = W_2(t)$ are independent so that $\psi W = \psi' W_1$ and $\psi_1 W = \psi_1' W_2$. The random variable Y_τ is independent of W_{2t} for any τ, t in this case. $\widehat{Y}_s$ is an optimal estimate of Y_{s+T} for infinitesimal $s - t$ if and only if the right member of (33) represents a linear mean square regression of $Y_{s+T} - Y_t$ on the random variables ξ_t and $\eta_t \varphi_{t1}(s - t) + \int_t^s \eta_\tau \psi_{1\tau} dW_\tau$ for infinitesimal $s - t$. Using standard method of linear regression theory and taking into account that $W_s - W_t$ is independent of $Y_t, Z_t, \widehat{Y}_t$ for any t and $s > t$ and that

$$M(W_s - W_t)(W_s^{\mathrm{T}} - W_t^{\mathrm{T}}) = \nu_t(s - t) + o(s - t)$$

we obtain the equations

$$M(Y_{s+T} - MY_{s+T} - \widehat{Y}_t + M\widehat{Y}_t)\, \xi_t^{\mathrm{T}} = \alpha M(\xi_t - M\xi_t)\, \xi_t^{\mathrm{T}}(s - t) + \\ + \beta M \eta_t \varphi_{1t}(\xi_t^{\mathrm{T}} - M\xi_t^{\mathrm{T}})(s - t) + o(s - t), \tag{40}$$

$$M(Y_{s+T} - MY_{s+T} - \widehat{Y}_t + M\widehat{Y}_t)\, \varphi_{1t}^{\mathrm{T}} \eta_t^{\mathrm{T}}(s - t) = \\ = \beta M \eta_t \psi_{1t} \nu_t \psi_{1t}^{\mathrm{T}} \eta_t^{\mathrm{T}}(s - t) + o(s - t), \tag{41}$$

$$M(Y_{s+T} - \widehat{Y}_t) = (\alpha M \xi_t + \beta M \eta_t \varphi_{1t} + \gamma)(s - t) + o(s - t). \tag{42}$$

Putting $s \downarrow t$ in (42) yields

$$MY_{t+T} = M\hat{Y}_t. \tag{43}$$

Using (31) and (43) and taking into account that $M\psi_{t+T}(W_{s+T} - W_{t+T}) = 0$ for any t, $s > t$ we rewrite (42) in the form

$$M\varphi_{t+T}(s-t) = (\alpha M\xi_t + \beta M\eta_t\varphi_{1t} + \gamma)(s-t) + o(s-t)$$

yielding

$$M\varphi_{t+T} = \alpha M\xi_t + \beta M\eta_t\varphi_{1t} + \gamma. \tag{44}$$

Putting $s \downarrow t$ in (41) we obtain in virtue of (43)

$$M(Y_{t+T} - \hat{Y}_t)\varphi_{1t}^{\mathrm{T}}\eta_t^{\mathrm{T}} = \beta M\eta_t\psi_{1t}\nu_t\psi_{1t}^{\mathrm{T}}\eta_t^{\mathrm{T}}. \tag{45}$$

Putting $s \downarrow t$ in (40) and using (43) we obtain

$$MY_{t+T}\xi_t^{\mathrm{T}} = M\hat{Y}_t\xi_t^{\mathrm{T}}. \tag{46}$$

Using (43) and (46) we may rewrite (40) in the form

$$M(Y_{s+T} - MY_{s+T} - Y_{t+T} + MY_{t+T})\xi_t^T = \alpha M(\xi_t - M\xi_t)\xi_t^T(s-t)$$
$$+ \beta M\eta_t\varphi_{1t}(\xi_t^{\mathrm{T}} - M\xi_t^T)(s-t) + o(s-t)$$

or using (31) and putting $s \downarrow t$

$$M(\varphi_{t+T} - M\varphi_{t+T})\xi_t^{\mathrm{T}} = \alpha M(\xi_t - M\xi_t)\xi_t^{\mathrm{T}} + \beta M\eta_t\varphi_{1t}(\xi_t^{\mathrm{T}} - M\xi_t^{\mathrm{T}}). \tag{47}$$

Equations (45), (47) and (44) give formulae (34) and (35) for optimal α, β, γ where

$$\varkappa_{01} = M(\varphi_{t+T} - M\varphi_{t+T})\xi_t^{\mathrm{T}},$$
$$\varkappa_{02} = M(Y_{t+T} - \hat{Y}_t)\varphi_{1t}^{\mathrm{T}}\eta_t^{\mathrm{T}}, \quad m_0 = M\varphi_{t+T} \tag{48}$$

and $\varkappa_{11}$, $\varkappa_{12} = \varkappa_{21}^{\mathrm{T}}$, $\varkappa_{22}$, m_1 and m_2 are determined by (36).

The espectations in (48) are determined by the two-dimensional distribution of the random process $[Y^{\mathrm{T}}Z^{\mathrm{T}}\hat{Y}^{\mathrm{T}}]^{\mathrm{T}}$. To find this distribution equation (7) for the set of stochastic differential equations (31) and (32) for $n = 2$ may be used. This equation is

$$\frac{\partial}{\partial s}g_2(\lambda_1, \lambda_2, \mu, \lambda', \lambda_2', \mu'; t, s) = M\{i\lambda_1'^{\mathrm{T}}\varphi(Y_s, s) + i\lambda_2'^{\mathrm{T}}\varphi_1(Y_s, Z_s, s) +$$
$$+ i\mu'^{\mathrm{T}}[\alpha_s\xi(Z_s, \hat{Y}_s, s) + \beta_s\eta(Z_s, \hat{Y}_s, s)\varphi_1(Y_s, Z_s, s) + \gamma_s] +$$
$$+ \chi(\psi(Y_s, s)^{\mathrm{T}}\lambda_1' + \psi_1(Y_s, Z_s, s)^{\mathrm{T}}\lambda_2' + \psi_1(Y_s, Z_s, s)^{\mathrm{T}}\eta(Z_s, \hat{Y}_s, s)^{\mathrm{T}}\beta_s^{\mathrm{T}}\mu'; s)\} \times$$
$$\times \exp\{i\lambda_1^{\mathrm{T}}Y_t + i\lambda_2^{\mathrm{T}}Z_t + i\mu^{\mathrm{T}}\hat{Y}_t + i\lambda_1'^{\mathrm{T}}Y_s + i\lambda_2'^{\mathrm{T}}Z_s + i\mu'^{\mathrm{T}}\hat{Y}_s\}. \tag{50}$$

The initial condition (9) has in this case the form

$$g_2(\lambda_1, \lambda_2, \mu, \lambda_1', \lambda_2', \mu'; t, t) = g_1(\lambda_1 + \lambda_1', \lambda_2 + \lambda_2', \mu + \mu'; t). \tag{51}$$

Equations (34), (35), (38), (39), (50) and (51) give an exact solution of the problem of conditionally optimal extrapolation of the process Y.

Similarly using equations (7)–(9) for $n = 1, 2, \ldots, N + 1$ the problem of simultaneous estimation of $Y_{t+T_1}, \ldots, Y_{t+T_N}$ for any $t > t_0$ may be solved $\hat{Y}$ in (32) representing a R^{nN}-valued random function in this case.

All the approximate methods for solving equations (7)–(9) discussed in Section 4 may be used to solve equations (38), (39), (50) and (51) while designing conditionally optimal estimating algorithms.

7. Conclusion

We obtained here the equations for the finite-dimensional characteristic functions of a random process determined by a stochastic differential equation. In the case of a linear stochastic differential equation these equations represent homogeneous linear equations in partial derivatives of the first order. Integrating these equations by standard techniques we obtained explicit expressions for the finite-dimensional characteristic functions of a process determined by a linear stochastic differential equation. For the case of a non-linear stochastic differential equation some approximate methods for solving the equations determining the finite-dimensional characteristic functions were discussed.

The equations for the finite-dimensional characteristic functions yield new methods for the statistical analysis of linear and non-linear systems. These methods are the first known methods which enable us to find the finite-dimensional distributions of the outputs of linear and non-linear systems.

The equations obtained give also the solution of problems of conditionally optimal estimation of the state vector and unknown parameters of non-linear systems. These methods may serve as a tool for designing mathematical models for a wide class of systems using experimental data and for solving a lot of control problems under uncertainty.

Using the equations for the finite-dimensional characteristic functions we developed a method for solving problems of conditionally optimal extrapolation (prediction) of the state vector of a system described by a stochastic differential equation. This is, as the author knows, the first constructive solution of prediction problems for non-linear systems.

The main characteristic feature of our new methods of estimation and extrapolation as compared with all other known methods is computational

simplicity of estimating algorithms. To obtain the estimates using observation only the integration of the differential equations determining the estimates is necessary. All complicated and tedious calculations necessary to find optimal coefficients in these equations deal only with prior data and can therefore be fulfilled beforehand while designing estimating and extrapolating algorithms and filters. This is due to the fact that the estimates given by our theory of conditionally optimal estimation and extrapolation are not absolutely optimal in general but only conditionally optimal constrained by the requirement that they satisfy certain differential equations. Owing to this the new methods are adequate for on-line identification processes and thus can be used directly in control systems.

References

1. *Pugachev, V. S.*, Random functions determined by ordinary differential equations. Trudy Akademii imeni Zhukovskogo, **118** (1944) (in Russian).
2. *Pugachev, V. S.*, Application of Markov processes theory to the analysis of accuracy of automatic systems. Izvestiya Akademii nauk SSSR, Energetika i avtomatika, **3**, 46–57 (1961) (in Russian).
3. *Pugachev, V.S.*, Statistical methods in automatic control. Trans. 2nd IFAC Congress, 27 Aug.–4 Sept., 1963. Butterworth (1965).
4. *Pugachev, V. S.*, A survey of statistical control theory. Problems of Control and Information Theory, Jubilee Number, 13–32 (1975).
5. *Pugachev, V. S.*, Finite-dimensional distributions of the processes determined by stochastic differential equations, and extrapolation of such processes. Doklady Akademii nauk SSSR, **251**, *1*, 40–43 (1980) (in Russian).
6. *Pugachev, V. S.*, Stochastic systems, No. *11*. "Nauka", Moscow, 1974.
7. *Ladas, G. E., Lakshmikantham, V.*, Differential Equations in Abstract Spaces. Academic Press, New York, London, 1972.
8. *Kazakov, I. Ye.*, An approximate probabilistic analysis of the accuracy of essentially non-linear systems. Avtomatika i telemekhanika, **17**, *5*, 385–409 (1956) (in Russian).
9. *Demukh, V. I.*, An approximate method for the analysis of the accuracy of non-linear systems. Avtomatika i telemekhanika, **6**, 1021–1025 (1965) (in Russian).
10. *Dashevskii, M. L.*, Algorithmization of the semiinvariant method of studying non-linear dynamic systems. Problems of Control and Information Theory, **7**, *5*, 305–316 (1978).
11. *Pugachev, V. S.*, Estimation of variables and parameters in stochastic systems described by differential equations. Doklady Akademii nauk SSSR, **241**, *5*, 1031–1034 (1978) (in Russian), English translation: Soviet Math., Dokl., **19**, *4*, 967–971 (1978).
12. *Pugachev, V. S.*, Estimation of the state and parameters of continuous non-linear systems. Avtomatika i telemekhanika, **6**, 63–79 (1979) (in Russian).
13. *Pugachev, V. S.*, Estimation of Markov processes. Time Series. Proc. Internat. Conference. Nottingham, March 1979, 389–400. North Holland, Amsterdam–New York–London, 1980.
14. *Liptser, R. S., Shiryayev, A. N.*, Non-linear filtering of diffusion Markov processes. Trudy Math. inst. imeni V. A. Steklova Akademii nauk SSSR, **104**, 135–180 (in (Russian).
15. *Liptser, R. S., Shiryayev, A. N.*, Statistics of Random Processes. "Nauka", Moscow, 1974 (in Russian).
16. *Siluyanova, I. D.*, Estimation of variables and parameters in stochastic systems described by differential equations in a singular case. Doklady Akademii nauk SSSR, **250**, *3*, 566–569 (1980) (in Russian).

Конечномерные распределения случайного процесса, определяемого стохастическим дифференциальным уравнением, и их приложение к задачам управления

В. С. ПУГАЧЕВ

(Москва)

Ставится задача нахождения всех конечномерных распределений случайного процесса $X(t)$, определяемого стохастическим дифференциальным уравнением Ито (1) при начальном условии $X(t_0) = X_0$, где $X = X(t)$ и $W = W(t)$ — случайные процессы со значениями в конечномерных пространствах R^p и R^q соответственно, причем W — процесс с независимыми приращениями, $a(x, t)$ и $b(x, t)$ — неслучайные функции, отображающие $R^p \times R$ в R^p и R^{pq} соответственно, X_0 — p-мерная случайная величина, независимая от $W(s) - W(t)$ при любых $t \geq t_0$ и $s > t$. Дается новый метод решения этой задачи, основанный на уравнениях для конечномерных характеристических функций процесса X [1–5]. В связи с этим доказывается теорема 1: *Если W — процесс с независимыми приращениями второго порядка с математическим ожиданием и ковариационной функцией вида* (2), *одномерная характеристическая функция которого $h_1(\mu; t)$ дифференцируема по t, а функции $a(x, t)$ и $b(x, t)$ удовлетворяют условиям* (3)–(6), *при некоторых $k_1, k_2, k_3, k_4, c_1, c_2 > 0$ для любых случайных величин Z и Z' с конечными вторыми моментами, то конечномерные характеристические функции $g_n(\lambda_1, \ldots, \lambda_n, t_1, \ldots, t_n)$ $(n = 1, 2, \ldots)$ процесса X в случае, когда X_0 имеет конечный момент второго порядка, удовлетворяют уравнением* (7) *при начальных условиях* (8) *и* (9), *где $\chi(\mu; t)$ определяется формулой* (10). Эта теорема обобщается на случай процесса W со случайными разрывами первого рода в фиксированных точках τ_k при выполнении условий (2)–(6) при всех t, отличных от τ_k (теорема 2).

В случае линейного относительно X, dX, dW стохастического дифференциального уравнения (21) уравнения (7) сводятся к одному и тому же линейному уравнению в частных производных первого порядка (23), которое интегрируется стандартным методом. В результате получаются явные выражения (28) для всех конечномерных характеристических функций процесса X (теорема 3).

В качестве первого приложения к теории управления рассматривается задача статистического анализа линейных и нелинейных систем, описываемых обыкновенными дифференциальными уравнениями, при условии, что все действующие на систему случайные возмущения определяются стохастическими дифференциальными уравнениями (таковы, например, все стационарные случайные функции с рациональными спектральными плотностями, не содержащими целой части). Для таких систем впервые дается метод определения конечномерных распределений выходных сигналов. В случае линейных систем формулы (28) дают точное решение этой задачи. В случае нелинейных систем указываются некоторые приближенные методы решения уравнений (7) для конечномерных характеристических функций выходных сигналов. Излагается метод аппроксимации конечномерных распределений нормальными, предложенный автором для одномерного распределения в [1], и обобщения этого метода, данные для одномерных распределений в [9, 10]. Показывается, что эти методы с соответствующими изменениями могут служить и для решения уравнения (7) для любых конечномерных характеристических функций. Указывается, что эти методы в принципе позволяют находить статистические характеристики выходных сигналов систем с любой степенью точности.

В качестве второго приложения рассматриваются проблемы нелинейной фильтрации. Дается краткое изложение разработанной автором теории условно оптимального оценивания вектора состояния и неизвестных параметров системы, описываемой дифференциальным уравнением, при условии, что оценки удовлетворяют дифференциальному уравнению данного вида (32) [11–13]. В случае линейных систем из этой теории вытекают как частные случаи хорошо известная теория линейной фильтрации Калмана и Бьюси и ее обобщение, данное Липцером и Ширяевым. Распространение новой теории нелинейной фильтрации на случай помехи, определяемой стохастическим дифференциальным уравнением и нелинейно входящей в уравнение наблюдения, дано Силуяновой [16].

В качестве третьего приложения рассматривается задача прогнозирования вектора состояния нелинейной системы. Впервые подробно излагается разработанная автором теория условно оптимальной экстраполяции вектора состояния системы, описываемой

дифференциальным уравнением, при условии, что оценка будущего состояния системы удовлетворяет дифференциальному уравнению данного вида (32) [5].

Отличительной особенностью новых методов нелинейной фильтрации и прогнозирования является простота вычисления оценок. Благодаря отказу от требования безусловной оптимальности оценок и ограничению их требованием условной оптимальности при условии, что они должны удовлетворять дифференциальному уравнению определенного вида, достигается возможность вычислять оценки по результатам наблюдений в натуральном масштабе времени непосредственно в процессе работы системы. А все сложные вычисления, необходимые для нахождения оптимальных коэффициентов дифференциальных уравнений, определяющих оценки, основаны только на априорной информации и не используют результатов наблюдений. Вследствие этого они могут выполняться заранее в процессе проектирования алгоритмов и фильтров.

Вследствие простоты вычисления оценок новая теория условно оптимальной фильтрации и прогнозирования может эффективно применяться для решения различных задач построения математических моделей сложных систем (идентификации) и задач адаптивного управления в условиях неопределенности.

В. С. Пугачев
Институт проблем управления
СССР, 117342 Москва ГСП-7,
Профсоюзная, 65

Problems of Control and Information Theory, Vol. 10 (2), pp. 115—127 (1981)

DYNAMICAL RECOGNITION OF RANDOM PROCESSES AND SYSTEMS IN CONTINUOUS TIME

E. K. SHPILEWSKI

(Vilnius)

(Received October 1, 1980)

On the basis of a unique approach, constructive methods for the recognition of processes in continuous time according to direct and indirect measurements are considered. The principal equations for statistics are obtained which enable us to construct sequential non-randomized decision rules for classification of fully and partially observable multivariate processes and systems in current time, described by stochastic differential equations. The classification of discontinuous Markov processes is considered. The classification of processes according to simple and composite observations under the condition noise is obtained. Partial differential equations for the Bayes error recognition are obtained and the Cauchy problem is formulated.

1. Introduction

A number of common problems connected with the development and functioning of automatic control systems can be successfully solved only if we know how to classify correctly or recognize the observable processes.

An essential peculiarity of process recognition is the presence of time and dynamics. In most of the practical problems there arises the necessity to recognize multivariate processes in current or real time. Signals characterizing the technical state of objects have a significant noise ratio. Processes, especially technological ones, are often almost inaccessible for direct measurement and check and therefore one has to make decisions by indirect observations. That is why data processing of remote measurement and recognition of partially observable objects is rather topical.

In the paper, on the basis of the theory of general and conditional Markov processes [1, 2] the following approach is being developed. Since observable processes are generated by somewhat material systems, the models of the latter can be used to describe signals. Dynamic stochastic systems described by stochastic differential, recurrent, difference equations are taken as such models.

1. Statement of the problem

Let us consider a general statement of the problem of recognition of partially observable random processes. Let $(\Omega, \mathfrak{F})$ be a measurable space,

given the $(k + l)$-dimensional measurable process

$$X, Y = (X_t, Y_t, 0 \leq t \leq T), \tag{1.1}$$

where X is an unobservable k-dimensional component and Y is an observable l-dimensional one. With respect to the process X, Y, N alternative hypotheses

$$H = \{h_1, h_2, \ldots, h_N\} \tag{1.2}$$

are introduced, forming a full group of events, i.e. if p_i is an a priori probability of the hypothesis h_i then

$$p_i > 0, \sum_{i=1}^{N} p_i = 1.$$

N probabilistic measures P_i on the measurable space $(\Omega, \mathfrak{F})$ are linked with the hypotheses. Then a random process

$$X_i Y_i = \{X_i(t), Y_i(t), 0 \leq t \leq T\} \tag{1.3}$$

corresponding to the hypothesis h_i is defined on the probabilistic space $(\Omega, \mathfrak{F}, P_i)$.

The problem of classification of random processes rests on decision making by the observable sample

$$y_0^t = \{y_\tau, 0 \leq t \leq T\}, \tag{1.4}$$

that the hypothesis h_i holds, i.e. on the reference of sample (1.4) to one of N classes (1.2). X being absent in (1.1) we have the classification problem of the fully observable process Y.

The Bayesian decision rule [3], minimizing the average risk, lies in the reference of the observable sample (1.4) to the class h_k for which

$$h_k: \sum_{i=1}^{N} l_{ik} p_i(t) = \min_j \sum_{i=1}^{N} l_{ij} p_i(t), \tag{1.5}$$

where l_{ij} is a given loss function, $p_i(t)$ is an a posteriori probability of the class h_i. Decision rules analogous to (1.5) can be obtained by using the statistics of the ratio of a posteriori probabilities of the classes $v_{ij}(t) = p_i(t)/p_j(t)$

$$h_k: \sum_{i=1}^{N} l_{ik} v_{i1}(t) = \min \sum_{i=1}^{N} l_{ij} v_{i1}(t),$$

the Radon–Nikodym derivative is $\Lambda_{ij}(t) = v_{ij}(t)\, p_j/p_i$ of the likelihood function $L_k(t)$, where $\Lambda_{ij}(t) = L_i(t)/L_j(t)$ [4].

For the criterion of maximum likelihood the decision rule rests on the determination of a maximal value of one of the statistics $L_i(t)$, $\Lambda_{ij}(t)$ or any of their monotonously increasing functions. In practical problems the statistics

of logarithms of either the likelihood function $\varphi_i(t) = \ln L_i(t)$ or of the likelihood ratio $u_{ij} = \ln \Lambda_{ij}(t)$ are used. The decision rule has the shape

$$h_k\colon\ u_{k1}(t) = \max_i u_{i1}(t). \tag{1.6}$$

The general statement and standard solution considered above simplify the problem excessively. The main difficulty in the solution of practical problems lies in the selection of parameters measurable in time and determination of the adequate mathematical model for their description.

In a general case the observable process y_t is the K. Ito process and in the case of the class h_i it is described by the equation

$$dy_t = A_i(\omega, y, t)\, dt + B(y, t)\, dw_t, \tag{1.7}$$

where A_i (ω, y, t) is an l-dimensional random vector-function of time, observation y_0^t, and random event ω, $B(y, t)$ is a matrix of dimensionality function $[l \times l]$, w_t is an l-dimensional standard Wiener process.

During the time of observation the jump of a process or a system from one class to another is possible, according to the Markov process with transition probability

$$P(h_i) = p_i,\ P_{ij}(t, t + \Delta t) = \delta_{ij} + \lambda_{ij}(t)\, \Delta t, \tag{1.8}$$

where δ_{ij} is the Kronecker symbol, and $\lambda_{ij}(t)$ is the intensity of the transition probability transforming the system from class h_i into h_j.

The aim of the paper is to obtain general differential equations for statistics sufficient to construct decision rules of type (1.5), (1.6) in the recognition of partially observable processes (1.7) by composite observations with transition probabilities (1.8). The problem of importance in practice is the investigation of the quality of the obtained algorithms of current classification, in particular, the error of recognition.

2. Main results

The basic equations for statistics sufficient to recognize processes by composite observations in continuous time will be obtained on the basis of the following statement.

Theorem. Let an observed process y_t be that of K. Ito and for each class h_i let it be determined by equation (1.7), whose coefficients comply with the usual restrictions adopted in the theory of optimal non-linear filtering

[2, Th. 8.1]. The class flow is that of Poisson with parameters (1.8). Then the a posteriori probability of the class $p_i(t)$ satisfies the equation*

$$dp_i(t) = \sum_{m=1}^{N} \lambda_{mi}(t)\, p_m(t)\, dt + p_i(t)\, [\bar{A}_i(t) - \bar{A}(t)]'(BB')^{-1}\, [dy_t - \bar{A}(t)\, dt] \quad (2.1)$$

with the initial conditions $p_i(0) = p_i$, where

$$\bar{A}_i(t) = E(A_i(\omega, y, t)/y_0^t, h_i) \quad (2.2)$$

$$\bar{A}(t) = \sum_{i=1}^{N} \bar{A}_i(t)\, p_i(t). \quad (2.3)$$

The ratio of a posteriori probabilities $\nu_{ij}(t)$ and Radon–Nikodym derivative (the likelihood ratio) $\Lambda_{ij}(t)$ satisfy the equation

$$d\nu_{ij}(t) = \nu_{ij}\left\{\sum_{m=1}^{N} \lambda_{mi}(t)\, \nu_{mi} dt - \sum_{m=1}^{N} \lambda_{mj}(t)\, \nu_{mj}\, dt + \right.$$
$$\left. + [\bar{A}_i(t) - \bar{A}_j(t)]'\, (BB')^{-1}\, [dy_t - \bar{A}_j(t)\, dt]\right\} \quad (2.4)$$

with the initial condition $\nu_{ij}(0) = p_i/p_j$ or $\Lambda_{ij}(0) = 1$.

Proof of the theorem. In order to get equation (1.1), let us introduce a Markov process ϑ_t with a finite set of states $(h_1, h_2, \ldots, h_N)$. The a priori probability of the state h_i at the initial time moment is equal to

$$P\{\vartheta_0 = h_i\} = p_i; \quad (2.5)$$

the transition probabilities from the state h_i at the moment t into the state h_j during the time Δt are equal to

$$P\{\vartheta_t = h_i, t;\ \vartheta_{t+\Delta t} = h_j, t + \Delta t\} = \delta_{ij} + \lambda_{ij}(t)\, \Delta t. \quad (2.6)$$

In this case the observable process y_t (1.7) satisfies the equation

$$dy_t = A(\omega, \vartheta_t, y, t)dt + B(y, t)\, dW_t,$$

where the dependence on ϑ_t can be either indicatorial or functional. The a posteriori probability of the class h_i is defined as

$$P_i(t) = E\,[\chi_{h_i}(\omega)/y_0^t] = E\,[\delta(h_i, \vartheta_t)/y_0^t], \quad (2.7)$$

where $\delta_{h_i}(\omega)$ is the indicator of the class h_i, $\delta(h_i, \vartheta_t)$ is the Kronecker symbol. Analogously to [2] it is possible to show the representability

$$\delta(h_i, \vartheta_t) = \delta(h_i, \vartheta_0) + \int_0^t \lambda_{\vartheta_s, h_i}(s)\, ds + x_t^{h_i},$$

* A' denotes the transposed matrix of the matrix A.

where $X_t^{h_i}$ is a square integrable martingale. By applying the equation of nonlinear filtering [2, eq. (8.10)] to $\delta(h_i, \vartheta_t)$ for $p_i(t)$ we get the equation

$$p_i(t) = p_i(0) + \int_0^t \pi_s^i(\lambda)\, ds + \int_0^t [\pi_i^s(\delta A) - p_i(t)\, \pi_s^i(A)] \times$$

$$\times [BB']^{-1} [dy_s - \pi_s^i(A)\, ds]. \tag{2.8}$$

Since

$$\pi_s^i(\lambda) = E[\lambda_{\vartheta_s, h_i}(s)/y_0^s] = \sum_{j=1}^N \lambda_{ji}(s)\, p_j(s),$$

$$\pi_s^i(\delta A) = \bar{A}_i(s)\, p_i(s), \quad \bar{A}_i(t) = E[A(\omega, h_i, y, t)/y_0^s, h_i],$$

$$\pi_s^i(A) = E[A(\omega, \vartheta_t, y, s)/y_0^s] = \sum_{i=1}^N \bar{A}_i(s)\, p_i(s) = \bar{A}(t).$$

By putting the obtained values of conditional means into (1.8), we get equation (1.1).

The equation for a posteriori probability ratios will be obtained from (1.1) by applying the formula of a change of K. Ito's variables [2] to the function

$$\nu_{ij}(t) = p_i(t)/p_j(t)^*.$$

In view of the basic equation (1.4), by applying the K. Ito formula to the function $\ln \nu_{ij}(t)$, we get a stochastic differential equation both for the logarithm of the a posteriori probability ratio as well as for that of the likelihood ratio $u_{ij}(t) = \ln \Lambda_{ij}(t)$

$$d \ln \Lambda_{ij}(t) = \sum_{m=1}^N \lambda_{mi}(t)\, \Lambda_{mi}(t)\, dt - \sum_{m=1}^N \lambda_{mj}(t)\, \Lambda_{mj}(t) dt +$$

$$+ (\bar{A}_i(t) - \bar{A}_j(t))' (BB')^{-1} \left[dy_t - \frac{1}{2} (\bar{A}_i(t) + \bar{A}_j(t))\, dt \right] \tag{2.9}$$

with the initial conditions

$$\ln \Lambda_{ij}(0) = 0.$$

Hence the equation for the logarithm of the likelihood function $\ln L_i(t)$ taking into account the relation $\ln \Lambda_{ij}(t) = \ln L_i(t) - \ln L_j(t)$ has the shape

$$d \ln L_i(t) = \sum_{m=1}^N \lambda_{mi}\, \Lambda_{mi}(t)\, dt + \bar{A}_i(t)' (BB')^{-1} \left(dy_t - \frac{1}{2} \bar{A}_i(t)\, dt \right). \tag{2.10}$$

By using the K. Ito formula for the compound function $L_i(t) = \exp\{\ln L_i(t)\}$ we obtain the equation

$$dL_i(t) = \sum_{m=1}^{N} \lambda_{mi}(t) L_m(t)\, dt + L_i(t)\, \bar{A}_i'(t)\, (BB')^{-1}\, dy_t \tag{2.11}$$

for the likelihood functions.

The differential equation (1) for $p_i(t)$ taking into consideration (1.3) is non-linear. For the classification it is more convenient to use the linear equations (1.4)–(2.4). The main difficulty, however, lies in the computation of a posteriori expectation (1.2). Further the special cases are considered for which the expressions $\bar{A}_i(t)$ for exact or approximate calculations are rewritten in an explicit way. On the basis of these equations constructive algorithms for the recognition in current time are formed.

3. Special cases. A constructive solution

Note that there is no need to calculate a posteriori expectation (2.2) for the fully observable processes in the case of class h_i described by the equations

$$dy_t = A_i(y_t, t)\, dt + B(y_t, t)\, dw_t. \tag{3.1}$$

The statistics used for the recognition of processes (1.1) satisfy equations (2.1)–(2.5) as $\bar{A}_i(t) = A_i(y_t, t)$.

When no passage between classes takes place, i.e. while identifying samples, belonging to the same class, the statistics necessary to recognize both partially and fully observable processes are defined by equations (2.1)–(2.11) as $\lambda_{ij}(t) = 0$, $(i, j = 1, 2, \ldots, N)$.

In this case, equations (2.4)–(2.7) can be solved and explicit expressions for the calculations of the statistics $\nu_{ij}(t)$, $\Lambda_{ij}(t)$, $\varphi_i(t)$, $u_{ij}(t)$ may be obtained. Thus $\varphi_i(t)$ is determined by the expression

$$\varphi_i(t) = \int_0^t \bar{A}_i'(t)(BB')^{-1}\left[dy_t - \frac{1}{2}\bar{A}_i(t)\, dt\right]. \tag{3.2}$$

Making use of (1.2) the expression for the calculation of the above statistics is easily obtained.

For partially obsrvable processes in the overwhelming majority of practical problems a random function $A_i(\omega, y, t)$ is defined as a function from the unobservable component X and then equation (1.7) will take the shape as follows

$$dy_t = A_i(x_t, y, t)\, dt + B(y, t)\, dw_t. \tag{3.3}$$

The conditional expectation (2.2) is defined by the expression

$$\bar{A}_i(t) = \int A_i(x_t, y, t)\, P_i(dx_i/y_0^t). \tag{3.4}$$

The problem is reduced to the determination of the a posteriori distribution of an unobservable component $P_i(dx/y_0^t)$ for each of the hypotheses h_i. The a posteriori distribution of an unobservable component for a wide class of practically important processes was investigated in connection with the filtering problem in [2].

If the unobservable component X is included into (1.3) linearly

$$dy_t = A_0^i(y, t)\, dt + A_1^i(y, t)\, x_t dt + B(y, t)\, dw_t\,, \tag{3.5}$$

then (1.4) is equal to

$$\bar{A}_i(t) = A_0^i(y, t) + A_1^i(y, t)\, m_t,\ \ m_t = E\,[x_t/y_0^t, h_i] \tag{3.6}$$

and is reduced to the calculation of the a posteriori expectation $m_i(t)$ of the unobservable component x_t for each of the hypotheses h_i. In case the unobservable process is described by a stochastic differential equation, its current estimates optimal in the definite class of the function can be ontained on the basis of general methods worked out in [5]. For diffusion processes on the basis of exact and approximate estimates of unobservable components dealt with in the filtering problem [2], an optimal and suboptimal classification can be obtained.

We present the examples of the application of general equations (2.1)–(2.11) to obtain the constructive solution of the problem of classifying different processes. The classification of conditionally Gaussian processes for which an unobservable component of class h_i is defined by the equation

$$dx_i = a_0^i(y, t)\, dt + a_1^i(y, t)\, x_t dt + b_1^i(y, t)\, dw_1(t) + b_2^i(y, t)\, dw_2(t) \tag{3.7}$$

in the case of compound observations (1.5) is determined by the equations (2.1)–(2.11) taking into consideration (1.6) where for $m_i(t)$ a suboptimal estimation can be obtained on the basis of conditional cumulants [6]. For simple observations optimal methods of classification are worked out in [7].

Let an unobservable component X, vectorial in a general case, be a step-wise Markov process with a discrete (finite or denumerable) set of states $E(\alpha, \beta, \ldots)$. Transition probabilities of the process X from state α at the time moment t into β at the moment $t + \Delta t$ for the class h_i are defined by the probability density functions of the transition $\mu_{\alpha\beta}$ in the form

$$P_{\alpha\beta}^i(t, t + \Delta t) = \delta_{\alpha\beta} + \mu_{\alpha\beta}^i(y, t)\, \Delta t + 0(\Delta t),\ \ \alpha, \beta \in E. \tag{3.8}$$

Since the process X takes a discrete set of values, we have

$$\bar{A}_i(t) = \sum_{\alpha \in E} A_i(\alpha, y, t)\, \pi_\alpha(t), \tag{3.9}$$

where $\pi_\alpha(t)$ is an a posteriori probability function of the state α for the class h_i

$$\pi_\alpha(t) = P\{x = \alpha / y_0^t, h_i\}, \quad \alpha \in E.$$

In the case of simple observations the a posteriori probability $\pi_\alpha(t)$ is determined by solving filtering equations for discontinuous Markov processes with a discrete set of states [2].

If the unobservable process X is included into (1.5) linearly, then $\bar{A}_i(t)$ is defined by expression (1.6) where

$$m_i(t) = \sum_{\alpha \in E} \alpha \pi_\alpha^i(t).$$

Let the unobservable component X be a discontinuous Markov process with a continuous set of states. In the case of class h_i the process X is characterized by the intensity of the Poisson flow $\lambda(x, t, u_i)$ and the two-dimensional density $Q(x', x, t, u_i)$ with the help of which the transition probability from state into a set of states at the interval $(x, x + \Delta x)$ is expressed on condition that a jump occurred

$$\int Q(x', x, t, u_i)\, dx = 1. \tag{3.10}$$

If the observable process y_t is defined by the stochastic differential (1.3), then in the classification of processes by using basic equations for the statistics (2.1)–(2.11) the a posteriori expectation (2.2) is determined in the form $A_i(t) = \int A_i(x, y_t, t)\, \pi_i(x, t)\, dx$. The a posteriori probability density function

$$\pi_i(x, t) = \frac{\partial}{\partial x} P\{\xi < x / y_0^t, h_i\}$$

satisfies the integro-differential equation of filtering [8].

Analogously, on the basis of the general approach considered the equations for sufficient statistics can be obtained, when the observable process Y is defined by stochastic differential (1.3) and the unobservable X consists of two components (x_1, x_2) where X is a diffusion Markov process

$$dx_1(t) = a_i(x_t, y_t, t)\, dt + b_1^i(x_t, y_t, t)\, dw_1(t) + b_2^i(x_t, y_t, t)\, dw_2(t) \tag{3.11}$$

and $x_2(t)$ is a discontinuous Markov process with transition probabilities (1.8). In this case the a posteriori distribution x_t satisfies the equation of non-linear filtering [2].

An example: It is required to recognize N telegraphic signals, taking values 0, 1 with the passage intensity

$$\mu_{01}^{i} = \alpha_i, \quad \mu_{10}^{i} = \beta_i, \qquad (i = 1, 2, \ldots, N) \tag{3.12}$$

by the observation under white noise

$$dy_t = x_t dt + B\, dw_t. \tag{3.13}$$

In order that we can use one of the expressions (2.1)–(2.11), it is necessary to determine $\bar{A}_i(t)$ (1.6). Since $\bar{A}_i(t) = m_i(t)$, $m_i(t) = \pi_1^i(z)$, from filtering equation [6] we have

$$dm_i(t) = -\beta_i m_i(t)dt + \alpha_i(1 - m_i(t))\,dt + (m_i(t) - m_i^2(t))\,B^{-2}(dy_t - m_i(t)dt). \tag{3.14}$$

In solving equations (2.7) by analogue computers it is necessary to go over to P. Stratonovich's symmetrization equations of the type

$$dm_i(t) = -\beta m_i(t)dt + \alpha_i(1 - m_i(t))\,dt + (m_i(t) - m_i^2(t))\,B^{-2}\left(dy_t - \frac{1}{2}dt\right).$$

By putting $m_i(t)$ into one of the equations (2.1)–(2.11) dependent on the criterion used, we calculate the statistics necessary for the recognition. In the case of simple observation (2.6) in order to determine the logarithm of the likelihood function we have (2.10)

$$\varphi_i(t) = \int_0^t m_i(t)\,B^{-2}\left[dy_t - \frac{1}{2}m_i(t)\,dt\right]$$

which enables us to recognize telegraphic signals making use of the maximum likelihood criterion.

4. Evaluation of the Bayesian error of recognition

Investigation of the quality of algorithms of dynamic classification is important for their application to the solution of practical problems, since it allows to define the main parameters of the recognition device. When using one or another measure of the quality of the classification, it is important to estimate the probability of misclassification. The probability of error in the recognition depends essentially on the volume of a priori data used for the construction of a classifier, the time of learning and recognition. In a general case, determination of the probability of misclassification is a complicated problem.

Let us study the Bayes misclassification $P_B(t)$ when recognizing the processes described by the system of equation

$$h_k\colon\ dy_t = A_k(y_t, t)\,dt + B(y_t, t)\,dw_t, \quad (k = 1, 2) \tag{4.1}$$

in the case of two hypotheses, with a priori probabilities p_k. The probability of misclassification

$$P_B(t) = p_1 P(h_2/h_1) + p_2 P(h_1/h_2)$$

can be presented as follows

$$P_B(t) = p_1 P\{\ln \nu_{12}(t) < 0/h_1\} + p_2 P\{\ln \nu_{12}(t) \geq 0/h_2\} \tag{4.2}$$

where $\ln \nu_{12}(t)$ is the statistic of the logarithm of the ratio of a posteriori probabilities, satisfying differential equation (2.9).

Consider two processes $s_k(t)$ equal to $\ln \nu_{12}(t)$ on condition of observing the samples belonging to class h_k $(k = 1, 2)$. On the basis of (2.9) taking into account (1.1) for $s_k(t)$, we have a stochastic differential equation

$$ds_k(t) = (-1)^k \frac{1}{2}(A_1 - A_2)'\,(BB')^{-1}(A_1 - A_2)\,dt + (A_1 - A_2)'\,(B')^{-1}dw_t \tag{4.3}$$

with the initial conditions $s_k(0) = s_0 = \ln(p_1/p_2)$.

Then the probability of misclassification (1.2) is representable in the form

$$P_B(t) = p_1 P\{s_1(t) < 0\} + p_2 P\{s_2(t) \geq 0\} \tag{4.4}$$

or, by using a characteristic set function

$$\chi_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A \end{cases}$$

in the form of continual integrals

$$P_B(t) = p_1 E_{s_0} \chi_{s_1(t)<0} + p_2 E_{s_0} \chi_{s_2(t)\geq 0}. \tag{4.5}$$

Introducing the notation $u_1(s, t) = E_s \chi_{s_1(t)<0}$ and $u_2(s, t) = E_s \chi_{s_2(t)\geq 0}$ we shall present $p_B(t)$ in the shape

$$P_B(t) = p_1 u_1(s_0, t) + p_2 u_2(s_0, t). \tag{4.6}$$

Making use of the relation between continual integrals and partial differential equations [9], with regard to (1.1), (1.3) we get that $u_k(s, t) = Eu_k(s, y, t)$

where $u_k(s, y, t)$ is the solution of the Cauchy problem for the equation

$$\frac{\partial u_k(s,y,t)}{\partial t} = -\frac{(-1)^k}{2}(A_1 - A_2)'(BB')^{-1}(A_1 - A_2)\frac{\partial u_k}{\partial s} + \frac{\partial u_k}{\partial y}A_k + \\ + \frac{1}{2}(A_1 - A_2)'(BB')^{-1}(A_1 - A_2)\frac{\partial^2 u_k}{\partial s^2} + \frac{\partial^2 u_k}{\partial s\,\partial y}(A_1 - A_2) + \\ + \frac{1}{2}Sp\left[(BB')\frac{\partial^2 u_k}{\partial y^2}\right] \tag{4.7}$$

with the initial conditions

$$u_1(s, y, 0) = \begin{cases} 0, & s \geq 0 \\ 1, & s < 0, \end{cases} \qquad u_2(s, y, 0) = \begin{cases} 1, & s < 0 \\ 0, & s \leq 0. \end{cases} \tag{4.8}$$

If the coefficients of equations in (1.3) do not depend on y_t, then $u_k(s, t)$ in (1.6) are the solutions of the Cauchy problem for an infinite straight line

$$\frac{\partial u_k}{\partial t} = -\frac{(-1)^k}{2}\alpha(t)\frac{\partial u_k}{\partial s} + \frac{1}{2}\alpha(t)\frac{\partial^2 u_k}{\partial s^2} \tag{4.9}$$

with the initial conditions (1.8)

$$\alpha(t) = (A_1 - A_2)'(BB')^{-1}(A_1 - A_2). \tag{4.10}$$

The solution of the Cauchy problem for equation (2.1), in case $\alpha(t)$ does not depend on time, has the shape

$$u_k(s, t) = \frac{1}{\sqrt{2\pi\alpha t}}\int_{-\infty}^{0}\exp\left[-\left(\xi + (-1)^k s - \frac{1}{2}\alpha t\right)^2 \Big/ 2\alpha t\right]d\xi. \tag{4.11}$$

By putting the solution of equation (2.1) into (1.6) we obtain the Bayes probability of error in the shape

$$P_B(t) = p_1\Phi\left(-\frac{s_0 + \frac{1}{2}\alpha t}{\sqrt{\alpha t}}\right) + p_2\Phi\left(\frac{s_0 - \frac{1}{2}\alpha t}{\sqrt{\alpha t}}\right) \tag{4.12}$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{1}}\,dz.$$

Let us investigate the expression obtained. At the initial time moment we have the probability of misclassification of the simplest classifier

$$P_B(0) = \min(p_1, p_2),$$

with the increase of the time of recognition the Bayes error decreases monotonically. It is easily shown if we calculate a derivative of the function $P_B(t)$ with respect to time

$$\frac{d}{dt}(P_B(t)) = \frac{-\alpha p_1}{4\sqrt{\alpha t}} \exp\left\{-\frac{\left(s_0 + \frac{1}{2}dt\right)^2}{2\alpha t}\right\}$$

and consider that $\alpha(t) > 0$. As $t \to \infty$, then $P_B(t) \to 0$. As $p_1 = p_2$, the Bayes error (2.4) is equal to $P_B(t) = \Phi\left(\frac{\sqrt{\alpha t}}{2}\right)$. Then given the value of misclassification γ, using the tables for the Laplace function we obtain the lower bound of the time of classification

$$t_{KL} \geq t_\gamma = \frac{(2\beta)^2}{\alpha},$$

where β is defined by γ from the relation $\Phi(\beta) = \gamma$.

5. Conclusion

The developed approach to the recognition of stochastic processes and systems on the basis of the construction of dynamic models of real signals enabled us to obtain differential equations for statistics sufficient to construct the decision rules of classification in current time. Decision making in continuous, current time is particularly important in control systems operating in real-time on the basis of information processing obtained in the process of exploitation of an object.

Essential difficulties arising in the application of these methods to practical problems and connected with the estimation of unobservable signal and parameters of dynamic models can be eliminated by using a new method of evaluation of variables and parameters in non-linear stochastic systems [5].

References

1. *Stratonovich, R. L.*, Conditional Markov processes and their application to the theory of optimal control. Modern Analytic and Computational Methods in Science and Mathematics, No. 7, New York, 1968.
2. *Liptser, R. Sh., Shiryayev, A. T.*, Statistics of stochastic processes. "Nauka", Moscow, 1974.
3. *Sysoev, L. P.*, Estimation of parameters. Detection and discrimination of signals. "Nauka", Moscow, 1969.
4. *Shpilewski, E. K.*, The principles of dynamical classification of stochastic processes and systems. In: Statistical Problems of Control, issue *28*, Institute of Math. and Cybern. of the Academy of Sci. of the Lithuanian SSR, Vilnius, 1978.

5. *Pugachev, V. S.*, Estimation of variables and parameters in discrete time non-linear systems. Avtomatika i telemekhanika, No. *4*, 1979, pp. 39–50.
6. *Yashin, A. I.*, Conditional semivariants in problems of filtering of processes with intermittent components. Avtomatika i telemekhanika, No. *2*, 1970, pp. 20–26.
7. *Shpilewski, E. K.*, Recognition of dynamic systems in the presence of noise. Avtomatika i telemekhanika, No. *12*, 1974, pp. 60–71.
8. *Grigelionis, B.*, On stochastic equations for nonlinear filtering problem of stochastic processes. Litovsk. Mat. Sb., vol. **12**, No. *4*, 1972, pp. 39–51.
9. *Shpilewski, E. K.*, On the solution of partial differential equations by the Monte-Carlo method with the help of analog computers. In: Kibernetika i upravlenie. "Nauka", Moscow, 1967, pp. 131–140.

Динамическое распознавание случайных процессов и систем в непрерывном времени

Э. К. ШПИЛЕВСКИЙ

(Вильнюс)

На базе единого подхода рассмотрены конструктивные методы распознавания процессов в непрерывном времени по прямым и косвенным измерениям. Получены основные дифференциальные уравнения для статистик апостериорных вероятностей классов $p_i(t)$; отношения апостериорных вероятностных классов $v_{ij}(t)$; функции правдоподобия $L_i(t)$; отношения правдоподобия $\Lambda_{ij}(t)$ и их логарифмов $\varphi_i(t)$ и $u_{ij}(t)$, которые позволяют построить последовательные нерандомизированные решающие правила классфикации в текущем времени полностью и частично наблюдаемых многомерных процессов и систем. Рассмотрено распознавание непрерывных марковских процессов, описываемых стохастическими дифференциальными уравнениями, разрывных марковских процессов с конечным и непрерывным множеством значений. Получена классификация по простым и составным наблюдениям в обстановке помех. Приведен пример распознавания телеграфных сигналов в обстановке шумов.

Для определения вероятности ошибочной классификации выведены дифференциальные уравнения в частных производных и сформирована задача Коши. В частном случае получена формула для вероятности ошибки с учетом времени классификации. Показано, что вероятность ошибки является монотонно убывающей функцией времени распознавания и в пределе стремится к нулю.

Э. К. Шпилевский
Институт математики и кибернетики АН Литовской ССР
СССР, 232600, г. Вильнюс, ул. К. Пожелос, 54

Problems of Control Information and Theory, Vol. 10 (2), pp. 129—136 (1981)

ON A RE-SAMPLING PROBLEM IN THE THEORY OF INFORMATION

N. KUSOLITSCH

(Wien)

(Received May 1, 1980)

A sample of size N is taken with uniform, independent distribution, with replacement from a finite population of cardinality k. Then a re-sampling is made: from the original sample one chooses n elements and considers a given statistic S of them. The re-sampling is repeated independently to produce a sequence of statistics $S_1, S_2, \ldots$, The paper investigates the statistical properties of this latter sequence as a function of the parameters N, k, n. In particular, bounds are given on the entropy-rate of the sequence $S_1, S_2, \ldots$.

1. Introduction

In the practice of information transmission there are a number of cases where additional randomness is artificially introduced into the problem. This is usually done in a way that sender and receiver jointly choose a random sequence, i.e. they both know it perfectly, and then they exploit some-how its randomness properties to attain their specific goals. An example for this kind of application is the rate-one error detecting algorithm proposed by M. Hellman [2]. Another application is the one-time-pad method in cryptography, see Kahn [3]. Here the length of the random sequence has to equal the message length. In the first application however, a much shorter random sequence would do: The message is partitioned into blocks of equal length, and the very random sequence is applied to all of these blocks. The method is constructed in a way that transmission errors result in adding modulo-sums of some randomly chosen elements of the random sequence to the message and, by that, roughly speaking, turning the redundant message into a "random" sequence.

The economic utilization of the random sequence is that the same sequence is applied to the subsequent blocks of the message. Obviously, the price for this economic utilization is that the additive random sums corresponding to the error patterns of the individual blocks will not be independent. There is a need for investigating the consequences of this dependence, and this is that motivates our investigations described in the abstract. Luckily, it turns out, that the dependence structure is weak enough, provided the parameters are properly chosen.

2. Problem statement

Let $X = \{0, \ldots, k-1\}$ be a finite set and $\xi_1, \xi_2, \ldots$ a sequence of i.i.d. random variables on $(X, 2^X)$. The elements of X are called characters.

Given two natural numbers n, N, $n \leq N$ we will construct a sequence of statistics $(S_l^{n,N})_{l \in \mathbf{N}}$ in the following manner: We choose a subset $M_l = \{i_1^l, \ldots, i_n^l\}$ of $\{1, \ldots, N\}$. for every $l \in \mathbf{N}$ uniformly and independently of the earlier choices. Then we consider the statistic $S_l^{n,N}$ of the random variables $\xi_{i_1^l}, \ldots, \xi_{i_n^l}$, i.e. $S_l^{n,N} = S_l^{n,N}(\xi_{i_1^l}, \ldots, \xi_{i_n^l})$ for every $l \in \mathbf{N}$.

For example, $S_l^{n,N}$ may be the minimum of $\xi_{i_1^l}, \ldots, \xi_{i_n^l}$ or as well, the modulo k-sum $\sum_{j=1}^{n} k\, \xi_{i_j^l}$ of the chosen random variables.

In the sequel, indices will be omitted if not necessary.

Here and in the following $M^{n,N}$ (or $M_l^{n,N}$) will always denote a random sample which describes the uniform choice of a subset of size n out of $\{1, \ldots, N\}$.

Our problem is to investigate the finite dimensional distributions and the dependence structure of the sequence $S_1, S_2, \ldots$. The one-dimensional (marginal) distributions are easy to calculate: If $S := \xi_1 \oplus \xi_2$ then $P(S = x) = \sum_{y=0}^{k-1} P(\xi_1 = y)\ P(\xi_2 = x \ominus y) = \frac{1}{k}$ for every $x \in X$. By a simple induction argument it can be shown that the modulo k-sum is always uniformly distributed on X.

In a similar simple manner the distribution of the minimum statistics can be derived. Obviously if $\{i_1, \ldots, i_n\} = \{1, \ldots, n\}$ then

$$P(\min\{\xi_1, \ldots, \xi_n\} = x) = \frac{(k-x)^n - (k-x-1)^n}{k^n} \qquad x \in X$$

and

$$P(S^{n,N} = x) = \sum_{\{i_1, \ldots, i_n\} \subseteq \{1, \ldots, N\}} P(S^{n,N} = x \mid M^{n,N} = \{i_1, \ldots, i_n\}) \cdot$$

$$\cdot\, P(M^{n,N} = \{i_1, \ldots, i_n\}) = \frac{(k-x)^n - (k-x-1)^n}{k^n} \qquad x \in X.$$

As a measure of dependence, we will use the entropy of a finite number, say L, of the statistics S. Since $S_1, S_2, \ldots$ consists of equivalent random variables, it always suffices to consider the first L statistics. If the entropy-rate $\frac{1}{L} \cdot H(S_1, \ldots, S_L)$ is near to the value $H(S_1)$ corresponding to the independent case, then the dependence is weak enough.

3. Results

The statistics $S_l^{n,N}\, l \in \mathbf{N}$ are generally not independent as Example 1 shows. They are, however, independent conditionally, i.e. if the original sequence, $\xi_1, \ldots, \xi_N$ is given, because then the values of statistics depend only on the choices of subsets $M_l^{n,N}$ and these choices are independent by assumption. Let us formulate this fact as a lemma.

Lemma 1. The statistics $S_l^{n,N}$, $l \in \mathbf{N}$ form an i.i.d. sequence of random variables if $\xi_1, \ldots, \xi_N$ rea given.

Example 1. $X = \{0, 1\}$, $n = 2$, $N = 3$, $S_l = \xi_{i_1^l} + \xi_{i_2^l})$ $l = 1, 2$. As we know already

$$P(S_l = 0) = P(S_l = 1) = \frac{1}{2} \quad l = 1, 2.$$

But

$$P(S_1 = 0, S_2 = 0) = P(S_1 = 0, S_2 = 0 \,|\, M_1 = M_2)\, P(M_1 = M_2) +$$

$$+ P(S_1 = 0, S_2 = 0 \,|\, M_1 \neq M_2)\, P(M_1 \neq M_2) = P(S_1 = 0)\, P(M_1 = M_2) +$$

$$+ \frac{1}{4} P(M_1 \neq M_2) = \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{2}{3} = \frac{1}{3} > \frac{1}{4} = P(S_1 = 0)\, P(S_2 = 0).$$

In what follows, Lemma 2 will turn out to be very useful. It concerns the probability of typical sequences.

Definition. A sequence $x_1, \ldots, x_N$, $x_i \in X$, $|X| = k$ is called ε-typical, if $\left| \frac{N_x}{N} - \frac{1}{k} \right| \leq \varepsilon$ for all $x \in X$, where N_x denotes the frequency of the character x in $(x_1, \ldots, x_N)$.

Lemma 2. Let $T_{N,\varepsilon}$ $\varepsilon \geq 0$ denote the set of ε-typical N-tuples derived from $\xi_1, \xi_2, \ldots$.

If a sequence (ε_N) tends to zero such that $\lim_{N \to \infty} N \varepsilon_N = \infty$ then $P(T^c_{N,\varepsilon_N})$ tends to zero exponentially, more precisely, $P(T^c_{N,\varepsilon_N}) \leq 2ke^{-N\varepsilon_N}$.

Proof. Chernoff's inequality (see, e.g. [1]) yields

$$P\left(\frac{N_x}{N} > \frac{1}{k} + \varepsilon\right) \leq \inf_{t \geq 0} \left\{ e^{-tN\left(\frac{1}{k} + \varepsilon\right)} \cdot \left(\frac{e - 1}{k} + 1 \right)^N \right\} \quad \forall\, x \in X.$$

This infimum is achieved at $t_0 = \ln \frac{(k-1)(1+\varepsilon k)}{k-1-\varepsilon k}$. *

Therefore

$$P\left(\frac{N_x}{N} > \frac{1}{k} + \varepsilon\right) \leq \exp\left(-N\left[\frac{1+k\varepsilon}{k} \cdot \ln \frac{(k-1)(1+k\varepsilon)}{k-1-k\varepsilon} + \ln \frac{k-1}{k-1-k\varepsilon}\right]\right).$$

Since there exists an $s_0 > 0$ such that $\frac{s}{2} \leq \ln(1+s)$, $0 \leq s \leq s_0$ the following holds true.

$$\frac{1+k\varepsilon}{k} \ln\left(1 + \frac{k^2\varepsilon}{k-1-k\varepsilon}\right) + \ln\left(1 + \frac{k\varepsilon}{k-1-k\varepsilon}\right) \geq \frac{1+k\varepsilon}{2k} \cdot \frac{k^2\varepsilon}{k-1-k\varepsilon} +$$

$$+ \frac{k\varepsilon}{2(k-1-k\varepsilon)} > \frac{k\varepsilon}{k-1-k\varepsilon} > \varepsilon \text{ if } 0 \leq \varepsilon < \frac{s_0 k - s_0}{s_0 k + k^2} =: \varepsilon_0.$$

This yields

$$P\left(\frac{N_x}{N} > \frac{1}{k} + \varepsilon\right) \leq \exp(-N\varepsilon) \text{ if } 0 \leq \varepsilon \leq \varepsilon_0.$$

By means of analog considerations, for the case $\frac{N_x}{N} < \frac{1}{k} - \varepsilon$ we get finally

$$P(T^c_{N,\varepsilon}) \leq \sum_{x=0}^{k-1} P\left(\left|\frac{N_x}{N} - \frac{1}{k}\right| > \varepsilon\right) \leq 2k \exp(-N\varepsilon).$$

Remark: The sequence $(\varepsilon_N) = \left(\frac{\ln N}{N}\right)$ yields a speed of convergence of $\frac{2k}{N}$.

If Y, Z are random variables then $H(Y)$, $H(Z)$ will denote the entropies of Y and Z, respectively, and $I(Z \mid Y)$ the information on Z provided by Y or vice versa.

We can state our first main result now.

Theorem 1. For all natural numbers n, N and every alphabetsize k there exists a bound $A(n, N, k)$ depending just on n, N and k such that for every natural number L

$$H(S_1^{n,N}) \geq L^{-1} H(S_1^{n,N}, \ldots, S_L^{n,N}) \geq H(S_1^{n,N}) - A(n, N, k).$$

If $n = o(\sqrt{N})$ then $\lim_{N \to \infty} A(n, N, k) = 0$.

* By ln we denote the natural logarithm, while log stands for logarithm of base 2.

Proof. Lemma 1 yields for every $L \in \mathbf{N}$

$$L^{-1} H(S_1^{n,N}, \ldots, S_L^{n,N}) \geq L^{-1} H(S_1^{n,N}, \ldots, S_L^{n,N} \mid \xi_1 \ldots, \xi_N) =$$
$$= H(S_1^{n,N} \mid \xi_1, \ldots, \xi_N) = H(S_1^{n,N}) - I(S_1^{n,N} \mid \xi_1, \ldots, \xi_N).$$

According to the data processing theorem (see, e.g. [1]) with $(\xi_1, \ldots, \xi_N) \to$ $\to (\xi_1, \ldots, \xi_N) \to (\xi_{i_1}, \ldots, \xi_{i_n}) \to S_1^{n,N}(\xi_{i_1}, \ldots, \xi_{i_n})$ one gets

$$L^{-1} H(S_1^{n,N}, \ldots, S_L^{n,N}) \geq H(S_1^{n,H}) - I((\xi_{i_1}, \ldots, \xi_{i_n}) \mid (\xi_1, \ldots, \xi_N))$$

$$I((\xi_{i_1}, \ldots, \xi_{i_n}) \mid (\xi_1, \ldots, \xi_N)) \leq H(\xi_{i_1}, \ldots, \xi_{i_n}) -$$
$$- \sum_{(x_1, \ldots, x_n) \in T_{N, \varepsilon_N}} k^{-N} H(\xi_{i_1}, \ldots, \xi_{i_n} \mid x_1, \ldots, x_N).$$

Since $(x_1, \ldots, x_N)$ is ε_N-typical, we have

$$P(\xi_{i_1} = y_1, \ldots, \xi_{i_n} = y_n \mid \xi_1 = x_1, \ldots, \xi_N = x_N) =$$
$$P(\xi_{i_1} = y_1 \mid \xi_1 = x_1, \ldots, \xi_N = x_N) \ldots P(\xi_{i_n} = x_n \mid \xi_1 = x_1, \ldots, \xi_N =$$
$$= x_N, \xi_{i_1} = y_1, \ldots, \xi_{i_{n-1}} = y_{n-1}) \leq$$
$$\leq \frac{N y_1}{N} \cdot \frac{N y_2}{N-1} \cdots \frac{N y_n}{N-n+1} \leq \left(\frac{1}{k} + \varepsilon_N\right)^n \left(\frac{N}{N-n}\right)^n.$$

Suppose now that the probabilities p_j of a discrete random variable Y are bounded from above by a constant U, i.e. $p_j \leq U$, then

$$H(Y) = -\Sigma p_j \log p_j \geq -(\log U) \Sigma p_j = \log \frac{1}{U}.$$

Therefore we may estimate $H(\xi_{i_1}, \ldots, \xi_{i_n} | x_1, \ldots, x_N)$ from below by

$$H(\xi_{i_1}, \ldots \xi_{i_n} \mid x_1, \ldots, x_N) \geq n \left(\log \frac{k}{1 + k\varepsilon_N} + \log \frac{N-n}{N}\right).$$

From these it follows finally that

$$L^{-1} H(S_1^{n,N}, \ldots, S_L^{n,N}) \geq H(S_1^{n,N}) - H(\xi_{i_1}, \ldots . \xi_{i_n}) +$$
$$+ nP(T_{N, \varepsilon_N}) \left[\log \frac{k}{1 + k\varepsilon_N} + \log \left(1 - \frac{n}{N}\right)\right].$$

Of course, $H(\xi_{i_1}, \ldots, \xi_{i_n}) = n \log k$, so $A(n, N, k)$ can be written in the form

$$A(n, N, k) = nP(T^c_{N, \varepsilon_N}) \log k + nP(T_{N, \varepsilon_N}) \left[\log (1 + k\varepsilon_N) - \log \left(1 - \frac{n}{N}\right)\right] \leq$$
$$\leq nP(T^c_{N, \varepsilon_N}) \log k + nk\varepsilon_N + \frac{n^2}{N}.$$

Choosing $\varepsilon_N = \dfrac{\ln N}{N}$ and applying Lemma 2, we get

$$A(n, N, k) \leq c\frac{n}{N} + \frac{\overline{n}^2}{N} + \frac{nk \cdot \ln N}{N} \quad \text{with } c = \text{constant}.$$

From here it follows that $\lim\limits_{N} A(n, N, k) = 0$ if $n = o(\sqrt{N})$ with a speed of convergence $\dfrac{n^2}{N}$.

Corollary 1. Under the assumptions and in the notations of Theorem 1 the entropy-rate $\lim\limits_{L\to\infty} H(S_L^{n,N} \mid S_1^{n,N}, \ldots, S_{L-1}^{n,N})$ of $(S_L^{n,N})_L$ is at least $H(S_1^{n,N}) - A(n, N, k)$, i.e.

$$\lim_L H(S_L^{n,N} \mid S_1^{n,N}, \ldots, S_{L-1}^{n,N}) \geq H(S_1^{n,N}) - \left(\frac{2nk \log k}{N} + \frac{n^2}{N} + \frac{nk \cdot \ln N}{N}\right).$$

Proof. The sequence $(S_L^{n,N})_{L \in \mathbf{N}}$ forms a stationary source, thus, as is well known,

$$\lim_L H(S_L^{n,N} \mid S_1^{n,N}, \ldots, S_{L-1}^{n,N}) = \lim_L L^{-1} H(S_1^{n,N}, \ldots, S_L^{n,N}).$$

Remark. The entropies $H(S_L^{n,N} \mid S_1^{n,N}, \ldots, S_{L-1}^{n,N})$ of $S_L^{n,L}$, given $S_1^{n,N}, \ldots, S_{L-1}^{n,N}$ are nonincreasing so they cannot be smaller than $H(S_1^{n,N}) - A(n, N, k)$. In other words, the information on $S_L^{n,N}$ provided by $S_1^{n,N}, \ldots, S_{L-1}^{n,N}$ never exceeds $A(n, N, k)$, however large L may be and $A(n, N, k)$ tends to zero if $n = o(\sqrt{N})$.

For the modulo k-sum statistics the condition $n = o(\sqrt{N})$ reduces to $n < N$ as the following lemma shows. From now on $S^{n,N}$ will denote the modulo k-sum statistics, i.e.

$$S_l^n = \sum_{j=1}^{n}{}_k \xi_{i^l}, \quad l \in \mathbf{N}.$$

Lemma 3. $H(S^{n,N} \mid \xi_1, \ldots, \xi_N) = H(S^{N-n,N} \mid \xi_1, \ldots, \xi_N)$.

Proof. $H(\xi_{i_1} \oplus \ldots \oplus \xi_{i_n} \mid x_1, \ldots, x_N) = H\left(\sum\limits_{j=1}^{N}{}_k x_j \ominus \sum\limits_{j=1}^{n}{}_k \xi_{i_j} \mid x_1, \ldots, x_N\right) =$

$= H\Big(\sum\limits_{j \notin \{i_1, \ldots, i_n^k\}} \xi_{i_j} \mid x_1, \ldots, x_N\Big) = H(S^{N-n,N} \mid x_1, \ldots, x_N) \; \forall\, (x_1, \ldots, x_N) \in X^N.$

Theorem 2. If $S_l^{n,N} = \sum\limits_{j=1}^{n}{}_k \xi_{i_j^l}$, $l \in \mathbf{N}$ and $n < N$ then

$$\log k \geq \lim_{L\to\infty} H(S_L^{n,N} \mid S_1^{n,N}, \ldots, S_{L-1}^{n,N}) \geq \log k - A(N, k)$$

where $A(N, k)$ depends only on N and k and tends to zero with $N \to \infty$, more precisely $A(N, k) \leq \dfrac{4k \log k}{N} + \dfrac{2 \ln N/2}{N} \; \forall N \in \mathbf{N}$.

Proof. Like in the first part of Theorem 1 it is seen immediately that

$$\log k \geq \lim_{L \to \infty} H(S_L^{n,N} \mid S_1^{n,N}, \ldots, S_{L-1}^{n,N}) = \lim_{L \to \infty} L^{-1} H(S_1^{n,N}, \ldots, S_L^{n,N}) \geq$$

$$\geq H(S_1^{n,N} \mid \xi_1, \ldots, \xi_N) \geq H(S_1^{n,N} \mid \xi_1, \ldots, \xi_N, M^{n-1,N}) =$$

$$= \sum_{\substack{\{i_2, \ldots, i_n\} \subseteq \{1, \ldots, N\} \\ (x_1, \ldots, x_N) \in X^N}} P(M^{n-1,N}\{i_2, \ldots, i_n\})\, P(\xi_1 = x_1, \ldots, \xi_N = x_N) \times$$

$$\times H(S_1^{n,N} \mid x_1, \ldots, x_N; \{i_2, \ldots i_n\}).$$

For every $\{i_2, \ldots, i_n\} \subseteq \{1, \ldots, N\}$ and $(x_1, \ldots, x_N) \in X^N$

$$H(S_1^{n,N} \mid x_1, \ldots, x_N; \{i_2, \ldots, i_n\}) = H(\xi_{i_1} \mid x_1, \ldots, x_{i_2-1}, x_{i_2+1}, \ldots, x_{i_n-1}, x_{i_n+1}, \ldots, x_N) =$$

$$= H(\xi_{i_1} \mid \bar{x}_1, \ldots, \bar{x}_{N-n+1}) \text{ with } (\bar{x}_1, \ldots, \bar{x}_{N-n+1}) =$$

$$= (x_1, \ldots, x_{i_2-1}, x_{i_2+1}, \ldots, x_{i_n-1}, x_{i_n+1}, \ldots, x_N).$$

If $(\bar{x}_1, \ldots, \bar{x}_{N-n+1}) \in T_{N-n+1,\varepsilon}$ then similarly to the proof of Theorem 1, $H(\xi_{i_1} \mid \bar{x}_1, \ldots, \bar{x}_{N-n+1})$ can be estimated from below by

$$-\log\left(\frac{1}{k} + \varepsilon\right) = \log \frac{k}{1 + k\varepsilon}.$$

Therefore

$$H(S_1^{n,N} \mid \xi_1, \ldots, \xi_N) \geq \log \frac{k}{1 + k\varepsilon} \sum_{(\bar{x}_1, \ldots, \bar{x}_{N-n+1}) \in T_{N-n+1,\varepsilon}} \times$$

$$\times P(\xi_1 = \bar{x}_1, \ldots, \xi_{N-n+1} = \bar{x}_{N-n+1}) = \log \frac{k}{1 + k\varepsilon} P(T_{N-n+1,\varepsilon}).$$

We may assume $n \leq \left[\dfrac{N}{2}\right]$ because of Lemma 3. Here $\left[\dfrac{N}{2}\right] = \dfrac{N}{2}$ if N is even and $\left[\dfrac{N}{2}\right] = \dfrac{N}{2} - 0.5$ if N is odd.

It can be seen easily that $T_{\left[\frac{N}{2}\right],\varepsilon} \subseteq T_{N-n+1,\varepsilon}$ if $\varepsilon > \dfrac{1}{\left[\frac{N}{2}\right]}$ and $n \leq \left[\dfrac{N}{2}\right]$. From that is follows:

$$H(S_1^{n,N} \mid \xi_1, \ldots, \xi_N) \geq P\left(T_{\left[\frac{N}{2}\right],\varepsilon}\right) \log k - P\left(T_{\left[\frac{N}{2}\right],\varepsilon}\right) \log (1 + k\varepsilon) =$$

$$= \log k - A(N, k, \varepsilon) \text{ with } A(N, k, \varepsilon) = P\left(T^c_{\left[\frac{N}{2}\right],\varepsilon}\right) \log k + P\left(T_{\left[\frac{N}{2}\right],\varepsilon}\right) \log (1 + k\varepsilon),$$

$$\text{provided } \varepsilon > \frac{1}{\left[\frac{N}{2}\right]}.$$

If we choose $\varepsilon = \dfrac{\ln\left[\frac{N}{2}\right]}{\left[\frac{N}{2}\right]} > \dfrac{1}{\left[\frac{N}{2}\right]}$ then the bound $A(N, k, \varepsilon)$ depends only on N and k and tends to zero with a speed of convergence

$$\frac{4k \log k}{N} + \frac{2k \ln \frac{N}{2}}{N}.$$

4. Conclusion

There occurs no n in Theorem 2 which means that the redundancy of the source producing the modulo k-sum statistics based on the re-sampling tends to zero if the size of the primary sample is going to infinity. Also, for given primary sample size our theorems give bounds on the redundancy of the source. These bounds may provide insight into the effects of the repeated utilization of the very sample in different applications.

References

1. *Gallager, R. G.*, Information Theory and Reliable Communication, Wiley (1968), New York.
2. *Hellman, M. E.*, On Using Natural Redundancy for Error Detection, IEEE Trans. on Communication, COM **22** (1974) 1690–1693.
3. *Kahn, D.*, The Codebrakers, 5th ed., Mac Millan Comp. (1969), Toronto.
4. *Rényi, A.*, Foundations of Probability, Holden-Day (1970), San Francisco.

N. Kusolitsch
Institut für Statistik
Technische Universität Wien
Argentinierstrasse 8/7
A-1040 Wien
Austria

Problems of Control and Information Theory, Vol. 10 (2), pp. 137–147 (1981)

REAL-TIME DIGITAL FILTERING OF ELECTROCARDIOGRAPHIC SIGNALS BY MEANS OF AN INTEL 8080 MICROPROCESSOR

Z. GYŐRFI, L. MOLNÁR, L. OSTVÁTH, A. POMOZI

(Budapest)

(Received January 28, 1980)

In this paper digital filtering of ECG signals is considered, and also a specific use of an approach, actually implemented for ECG population screening by an I-8080 based microcomputer. First, the authors show that the output of the algorithm, measuring the parameters of the medically important P, Q, R, S, and T waves, does not change significantly, provided the ECG curve is passed through a low-pass digital filter, having a cut-off frequency of 40 Hz. Next, an approach for devising the filtering algorithm is proposed for real-time processing. This algorithm has been actually implemented for ECG analysis with 300 Hz sampling rate. The program computes the filtered value of each ECG sample within 1.5 ms.

1. The statement of the problem

In automatic systems analysing ECG signals (see e.g. [1], [2]) naturally arises the problem of filtering. First of all the noise component originating from the mains supply voltage must be filtered [5], [6], especially in cases when the ECG records are made under relatively poor conditions in terms of electronic measuring technique. Things are like this, for example, at the so-called mass-screening, i.e. when a great many of patients have to be examined in quick succession. The system in which the filtering problem has been raised, is shown schematically in Fig. 1.

The microcomputer is assigned the duty of compressing the sampled and quantized signal (this task may be considered as source coding) and transmitting this compressed data sequence to a digital magnetic tape recorder which functions as a channel in the information transmission scheme. The final processing of the ECG signal stored in digital form on the magnetic tape is made on a medium-size general-purpose computer, in our case on an IBM 370/115 model. This processing essentially consists of measuring a number of important parameters of the ECG curve (e.g. the amplitudes and the durations of the Q, R, S, and T waves) on the one hand, and of the making of a standard medical ECG certificate on the basis of these parameters on the other hand.

The afore-mentioned task of source coding and the measuring of the parameters requires ECG records, from which all the noise-components originating from the mains supply voltage and other sources are removed,

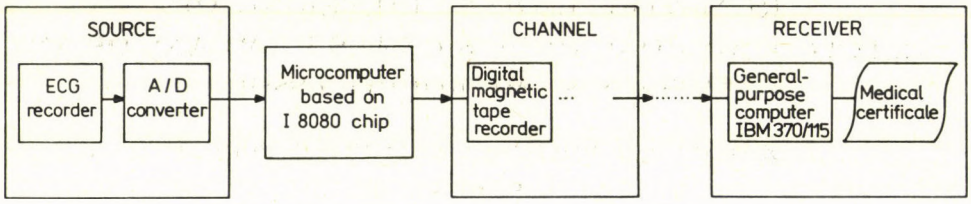


Fig. 1. The scheme of the ECG processing

while preserving the essential biophysical information content at the same time. Accordingly, the microcomputer has to perform a filtering function, too. Moreover, this filtering is to be performed within such a short time so as to secure the absolute continuousness of the mass-screening process.

Hence, the main problems to be solved are the following:

(a) the design of the frequency response characteristics of the filter,
(b) the realization of a filter which has this frequency characteristics.

In the rest of this paper these two problems will be treated in some detail.

2. The design of the frequency response characteristics

In order to prepare the adequate determination of the frequency response characteristics of the filter, spectral analysis was made on a number of typical ECG curves, which were selected out of archives consisting of about 2000 normal and abnormal records. The power-density spectrum of the chosen ECG records were computed. A few of these are shown in Figs 2b, 3b and 4b. These power-density vs. frequency curves were determined in such a way that the respective ECG records (the most important intervals of these ECG's, i.e. the so-called QRS complexes, are shown in Figs 2a, 3a, and 4a) were successively filtered by means of a digital band-pass filter having a band-width of 2 Hz, increasing the position of the pass-band in each step of the procedure by 2 Hz from zero to 150 Hz, and the squared mean of the output was computed.

In Fig. 2a an abnormal ECG record having a depressed ST-junction can be seen. Figure 3a shows a period of a normal, relatively noiseless ECG record. Finally, in Fig. 4a a medically normal ECG is shown which is contaminated with a strong 50 Hz (mains-supply) noise component. It appears from the demonstrated typical ECG's that the essential components of the ECG curves, in terms of the power-density function, are located in the low-frequency band: from zero to about 40 Hz.

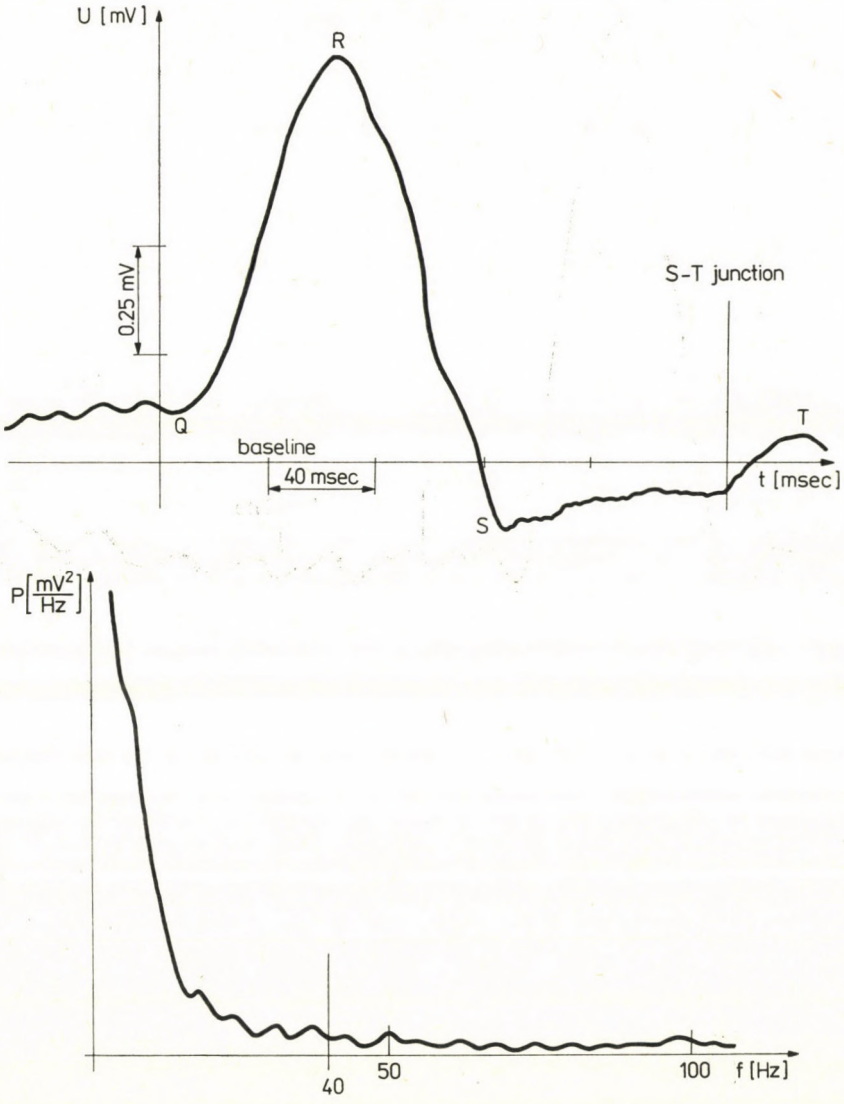
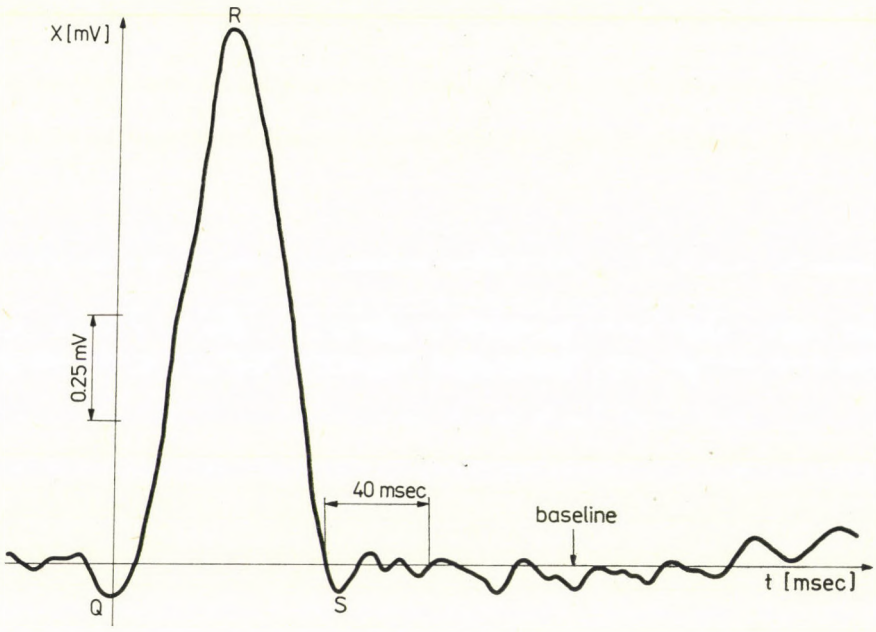


Fig. 2. Abnormal ECG with a depressed ST-junction

With a view to support this assertion, the variation of the output values of an algorithm developed previously for the measuring of the parameters of the Q, R, S and T waves, vs. the cut-off frequency of the low-pass filter was investigated. It became evident from this investigation that the ECG signal, in terms of the above-mentioned, medically important parameters,



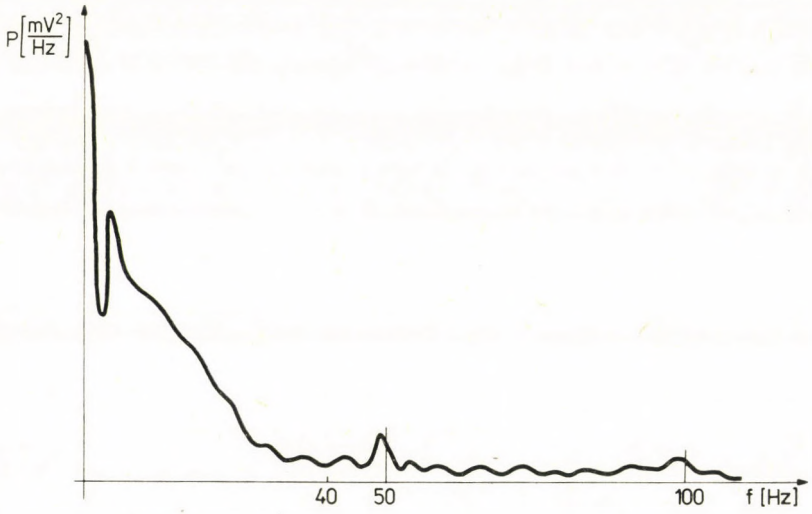


Fig. 3. Relatively noiseless, normal ECG

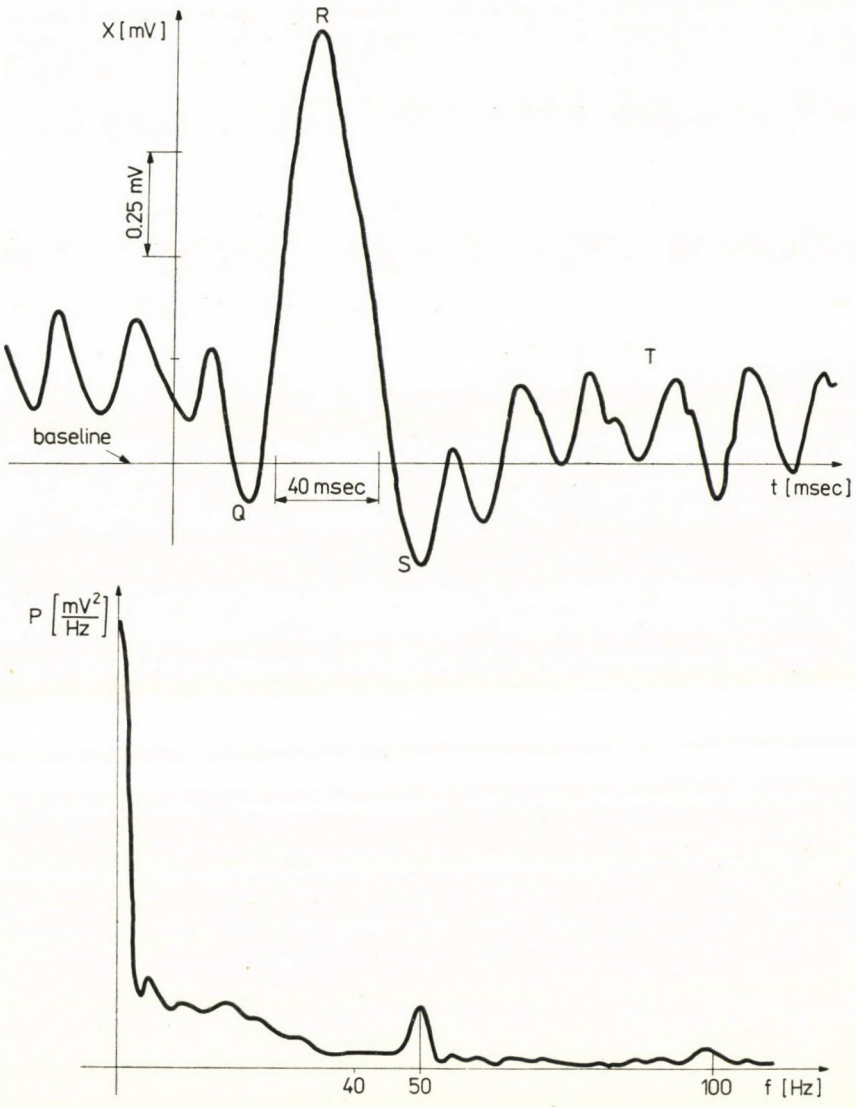


Fig. 4. Normal ECG contaminated with 50 Hz noise

can be appropriately described by its low-frequency components occupying the 0–40 Hz range. Whether this bandwidth is really sufficient for preserving the total information content of the ECG signal, is an open question (see e.g. [2]). After all, a digital low-pass filter, having or approximating a frequency-response characteristics of Fig. 5, must be realized.

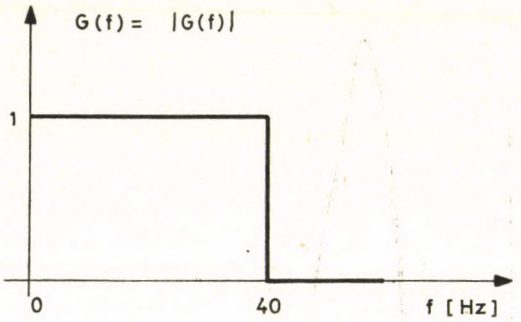


Fig. 5. Characteristics of an ideal low-pass filter

3. The symmetric, transversal digital filter

Let $\{x_n\}_{n=1,2,\ldots}$ denote the sampled input data-sequence to be filtered, and $\{y_n\}_{n=1,2,\ldots}$ the corresponding output sequence. Then

$$(*) \qquad y_n = \sum_{i=-N}^{N} c_i x_{n-i}\,; \qquad n = N+1,\ N+2, \ldots;\ c_i = c_{-i}$$

where the numbers $\{c_i\}_{i=1,2,\ldots,N}$ are called the coefficients of the symmetric, transversal digital filter.

If the signal to be filtered is band-limited to $[0, F]$, and its values, sampled at Δ intervals, can be described in the form of the inverse Fourier–Stieltjes transform

$$x_n = \int_{-F}^{F} e^{in\Delta 2\pi f} \cdot d\mathfrak{F}(f)\,; \qquad n = 1, 2, \ldots$$

then

$$y_n = \int_{-F}^{F} e^{in\Delta 2\pi f} \cdot G(f)\, d\mathfrak{F}(f)\,; \qquad n = N+1,\ N+2, \ldots$$

where $\mathfrak{F}$ stands for the Fourier–Stieltjes transform of the ECG record to be processed, and G denotes the transfer characteristics of the filter:

$$G(f) = 2\left[\frac{c_0}{2} + \sum_{i=1}^{N} c_i \cos\left(i\Delta 2\pi f\right)\right].$$

It can be seen that the transfer characteristics of the filter, G, is determined by the coefficients $\{c_i\}_{i=1,2,\ldots,N}$, which can be considered, in this

context, as Fourier-coefficients. For this reason, on the other hand, these coefficients can be computed, if G is given, by means of the formula

$$c_i = 2\Delta \int_0^{1/2\Delta} G(f) \cos(if\, 2\pi\, \Delta)\, df; \qquad i = 0, 1, \ldots, N.$$

Considering, furthermore, the relation $2F = 1/\Delta$, which is prescribed by the Sampling Theorem, and taking into account the physical value of 150 Hz, which is determined by the cut-off frequency of the analogue low-pass filter built into the ECG processing system, then (recalling our previous statement concerning the function G) the values contained in column A of Table 1 will be derived for the coefficients c_i (the value of N was set at 6).

Table 1

Precise and rounded-off values for the coefficients of the digital filter

	A	B
c_0	0.266667	68
c_1	0.236550	60
c_2	0.158283	41
c_3	0.062366	16
c_4	−0.016502	− 4
c_5	−0.055133	−14
c_6	−0.050455	−13

Concerning this computational procedure for the determination of the coefficients of the filter, the following comments must be made. If N is fixed, the method gives the best approximation of the function G in mean-square sense. It is far from evident that this very approximation is the best, in terms of the given algorithmic and medical aims, i.e. the efficiency of the data compressing algorithm, the sensitivity of the parameter measuring algorithm, and the final medical diagnosis, which is based on the previous two procedures. In other words, it is not sure at all that these final aims can be achieved by using the mean-square approximation for G. In this complex situation, however, it cannot be even formulated in exact mathematical terms what to consider as the best approximation.

Hence, the precise filter coefficients computed by this method can be regarded only as a first approximation, and the final setting of the coefficients needs further, subjective considerations. This is all the more true, because

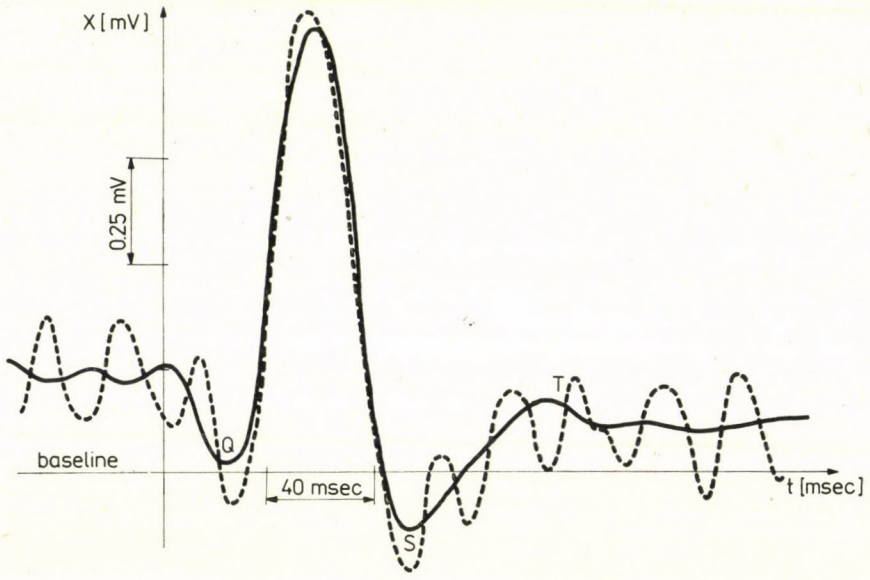


Fig. 6. The ECG curve before (dotted line) and after the digital filter

(as it will be outlined in the following section) the realization of the algorithm by means of a microprocessor, in which the numbers are represented in one *byte* (i.e. in an 8-bit digit), automatically sets limits to the choice of the coefficient values. Nevertheless, as it is well known, there are methods for the computation of the coefficients under special constraints, on the basis of various criteria [4].

Finally, as an illustration, the dotted-line curve in Fig. 6 (which is the same as Fig. 4a) shows the ECG curve, which has the power-density spectrum of Fig. 4b. The filtered equivalent of the very same ECG record is shown by the continuous line curve in Fig. 6. This filtered ECG was computed with $N = 50$, which gives "absolute precise" coefficient values for practical purposes.

4. The realization of the filter by means of a microprocessor

As the formula (*) of Section 3 shows, the algorithm of the filter is very simple: only multiplications and additions have to be made. Because of the real-time realization of the algorithm, one has, however, only very short time to carry out these arithmetic operations, considering the value $\Delta = 3.3$ ms, and the fact that the time-demand of the other operations, i.e. sampling,

quantizing, compressing of the data (which must be carried out also within this 3.3 ms interval) is about 1.5 ms.

First of all we had to investigate the problem whether N can be decreased or not, and (if the answer is positive) in which extent. As one can see from the basic formula (*), the time-demand of the computation increases linearly with N. The choice $N = 6$ proved to be appropriate. Figure 7 shows two filtered versions of the same original ECG curve with $N = 20$ and $N = 6$, respectively. The difference between the two curves seems negligible.

Besides, if the parameters of the Q, R, S, and T waves are measured on these two filtered ECG's, the resulting difference between the two respective parameter sets is much less than the accuracy with which a medical doctor can subjectively read the same parameters, surveying the paper-tape output of a standard ECG apparatus.

The precise values of the filter coefficients have been already listed in column A of Table 1. These numbers, however, cannot be directly represented on the I-8080 microprocessor, which has a byte-oriented memory organization. Because of this hardware limitation, the integer part of a product (the original values multiplied by 2^8) must be chosen as the real coefficients (see column B of Table 1), instead of the precise values.

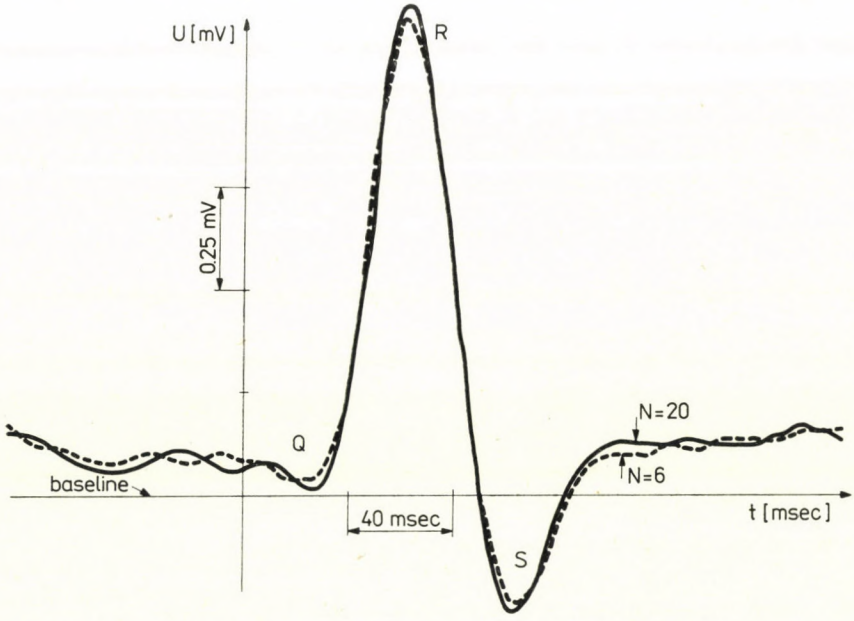


Fig. 7. Two filtered versions of the same ECG

The second computational problem is the realization of the multiplications. It seems that the problem cannot be solved within the given short period of time, using exclusively software methods. Although it is technically possible to use a hardware multiplier which operates quickly enough, this is a somewhat costly solution to the problem: the price of such a multiplier is about three times as much as the price of the I-8080 chip itself.

In terms of cost and time-demand, the application of a multiplication table gives the optimal solution. The multiplication table contains the products of the rounded-off values for the filter coefficients (see column B of Table 1) and all the 4- or 7-bit integers (or possibly the two-complement codes of the numbers, which can be represented in 8 bits). This method, however, requires a pretty large amount of memory capacity.

The third problem arose at the programming of the filter algorithm. That is to say, it turned out that using a hardware- or a multiplication-table method, the time-demand of the frame program organizing the due sequence of the various arithmetic operations well surpasses that of the multiplications itself. Let us carefully examine the situation! During the entire execution of the filtering procedure, $2N + 1$ samples must be stored, the actual stored samples must be shifted in the memory to reach their assigned locations within

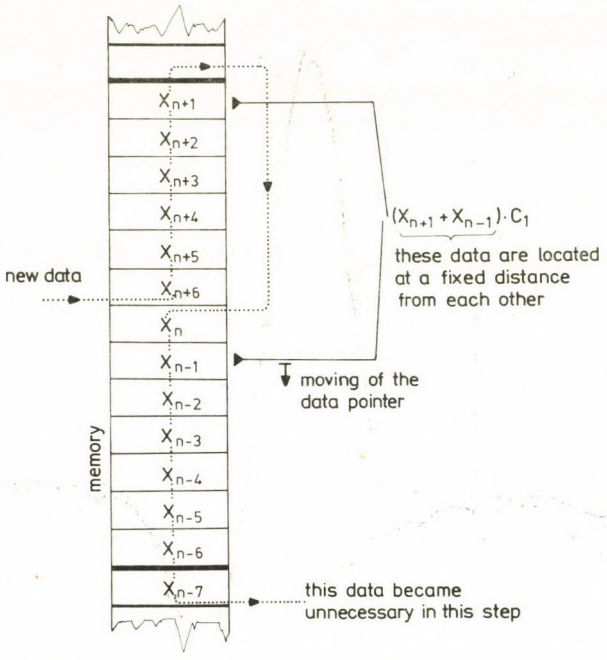


Fig. 8. The optimal organization of the data-flow

each sampling period, and the new sample must be placed in the appropriate memory cell. Meanwhile, the sample pairs, located symmetrically to the right and to the left of the actual instant of time, respectively (for which the filtered value will be computed at the present step), and the filter coefficients must be read out. The operation of the frame program will be the quickest possible, if the flow of the data is according to Fig. 8.

Naturally, the frame program can be organized in several different ways, too. But in this case we are either forced to give up the addition of the symmetrically located sample pairs before the multiplication, or we must determine the memory locations of the addable sums by means of a separate calculation. In both cases a considerable amount of time will be lost.

References

1. *Pipberger, H. V.*, Computer analysis of the electrocardiogram. Computers and Biomedical Research. Academic Press, 1965.
2. *Berson, A. S.*, Bandwidth, sampling and quantizing for automated ECG processing, "Computers in Cardiology" conference, 1965.
3. *CAC: Computer Assisted Cardiology System on a R-10 minicomputer* (developed by a team of the "Research Group for Informatics and Electronics", Hungarian Academy of Sciences), demonstrated at the *4th International Conference on Electrocardiology*, Balatonfüred (Hungary), 1977.
4. *Rabiner, L., Gold, G. V.*, Theory and application of digital signal processing. Prentice Hall, 1975.
5. *Macfarlane, P. W., Lawrie, T. D. V.*, An introduction to automated electrocardiogram interpretation. Butterworths, 1974.
6. *Lynn, P. A.*, Recursive digital filters for biological signals. Medical and biological engineering, vol. **9**, pp. 37–43, 1971.
7. *Richard, R. K.*, Arithmetic operations in digital computers. Van Nostrand, 1976.

Фильтрация электрокардиологических сигналов с помощью микропроцессора I-8080 в реальном масштабе времени

З. ДЕРФИ, Л. МОЛНАР, Л. ОШВАТ, А. ПОМОЗИ

(Будапешт)

Цифровая фильтрация электрокардиологических сигналов (ЭКГ) обсуждается в статье. Алгоритм фильтрации осуществляется с помощью одной микро — ЭВМ тип INTEL 8080. Доказывается чтобы принять во внимание только компоненты между 0 и 40 Hz ЭКГ сигналов достаточно. Иначе говоря: вывод алгоритма для измерении параметров P, Q, R, S и T (которые важные с медицинской точки зрения) не изменяется значительно если сигнал ЭКГ фильтровается фильтром низких частот который имеет предельную частоту 40 Hz. Наконец предлагается метод программирования который возможно применять в случае обработки в реальном масштабе времени при 300 Hz частоты развертки потому программа ЭВМ вычисляет одну величину развертки ЭКГ сигнала во 1,5 ms.

Z. Győrfi
L. Molnár
L. Osváth
A. Pomozi

Institute for Electronics of Communication
Technical University of Budapest
H—1111 Budapest, Stoczek u. 2.
Hungary

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

L. Györfi

Technical University of Budapest

H–1111 Budapest XI., Stoczek u. 2, 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 (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 th paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) 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 статей в четырех экземплярах на русском языке или в двух на английском и в двух на русском языке следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в двух экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Outrata, J. V., Schindler, Z.: An augmented Lagrangian method for a class of convex continuous optimal control problems (*Оутрата И. В., Шиндлер З.* Метод обобщенных функций Лагранжа для решения одного класса выпуклых непрерывных задач оптимального управления) 67

Tuzar, A., Beran, Z.: Optimization of the discrete description for continuous systems (*Тузар А., Беран З.* Оптимизация дискретного представления непрерывных систем) 83

Pugachev, V. S.: The finite-dimensional distribution of a random process determined by a stochastic differential equation and their application to control problems (*Пугачев В. С.* Конечномерные распределения случайного процесса, определяемого стохастическим дифференциальным уравнением, и их приложение к задачам управления) 95

Shpilewski, E. K.: Dynamical recognition of random processes and systems in continuous time (*Шпилевский Э. К.* Динамическое распознавание случайных процессов и систем в непрерывном времени) 115

Kusolitsch, N.: On a re-sampling problem in the theory of information (*Кусолитш Н.* О повторных выборках в задачах теории информации) 129

Győrfi, Z., Molnár, L., Osváth, L., Pomozi, A.: Real-time digital filtering of electrocardiographic signals by means of an INTEL 8080 microprocessor (*Дерфи З., Молнар Л., Ошват Л., Помози А.* Фильтрация электрокардиологических сигналов с помощью микропроцессора I-8080 в реальном масштабе времени) 137

HU ISSN 0370—2529

Index: 26 660

316.920

VOL. ТОМ 10 * NUMBER НОМЕР 3

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1981

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

Is an international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Roumania, Viet-Nam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H–1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1981 Subscription Rate US Dollars 110.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)

G. BOGNÁR (Budapest)

COORDINATING EDITORS

S. V. EMELYANOV (Moscow)
N. N. KRASOVSKII (Sverdlovsk)
E. P. POPOV (Moscow)
V. S. PUGACHEV (Moscow
V. I. SIFOROV (Moscow)
E. D. TERYAEV (Moscow)
T. VÁMOS (Budapest)
L. VARGA (Budapest)
A. PRÉKOPA (Budapest)
S. CSIBI (Budapest)
I. CSISZÁR (Budapest)
J. KOCSIS (Budapest)
J. BENEŠ (Prague)
V. STREJC (Prague)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)

Г. БОГНАР (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

С. В. ЕМЕЛЬЯНОВ (Москва)
Н. Н. КРАСОВСКИЙ (Свердловск)
Е. П. ПОПОВ (Москва)
В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Е. Д. ТЕРЯЕВ (Москва)
Т. ВАМОШ (Будапешт)
Л. ВАРГА (Будапешт)
А. ПРЕКОПА (Будапешт)
Ш. ЧИБИ (Будапешт)
И. ЧИСАР (Будапешт)
Я. КОЧИШ (Будапешт)
И. БЕНЕШ (Прага)
В. СТРЕЙЦ (Прага)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

Problems of Control and Information Theory, Vol. 10 (3), pp. 149—160 (1981)

SYNTACTIC RECOGNITION OF NONSTYLIZED HANDWRITTEN NUMERALS

J. PIK

(Prague)

(Received June 1, 1980)

A syntactic recognition system of nonstylized handwritten numerals 0–9 is considered in this paper. The proposed method includes description of a two-dimensional two-valued input array containing a number of the presented numerals, extraction of a single numeral description and its useful interpretation to enable the use of one-dimensional grammatical forms, and specifications of character classes using three-level hierarchical description systems. Parallel processing can be performed at some method stages. In addition to the recognition of perfectly specified numerals, recognition of ill-formed ones is also possible as mentioned.

1. Introduction

Automatic recognition of unconstrained handwritten characters is counted among the most important recognition problems. The high variability of handwritings and the noise in input patterns considerably complicate the problem. To solve it, many pattern recognition techniques have been proposed [1]–[3].

This paper deals with recognition of nonstylized handwritten arabic numerals using the syntactic approach [4]. The considered numerals and their positions in the input array, roughly speaking, are subjected to the same requirements as there are those for their reasonable human recognition.

The input pattern of the proposed system is a two-dimensional two-valued array where the mutual positions of the elements of the same value form the presented numerals. An unknown number of the digitalized characters 0–9 is supposed to be in this array. As we do not deal with real time recognition we cannot utilize any time or position function feasible during character generation. Because of the possibility of arbitrary numeral locations in the input array, numerals are required to be represented by connected digital patterns (with an exception mentioned below) and they do not intersect each other. The processing of the input array using the empirical feature system is easy and, moreover, parallelism can be realized. The resulting description of the whole input array contains the descriptions of all presented numerals. To extract the single ones, the proposed algorithm utilizes an ability of feature

ystem to reflect the connectivity of digitalized numerals. The extracted two-sdimensional single numeral description is interpreted as a one-dimensional string which is processed by a syntactic recognition system. A decision procedure based on the use of hierarchical description systems [5], [6] has been proposed. Three-level hierarchical systems specify separate classes of the characters 0–9 and their use can simplify and speed up the recognition of some numerals.

The considered method has originated in an example of applications of hierarchical description systems. A high degree of correct experimental results has led to the continuation of research of this problem.

2. Recognition system

A block diagram of the considered recognition system is shown in Fig. 1, To clarify the functions of the separate blocks in a more detailed manner. the feature system, the extraction of the single numeral representation, and the specification of the character classes are described below.

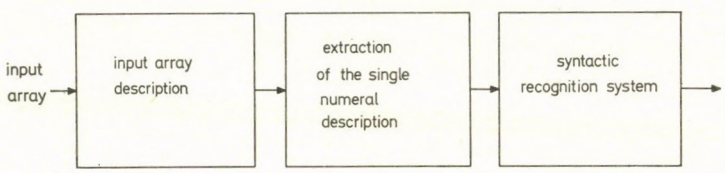


Fig. 1. Block diagram of the considered recognition system

(i) *The feature system*

The input pattern of the system is a two-dimensional two-valued array containing an unknown number of connected non-intersecting digitalized patterns that correspond to the presented nonstylized handwritten numerals 0–9. In what follows we shall represent these numerals by connected sequences of 1's in neighbourhoods of 0's as depicted in Fig. 2. From the point of view of single input array rows, the numerals are thus formed by row clusters of 1's separated by 0's.

The simple feature system is used to obtain the description of the input array. The basic idea of the system follows from the analysis of the considered numerals and consists in their suitable representation by means of the occurrences and the configurations of some topological features. To find them, those operations are used which take into account one or two horizontal rows of the input array.

The input description contains two parts. The first one consists in the characterizations of the row clusters by their lengths, i.e., by the number of their 1's. One-element horizontal gaps are bridged so that the only cluster is considered when clusters are separated by single 0's; the length of the resulting cluster is the sum of the partial lengths. The obtained lengths are arranged in

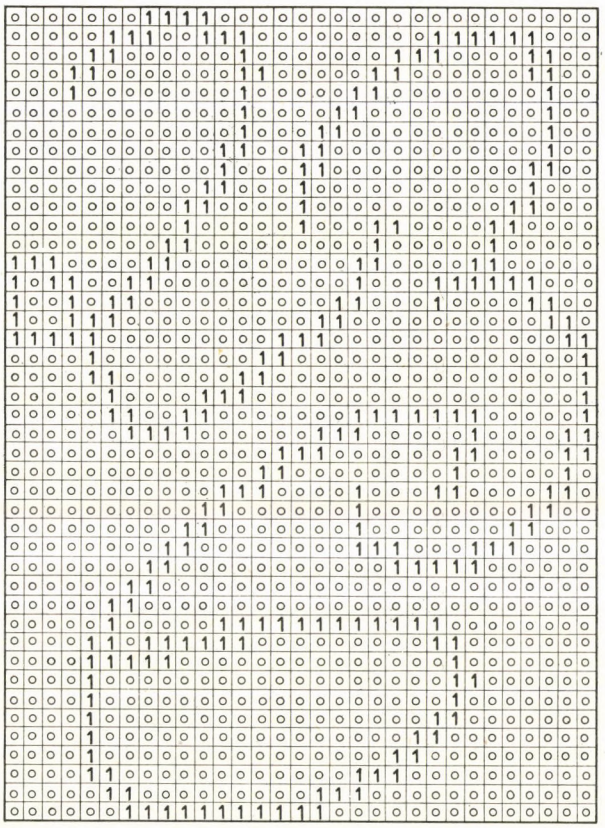


Fig. 2. Example of the input array with numerals 2, 3, and 6

a two-dimensional matrix like locations of the corresponding clusters in the input array, see Fig. 4. An effective input data reduction results from the proposed way of description; while the vertical size of this matrix can equal that of the input array, the number n_m of the non-zero matrix columns is always smaller (usually much smaller) than the corresponding n_i and depends principally on the input array complexity ($n_m < \frac{n_i}{3} + 1$, where n_i denotes the number of the considered input columns).

To obtain the second part of the description of the input array, the location of every row cluster is tested with respect to locations of clusters of the foregoing row. In this process, the input array is completed by the single rows of 0's at the top and the bottom, and two rows of this extended array are always taken into consideration. Apart from cluster

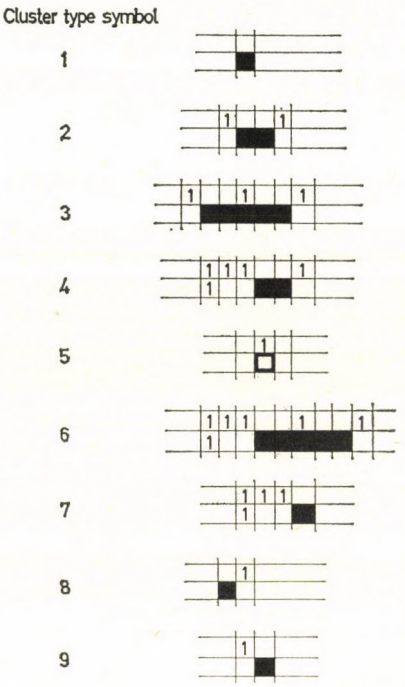


Fig. 3. The basic forms of the cluster configurations

lengths, the location of the outer 1's of every cluster of the lower row is compared with those of the upper one. On the basis of this, every cluster of the lower row is classified into nine groups of mutual configurations. The basic considered forms are shown in Fig. 3, where the classified clusters are filled and the cluster representations of the minimal lengths are used. Cases of clusters that are longer than depicted in Fig. 3 are viewed analogously. The separate considerations of the horizontal two rows strips make possible the parallel processing of the input array. Using the basic cluster configurations, difficulties may be encountered in the cluster classifications, because some clusters can be classified simultaneously both as left (a type-8 cluster) and right (a type-9 cluster) continuation. The following rules exclude this ambiguity. Such a

cluster is classified as a) a type-8 cluster if the corresponding cluster of the foregoing row is classified as a type-1, 2, 3 or 8 one, b) a type-9 cluster in the other cases. The cluster type symbols are arranged in a two-dimensional matrix in the same way as the first part of the description, i.e., as the cluster locations in the input array.

4	0	0	0	0
3	3	6	0	0
2	1	3	2	0
2	2	2	2	0
1	1	2	1	0
0	1	2	1	0
1	2	1	0	0
2	2	1	0	0
1	2	2	0	0
2	1	1	0	0
2	1	2	0	0
1	1	2	2	0
2	0	1	1	0
3	2	2	2	0
4	2	1	6	0
1	4	2	1	2
1	3	2	0	2
5	3	2	0	0
1	2	1	0	0
2	2	1	0	0
1	3	1	0	0
2	2	7	1	0
4	3	1	2	0
0	3	2	2	0
2	1	1	0	0
3	1	2	2	0
2	1	0	2	0
2	1	2	0	0
2	3	3	0	0
2	5	0	0	0
2	0	0	0	0
2	0	0	0	0
1	12	0	0	0
9	2	0	0	0
5	1	0	0	0
1	2	0	0	0
1	1	0	0	0
1	2	0	0	0
1	2	0	0	0
1	2	0	0	0
2	3	0	0	0
2	3	0	0	0
11	0	0	0	0
0	0	0	0	0

a)

1	0	0	0	0
8	7	1	0	0
8	9	8	7	0
8	9	8	9	0
8	8	8	9	0
5	8	8	9	0
8	8	9	0	0
8	8	9	0	0
8	8	8	0	0
8	8	8	0	0
8	8	8	0	0
8	8	1	8	0
8	5	8	8	0
1	8	8	8	0
9	8	8	8	0
9	4	8	8	7
9	8	8	5	9
2	8	9	0	0
9	8	9	0	0
9	8	9	0	0
9	8	9	0	0
9	8	1	9	0
2	8	7	8	0
5	8	8	8	0
8	8	8	0	0
8	1	8	8	0
8	8	5	8	0
8	8	8	0	0
8	9	8	0	0
8	2	0	0	0
8	5	0	0	0
8	0	0	0	0
8	1	0	0	0
2	7	0	0	0
8	9	0	0	0
8	9	0	0	0
8	8	0	0	0
8	8	0	0	0
8	8	0	0	0
8	8	0	0	0
9	8	0	0	0
9	8	0	0	0
2	0	0	0	0
5	0	0	0	0

b)

Fig. 4. The first (a) and the second (b) parts of the description of the input array from Fig. 2

(ii) *The extraction of the description of the single numeral*

As the two resulting matrices contain the whole input array description we are faced with the requirements of the extractions of their parts corresponding to the single presented numerals. The connectivity of the digitalized numerals reflected by the feature system results in simple rules of the cluster type description syntax. Obviously, if there exists a type-4, 6, 7 cluster, then it

necessitates a cluster of the same numeral on the same row according to Fig. 3. Further, type-2, 4 and type-3, 6 clusters necessitate two and three clusters on the foregoing row, respectively.

Beginning from the left top entry of the cluster type matrix, the extraction algorithm takes successively into consideration the elements of the first column and removes them in an auxiliary one-column matrix. If these elements are different from 1's or 5's then the next elements on the same row are tested to discover a line disconnection; the symbols 4, 6, 7 indicate it. If the symbol 1 is discovered, the algorithm scans the column section headed with this one and a new auxiliary matrix is used to store this section. The scanning of the

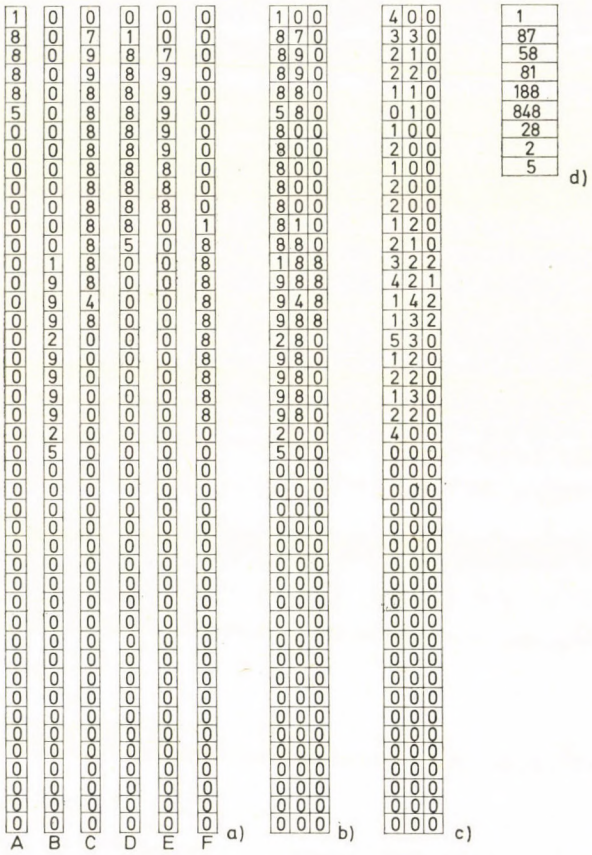


Fig. 5. a) The auxiliary one-column matrices obtained during the extraction of the first numeral (namely "2" in Fig. 2) description; b) The second part of the description resulted from the matrices A, B, C and F; c) The corresponding first part of the description; d) The first level word that is quite sufficient to the recognition of the considered numeral "2"

just considered column section terminates when the symbol 5 or a line connection (with an exception of its first cluster) are found. Now, the rows of the cluster type matrix are shifted to the left to replace the extracted first column section and the algorithm starts the scanning of the newly formed first column. However, before it returns at the column top, it extracts the sections corresponding to the discovered line disconnections and the sections whose extractions were interrupted owing to the occurrences of 1's. In this way, the column sections representing the numeral parts are successively removed from the original matrix and stored in a set of auxiliary matrices, see Fig. 5.

Some of these matrices are merged into a new one to obtain the cluster type description of the single numeral. This is a case, when the mentioned syntactic rules concerning elements of resulting matrix are satisfied and all discovered parts of this numeral are considered. The extraction of lengths of clusters corresponding to the numeral is performed utilizing the same arrangement of both parts of the description. To enable the use of one-dimensional grammatical forms for the representation of the character classes, the two-dimensional cluster type description is transformed into a one-dimensional one. For this purpose, each row of the cluster type description of the single numeral is considered as some symbol of a new alphabet, e.g., the sequence of cluster type symbols of each row can be interpreted as the corresponding decimal number. Every numeral submitted to recognition is thus represented by a vertically arranged word over the alphabet that reflects the horizontal numeral structure. Vertical sequences of the same row symbol may appear in the resulting word. The only symbol is used in the place of each sequence and the first part of the description is completed by the numbers of the symbol occurrences.

(iii) *Specifications of the character classes*

The hierarchical description systems are used for the specifications of classes of the numerals 0–9. From the point of view of formal language theory, the basic idea of the considered systems consists in the repeated application of substitution, possibly of a context-dependent one. In these systems, the generation of words proceeds successively on separate levels by substituting symbols of the currently processed word. The notion of these systems is very similar to that of controlled iteration grammars and, in some cases, very closely related to subgrammars of context-free grammars [5], [6].

The following definition can be given for hierarchical description systems used in this recognition experiment (the systems without interaction [5], [6]).

Definition. A hierarchical description system (without interaction), denoted by HDS, is a system HDS $= (h, \Sigma, S, F)$, where

h is the order of this system, $h > 0$,

Σ is an alphabet (of terminals)

$\Sigma = \bigcup_{i=1}^{h} \Sigma_i$, the alphabets Σ_i, Σ_j are nonempty but not necessarily disjoint for $i, j = 1, \ldots, h$

S is an initial alphabet, also denoted by Σ_0

F is a finite nonempty set of relations

$$F = \{\varphi_1, \ldots, \varphi_h\},$$

where for each $i = 1, \ldots, h$ there exists $r_i \geq 0$ such that $r_i < i$ and φ_i is a finite nonempty relation $\varphi_i \subseteq B_i \times C_i$, where

$$B_i = \Sigma_{r_i} \quad C_i \subseteq 2^{\Sigma_i}$$

and for every $b \in B_i$ there exists $c \in C_i$ such that $(b, c) \in \varphi_i$.

Rewriting of words in HDS is defined as follows.

Definition. Let HDS $= (h, \Sigma, S, F)$ be a hierarchical description system, let for some $i = 1, \ldots, h$ there exist $X \in \Sigma_{r_i}^+$ and $Y \in \Sigma_i^*$, where $X = x_1 \ldots x_l$ with $l \geq 1$ (and $l = 1$ if $r_i = 0$) and $x_j \in \Sigma_{r_i}$ for $j = 1, \ldots, l$. Then Y is directly derived from X in this HDS, in symbols $X \underset{\text{HDS}}{\Longrightarrow} Y$, if the following conditions are satisfied for each $k = 1, \ldots, l$:

$$X = x_1 \ldots x_k \ldots x_l \text{ and } Y = Y_1 \ldots Y_k \ldots Y_l,$$

where

$$Y_k \in c_k, \; c_k \in C_i \text{ such that } (x_k, c_k) \in \varphi_i.$$

The application of HDS is based on a suitable formulation of the pattern recognition problem.

Using the proposed feature system, the words of the first level result from one-dimensional cluster descriptions when the only rows containing at least a one cluster type symbol different from 8 or 9 are considered and the symbols 8 and 9 are not distinguished, the second level ones are just the complete one-dimensional cluster type descriptions, and the third level ones correspond to the first part of descriptions.

Numeral analyses have led to the use of the following HDS for the character classes specifications

$$\text{HDS} = (h, \Sigma, S, F),$$

where

$$1 \leq h \leq 3,$$

$$\Sigma = \Sigma_1 \cup \ldots \cup \Sigma_h$$

$$\Sigma_1 \subseteq \{a_1 a_2 a_3 a_4 a_5 \colon\ a_1, a_2, a_3, a_4, a_5 \in \{0, 1, \ldots, 8\}\},$$

$$\Sigma_2 \subseteq \{b_1 b_2 b_3 b_4 b_5 \colon\ b_1, b_2, b_3, b_4, b_5 \in \{0, 1, \ldots, 8, 9\}\},$$

$$\Sigma_3 \subseteq \{l_1 \ldots l_{10} v_1 v_2 \colon\ l_1, \ldots, l_{10}, v_1, v_2 \in \{0, 1, \ldots, 8, 9\}\},$$

$$S \in \{S_0, S_1, \ldots, S_8, S_9\},$$

$$F = \{\varphi_1, \ldots, \varphi_h\}.$$

The considered recognition system has a partial learning capability; namely, the orders h, the alphabets Σ_1 and Σ_2 and the hierarchical relations φ_1 and φ_2 result from the system's learning. Although the relations of the first levels and the second ones are constructed on the basis of training samples, the type of the corresponding languages has been a priori determined in view of the effective grammatical inference. The sets C_1 and C_2 are supposed to be the sets of finite and regular languages over alphabets Σ_1 and Σ_2, respectively. These assumptions are based on the numeral analyses and make possible the specification of these formal languages by means of finite automata or transition graphs. The third levels utilize the first parts of the numeral descriptions and have been entirely proposed in advance.

The first HDS levels

As the only finite languages are supposed on the first levels, the finite directed labelled trees can specify them. Because of the nonempty intersection of these languages, the only corresponding tree with the labelled final states can be constructed for the representation of their union. The labels of the final states of this tree serve to distinguish a) the character class (including a possibility of the numeral rejection) when the first level description of the submitted numeral is sufficient to its recognition, i.e., when the word accepted due to the considered final state corresponds only to the numerals of the same character class, b) the requirement of the next HDS level consideration owing to the same first level descriptions of numerals of the different character classes.

The second HDS levels

In accordance with the definition of HDS, the words of the second levels are obtained using their first level representations and the hierarchical relations φ_2. The symbols of the corresponding word of the first level are substituted by words of regular languages over the alphabet Σ_2. State graph specifications of those have been used. Every word of the second level is segmented by means of the symbols of the alphabet Σ_1. Since the obtained subwords represent numeral parts contained in horizontal strips of the input array, the languages of this level are simple and can be processed simultaneously, and some of those can be shared by several character classes. It is just the case of descriptions of numeral parts that are unimportant from the view-point of assignment to the classes 0–9, e.g., the only language can correspond to the loops of numerals 6, 8, 9, 0, etc. The unrecognized numerals are submitted to the next HDS level which is the last one in this recognition system.

The third HDS levels

As it follows from the meaning of the symbols of the alphabets Σ_1 and Σ_2, the first and second level decisions take into consideration the occurrences of loops, beginnings and ends of lines, etc., including their mutual locations, and the line direction information extracted from the cluster type description, respectively. In this way, as this third level utilizes the first part of description, the horizontal lengths of row clusters and multiple adjoining occurrences of the same cluster type symbol ("vertical lengths") are considered now.

The numerals that are submitted to third level are, in fact, pre-classified into equivalence groups corresponding to the same descriptions of the preceding HDS levels. In consequence of that, only some numeral parts are sufficient for recognition. Their choice depends on the considered equivalence group and is aimed at an effective partition of these groups into character classes in the teacher's view. A heuristic decision procedure, where the horizontal and vertical lengths are taken into account, has been proposed for this partition.

3. Concluding remarks

Although the considered handwritten numerals are stylized by means of the feature system, another effective stylization is possible to reduce the training set. As forms and lengths of some numeral parts are informationless from the point of view of the recognition of 0-9, these parts can be replaced by their simplest modifications before the corresponding words of HDS are

extracted. To this purpose, the decomposition of the numerals into vertical or verticalwise parts using the mentioned one-column matrices can be utilized. For instance, the open loops with the free ends of minimal lengths are considered for the numerals 1, 2, 3, 5, 6, 7, 9, etc.

Using the knowledge of the proposed HDS, additional requirements can be placed on the presented numerals to ensure their recognition at most on a priori determined level. For example, the processing on the only first HDS level is sufficient if the numerals of the different character classes correspond to the different first level words, i.e., to the different feature configurations. The common requirement of the complex numeral form follows easily from it.

Since a number of the presented digitalized numerals is contained at the input array, the numeral connectivity is required with an exception of one-element row gaps. To make possible such an exception also in the vertical direction, a new feature system with at least three considered rows would be necessary. If the input array contains, however, always the only numeral, the used feature system and HDS are sufficient to be completed by features and substitutions corresponding to the line interruptions, respectively. The constraint of the connectivity can be omitted and, moreover, numerals with partly missing parts can be considered. The only perfectly specified numerals are enough to the system learning if the substitutions corresponding to a priori supposed errors are determined in advance. Owing to possible ambiguities in numeral completion, a set of character classes, instead of just one class, may result from the recognition. Additional criteria are to be used in this case.

The considered recognition system has been simulated on the computer using PL/1 language. Encouraging experimental results have stimulated further research efforts in the field.

Reference

1. *Genchi, H., Mori, K. I., Watanabe, S., Katsuragi, S.*, Recognition of Handwritten Numerical Characters for Automatic Letter Sorting, Proc. of the IEEE **56,** 1292–1301 (1968).
2. *Tou, J. T., Gonzales, R. C.*, Recognition of Handwritten Characters by Topological Feature Extraction and Multilevel Categorization, IEEE Trans. on Comp. **21,** 776–785 (1972).
3. *Ali, F., Pavlidis, T.*, Syntactic Recognition of Handwritten Numerals, IEEE Trans. on Systems, Man, and Cybernetics **7,** 537–541 (1977).
4. *Fu, K. S.*, Syntactic Methods in Pattern Recognition, Academic Press, New York and London (1974).
5. *Pik, J.*, Hierarchical Description Systems and their Use in Pattern Recognition (in Czech), Report of the Institute of Information Theory and Automation, Prague (1976).
6. *Pik, J.*, A Hierarchical Pattern Description in the Syntactic Approach to Pattern Recognition, Pattern Recognition (in press).

Синтаксическое распознавание нестилизованных рукописных цифр

И. ПИК

(Прага)

В статье приводится синтаксическая система распознавания нестилизованных рукописных цифр 0–9. Система представлена в виде нескольких отдельных частей, которые выполняют следующие задачи:

— описание входного двухразмерного поля, содержащего неизвестное количество цифр
— отбор описания одной цифры
— одноразмерная интерпретация полученного описания
— репрезентация классов цифр 0–9 при помощи иерархических систем описания и синтаксическая классификация.

Приводятся основные принципы этих частей и отмечается возможность распознавния цифр с некоторыми ошибками.

J. Pik
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
182 08 Praha 8
Pod vodárenskou věží 4
Czechoslovakia

Problems of Control and Information Theory, Vol. 10 (3), pp. 161–174 (1981)

DEVELOPMENT OF FEEDBACK TYPES AND THEIR APPLICATION TO DESIGN OF CLOSED-LOOP DYNAMIC SYSTEMS

S. V. YEMEL'YANOV, S. K. KOROVIN

(Moscow)

(Received June 10, 1980)

Development of various types of feedback leads to the design of new kinds of closed-loop dynamic systems. This approach is applied to the design of control systems whose properties are invariant for parametric or coordinate disturbances. The approach does not require sliding modes or infinitely large gains in the main control loop. The prospects of the approach are discussed. Examples are given.

1. Introduction. Statement of the problem

Our approach to the structural design of closed-loop dynamic systems requires an expansion of the set of feedback types. The potential of the approach will become obvious in the design of a control system whose properties do not depend much on coordinate or parametric disturbances. Many approaches to the design of such systems have been developed in control engineering. If complete information is available on the system and on exogenous influences, the system can be designed by compensation of disturbances [1–4]. If it is not, the principle of feedback is used in an arrangement shown in Fig. 1 [1, 2, 5]. The need in better ways to compensate parametric disturbances have stimulated development of adaptive control methods to be employed in systems such as that of Fig. 2 [6–11]. Their use instead of those of Fig. 1 has significantly enhanced the scope of tasks which control systems can tackle. Nevertheless, the methods of adaptive control of today presume a quasista-

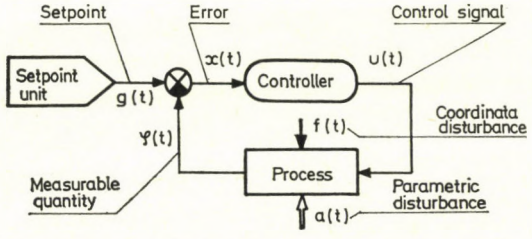


Fig. 1

tionary variation of process parameters, which is not necessarily true [6, 9, 10]. One exception is variable structure systems (VSS) which are made independent of disturbances in sliding modes but these may prove hard to come by [8].

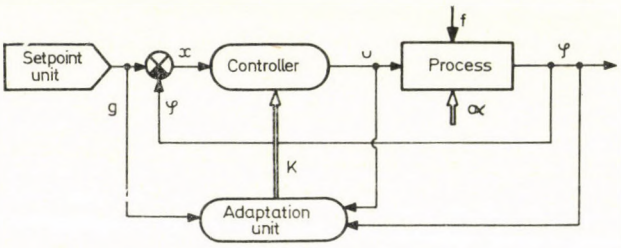


Fig. 2

This is why new approaches to the design of closed-loop dynamic systems are tried; such systems should not depend much on coordinate or parametric disturbances which vary arbitrarily in a wide range. Furthermore, this effect should be obtained with finite gains and without sliding modes or any searching signals in the main control loop. One such approach will be described.

Our reasoning is as follows. Every dynamic process can be represented as a block-diagram of many dynamic elements. In the conventional control theory only one type of these, shown in Fig. 3a, was used. Later, adaptive control theory relied also on another type, Fig. 3b. In this way the variety of

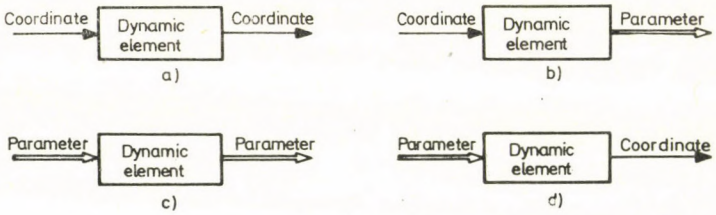


Fig. 3

structural diagrams and potential of control systems were improved. What we propose to do is to give the input and output variables a meaningful interpretation to be used in further expansion of the set of dynamic elements by introducing the types of Figs. 3c and 3d to increase the potential of control systems. The systems to be designed offset deflection from the rated mode. Its simplest form, Fig. 4, has led to numerous types of feedbacks, Fig. 4a will be referred to as coordinate; the second, Fig. 4b as coordinate-parametric; the **third,**

Fig. 4c as parametric; and the fourth, Fig. 4d as parametric-coordinate. Though largely conditional, this classification permits a constructive approach to structural design of systems. Any part of the system, be it a setpoint unit,

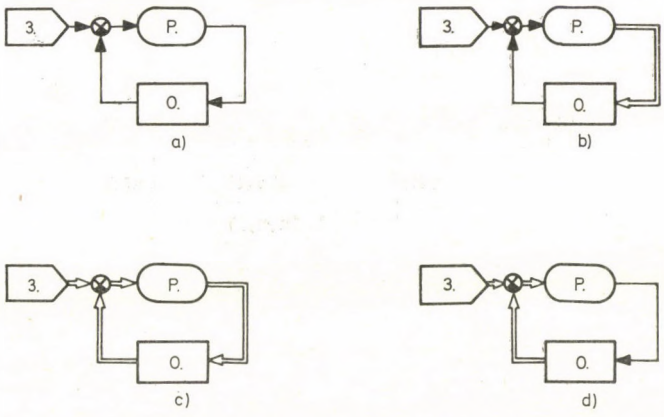


Fig. 4

a controller, or a process, can be regarded as a dynamic element of one of these types and for purposeful change of its properties the necessary type of feedback is used. This approach to structural design of systems calls for investigation of properties and potential of each feedback type and development of ways to apply them in the design of control structures best suited to the task at hand. In Section 2 the above notions will be used to design a generalized structure of a control system.

In conclusion of this Introduction let us note that some types of feedbacks have been used in the design of closed-loop dynamic systems [9, 10] but, as a rule, in the framework of pre-designed structures for improvement of some system properties. Furthermore, there is no terminological unity in as to feedback types [9, 10].

2. Generalized structures of control systems with use of all types of feedback

Let the process be specified by an operator

$$\varphi(t) = L_{\varphi}(U, a(t), f(t)), \tag{2.1}$$

where $\varphi(t)$ is the coordinate to be controlled; U is control, $f(t)$ is coordinate disturbance; and $a(t)$ are process parameters which can vary in time and take on arbitrary values from a certain limited set A. The coordinate $\varphi(t)$ should

reproduce a certain time function $g(t)$ generated by the setpoint unit. A feedback loop, Fig. 5, is usually designed in which the controller operator is to be found. As a rule, that operator is assumed dynamic and from the data on measurable quantities it generates a coordinate signal

$$U = L_u(x, k, f). \tag{2.2}$$

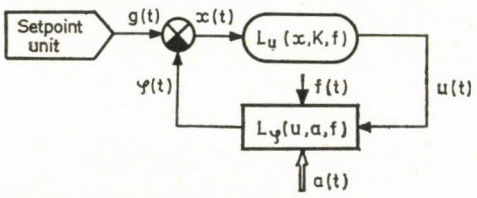


Fig. 5

The type of the operator L_u and its parameters should be chosen so as to obtain a desired law, independent of $a(t)$ and $f(t)$ for the variation of the signal $x(t)$ in a closed-loop system

$$\left.\begin{aligned} \varphi &= L_\varphi(U, a, f) \\ U &= L_u(x, k, f) \end{aligned}\right\}. \tag{2.3}$$

L_u can be chosen by some method of control engineering. But if the parameters k belong to some limited set K, there are no sliding modes in (2.3), and $a(t)$ is substantially variable, conventional methods are not easy to apply. Considering k as free variables, let us use a coordinate parametric feedback which responds to offset the deflection from the rated mode. For this purpose assume that the signal $x(t)$ varies in the desired way if at each time it is close to the signal

$$g^1(t) = L_g(x, c^0), L_g(0, c^0) = 0, \tag{2.4}$$

where the parameters c^0 of the dynamic operator L_g embody the requirements considering dynamic and static properties of system (2.3). Regarding (2.4) as a setpoint unit and introducing a new error

$$x^1(t) = g^1(t) - x(t), \tag{2.5}$$

construct a control loop of Fig. 4b designed to reduce the error $x^1(t)$ to zero

$$k(t) = L_k(x^1, k^1, x), \quad k \in K \tag{2.6}$$

where k^1 are parameters of the controller in the new control loop. What we obtain is a structure (Fig. 6) which contains two control loops of which the one performs coordinate feedback (2.2) and the other performs coordinate parametric feedback (2.6) functions. Let us note two properties of the latter loop which will be important in the further discussion. First, with a linear

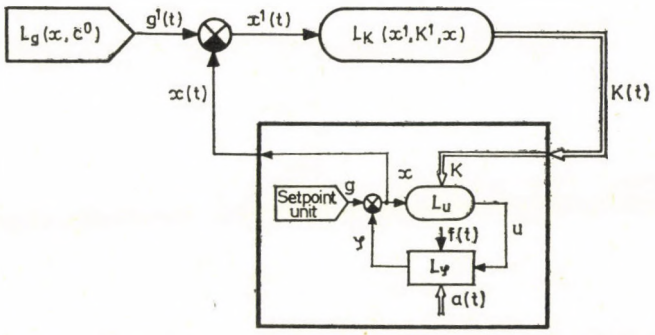


Fig. 6

operator L_u stability in the coordinate parametric loop for controllers (2.6) which has constant structure is essentially impossible because of the changing sign of this feedback. Consequently, (2.6) should allow for appropriate structural changes. Second, for consistency with the problem statement, the parameters k are constrained, $k \in K$, whereas no constraints at all are imposed on the set of values of k^1. This permits making the rates at which the loops operate very different and, before all, change $k(t)$ swiftly to make the equality $x^1(t) = 0$ true. If, starting with a certain time, this condition is met, then the further motion in (2.3) is described by the equation

$$x = L_g(x, c^0)$$

with any $a(t) \in A$ and $f(t)$, which would imply solution of the problem.

Note that the diagram of Fig. 6 permits the following interpretation of motion independence in variable structure systems [8] of parametric or coordinate disturbances. In point of fact, that independence was obtained by using coordinate parametric feedback in the relay law (2.6), which insured an infinitely high rate of change for the parameters k in (2.2) and maintaining the equality $x^1(t) = 0$ in sliding mode. From Fig. 6 the independence is seen to be also obtainable with other controls in a coordinate parametric loop such as linear and so the use of sliding modes is not compulsory.

In the equation $x = L_g(x, c^0)$ the parameters c^0 univalently determine the system properties. There is often no need in this and the system performance can be assumed acceptable for any $c(t) \in C$ where C is a certain limited set. In that case the quantities a can be regarded as controlled variables and an additional control loop can be constructed for their manipulation. In doing so the reasoning is as follows. In variation of the parameters a the process properties vary in a wide range and so with fixed c^0 which univalently specify the properties of the closed-loop system and with the constraints $k \in K$ the performance of the coordinate parametric loop will, generally speaking, vary.

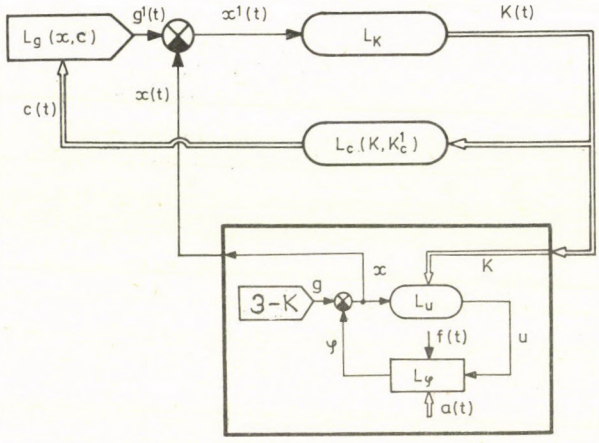


Fig. 7

Since in the vicinity of the desired mode $x^1(t) = 0$ the parameters $k(t)$ contain data on process parameters the parameters c of the setpoint. Unit (2.4) should be changed as a function of the parameters k in order to make the dynamic properties of the system and the setpoint unit similar and facilitate in this way stabilization of the error $x^1(t)$ around zero.

Let, for instance, this relation be given by the operator

$$c(t) = L_c(k, k_c^1), \; c(t) \in C \tag{2.7}$$

where k_c^1 are certain parameters. The structure of Fig. 7 results, in which as an addition to that of Fig. 6, a parametric feedback (2.7) is used. Of course, this feedback makes the properties of system (2.3) dependent on $a(t)$ and $f(t)$. What is important, however, is that their effect stays within admissible limits since $c(t) \in C$.

Introduction of a parametric feedback (2.7) leads generally to the deflection of c from their specified values c^0, which can be interpreted as statism in the coordinate parametric loop. There is a way in which that statism can be eliminated or, at least, reduced. The structure of Fig. 7 can be supplemented with a new parametric coordinate feedback. The data needed for this is contained in the difference $c^0 - c(t)$, and therefore introduction of a parametric coordinate feedback of the form

$$v(t) = L_v(c^0 - c^*(t), k_v^1, x), \tag{2.8}$$

where k_v^1 are certain parameters in the structure of Fig. 8 essentially makes this feasible. Signal (2.8) at the process input makes the dynamic properties of system (2.3) and the setpoint unit (2.5) similar and so reduction of the difference $c^0 - c^*(t)$ in steady state seems feasible.

The structural diagram of Fig. 8 is a natural extension of those shown in Figs 1 and 2 in the direction of using supplementary types of feedbacks. Therefore solution of the problem stated in this article reduces to a proper choice of the operators L_u, L_k, L_g, L_c, and L_v of the feedbacks. The examples which are to follow will show that this choice is possible.

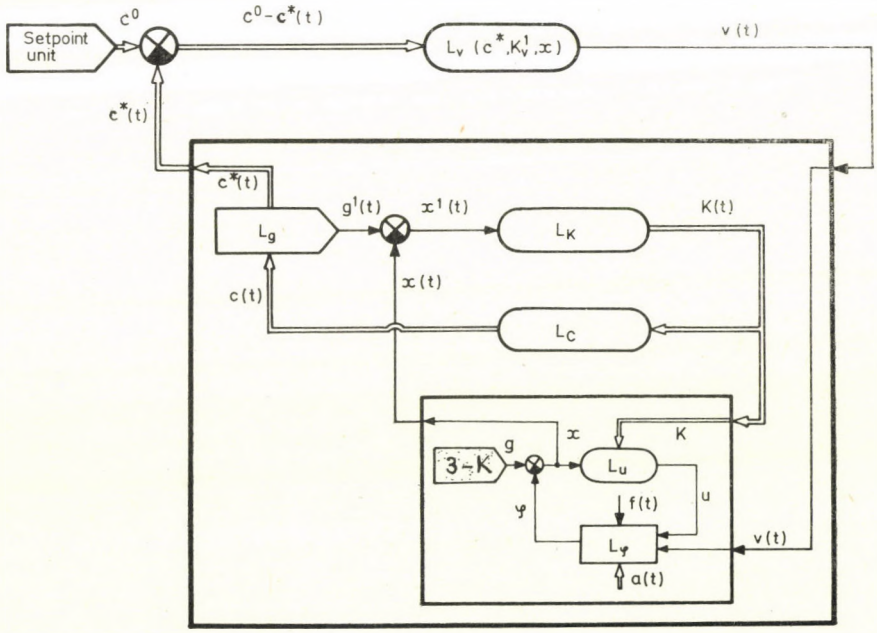


Fig. 8

3. Design of second-order systems

Let the process equation have the form

$$\ddot{\varphi} = aU \tag{3.1}$$

where $a(t) \in [a_{\min}, a_{\max}]$, $a_{\min}$, and $a_{\max}$ are constant positive and known quantities. It is required to choose the control U so as to insure that the controlled coordinate will reproduce the setpoint $g(t)$ which for specificity is assumed to be equal to zero. Let us try to solve this problem by using a linear feedback which responds to variations in the error coordinate x and has a limited gain. The operator L_u in (2.2) is thus linear and the control function is equal to

$$U = kx,\ k \in [-k_0, k_0],\ k = \text{const} > 0. \tag{3.2}$$

From (3.1) and (3.2) it follows that equations of the closed loop system (2.3) have in this case the form

$$\ddot{x} = -akx. \tag{3.3}$$

The system (3.3) is obviously not asymptotically stable on the whole with any fixed k from the range $[-k_0, k_0]$. To solve this problem, let us resort to coordinate parametric feedback and determine the requirements concerning the performance of transient processes in (3.3) by specifying the operator L_g in (2.4) in the form

$$L_g = \frac{d}{dt} + (c^0 + 1),\ c^0 = \text{const} > 0. \tag{3.4}$$

Then the error x^1 determined in (2.5) is equal to

$$x^1 = \dot{x} + c^0 x$$

and its vanishing would imply that since that time system (3.3) behaves as

$$\dot{x} + c^0 x = 0. \tag{3.5}$$

Since solutions of (3.5) are asymptotically stable, asymptotic stability of solutions to system (3.3) follows.

The equality $x^1(t) = 0$ can be strictly insured, for instance, in sliding mode by using relay control laws of the form $k = k_0 \operatorname{sign}(x^1 x)$ in the coordinate parametric loop, which, as noted above, is consistent with methods of the VSS theory. Since we reject the use of sliding modes in the main loop of (3.3) let us consider the potential of a law in the form

$$\dot{K} = \begin{cases} k^1 \operatorname{sign} x^1 & \text{with} \quad k \in [-k_0, k_0], \\ 0 & \text{with} \quad k \notin [-k_0, k_0]. \end{cases} \tag{3.6}$$

It is obvious from, for instance the phase portrait of the system (3.3), (3.5), and (3.6) on the plane (x, x') that the law (3.6) insures asymptotic stability only under certain initial conditions. This is so because the sign of the coordinate parametric feedback changes when that of the signal does. Asymptotic stability would be absolute if (3.6) is replaced by

$$\dot{\varkappa} = \begin{cases} k^1 \operatorname{sign} x^1 & \text{with} \quad \varkappa \in [-k_0, k_0], \\ 0 & \text{with} \quad \varkappa \notin [-k_0, k_0]; \end{cases} \tag{3.7}$$

$$k(t) = \varkappa(t) \operatorname{sign} x,$$

under which the coordinate parametric feedback does not change sign.

A system built in this way will be investigated in the $(x, \dot{x}^1, \varkappa)$ frame of reference. Elementary transformations lead to an equation set

$$\begin{cases} \dot{x} = -c^0 x + x^1 \\ \dot{x}^1 = c^0 x^1 - (a\varkappa + (c^0)^2 \operatorname{sign} x)\,|x|\,, \\ \dot{\varkappa} = \begin{cases} k^1 \operatorname{sign} x^1 & \text{with} \quad \varkappa \in [-k_0, k_0], \\ 0 & \text{with} \quad \varkappa \notin [-k_0, k_0]. \end{cases} \end{cases} \tag{3.8}$$

In system (3.8) the equality $x^1(t) = 0$ cannot obviously be stringently adhered to and in the vicinity of $x^1 = 0$ oscillations may arise whose amplitude would be proportional with x. Determine on the plane $(x, \dot{x})$ a certain vicinity $G(x)$ of the straight line $x^1 = 0$ by specifying its boundaries

$$\begin{aligned} s^+ &= \dot{x} + (c^0 + \Delta c)\, x = 0, \\ s^- &= \dot{x} + (c^0 - \Delta c)\, x = 0, \end{aligned} \qquad G(x) = \{x\colon s^+ s^- < 0\}, \tag{3.9}$$

where Δc is a certain positive quantity. Direct computation reveals that if the inequalities

$$\left.\begin{aligned} k_0 &> \frac{(c^0 + \Delta c)^2}{a_{\min}}, \\ k^1 &> 4\,\frac{k_0^2\, a_{\max}}{\Delta c} \end{aligned}\right\} \tag{3.10}$$

hold, then any system motion which starts with $x^1(0) = 0$ for all $t \geq 0$ and any kind of change in $a(t) \in [a_{\min}, a_{\max}]$ would belong to the region $G(x)$. This implies that even if they do arise in the system, oscillations will stay within the region $G(x)$. Consequently, for solutions of (3.8) the estimates

$$x(0)\, e^{-(c^0 + \Delta c)t} \leq x(t) \leq x(0)\, e^{-(c^0 - \Delta c)t}$$

will hold, which mean that $x(t)$ tends to zero as $t \to \infty$ even though not necessarily monotonously. To insure asymptotic stability on the whole, the equality $x^1 = 0$ should be true at some time. But existence of such a time is near-evident because with k^1 large enough in the region $xs^- > 0$, the system paths on the plane $(x, \dot{x})$ are elliptical at each time, and so $x^1 = 0$ can be reached from this region. On the other hand, in the region $xs^+ < 0$ the paths are near-hyperbolic

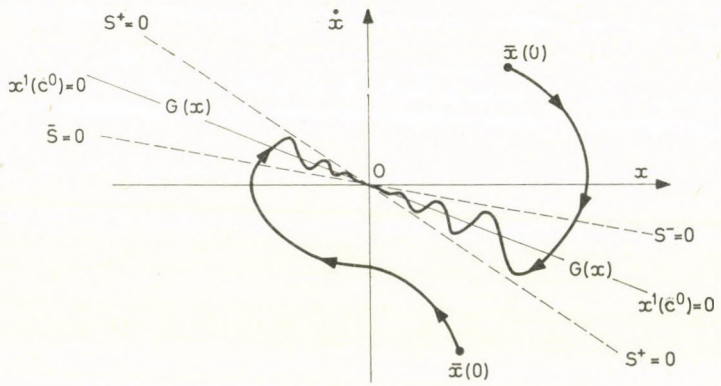


Fig. 9

and the motion along them leads either into $G(x)$ or into $xs^- < 0$. The possible form of phase trajectories of the system under the above conditions is shown in Fig. 9. What is important is that by increasing k^1 to infinity the dimensions of the region $G(x)$, which depend on the value of Δc, can be made as small as desired. This implies that in the limit (as $k^1 \to \infty$) the final phase of the transient process is described by the equation

$$x(t) = x(0)\, e^{-c^0 t}.$$

This effect entails an increase in the frequency of oscillations in the vicinity of $x^1 = 0$. which may be undesirable. These oscillations can be eliminated by existing methods of control engineering such as using a derivative of the signal $x^1(t)$ in the control law of the coordinate parametric loop. This approach calls for additional differentiation of the signal $x(t)$ and will not be discussed. The oscillations can also be eliminated by using an additional parametric feedback as shown in Fig. 7. Let us consider the simplest kind of this feedback by making the operator in (2.7) static so that

$$c(t) = c^0 - \frac{\Delta c}{k_0}\, k(t), \quad \left[k_c^1 = \frac{\Delta c}{k_0}\right]. \tag{3.11}$$

Since $k \in [-k_0, k_0]$ the control law (3.11) insures that the value of $c(t)$ will belong to the segment $[c^0 - \Delta c, c^0 + \Delta c] = C$. The parametric feedback (3.11) changes the system motion equations and (3.8) should be replaced by

$$\dot{x} = -c(t)\, x + x^1,$$

$$\dot{x}^1 = c(t)\, x^1 - \frac{\Delta c}{k_0} k^1 |x| \operatorname{sign} x^1 - (a\varkappa + c^2 \operatorname{sign} x)\, |x| \tag{3.12}$$

and x varies in compliance with the last equation in (3.8). A discontinuous component in (3.12) indicates that a sliding mode may arise in the region $G(x)$ along the straight line $x^1 = 0$. What is significant, however, is that a sliding mode arises in the parametric feedback loop rather than in the main loop (3.3). Sliding mode equations can be found by the method of equivalent control [11] through solving the equation $\dot{x}^1 = 0$ for a discontinuous control $\operatorname{sign} x^1$ under the condition $x^1 = 0$. For our purpose it is sufficient to determine

$$(\operatorname{sign} x^1)_{eq} = \frac{-k_0}{k^1 \Delta c} [a\varkappa + c^2 \operatorname{sign} x]. \tag{3.13}$$

From (3.11) it follows that $\operatorname{sign} \dot{c} = \operatorname{sign} \dot{\varkappa} \operatorname{sign} x$. Substituting the law of $\varkappa$ variation in the region $G(x)$ from (3.8) into this expression we have a relation

$$\operatorname{sign} \dot{c} = -\operatorname{sign} x^1 \operatorname{sign} x$$

which, with the use of (3.13), is easily transformable to the form

$$(\operatorname{sign} \dot{c})_{eg} = \operatorname{sign} [ak + c^2]. \tag{3.14}$$

Now, since according to (3.11), $k = k_0$ with $c = c^0 + \Delta c$ and $k = k_0$ with $c = c^0 - \Delta c$ and for k_0 the inequality (3.10) holds, it follows from (3.14) that

$$(\operatorname{sign} \dot{c})_{eq}\Big|_{c=c^0+\Delta c} < 0 \quad \text{and} \quad (\operatorname{sign} \dot{c})_{eq}\Big|_{c=c^0-\Delta c} > 0.$$

Validity of these relations implies that if a sliding mode arises in the region $G(x)$ the system paths cannot leave that region for all subsequent times and with any kind of variations in $a(t)$. Consequently, and this follows directly from (3.12), in sliding mode the coordinate x varies in compliance with the equality

$$\dot{x} = -c(t)x, \; c(t) \in [c^0 - \Delta c, \; c^0 + \Delta c] \tag{3.15}$$

or the system is asymptotically stable in the region $G(x)$. Whether the region $G(x)$ is reached is found by reasoning as above in analysis of a system which employs a coordinate parametric feedback. From all these facts follows asymptotic stability for all solutions to the system.

A more thorough analysis of the way in which solutions to (3.15) behave requires a study of the equation for variations of the parameter c in sliding mode. This equation is easily obtained from (3.11) and (3.13); omitting the detailed transformations, it can be written as

$$\dot{c} = \left[a \frac{k_0}{\Delta c} (c^0 - c) + c^2 \right]. \tag{3.16}$$

Its remarkable feature is that it is independent of the variables x and $\dot{x}$ and is completely defined by system parameters. Consequently, by a proper choice of the parameters the rates of equations (3.15) and (3.16) can be made to differ and so a faster decay of the transient process in (3.16) is obtained. Thus with the value of c^0 small enough and the ratio $\frac{k^0}{\Delta c}$ high enough* the parameter c can be assumed close to the steady state value c_* which is the root of the equation

$$c_*^2 = - a \frac{k_0}{\Delta c} (c^0 - c_*). \tag{3.17}$$

Then, once a sliding mode has arisen and the transient process in (3.16) has decayed, the solutions will be close to the form $x(t) = x(0)e^{-c_* t}$. The value itself of the parameter c_* will approach c^0 with the increase of $\frac{k_0}{\Delta c}$. A possible kind of motions in such a system is shown in Fig. 10.

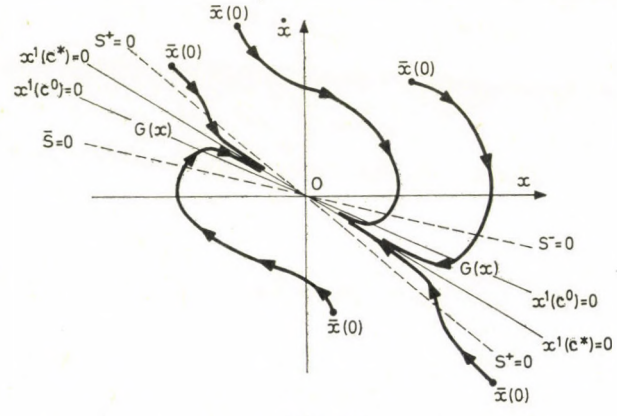


Fig. 10

* Increasing the ratio $k_0/\Delta c$ through increase of k_0 or reduction of Δc, inequalities (3.10) should, of course, be maintained.

Now let us try to make the steady state value of the parameter c_* in (3.17) approach the desired one c^0, instead of increasing the ratio $\frac{k_0}{\Delta c}$, rather by building a parametric coordinate feedback of Fig. 8 in addition to the feedback types which have been used earlier. If we succeed, the steady state solutions will obviously be less sensitive to the parameter a. During motion in the region $G(x)$ we have at each time the data on the difference $c^0 - c(t)$. Let us use this difference in generating a parametric coordinate signal of the form (2.8). For simplification of the reasoning assume that the operator L_v in (2.8) is linear for $c^0 - c(t)$ and for x; the parametric coordinate control v is then

$$v = k_v^1 x(c^0 - c(t)), \tag{3.18}$$

where k_v^1 is a constant coefficient. The behaviour of such a system is described by equations which differ from (3.12) only in that the second equation should incorporate a control

$$\begin{cases} \dot{x} = -c(t)\, x + x^1, \\ \dot{x}^1 = cx^1 - \dfrac{\Delta c}{k_0} k^1 |x| \operatorname{sign} x^1 - (ax + c^2 \operatorname{sign} x)\,|x| - av. \end{cases} \tag{3.19}$$

In (3.19) a sliding mode can also arise on $x^1 = 0$ with motion equations of the type:

$$\dot{c} = a \frac{k_0}{\Delta c}(c^0 - c) + c^2 - ak_v^1(c^0 - c), \tag{3.20}$$

which differs from (3.16) only in that it incorporates an addend $-ak_v'(c^0 - c)$. The equilibrium position of (3.20), c_{**}, can be seen to be the root of the equation $-a\left(1 + k_v^1 \frac{\Delta c}{k_0}\right) \frac{k_0}{\Delta c}(c - c_{**}) = c_{**}^2$ and with any $k_v^1 > 0$ and a fixed a the inequality $c^0 < c_{**} < c_*$ can be seen to hold. Moreover, with an increasing gain k_v^1 the parameter c_{**} approaches c^0. This implies that the relation (3.18) does the job. Whether the region $G(x)$ is reached is found as in the analysis of other cases.

Geometrically, the introduction of a control (3.18) can be interpreted in this way. The presence of a signal $v(t)$ in the right-hand side of the equation $\ddot{x} = -akx - av$ denotes bias of its instantaneous equilibrium position as a function of x and $c^0 - c(t)$ so that the phase velocity vector in the region $G(x)$ is always directed towards the desired path.

References

1. *Voronov, A. A.*, Fundamentals of Control Theory. Part I. Energiya. Moscow, 1965 (in Russian).
2. *Solodovnikov* (Ed.), Theory of Automatic Regulation. Mashgiz, Moscow, 1967 (in Russian).
3. *Petrov, B. N.*, On Application of Invariance Conditions. Proceedings of the Second All-Union Conference on Control Theory, Vol. **2**, Published by the USSR Academy of Sciences, Moscow, 1955 (in Russian).
4. *Kulebakin, V. S.*, High Performance Invariant Control Systems. Proceedings of the First Conference on Invariance Theory. Published by the USSR Academy of Sciences. Moscow, 1959 (in Russian).
5. *Meerov, M. V.*, Design of High Precision Control Systems, 2nd ed. Nauka, Moscow, 1967 (in Russian).
6. *Krasovskii, A. A.*, Dynamics of Continuous Self-Adjusting Systems. Fizmatgiz, Moscow, 1963 (in Russian).
7. *Tsypkin, Ya. Z.*, Adaptation and Learning in Automatic Systems. Nauka, Moscow, 1968 (in Russian).
8. *Yemel'yanov, S. V.* (Ed.), Theory of Variable Structure Systems. Nauka, Moscow, 1970 (in Russian).
9. *Petrov, B. N., Rutkovskii, V. Yu., Krutova, I. N., Zemlyakov, S. D.*, Design Principles of Self-adjusting Control Systems. Mashinostroenie, Moscow, 1972 (in Russian).
10. *Doganovskii, S. A.*, Parametric Control System. Energiya, Moscow, 1973 (in Russian).
11. *Utkin, V. I.*, Sliding Modes and Their Application in Variable Structure Systems. Mir Publishers, Moscow, 1978.

Пути развития типов обратных связей и их применение при построении замкнутых динамических систем

С. В. ЕМЕЛЬЯНОВ, С. К. КОРОВИН

(Москва)

Описывается подход к построению замкнутых динамических систем, основанный на систематическом применении принципа управления по отклонению и использовании его для синтеза обратных связей различных типов. В зависимости от вида входного и выходного сигналов регулятора, обратная связь интерпретируется, соответственно, как координатная, координатно-параметрическая, параметрическая или параметро-координатная. Показывается, что такой подход обеспечивает возможность конструктивного синтеза структуры динамической системы. Возможности подхода демонстрируются на примере синтеза структуры системы автоматического управления, свойства которой инвариантны по отношению к изменению в широком диапазоне параметров объекта и характеристик внешних сил. Показана принципиальная возможность решения этой задачи при конечных коэффициентах передачи в координатном контуре и без использования в нем скользящих режимов.

С. В. Емельянов
Всесоюзный НИИ системных исследований
СССР, Москва, ул. Рылеева 2.

Problems of Control and Information Theory, Vol. 10 (3), pp. 175—180 (1981)

ALGORITHM FOR TRANSFORMATION OF A POLYNOMIAL MATRIX TO CANONICAL FORM

N. K. SINHA, P. RÓZSA

(Hamilton, Budapest)

(Received January 10, 1980)

(Revised version received June 5, 1980)

A simple and straightforward algorithm is presented for the transformation of a polynomial matrix to a canonical form. This transformation has many applications in the theory and design of linear multivariable systems.

The use of polynomial matrices for representing linear multivariable systems has attracted considerable attention recently. As shown by Wolovich [1], this representation has many applications to the minimal realization of transfer function matrices as well as to the design of compensators for multivariable control systems in the frequency domain. Starting with the transfer function matrix of the fixed part of the system, one can utilize this approach to obtain the transfer function matrices of physically realizable series and/or feedback compensators in order that the closed-loop system may have the desired transfer function matrix. Another application to system identification has been developed [2, 3] which is based on this representation. Guidorzi [2] has shown that with a canonical form of the polynomial matrix representation, one may utilize the input-output data from a multivariable system to estimate the system parameters in a manner which essentially decouples the system to several single-output multi-input systems. A recursive technique for on-line identification proposed since then [3] also utlilizes this approach.

The basic problem of representation may be stated as follows. Consider a linear multivariable system represented by

$$Y(s) = G(s)\, U(s) \tag{1}$$

where $Y(s)$ is the Laplace transform of $y(t)$, the output vector, $U(s)$ is the Laplace transform of the input vector $u(t)$, and $G(s)$ is the transfer function matrix, the elements of which are rational functions of the complex frequency variable s. It will be assumed that $y(t) \in R^n$ and $u(t) \in R^m$. The problem of representation in the other form is to express equation (1) as

$$F(s)\, Y(s) = H(s)\, U(s) \tag{2}$$

where $F(s)$ and $H(s)$ are polynomial matrices, and $F(s)$ is nonsingular, so that

$$G(s) = F^{-1}(s)\, H(s)\,. \tag{3}$$

It is also desired that the degrees of the polynomials of $F(s)$ be as small as possible.

One approach to obtaining the minimal polynomial matrices is to first transform $F(s)$ into a canonical form using elementary row and column operations.

Given a matrix $P(s)$, the elements of which are polynomials in s, Gantmacher [4] has shown that by means of elementary transformations, it is possible to obtain a canonical matrix, $\hat{P}(s)$, which has nonzero elements only on its main diagonal, i.e.,

$$\hat{P}(s) = Q(s)\, P(s)\, R(s) \tag{4}$$

where the matrices $Q(s)$ and $R(s)$ are polynomial matrices with constant nonzero determinants and are called "unimodular" matrices.

The object of this note is to present a simple and straightforward algorithm for performing this transformation.

Let the matrix $P(s)$ be as given below:

$$P(s) = \begin{bmatrix} p_{11}(s) & p_{12}(s) & \dots & p_{1n}(s) \\ \vdots & & & \\ p_{m1}(s) & & & p_{mn}(s) \end{bmatrix} \tag{5}$$

where $p_{ij}(s)$ are polynomials of degree d_{ij}.

The transformation will proceed in the following steps.

(1) Select the polynomial of the lowest degree in the matrix, say $p_{i_1 j_1}$. Divide each element of the j_1th column by $p_{i_1 j_1}(s)$ and express these as

$$p_{ij_1}(s) = p_{i_1 j_1}(s)\, q_{ij_1}(s) + r_{ij_1}(s), \quad i \neq i_1. \tag{6}$$

From the ith row of P subtract $q_{ij_1}(s)$ times the i_1th row, and do this for each row except the i_1th row. This is equivalent to premultiplication by the following unimodular matrix

$$Q_1 = \begin{bmatrix} & & & \overset{i_1}{\downarrow} & & \\ 1 & 0 & \dots & -q_{1j_1} & \dots & 0 \\ 0 & 1 & \dots & -q_{2j_1} & \dots & 0 \\ & & & \vdots & & \\ 0 & 0 & \dots & 1 & \dots & 0 \leftarrow i_1. \\ & & & \vdots & & \\ 0 & 0 & \dots & -q_{mj_1} & \dots & 1 \end{bmatrix} \tag{7}$$

The resulting matrix $P^{(1)} = Q_1 P$ will have the following j_1th column

$$\begin{bmatrix} r_{1j_1}(s) \\ r_{2j_1}(s) \\ \vdots \\ p_{i_1 j_1}(s) \\ \vdots \\ r_{mj_1}(s) \end{bmatrix} \tag{8}$$

where each remainder $r_{ij_1}(s)$ will be of degree lower than $p_{i_1 j_1}(s)$.

(2) Repeat step 1 after selecting the polynomial of the lowest degree in the j_1th column until either (i) at least one element of this column is reduced to a nonzero constant, or (ii) all elements except one are zero.

In case (i) the j_1th column of the resulting matrix will be of the form

$$\begin{bmatrix} r_{1j_1}^{(q)}(s) \\ r_{2j_1}^{(q)}(s) \\ \vdots \\ a_{i_q j_1} \\ r_{mj_1}^{(q)}(s) \end{bmatrix} \tag{9}$$

and can now be reduced to the form

$$\begin{bmatrix} 0 \\ \vdots \\ a_{i_q j_1} \\ \vdots \\ 0 \end{bmatrix} \tag{10}$$

by subtracting a suitable multiple of the i_qth row.

Now we can reduce the remaining elements of the i_qth row to zero by subtracting suitable multiples of the j_1th column from every other column. This is equivalent to post-multiplication by a unimodular matrix of the form

$$R_1 = \begin{bmatrix} 1 & & & & \\ \vdots & & & & \\ -\frac{p_{i_q 1}^{(q)}}{a_{i_q j_1}} & -\frac{p_{i_q 2}^{(q)}}{a_{i_q j_1}} & \dots & -\frac{p_{i_q n}^{(q)}}{a_{i_q j_1}} & \leftarrow j_1 \\ & & & 1 & \end{bmatrix} \tag{11}$$

In case (ii) the j_1th column of the resulting matrix will have the form

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ p^{(q)}_{i_q j_1}(s) \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{12}$$

Now we perform column operations, to reduce each element in the i_qth row to degree less than that of $p^{(q)}_{i_q j_1}(s)$. This is, again, equivalent to post-multiplication by a suitable unimodular matrix. This process of column operations may be repeated until we either get a constant in the i_qth row, or all the elements except one are zero in this row. Performing row operations now and repeating this process, if necessary, we shall eventually get a matrix in which the j_rth column and i_rth row will have zero entries everywhere except the $i_r j_r$ element which will be a polynomial $p^{(r)}_{i_r j_r}(s)$.

If even one of the elements $p^{(r)}_{ij}(s)$ $(i \neq i_r, j \neq j_r)$ is not divisible by $p^{(r)}_{i_r j_r}(s)$ then the column containing this element has to be added to the j_rth column. Repeating steps (1) and (2), the degree of $p^{(r)}_{i_r j_r}(s)$ can be reduced until this element becomes a common divisor of all other elements.

(3) Performing row and column permutations, the resulting matrix for both cases may be written as

$$P^{(r)} = \begin{bmatrix} p^{(q)}_{11}(s) & 0 & 0 & \dots & 0 \\ 0 & & & & \\ \vdots & & & & \\ 0 & & \tilde{P}(s) & & \\ 0 & & & & \end{bmatrix}. \tag{13}$$

Hence we may now carry out steps (1) and (2) with $\tilde{P}(s)$, and so on, until the entire matrix has been brought to the diagonal canonical form.

The method will be illustrated through an example. Let

$$P(s) = \begin{bmatrix} s^2 + 3s + 3 & s + 3 \\ s + 2 & s + 1 \end{bmatrix}.$$

Subtracting $(s + 1)$ times the second row from the first is equivalent to pre-multiplication by

$$Q_1(s) = \begin{bmatrix} 1 & -(s+1) \\ 0 & 1 \end{bmatrix}$$

and gives

$$P^{(1)}(s) = Q_1(s)\, P(s) = \begin{bmatrix} 1 & -s^2 - s + 2 \\ s + 2 & s + 1 \end{bmatrix}.$$

Subtracting $(s + 2)$ times the first row from the second gives

$$Q_2(s) = \begin{bmatrix} 1 & 0 \\ -(s+2) & 1 \end{bmatrix}$$

and

$$P^{(2)}(s) = Q_2(s)\, Q_1(s)\, P(s) = \begin{bmatrix} 1 & -s^2 - s + 2 \\ 0 & s^3 + 3s^2 + s - 3 \end{bmatrix}.$$

Finally adding $(s^2 + s - 2)$ times the first column to the second gives

$$R(s) = \begin{bmatrix} 1 & s^2 + s - 2 \\ 0 & 1 \end{bmatrix}$$

and

$$\hat{P}(s) = Q_2(s)\, Q_1(s)\, P(s)\, R(s) = \begin{bmatrix} 1 & 0 \\ 0 & s^3 + 3s^2 + s - 3 \end{bmatrix}.$$

Also,

$$Q(s) = Q_2(s)\, Q_1(s) = \begin{bmatrix} 1 & -(s+1) \\ -(s+2) & s^2 + 3s + 3 \end{bmatrix}.$$

References

1. *Wolovich, W. A.*, "Linear Multivariable Systems", Springer-Verlag, 1974.
2. *Guidorzi, R.*, "Canonical Structures in the Identification of Multivariable Systems", Automatica, vol. **11**, 1975, pp. 361–367.
3. *Sinha, N. K., Kwong, Y. H.*, "Recursive Estimation of the Parameters of Linear Multivariable Systems", Automatica, vol. **15**, 1979, pp. 471–475.
4. *Gantmacher, F. R.*, "Theory of Matrices", Chelsea, 1959. vol **1**.
5. *Rosenbrock, H. H.*, "State-space and Multivariable Theory", Wiley, New York, 1970.
6. *Kučera, V.*, "Discrete Linear Control – The Polynomial Equation Approach", Wiley, Chichester, 1979.
7. *Kailath, T.*, "Linear Systems", Prentice-Hall, Englewood Cliffs, N. J., 1980.

Алгоритм для преобразования многочленных матриц к каноническому виду

Н. К. СИНХА, П. РОЖА

(Гамильтон), (Будапешт)

Как известно, произвольная прямоугольная матрица эквивалентна некоторой канонической матрице, т. е. с помощью левых и правых элементарных операций она может быть преобразована к диагональному виду. В настоящей работе предлагается простой алгоритм для преобразования многочленных матриц к каноническому виду. Это преобразование имеет различные приложения в теории линейных систем с многими переменными.

N. K. Sinha
Group on Simulation, Optimization and Control
Faculty of Engineering
McMaster University
Hamilton, Ontario, Canada L8S 4L7

P. Rózsa
Department of Mathematics
Faculty of Electrical Engineering
Technical University of Budapest
H—1111 Budapest, Stoczek u. 2.
Hungary

Problems of Control and Information Theory, Vol. 10 (3), pp. 181—189 (1981)

LINEAR OPTIMAL DYNAMIC REGULATORS

V. KUČERA, M. ŠEBEK

(Prague)

(Received June 10, 1980)

A new method of designing linear-quadratic regulators for discrete-time multivariable systems is presented. These regulators process just the measurable data and hence are dynamical in nature. However, they are independent of the system's intitial state.

The design procedure consists of spectral factorization and the solution of linear equations in polynomial matrices. The optimal regulator is obtained in terms of its input-output relation; there is no need for explicit state reconstruction.

1. Introduction

Linear optimal regulators have obtained much attention in the control literature. Given a discrete-time, linear, l-output, m-input, n-dimensional system

$$\begin{aligned} x_{t+1} &= Fx_t + Gu_t \\ y_t &= Hx_t \end{aligned} \tag{1}$$

which is stabilizable and detectable as well, the standard objective is to minimize the quadratic cost

$$J = \sum_{t=0}^{\infty} u_t' \varphi u_t + y_t' \psi y_t \tag{2}$$

with nonnegative definite φ and ψ for every x_0.

If the entire state x_t is available for direct measurement, this objective can be accomplished by a linear state-variable feedback

$$u_t = -Kx_t.$$

Standard methods for the calculation of such a control law are based on solving algebraic Riccati equations. Original results are due to Kalman and Koepcke [1] while modern account of this subject can be found in Kwakernaak and Sivan [4].

If, however, some state variables are not accessible, this control law cannot be implemented. A common trick is to introduce an observer which generates a reconstruction $\hat{x}_t$ of the actual state x_t from measurable data, and then to apply the control law

$$u_t = -K\hat{x}_t.$$

Various techniques of designing observers are discussed by Luenberger [5] or Kwakernaak and Sivan [4].

In a recent paper [2], the first author proposed an alternative method for calculating state-variable control laws. The method is based on spectral factorization and the solution of linear equations in polynomial matrices. The purpose of this paper is to extend the polynomial design technique so as to cope with unmeasurable states.

2. Preliminaries and notation

The mathematical objects studied in this paper are real rational and polynomial matrices. They are treated in detail by Kučera [3]; here we just introduce some notions briefly.

Given a rational matrix of the form

$$R = R_0 + R_1 d + R_2 d^2 + \dots$$

in an indeterminate d, we say that R is Hurwitz if the sequence of matrices R_k, $k = 0, 1, \dots$ has a polynomial rate of growth and we say that R is stable if the sequence of matrices R_k, $k = 0, 1, \dots$ converges to zero. The conjugate matrix R_* is defined by

$$R_* = R_0' + R_1' d^{-1} + R_2' d^{-2} + \dots$$

where the prime denotes transposition, and we shall write $\langle R \rangle$ for the constant coefficient of R, that is

$$\langle R \rangle = R_0.$$

A polynomial matrix is said to be Hurwitz (stable) if its inverse is a Hurwitz (stable) rational matrix. Further, a polynomial matrix whose inverse is again polynomial is called unimodular.

A greatest common left (right) divisor of two polynomial matrices is such a common left (right) divisor which is a right (left) multiple of every other common left (right) divisor. Two polynomial matrices are said to be left (right) coprime if every square common left (right) divisor of theirs is unimodular.

Finally, the symbol I_n is reserved for the $n \times n$ identity matrix.

3. Problem formulation

We now formulate the problem in terms of input-output data. Write $u = (u_t)$ and $y = (y_t)$ for the input and output sequences, respectively, and introduce the delay operator d by the relation $dx_{t+1} = x_t$ for any sequence (x_t), $t = 0, 1, \ldots$

System (1) will be described by the equation

$$Ay = Bu + C, \tag{3}$$

where A, B and C is a triple of left coprime polynomial matrices in the delay operator d. A is $l \times l$ with $\langle A \rangle$ invertible, B is $l \times m$ with $\langle B \rangle = 0$, and the $l \times 1$ matrix C represents the effect of the initial state, x_0, on the system output. Passing from (1) to (3), we have the relationships

$$A^{-1}B = H(I_n - dF)^{-1}dG$$

$$A^{-1}C = H(I_n - dF)^{-1}x_0.$$

Thus letting C vary we can account for any initial state in the system.

Note that stabilizability of our system implies that the greatest common left divisor of A and B is a stable polynomial matrix. The detectability of our system then implies that the greatest common left divisor of A and C for every C is also a stable polynomial matrix. Finally, the assumption that $\langle A \rangle$ is invertible guarantees causality of the system whereas $\langle B \rangle = 0$ entails strict causality.

The task is to find a regulator in the form

$$Pu = -Qy \tag{4}$$

where P and Q are polynomial matrices in d, respectively $m \times m$ and $m \times l$, such that

1° closed-loop system (3)–(4) is asymptotically stable,

2° cost (2) is minimized,

3° P and Q are independent of C.

Clearly, the last requirement calls for a regulator which does not depend upon the system's initial state x_0. This feature is of considerable practical significance and implies that in fact our approach provides an observer-based control law alluded to in the Introduction.

4. Design procedure

The purpose of this section is to give the entire design procedure step by step in a concise and easy-to-follow way. Setting aside all questions of solvability for the moment, we start with the primary data A, B and φ, ψ and wish to construct P and Q. The related statement on existence is relegated to the next section.

The optimal design can be carried out in the following steps:

1) Calculate any right coprime polynomial matrices A_1, B_1 such that

$$B_1 A_1^{-1} = A^{-1} B.$$

2) Perform the spectral factorization to obtain a Hurwitz polynomial matrix D satisfying

$$A_{1*} \varphi A_1 + B_{1*} \psi B_1 = D_* D. \tag{5}$$

3) Solve the pair of equations

$$\begin{aligned} D_* X - Z_* B &= A_{1*} \varphi \\ D_* Y + Z_* A &= B_{1*} \psi \end{aligned} \tag{6}$$

for polynomial matrices X, Y, and Z such that $\langle Z \rangle = 0$.

Then

$$P = X, \quad Q = Y \tag{7}$$

defines an optimal regulator.

5. Solvability

The preceding design procedure produces an optimal regulator whenever one exists. It is the purpose of this section to prove this claim as well as the related result on existence.

Theorem. Let

$$\operatorname{rank} \begin{bmatrix} \varphi A_1 \\ \psi B_1 \end{bmatrix} = m. \tag{8}$$

Then the regulator problem is solvable if and only if the rational matrices

$$D^{-1} X, \quad D^{-1} Y$$

are stable and an optimal regulator is given by (7).

Proof. The proof will be divided into two parts. First we shall construct the regulator (7) provided it exists, and then we shall discuss its existence.

Regarding the first part, let a regulator (4) exist which minimizes cost (2) while making the closed loop system asymptotically stable; write the regulator in the form

$$u = -Q_1 z, \quad y = P_1 z \tag{9}$$

where z is an internal variable and

$$P^{-1}Q = Q_1 P_1^{-1}. \tag{10}$$

The cost can then be expressed as

$$J = \langle u_* \varphi u + y_* \psi y \rangle \tag{11}$$

and the design proceeds as follows.

Define the matrices

$$M = (PA_1 + QB_1)^{-1}P, \quad N = (PA_1 + QB_1)^{-1}Q$$
$$M_1 = P_1(AP_1 + BQ_1)^{-1}, \quad N_1 = Q_1(AP_1 + BQ_1)^{-1}$$

and observe that

$$\begin{aligned} AM_1 + BN_1 &= I_l \\ MA_1 + NB_1 &= I_m \end{aligned} \tag{12}$$

and

$$\begin{aligned} A_1 N &= N_1 A \\ B_1 M &= M_1 B. \end{aligned} \tag{13}$$

Inserting (9) into (3) one obtains

$$(AP_1 + BQ_1)z = C.$$

Substituting for z in (9) and then employing (12)–(13) we get

$$u = -N_1 C = -A_1 N A^{-1} C$$
$$y = M_1 C = -B_1 N A^{-1} C + A^{-1} C.$$

Hence (11) reads

$$J = \langle C_* V C \rangle, \tag{14}$$

where

$$V = A_*^{-1}(\psi - N_* B_{1*}\psi - \psi B_1 N + N_* B_{1*}\psi B_1 N + N_* A_{1*}\varphi A_1 N) A^{-1} =$$
$$= A_*^{-1}(\psi - N_* B_{1*}\psi - \psi B_1 N + N_* D_* D N) A^{-1}$$

with D defined in (5). Completing the squares, we obtain

$$V = V_0 + V_{1*} V_1, \tag{15}$$

where

$$V_0 = A_*^{-1}(\psi - \psi B_1 D^{-1} D_*^{-1} B_{1*} \psi) A^{-1}$$

$$V_1 = D_*^{-1} B_{1*} \psi A^{-1} - DNA^{-1}.$$

Let X, Y and Z with $\langle Z \rangle = 0$ be polynomial matrices satisfying the pair of equations (6). As shown by Šebek and Kučera [6], such matrices always exist whenever the greatest common left divisor of A and B is stable, and they ore unique. Moreover, postmultiplying the first equation by A_1, the second ane by B_1 and adding them up, we obtain

$$XA_1 + YB_1 = D. \tag{16}$$

The second equation of (6) now yields

$$D_*^{-1} B_{1*} \psi A^{-1} = D_*^{-1} Z_* + YA^{-1}$$

so that

$$V_1 = D_*^{-1} Z_* + W, \tag{17}$$

where

$$W = YA^{-1} - DNA^{-1} = YA^{-1}(AM_1 + BN_1) - DNA^{-1} =$$

$$= YM_1 - (D - YB_1)A_1^{-1} N_1 = YM_1 - XN_1 \tag{18}$$

on employing (12), (13) and (16) successively.

Write

$$C = C_0 + C_1 d + \ldots + C_k d^k$$

$$ZD^{-1} = D_1 d + D_2 d^2 + \ldots$$

$$W = W_0 + W_1 d + W_2 d^2 + \ldots$$

and make use of (11), (14) ,(15) and (17) to arrange the cost as follows:

$$J = J_1 + J_2 + J_3,$$

where

$$J_1 \langle C_* V_0 C \rangle + \sum_{j=-1}^{-\infty} \left(\sum_{i=0}^{k} D'_{i-j} C_i \right)' \left(\sum_{i=0}^{k} D'_{i-j} C_i \right),$$

$$J_2 = \sum_{j=0}^{k-1} \left(\sum_{i=0}^{j} W_{j-i} C_i + \sum_{i=j+1}^{k} D'_{i-j} C_i \right)' \times \left(\sum_{i=0}^{j} W_{j-i} C_i + \sum_{i=j+1}^{k} D'_{i-j} C_i \right),$$

$$J_3 = \sum_{j=k}^{\infty} \left(\sum_{i=0}^{k} W_{j-i} C_i \right)' \left(\sum_{i=0}^{k} W_{j-i} C_i \right).$$

First note that J_1 is independent of W (and hence of the regulator). For a given C both J_2 and J_3 can be made zero by an appropriate choice of W; however, such a regulator would depend upon C. In order to minimize J by means of a regulator which is indpendent of C, the best we can do is to put $W = 0$ so that J_3 can vanish.

Taking the definition of M_1 and N_1 into account, relation (18) then calls for

$$YP_1 = XQ_1$$

and the optimal regulator (7) follows from (10).

Concerning the second part of the proof, observe that condition (8) implies the existence of a spectral factor D defined in (5). Further invoke the result of Kučera [3] that the closed loop system (3)–(4) is asymptotically stable if and only if the rational matrices M and N (or equivalently M_1 and N_1) are stable. Since (12) and (16) combined yield

$$M = D^{-1}X, \quad N = D^{-1}Y$$

the existence condition results.

6. Example

In order to illustrate the design procedure, consider a single-input-output system described by model (3) with

$$A = 1 - d^2, \quad B = d, \quad C = c_0 + c_1 d$$

where c_0, c_1 are any reals representing the system's initial state. Let the cost (2) be defined by the weighting coefficients

$$\varphi = 2, \quad \psi = 1.$$

The first step in the design is to calculate A_1 and B_1. In this case we have

$$A_1 = A, \quad B_1 = B$$

by the commutativity of polynomials. The second step is to perform the spectral factorization

$$(1 - d^{-2})\,2(1 - d^2) + 1 = D_* D.$$

Clearly,

$$D = 2 - d^2.$$

Now we are to solve the pair of equations

$$(2 - d^{-2})X - Z_* d = 2(1 - d^{-2})$$

$$(2 - d^{-2})Y + Z_*(1 - d^2) = d^{-1}$$

for the polynomials X, Y, and Z such that $\langle Z \rangle = 0$. We get

$$X = 2, \quad Y = d, \quad Z = 2d$$

and hence the optimal regulator (7) is given by

$$P = 2, \quad Q = d.$$

A realization of this control law is shown in Fig. 1.



Fig. 1. An optimal regulator

Of course, this control law can also be implemented using an observer plus state-variable feedback. Let the state-variable description (1) of our system be given by the matrices

$$F = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$H = [1 \; 0]$$

Then the full-order observer with both poles at the origin, combined with feedback gain matrix as depicted in Fig. 2, gives the same input-output behaviour as our regulator. Note that it is the reconstruction of the entire state vector which makes this regulator unnecessarily complicated.

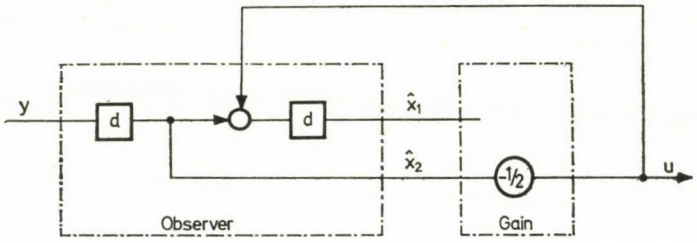


Fig. 2. Observer-based regulator

7. Discussion

The solution of the regulator problem advanced in this paper has several interesting features. The design procedure consists of three steps — the calculation of a matrix fraction, spectral factorization, and the solution of matrix polynomial equations — and it is simple and computationally efficient. For details on the numerical algorithms see Kučera [3].

This is an attractive alternative to the standard state-space methods based on the solution of algebraic Riccati equations and the design of an observer. In fact, our solution is equivalent to taking an observer with all poles at the origin. Clearly, the approach presented here is more compact and often leads to a more economical realization of the control law.

References

1. *Kalman*, R. E., *Koepcke, R. W.*, Optimal synthesis of linear sampling control systems using generalized performance indexes. Trans. ASME Ser. D. J. Basic Eng. **80**, 1812–1817, 1958.
2. *Kučera, V.*, Discrete linear regulator revisited. to Kybernetika **17**, *1*, 62–79, 1981.
3. *Kučera, V.*, Discrete Linear Control: The Polynomial Equation Approach. Wiley, Chichester, 1979.
4. *Kwakernaak, H., Sivan, R.*, Linear Optimal Control Systems. Wiley, New York, 1972.
5. *Luenberger, D. G.*, Observers for multivariable systems. IEEE Trans. Automatic Control **AC-11**, 190–197, 1966.
6. *Šebek, M., Kučera, V.*, Matrix equations arising in regulator problems. Kybernetika **17**, *2*, 1981.

Линейные оптимальные динамические регуляторы

В. КУЧЕРА, М. ШЕБЕК

(Прага)

В статье предлагается новый метод синтеза линейных регуляторов для дискретных многосвязных систем и квадратичных критериев качества. На вход регулятора подаются только измеримые величины и поэтому регуляторы по существу динамические. Однако они не зависят от начального состояния системы.

Синтез регулятора сведен к спектральной факторизации и решению линейных уравнений с матрицами полиномов. Результат получается в форме уравнения, связывающего вход и выход системы.

V. Kučera
M. Šebek
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
182 08 Praha 8
Pod vodárenskou věží 4
Czechoslovakia

Problems of Control and Information Theory, Vol. 10 (3), pp. 191—198 (1981)

MODELING OF MORPHOLOGICAL AND FUNCTIONAL STRUCTURE IN THE SYMBOLICAL SYSTEM FOR THE AUTOMATIC SYNTHESIS OF TECHNICAL OBJECT

J. BÍLA

(Prague)

(Received June 4, 1980)

General concepts and characteristics of the morphological and the functional descriptions are introduced. The essential ideas of the conception of the symbolical system for the automatic synthesis of technical object structure [1] are recapitulated. The primary role of the functional description in the synthesis is emphasized. The main part of the paper deals with the problems of extension of the symbolical model for the functional structure synthesis as it has already been introduced with more details in [3]. The main interest is concentrated in the features of functional products emergence and in the formal problems of indeterminism of the functional structure synthesis.

Introduction

Under morphology of an object or of a phenomenon we understand more than shape or its generalizations.

Morphological description contains all parameters, characteristics and features which enable identification of the described object and phenomenon in all supposed problem environments [4]. We can say that the morphological description is designed with the effort to achieve the bounds of objectivity.

The functional description on the contrary strongly depends on the subjective characteristics of its designer (user). The typical assertions which we can expect in the functional description are as follows:

"The action A_1 supports the action A_2" or "the object O_1 cooperates with the set of objects O_2", etc.

As for the modeling of morphological and functional qualities in the symbolical model it is necessary to emphasize that objects of model — symbolical formations — contain and transfer information by means of their morphology. From this point of view the morphological equalities are easier to model than the qualities characterizing the function.

Essential ideas of an approach to automation of the synthesis of technical object

In paper [1] a basic conception has been introduced for the design of a symbolical system which could enable to describe morphological and functional qualities needed for the technical object synthesis modeling. This symbolical system has been formed within the frameworks of the language of the type *RX*-code [11] (used for the morphological description) and the language of the type of Universal semantical code (USC) [10] (used for the functional description).

The central object of the symbolical systems was a complicated symbolical mechanism labeled as the "morphism" *int* — (interpreter). An "elementary step" of the interpreter has been constructed as follows:

$$\begin{aligned} &s_1, \ldots, s_n \in D_S^*, \quad \mu_1, \ldots, \mu_n \in D_F^*. \\ &\forall\, s_i, \mu_i, i = 1, n, \quad \text{int}\,(s_i) = \mu_i, \\ &f(s_1, \ldots, s_n) \in D_S^*, \; g(\mu_1, \ldots, \mu_n) \in D_F^*, \\ &\text{int}\,(f(s_1, \ldots, s_n)) = g(\mu_1, \ldots, \mu_n), \end{aligned} \tag{1}$$

where $s_1, \ldots, s_n$ are the elements of the carrier D_S^* of the morphological description model $\mathfrak{D}_S$, $\mu_1, \ldots, \mu_n$ are the elements of the carrier of the functional description model $\mathfrak{D}_F$, f and g are the function in the model $\mathfrak{D}_S$ resp. $\mathfrak{D}_F$. As a global strategy of the synthesis using the interpreter could serve e.g. the method of GPS [7].

However it is possible to illustrate by many examples that the strategy of design and synthetical actions of man is more complicated than the method of GPS [5, 6]. In accordance with the conclusions of paper [1] the interpreter *int* is not a direct means of the synthetical strategy but a result of complicated analysis which would have been done before its construction.

We introduce a few essential ideas which may be a motivation of analysis:

(A) The symbolical models $\mathfrak{D}_S$ and $\mathfrak{D}_F$ we can represent by the triple of the type:

$$\langle \mathfrak{L}_m, C, Ch \rangle, \tag{2}$$

where the symbols $\mathfrak{L}_m$, C, Ch represent a *language* of the model, a *calculus* of the model and a *method of choice*.

The method of choice contains not only the tests for estimation of products reached in the frame of $\mathfrak{L}_m$ and C but also the criteria of estimation of effectiveness of the realized steps of the synthesis.

Designing the symbolical model we cannot construct the components $\mathfrak{L}_m$, C, Ch separately but as a whole.

(B) The roles of morphological and functional descriptions are different.

The *functional description* is *total* and *undecomposable*. In the process of the technical object synthesis it acts as a primary description.

The *morphological description* is a secondary description in a synthetical process; it is *decomposable*; the synthesis of morphological structure we can represent in most cases by means of *compositions* and *bond constructions*, which are *operations* performed on the set of basic elements.

(C) The concept of morphism is interpretable in an ordinary algebraic sense only for objects of morphological description. Transformations of functional description objects are necessary to represent by means of another object. We shall use the concept of "functional morphism" introduced in [3] which is an analogous notion of functor defined in the theory of categories.

(D) The symbolical model for the synthesis of the functional structure must be equipped by an apparatus for the description of context properties of symbolical formations, modal features of actions as well as by methods of temporal logic [8, 9].

(E) The interpreter int : $\mathfrak{D}_S \to \mathfrak{D}_F$ there is designed as a result of an inductive generalization of the analysis. From this point of view the interpreter represents the goal of our theoretical work when the design of models $\mathfrak{D}_S$ and $\mathfrak{D}_F$ is transferred into region of constructive mathematics and algebra.

Some general features of the functional structure synthesis

Nature of the functional objects emergence

Let us consider the configurations of technical elements which are illustrated on Fig. 1.

As for the synthesis of the morphological structure of *E-* — we understand — is trivial. The topological and geometrical description as well as the

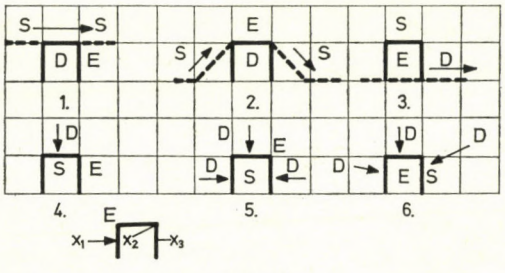


Fig. 1

description of material properties of the elements x_1, x_2, x_3 implies the description of product E and v.v.

The synthesis of the morphological structure of E we can perform as an "additive process" [2]. (Additive processes — resp. the notion "additive process" — has been introduced in this sense e.g. by H. Simon [12].)

The synthesis of *functional structure* is on the contrary a "non-additive process". The illustration of "non-additivity" is very simple in this case:

We apply the analogous strategy of the morphological structure synthesis and we try to construct the functional description of E by means of the functional descriptions of elements x_1, x_2, x_3.

As we quickly find out — it is a problem with no sense. The functional descriptions of x_1, x_2, x_3 we can construct with the knowledge of the function of E (after interpretation of E).

If we take into consideration also the objects D and S which represent an environment and a subject of the situations with the product E we can analyse the emergence of the functional structure of E in a more detailed way.

We introduce only the most evident functions of E and the corresponding functions of the elements x_1, x_3:

Situations:

The function of E	The functions of x_1, x_3
1. To bridge a region with D	To support x_2
2. To overbridge a region with D	To support and hold up x_2
3. To fix x_2	To rest x_2 on x_1 and x_3
4. To separate S and D	To support x_2
5. To separate S and D	To support x_2 and to separate S and D
6. To serve as a skeleton of S in an environment D	To behave as a monolithic form with x_2

Let us analyse the transformation of the functional structure of situation 1 to the functional structure of situation 4.

In order to make clear our operations we introduce first a general scheme of action dependences — Fig. 2.

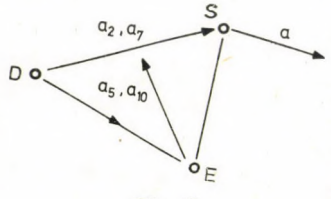


Fig. 2

In the scheme the environment D acts on the subject S and product E. The subject S has an external action a. The product E acts on the effects of D on S. The functional structure of situation 1:

The actions $a_{\hat{i}}$:

a_1: The subject S moves on a trajectory $m_1 T_1 : a$.

a_2: The effects of D destruct the trajectory $m_1 T_1$ in a place m_2. The effects of D destroy the movement of S on the trajectory $m_1 T_1$.

a_3: The element x_2 substitutes the trajectory $m_1 T_1$ in the place m_2. (The trajectory $m_1 T_1$ in the place m_2 exists; the element x_2 is identical with the trajectory $m_1 T_1$ in the place m_2 and inputs in a relevant relation with the environment D).

a_4: The elements x_1, x_3 enable to realize the action a_3.

a_5: The product E enables to realize the action a_1.

(The action a_5 represents the functional description of E which could be used in some other context as a requirement for the goal product. The actions a_1, a_2, a_5 represent "material of the functional structure synthesis problem", the actions a_3, a_4 represent the result of the problem solution.) In order to describe the functional structure of the situation 1; by means of USC we should introduce the grammar of a high type language which enables to form structures of purport chains with context properties and with means of temporal logic. Such a grammar has been introduced in [3] and in this paper we used only one context property Caus, which has already been defined in [1].

The functional structure of situation 1:

a_1: $/S_1/\bar{S}_1 \sqsubset A \bar{O}_1 m_1 T_1 : a$

a_2: $(\overline{m}_3 D \sqsubset A \overline{m_1 m_2 T_1}\, m_1 m_2 T_{1\rangle} \urcorner A \overline{m_1 m_2 T_1}\ \overline{m_1 m_2 T_1})_{\sqsubset}$ Caus
$(m_3 D \sqsubset A /S_1/ \bar{S}_{1\sqsubset} \urcorner A \bar{O}_1 m_1 T_1)$

a_3: $\overline{x}_{2\sqsubset} A \overline{m_1 m_2 T_1}\, m_1 m_2 T_{1\rangle} A \overline{m_1 m_2 T_1}\ \overline{m_1 m_2 T_1}$
$\Vdash$
$\overline{x}_{2\sqsubset} A \{\overline{\overline{x}_2 m_1 m_2 T_1}\}\ \{\overline{x}_2 m_1 m_2 T_1\}_{\rangle} A \{\overline{\overline{x}_2 m_1 m_2 T_1}\}\ \{\overline{\overline{x}_2 m_1 m_2 T_1}\}$
$\Vdash$
$\overline{x}_{2\sqsubset} A \{\overline{\overline{x}_2 D}\}\ \{\overline{x}_2 D\}_{\rangle} A \{\overline{\overline{x}_2 D}\}\ \{\overline{\overline{x}_2 D}\}$

a_4: $\left[\begin{matrix}\overline{x}_1 \\ \overline{x}_3\end{matrix}\right]_{\sqsubset} A \overline{x}_{2\sqsubset} \Diamond A \overline{m_1 m_2 T_1}\, m_1 m_2 T_{1\rangle} A \overline{m_1 m_2 T_1}\ \overline{m_1 m_2 T_1}$ (3)
$\Vdash$
$\left[\begin{matrix}\overline{x}_1 \\ \overline{x}_3\end{matrix}\right]_{\sqsubset} A \overline{x}_{2\sqsubset} \Diamond A \{\overline{\overline{x}_2 m_1 m_2 T_1}\}\ \{x_2 m_1 m_2 T_1\}_{\rangle} A \{\overline{\overline{x}_2 m_1 m_2 T_1}\}\ \{\overline{\overline{x}_2 m_1 m_2 T_1}\}$
$\Vdash$
$\left[\begin{matrix}\overline{x}_1 \\ \overline{x}_3\end{matrix}\right]_{\sqsubset} A \overline{x}_{2\sqsubset} \Diamond A \{\overline{\overline{x}_2 D}\}\ \{\overline{x}_2 D\}_{\rangle} A \{\overline{\overline{x}_2 D}\}\ \{\overline{\overline{x}_2 D}\}$

a_5: $(\bar{\bar{E}}_{\square} A\, m_3 D_{\square} \urcorner A\, \overline{m_1 m_2 T_1}\, m_1 m_2 T_{1\aleph} \urcorner A\, \overline{m_1 m_2 T_1}\, \overline{m_1 m_2 T_1})_{\square}$ Caus
$(\bar{\bar{E}}_{\square} A\, /S_1/\, \bar{S}_{1\square} \diamond A\, \bar{O}_1 m_1 T_1)$

The synthesis of the functional structure of situation 4 we illustrate by the following steps:

a) General functional pattern of "the material of the functional structure synthesis problem" —

a_6: $S_1 \bar{S}_{1\aleph} A \bar{O}_1 \bar{O}_1 : a$

a_7: $(m_4 D \bar{D}_{\square} A\, \overline{m_4 D}\, O_1)$ Caus $(m_4 D_{\square} A\, S_1 \bar{S}_{1\aleph} \urcorner A\, \bar{O}_1 \bar{O}_1)$

a_{10}: $(\bar{\bar{E}}_{\square} A\, m_4 D_{\square} \urcorner A\, S_1 \bar{S}_{1\aleph} \urcorner A\, \bar{O}_1 \bar{O}_1)_{\square}$ Caus $(\bar{\bar{E}}_{\square} A\, S_1 \bar{S}_{1\aleph} \diamond A\, \bar{O}_1 \bar{O}_1)$,

(the biological existence of the subject S (a_6) being destroyed by effects of D (e.g. by γ-radiation) (a_7) is saved by the actions of the goal product E (a_{10})) has been analysed as the same as one for situation 1:

General functional pattern of the functional structure synthesis problem of the situations 1 and 4 (and of situations 2, 3, 5, 6, too):

$$\begin{aligned} &-S_1 A O \\ &-(mD_{\square} A\, O_1)_{\square} \text{ Caus } (mD_{\square} A\, S_1 \urcorner A\, O) \\ &-(\bar{\bar{E}}_{\square} A\, mD_{\square} \urcorner A\, S_1 \urcorner A\, O)_{\square} \text{ Caus } (\bar{\bar{E}}_{\square} A\, S_1 \diamond A\, O) \end{aligned} \tag{4}$$

b) By means of generalization of the functional structure of situation 1 is constructed the result of the functional structure synthesis problem solution for situation 4:

$$\begin{aligned} a_8\text{:}\quad & m_4 D \bar{D}_{\square} A\, \overline{m_4 D}\, x_2 \\ & m_4 D \bar{D}_{\square} \urcorner A\, \overline{m_4 D}\, \{\urcorner x_2\} \\ a_9\text{:}\quad & \begin{bmatrix} \bar{x}_1 \\ \bar{x}_3 \end{bmatrix}_{\square} A\, m_4 D \bar{D}_{\square} \diamond A\, \overline{m_4 D}\, x_2 \\ & \begin{bmatrix} \bar{x}_1 \\ \bar{x}_3 \end{bmatrix}_{\square} A\, m_4 D \bar{D}_{\square} \diamond A\, \overline{m_4 D}\, \{\urcorner x_2\}. \end{aligned} \tag{5}$$

The effects of D input into x_2 and not output from x_2 (the effects of D are absorbed by x_2).

(The operation of generalization takes in USC the features of an operation with morphology of purport chains and works very effectively. The construction of symbolical formation which represents the action of absorption (a_8) performed e.g. in a language of predicates or in a language of RX-code type has been analysed as an operation of "pure" substitution on a preformed set of symbolical formations.)

The indeterminism of the functional structure synthesis

In the last section we have analysed the general conditions of the transformations of functional structures (e.g. the functional structure of situation 1 to the functional structure of situation 4). The transformations we now introduce into the context of the scheme of the functional structure synthesis using the set of "possible worlds". The method of the synthesis we illustrate in Fig. 3:

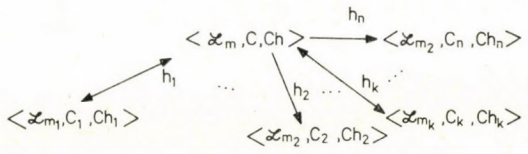


Fig. 3

The method of the functional structure synthesis is described as follows:

The description of "material of the functional structure synthesis problem" is constructed in one of the symbolical models (e.g. in $\langle \mathfrak{L}_{m_1}, C_1, Ch_1 \rangle$) that represent various different "possible worlds". To this functional description is constructed a generalized functional description (general functional pattern) in a "generalized" world $\langle \mathfrak{L}_m, C, Ch \rangle$.

The general functional pattern is gradually transformed by means of $h_2, \ldots, h_k, \ldots, h_n$ into the symbolical models $\langle \mathfrak{L}_{m_2}, C_2, Ch_2 \rangle, \ldots, \langle \mathfrak{L}_{m_k}, C_k, Ch_k \rangle, \ldots$ $\ldots, \langle \mathfrak{L}_{m_n}, C_n, Ch_n \rangle$ in order to find a functional description which would be an adequate specification of the general functional pattern. When discovering such a description with a correspondent functional structure (e.g. in the model $\langle \mathfrak{L}_{m_k}, C_k, Ch_k \rangle$) we construct the description of the goal product functional structure by means of the inverse transformations h_k and h_1 leading through the model $\langle \mathfrak{L}_m, C, Ch \rangle$ to the model $\langle \mathfrak{L}_{m_1}, C_1, Ch_1 \rangle$. (The transformations $h_1, \ldots, h_n$ we have called in [3] *the functional morphisms*).

The functional structure of the goal product has been designed not by means of synthetical operations producing compositions of elements but with the help of special transformations of a "similar" functional structure which had been preformed in a symbolical model of the set of "possible worlds".

The indeterminism of the functional structure synthesis is one of its principal features with all consequences which are common to every direct synthetical approach.

The problem of formal overcoming of this indeterminism shifts the investigation of the synthesis nature to analytical part lying out of our interest at present.

Conclusions

The relations of the morphological and functional structure have been discussed in the context of the technical object structure synthesis.

By means of a simple example an alternative approach to functional structure synthesis method has been explored. The indeterminism that we have encountered in the synthesis we have understood as one of the principal features of the informational processes which operate with the intensions of notions.

References

1. *Balda, M., Bíla, J.*, Symbolical system for the automatic synthesis of technical object structure; Problems of Control and Information Theory, Vol. **8**, pp. 353–362, 1979
2. *Beneš, J.*, Automatic Shaping in Large Systems: A Renaissance of Self-Organizing Systems; Third Formator Symposium on Mathematical Methods for the Analysis of Large Scale Systems. Academia, Prague, 1979
3. *Bíla, J.*, Introduction to Problems of Automation of Design Operations; Department of Automatic Control, Faculty of Mechanical Engineering, Prague, 1980
4. *Thom, R.*, Structural Stability and Morphogenesis; W. A. Benjamin, INC., Advanced Book Program Reading, Massachusetts, 1975
5. *Miller, G. A.*, A psychological method to investigate verbal concepts; J. Mathematical Psychology, Vol. **6**, No. 2, 1969, pp. 169–192
6. *Kolman, L.*, Memory and Recognition, Academia, Prague, 1976
7. *Newell, A., Shaw, J., Simon, H.*, Report on General Problem Solving Program; Proc. Int. Conf. Information Processing (UNESCO), Paris, 1959
8. *Rescher, N., Urquhart, A.*, Temporal Logic; Springer-Verlag, Wien, New York, 1971
9. *Kostuk, V. N.*, Elements of Modal Logic; Science Lectures, Kiev, 1978
10. *Martynov, V. V.*, The Universal Semantical Code; Science and Technology, Minsk, 1977
11. *Kozachkov, L. C.*, Informational Analysis in Control, Technology, Kiev, 1977
12. *Simon, H.*, The Sciences of the Artificial, The MIT Press, Cambridge, Mass., 1969

Моделирование морфологической и функциональной структуры в символической системе для автоматического синтеза технического объекта

И. БИЛА

(Прага)

Рассмотрены общие основы морфологического и функционального описания объектов в направлении идей автоматического синтеза структуры технического объекта. Полагаются важные вопросы отношения морфологических и функциональных свойств объектов; предполагается примерная роль функционального описания. Изучается: природа синтеза функционального описания, возможности развития знаковой модели для синтеза функциональной структуры и формальные свойства недетерминисма процесса синтеза функциональной структуры.

J. Bíla
Technical University of Prague
Faculty of Mechanical Engineering
Department of Automatic Control
166 07 Suchbatárova 4, Prague 6
Czechoslovakia

Problems of Control and Information Theory, Vol. 10 (3), pp. 199–212 (1981)

THE COMPENSATION OF UNMEASURABLE DISTURBANCES VIA PROPORTIONAL-PLUS-INTEGRAL REGULATORS WITH A PRESCRIBED DEGREE OF STABILITY

D. PETKOVSKI*, M. RAKIĆ**

(Novi Sad) (Belgrade)

(Received April 24, 1980)

The formulation of the multivariable, time-invariant, optimal regulator problem with a prescribed degree of stability is extended so that the solution includes multivariable integral feedback. The integral action is produced by augmenting the state vector, and eliminates offset in the selected state variables due to the unmeasurable constant or slowly varying disturbance inputs. A design technique producing output constrained regulators, simple enough for practical realization, is proposed. This structure includes local as well as hierarchical types of control systems. An illustrative example is worked out.

1. Introduction

Optimally regulated linear systems possess a number of properties attractive to the designer of feedback control loops. Due to these properties both, the optimal controller and the compensated system have been the topic of extensive research in recent years. However, in the presence of persistently acting disturbances, an optimal proportional controller cannot eliminate the steady-state error, as needed in many practical applications.

The multivariable proportional-plus-integral control can, with very little increase in the complexity of the implementation, eliminate the steady-state offsets which arise when proportional control alone is used on systems subject to constant disturbances.

There are a number of papers in the literature which deal with multivariable integral control (Power and Porter 1970a, b, Porter 1971, Davison and Smith 1971, Smith and Davison 1972, Park and Seborg 1974). However, these multivariable proportional-plus-integral regulators require the complete state feedback, either directly or via an observer. In many practical applications this is not feasible because of the inaccessibility for measuring some of the system states, or to the high cost of state reconstruction. These facts justify the demands for simpler or suboptimal control policies. Practical constraints on the feedback system have to be considered, too. A realistic con-

* Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA 02139 U.S.A.; On leave from: Faculty of Technical Sciences, Novi Sad, Yugoslavia.

** Electrical Engineering Faculty, Belgrade, Yugoslavia.

trol scheme in such cases involves the use of incomplete state measurements to form a control feedback.

In this paper, the procedure for designing the output feedback regulator with a prescribed degree of stability is proposed. Computationally, the method is very feasible; it involves only the solution of the usual Riccati matrix equation. This regulator is much more appropriate for a practical realisation than the aforementioned regulators.

A state feedback control is used as the starting point in this paper. The step taken is to fit this control into another "similar" control with a predefined structure. The proposed structure is quite general and can be used to describe local as well as hierarchical types of control systems, i.e., it should cover a large variety of practical constraints that might be imposed on the structure of a feedback system.

The paper is organized as follows: in Section 2, a multivariable proportional-plus-integral optimal state regulator with a prescribed degree of stability is derived. Using this regulator as the starting point, a procedure for designing proportional-plus-integral regulators, where regulators act directly on the available system outputs is suggested in Section 3. Finally a numerical example is given to illustrate the ideas of the previous sections in Section 4.

2. A proportional-plus-integral optimal regulator with a prescribed degree of stability

The results in this section represent an extension of the results of Porter (1971). In the formulation of the optimal control problem, considered in this paper, it is assumed that the process to be controlled can be adequately represented in the region of normal operation by the following linear time-invariant model

$$\dot{x}(t) = Ax(t) + Bu(t) + w(t); \quad x(t_0) = x_0 \tag{1}$$

where

$x(t) \in R^n$, $x(\cdot)$ unconstrained state vector

$u(t) \in R^m$, $u(\cdot)$ control vector

$w(t) \in R^n$, $w(t) = w$ vector of unmeasurable constant of slowly varying input disturbances

$A \in R^{n \times n}$, $B \in R^{n \times m}$.

The objective is to design a constant optimal state regulator, incorporating an integral action to compensate the effect of constant or slowly varying

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

disturbance, $w(t) = w \simeq$ constant, i.e. to eliminate the steady-state offset in the first r elements of $x(t)$*,

$$u(t) = F_1 x(t) + F_2 z(t) \tag{2}$$

where

$$\dot{z}(t) = C_1 x(t), \quad z(t_0) = z_0 \tag{3}$$

$$z(t) \in R^r, \quad r \leq n \quad \text{and} \quad C_1 \in R^{r \times n}.$$

The output vector $y_1 = C_1 x$ expresses not only the output in the strict sense but also the state variables whose integrals are fed back, and is defined as $y_1 = \text{col}\,(x_1, x_2, \ldots, x_r)$. Then the element C_{1ij} of the matrices C_1 is

$$C_{1ij} = \begin{cases} 1, & \text{for} \quad i = j \\ 0, & \text{for} \quad i \neq j. \end{cases} \tag{4}$$

The quadratic performance index that is adopted for the optimal regulatory control problem is of the form

$$J(t_0, x_0, u) = \int_{t_0}^{\infty} e^{2\alpha t} (||x(t)||_Q^2 + ||u(t)||)_R^2 dt \tag{5}$$

where $Q \in R^{n \times n}$ is a symmetric, nonnegative definite matrix, $R \in R^{m \times m}$ is a symmetric, positive definite matrix and $\alpha \in R^1$ is nonnegative number.

It is evident from Eqs. (1) and (3) that the augmented disturbance-free $(n + r)$th-order system will be governed by the state equation

$$\dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + \bar{B}u(t); \quad \bar{x}(t_0) = \bar{x}_0 \tag{6}$$

where

$$\bar{x} = \begin{bmatrix} x \\ z \end{bmatrix} \bar{A} = \begin{bmatrix} A & 0 \\ C_1 & 0 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B \\ 0 \end{bmatrix}. \tag{7}$$

Introducing the transformation

$$\hat{x}(t) = e^{\alpha t} \bar{x}(t) \tag{8}$$

$$\hat{u}(t) = e^{\alpha t} u(t) \tag{9}$$

Eq. (6) implies the equation

$$\dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \bar{B}\hat{u}(t), \quad \hat{x}(t_0) = \hat{x}_0 \tag{10}$$

where

$$\hat{A} = \begin{bmatrix} A + \alpha I & 0 \\ C_1 & \alpha I \end{bmatrix}. \tag{11}$$

* The elements of $x(t)$ may always be arranged so that those which are to be integrated occupy the first r rows of $x(t)$.

The optimal state-regulator problem for the augmented system (10) is that of determining, in the class of continuous function, a control $\hat{u}(t)$ which minimizes the functional

$$J(t_0, \hat{x}, \hat{u}) = \int_{t_0}^{\infty} (||\hat{x}(t)||_{\hat{Q}}^2 + ||\hat{u}(t)||_R^2)\, dt \tag{12}$$

subject to constraints (10). The constant weighting matrix $\hat{Q}$, $\hat{Q} \in R^{(n+r)\times(n+r)}$

$$\hat{Q} = \begin{bmatrix} Q & 0 \\ 0 & Q_1 \end{bmatrix} \tag{13}$$

is a symmetric, nonnegative definite matrix. As it is known (Anderson and Moore 1971), under the assumption that the pair $(\hat{A}, \bar{B})$ is completely controllable, there exists a unique optimal control law

$$u(t) = -F\bar{x}(t), \qquad F = R^{-1}\bar{B}'\hat{K} \tag{14}$$

where

$$\hat{K}, \hat{K} \in R^{(n+r)\times(n+r)}, \tag{15}$$

is a symmetric, positive definite matrix which is the solution of the Riccati equation

$$\hat{K}\hat{A} + \hat{A}'\hat{K} - \hat{K}\bar{B}R^{-1}\bar{B}'\hat{K} + \hat{Q} = 0 \tag{16}$$

such that $u(t)$ minimizes $J(t_0, \hat{x}_0, u)$ in (12). The optimal cost is $J^*(t_0, \hat{x}_0) = e^{2\alpha t_0} \left\| \begin{matrix} x_0 \\ z_0 \end{matrix} \right\|_{\hat{K}}$. Under the assumption that $\hat{Q}$ can be factored as $\hat{Q} = \hat{T}\hat{T}'$, where $\hat{T} \in R^{(n+r)\times(n+r)}$ is a constant matrix, so that the pair $(\hat{A}, \hat{T})$ is completely observable, the system

$$\dot{\bar{x}} = (\bar{A} - \bar{B}R^{-1}\bar{B}'\hat{K})\bar{x} \tag{17}$$

is globally exponentially stable with degree α.

This controllability requirement is crucial because it has been shown (Porter 1970) that the introduction of integral feedback may destroy the controllability of an initially controllable system, i.e. the complete controllability of the pair $\{A, B\}$ of the original system (1) is not sufficient to guarantee the controllability of $(\hat{A}, \bar{B})$ of the augmented system (10). The theorem in the Appendix provides the necessary and sufficient condition to retain the complete controllability of (10).

The theorem implies that the number of integrators must be smaller or equal to the number of control inputs, i.e., $r \leq m$.

If $\hat{K}$ is written in the partitioned form

$$\hat{K} = \begin{bmatrix} \hat{K}_{11} & \hat{K}_{12} \\ \hat{K}'_{12} & \hat{K}_{22} \end{bmatrix} \tag{18}$$

where $\hat{K}_{11} \in R^{n\times n}$, $\hat{K}_{12} \in R^{n\times r}$ and $\hat{K}_{22} \in R^{r\times r}$, it is evident from (10) that the required optimal input vector for the augmented system of equation (10) is given by

$$u(t) = -R^{-1}B'\left(\hat{K}_{11}x(t) + \hat{K}_{12}z(t)\right) \tag{19}$$

i.e.

$$u(t) = -R^{-1}B'\hat{K}_{11}x(t) - R^{-1}B'\hat{K}_{12}C_1\int_{t_0}^{t} x(t)\,dt + u_0 \tag{20}$$

where

$$u_0 = -R^{-1}B'\hat{K}_{12}z_0. \tag{21}$$

Now when a constant disturbance represented by the vector w is present, the optimal closed loop system with the proportional-plus-integral regulator is governed by the equation

$$\begin{bmatrix} \dot{x}(t) \\ \dot{z}(t) \end{bmatrix} = \begin{bmatrix} A - BR^{-1}B'\hat{K}_{11} & -BR^{-1}B'\hat{K}_{12} \\ C_1 & 0 \end{bmatrix}\begin{bmatrix} x(t) \\ z(t) \end{bmatrix} + \begin{bmatrix} w(t) \\ 0 \end{bmatrix}. \tag{22}$$

Following Porter (1971), it can be easily shown that the introduction of feedback of the integrals of the first r elements of $x(t)$ eliminates in these variables the steady-state offset which arises when proportional control alone is used. That is, the solution $x_i(t, t_0, x_0)$, $i = 1, 2, \ldots, r$ approaches the equilibrium at the origin at least as fast as $e^{-\alpha t}$ for all $(t_0, x_0) \in R^{n+1}$ and $t \in T$, where $T = [t_0, +\infty]$. Particularly, for $\alpha = 0$, this solution agrees with the results of Porter (1971).

Thus, the optimal closed loop system, with the proportional-plus-integral regulator is not only optimal with respect to a quadratic performance index but also possesses good steady-state response characteristics in the face of constant or slowly varying input disturbances.

The optimal closed-loop system is given by

$$\begin{bmatrix} \dot{x}(t) \\ \dot{u}(t) \end{bmatrix} = \begin{bmatrix} A & B \\ -R^{-1}B'(\hat{K}_{11}A + \hat{K}_{12}C_1) & -R^{-1}B'\hat{K}_1 B \end{bmatrix}\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + \\ + \begin{bmatrix} I \\ -R^{-1}B'\hat{K}_{11} \end{bmatrix} w(t) \begin{bmatrix} x(t_0) \\ u(t_0) \end{bmatrix} = \begin{bmatrix} x_0 \\ u_0 \end{bmatrix}. \tag{23}$$

3. A proportional-plus-integral regulator with multiple constrained feedback structure

The design of the proportional-plus-integral optimal regulator, described in the previous section requires a measurement of all the system states. Frequently, not every state variable is either physically available for measurement, or can be measured accurately or conveniently. A realistic control scheme in such cases involves the use of incomplete state measurements to form a control feedback. On the other hand, in the case of a large-scale system the more suitable concept of decentralised control has gained in popularity. So, practical constraints on the feedback system must be considered. In such cases a complex structure can be obtained if the ith input component vector is restircted to be a function of certain specified outputs, that is,

$$u_i(t) = L_i^1 y_i^1(t) + L_i^2 \int_{t_0}^{t} y_i^2(t)\, dt + u_{i0} \quad i = 1, 2, \ldots, t_1 \tag{24a}$$

if the integral action is incorporated in $u_i(t)$, or

$$u_j(t) = L_j^1 y_j^1(t) \quad j = 1, 2, \ldots, t_2 \quad t_1 + t_2 = k \tag{24b}$$

if the integral action is not incorporated in u_j, where

$$y_v^1(t) = C_v^1 x(t), \quad v = i, j \quad y_v^1 \in R^{p_v}, \; y_v^1(.) \tag{25}$$

vth measurable component vector $\sum_v p_v = p$,

$$y_i^2(t) = C_i^2 x(t); \quad y^2 \in R^{r_i}, \quad y_i^2(.) \tag{26}$$

ith measurable output component vector whose integrals are feedback, $\sum_i r_i = r$.

The partition of the control vector (24) is determined by a physical structure of the system. It is easily verified that decentralised control, i.e. control exercised by several controllers, each of them manipulating only a subset of the input or control variables, this manipulation being based on an incomplete knowledge of the system, is included in this formulation. If various controllers are arranged hierarchically, this approach leads to hierarchical control. This fact is of particular interest for large-scale systems (e.g. energy systems, environment systems, etc.) where the subsystems are geographically somewhat distant from each other and where the feedback of state information between the subsystem controllers can prove to be the major component

of the cost of implementation of the control. Therefore, there may be a desire to control the systems with local feedback on the different processes, eventually with the addition of a small number of interconnections.

In order to simplify the notation the asterisk (*) will be used to indicate the properties associated with the reduced control law. Rewrite Eqs (24) as

$$u_i^*(t) = E_i \bar{x}^*(t);\ E_i = L_i \bar{C}_i \tag{27}$$

where

$$L_i = \begin{bmatrix} L_i^1 & L_i^2 \end{bmatrix} \tag{28}$$

and

$$\bar{C}_i = \begin{bmatrix} C_i^1 & 0 \\ 0 & G_i \end{bmatrix};\quad G_i = [0\ 0\ \ldots\ I_i\ \ldots\ 0] \tag{29}$$

for Eq. (24a) and

$$L_i = L_i^1 \tag{30}$$

$$\bar{C}_i = [C_i^1\ 0] \tag{31}$$

for Eq. (24b).

Therefore, the objective is to design a constant linear feedback control that has a multiple structure constraint

$$u^*(t) = E\bar{x}^*(t) \tag{32}$$

where

$$E = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_k \end{bmatrix} \tag{33}$$

to compensate the effect of constant or slowly varying disturbance $w(t)$.

The augmented closed-loop system with a proportional plus-integral regulator with the multiple control structure constraint on $u^*(t)$ becomes

$$\dot{\bar{x}}^*(t) = \left(\bar{A} + \sum_{i=1}^{k} \bar{B}_i L_i \bar{C}_i\right) \bar{x}^*(t) + \bar{w}(t) \tag{34}$$

where

$$\bar{A} = \begin{bmatrix} A & 0 \\ C_1 & 0 \end{bmatrix},\ \bar{x}^*(t) = \begin{bmatrix} x^*(t) \\ z^*(t) \end{bmatrix},\ \bar{w}(t) = \begin{bmatrix} w(t) \\ 0 \end{bmatrix} \tag{35}$$

and

$$\bar{B} = [\bar{B}_1\ \bar{B}_2 \ldots \bar{B}_k]$$

is a partition of the input matrix compatible with the partition of the contro vector $u^*(t)$. On the other hand the augmented close-loop system with the complete feedback control can be written in the form

$$\dot{\bar{x}}(t) = \left(\bar{A} + \sum_{i=1}^{k} \bar{B}_i F_i\right) \bar{x}(t) + \bar{w}(t) \tag{36}$$

where

$$F = -R^{-1} B' \hat{K} \tag{37}$$

and

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{bmatrix}. \tag{38}$$

When some states are not fed back into the control system, it is impossible to obtain the optimal feedback control agreeing with the optimal one for the state feedback system. Then the trajectory of the output feedback system becomes different from that of the state feedback system. So the objective is to determine control (32) so that the trajectories of systems (34) and (36) will be "near" each other.

To quantify the aforementioned nearness, define the trajectory error between $\bar{x}^*$ and $\bar{x}$ as

$$r(t) = \bar{x}^*(t) - \bar{x}(t) \tag{39}$$

that is,

$$\dot{r}(t) = \left(\bar{A} + \sum_{i=1}^{k} \bar{B}_i L_i \bar{C}_i\right) r(t) + \sum_{i=1}^{k} \bar{B}_i (L_i \bar{C}_i - F_i) \bar{x}(t).$$

$$r(t_0) = 0. \tag{40}$$

It is obvious that if $\sum_{i=1}^{k} \bar{B}_i (L_i \bar{C}_i - F_i) \bar{x}(t) = 0$, then $r(t) = 0$, for $\forall t \in T$, where $T = [t_0, +\infty]$, since $r(t_0) = 0$. However, the case $L_i \bar{C}_i = F_i$ is excluded. There is a variety of quantities one can choose to minimize the influence of $\sum_{i=1}^{k} \bar{B}_i (L_i \bar{C}_i - F_i) x$. For instance, an obvious possibility is to choose $u^*(t)$, i.e. $L_i \bar{C}_i$ so as to minimize $\left\| \sum_{i=1}^{k} \bar{B}_i (L_i \bar{C}_i - F_i) \bar{x} \right\|$, where $||.||$ denotes the ordinary euclidian quadratic norm. So, the problem is reduced to choose the control $u^*(t)$ so that $\inf_{L_i} \left\| \sum_{i=1}^{k} \bar{B}_i (L_i \bar{C}_i - F_i) \bar{x} \right\|$ is achieved for all $\bar{x} \in R^{n+m}$. Remembering that

$$\min_{L_i} \sum_i ||\bar{B}_i (L_i \bar{C}_i - F_i) \bar{x}|| \leq \min_{L_i} \sum_i ||\bar{B}_i|| \, ||L_i \bar{C}_i - F_i|| \, ||\bar{x}|| \tag{41}$$

holds for $\forall \bar{x}(t) \in R^{n+m}$ the problem reduces to determine

$$\min_{L_i} ||L_i \bar{C}_i - F_i||. \tag{42}$$

From this condition the suboptimal controller L is

$$L_i = F_i \bar{C}_i'(\bar{C}_i \bar{C}_i')^{-1}, \quad i = 1, 2, \ldots, k \tag{43}$$

i.e., the linear feedback control with multiple structure constraints is

$$u_i^*(t) = F_i \bar{C}_i'(\bar{C}_i \bar{C}_i')^{-1} \bar{C}_i \bar{x}^*(t). \tag{44}$$

Partition the state feedback matrix F as

$$F = \begin{bmatrix} F_1^1 & F_1^2 \\ F_2^1 & F_2^2 \\ \vdots & \\ F_k^1 & F_k^2 \end{bmatrix}. \tag{45}$$

Then, $u_i^*(t)$ can be expressed as

$$u_i^*(t) = F_i^1 C_1^{1'} (C_i^1 C_i^{1'})^{-1} C_i^1 x^*(t) + F_i^2 G_i (G_i G_i')^{-1} G_i z(t) \tag{46}$$

for $u_i(t)$ when the integral action is incorporated in $u_i(t)$. When the integral action is not incorporated in $u_i(t)$ then (46) is reduced to

$$u_i^*(t) = F_i^1 C_i^1 (C_i^1 C_i^{1'})^{-1} C_i^1 x^*(t). \tag{47}$$

When this control problem has a solution, i.e., when the output constrained closed-loop system is asymptotically stable, the integral action produced by augmenting the state vector eliminates the steady-state offset in the first r elements of $x^*(t)$, i.e. the solution $x_i^*(t, t_0, x_0^*)$, $i = 1, 2, \ldots, r$ approaches the equilibrium at the origin "nearly" as fast as $e^{-\alpha t}$ for all $(t_0, x_0^*) \in R^{n+1}$ and $t \in T$, where $T = [t_0, +\infty]$.

Now, when a constant disturbance represented by vector w is present, the optimal closed-loop system with the proportional-plus-integral regulator with multiple constrained feedback structure, is governed by the equation

$$\begin{bmatrix} x^*(t) \\ z^*(t) \end{bmatrix} = \left[\begin{array}{c|c} A - \sum_{i=1}^{k} \bar{B}_i L_i \bar{C}_i' (\bar{C}_i \bar{C}_i')^{-1} \bar{C}_i & \\ \hline C_1 \quad | \quad 0 & \end{array}\right] \begin{bmatrix} x^*(t) \\ z^*(t) \end{bmatrix} + \begin{bmatrix} w(t) \\ 0 \end{bmatrix}. \tag{48}$$

The design of proportional — plus — integral regulators with multiple constrained feedback structure involves only the solution of the usual nonlinear

matrix Riccati equation. Besides very simple numerical calculations these regulators require neither all states measurement, nor state reconstruction. So these regulators are much more appropriated for practical realisation than the afore-mentioned proportional — plus — integral regulators.

4. Example

The design technique proposed is illustrated by a six-order linear time-invariant system defined by Lapidus and Amundson (1950)

$$\dot{x} = \begin{bmatrix} a_1 & a_2 & 0 & 0 & 0 & 0 \\ a_3 & a_1 & a_2 & 0 & 0 & 0 \\ 0 & a_3 & a_1 & a_2 & 0 & 0 \\ 0 & 0 & a_3 & a_1 & a_2 & 0 \\ 0 & 0 & 0 & a_3 & a_1 & a_2 \\ 0 & 0 & 0 & 0 & a_3 & a_2 \end{bmatrix} x + \begin{bmatrix} b_1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & b_2 \end{bmatrix} u + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} w.$$

The entries of the matrices A and B are

$$a_1 = -1.73058 \qquad a_1 = 0.634231 \qquad a_3 = b_1 = 0.538827$$
$$b_2 = 0.8809$$

and

$$x(t_0) = \begin{bmatrix} -0.0306 \\ -0.0568 \\ -0.0788 \\ -0.0977 \\ -0.1138 \\ -0.1273 \end{bmatrix}.$$

For the matrices, as performance index $Q = Q_1 = R = I$ is chosen. Numerical values for w is

$$w' = [0.01 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0].$$

Four types of regulator are considered

1) State feedback P regulator
2) State feedback PI regulator
3) Output constrained PI regulator with $\overline{C}_i = \overline{C}$ for $i = 1, 2$

$$\overline{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

4) Local type of control system

$$\bar{C}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}; \qquad \bar{C}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

A time response of the controlled P-regulator and uncontrolled system with and without external disturbance is given in Fig. 1 for the output variable $x_1(t)$. As can be seen the P-regulator cannot eliminate the steady state offset in $x_1(t)$.

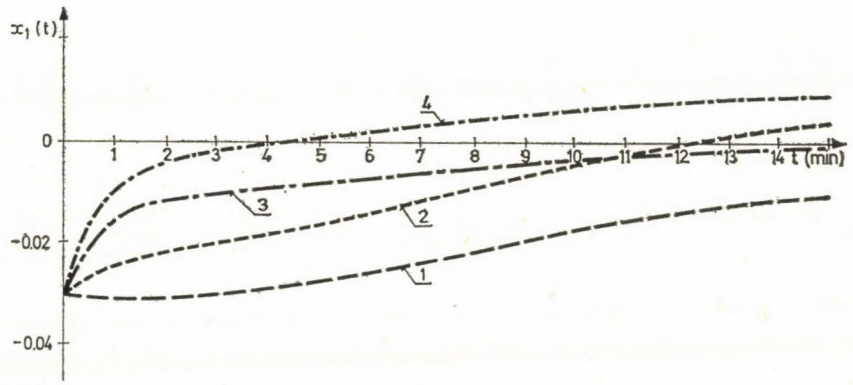


Fig. 1. Time response of the open and the closed loop system for the output of the plate ($x_1(t)$): (1) open loop response without disturbance; (2) open loop response with disturbance; (3) closed loop response with P-regulator, without disturbance; (4) closed loop response with P-regulator, with disturbance

Applying the control strategy proposed in this paper we obatin the control and trajectories for two different values of $\alpha(0, 0.01)$. Figure 2 shows the time

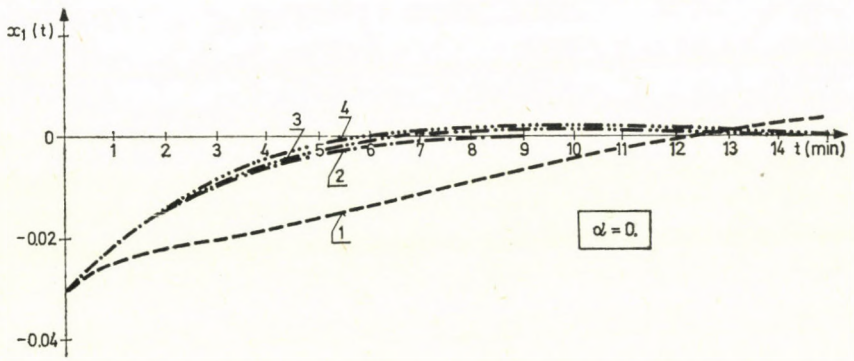


Fig. 2. Time response of the open and the closed loop system with disturbance for the output of the plate ($x_1(t)$: (1) open loop response; (2) closed loop response with state feedback *P1* regulator; (3) closed loop response with output constrained *P1* regulator; (4) closed loop response with local type *P1* regulator; $\alpha = 0$

response of closed loop systems with *PI* regulator for $\alpha = 0$, while Fig. 3 shows the time response of closed loop systems with *PI* regulator for $\alpha = 0.01$. From these curves we see that the time response of the closed loop system is a significant improvement over the open loop system. On the other hand the difference between the exact and the reduced control laws (even with local types of control) is astonishingly small, indicating that a control only using feedback from the measured variables will be sufficient in this case.

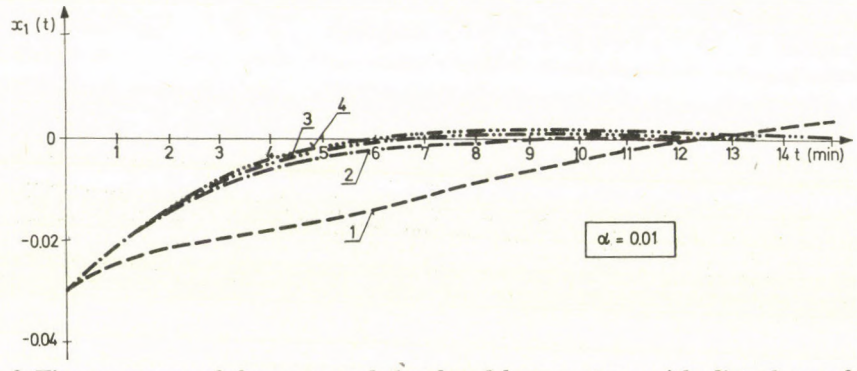


Fig. 3. Time response of the open and the closed loop system with disturbance for the output of the plate ($x_1(t)$): (1) open loop response; (2) closed loop response with state feedback *PI* regulator; (3) closed loop response with output constrained *PI* regulator; (4) closed loop response with local type *PI* regulator, $\alpha = 0.01$

The same conclusions for the effectiveness of the proposed method for the design of *PI* regulator with multiple constrained feedback structure, can be taken from the results given in Table 1, which gives the corresponding values of the performance index. The effect of $\hat{x}_0$ is averaged out by assuming that $\hat{x}_0$ is a random vector with known statistical property: $E\{\hat{x}_0\} = 0$ and $E\{\hat{x}_0\hat{x}_0'\} = I$ (Anderson and Moore, 1971)

Table 1. Comparison of effectiveness of reduced control law

	State feedback	Output feedback	Local type control
tr $V(\alpha = 0)$	9.710800	9.938550	9.991754
tr $V(\alpha = 0.01)$	9.987838	10.242786	10.304426

The degradation of the system performance due to the output constrained feedback structure is negligible. The maximum relative error is less than 3.2 per cent.

The excellent performance of the derived local type regulator is of special interest, as due to increasing complexity of modern technological processes, the local (as well as hierarchical) control schemes have become an essential requirement in the design of large-scale system.

Conclusions

A method was presented for designing the proportional-plus-integral regulators with multiple constrained feedback structure. The proposed design scheme, which includes local as well as hierarchical types of control systems, involves the solution of one nonlinear Riccati matrix equation.

The numerical example considered shows the practicality and good performance of the control system developed, and indicates that this method is a realistic design method for multivariable systems.

Appendix

Theorem 1. The pair $(\hat{A}, B)$ is controllable if, and only if, the following conditions are satisfied: (a) If A is non-singular, (A, B) must be a controllable pair and

$$\text{rank } TA^{-1}B = r. \tag{A.1}$$

(b) If A is singular, (A, B) must be a controllable pair and

$$\text{rank } T(A + BD)^{-1}B = r \tag{A.2}$$

where D is any matrix for which $(A + BD)$ is non-singular.

Proof. Matrix $\hat{A}$ can be equivalently written as

$$\hat{A} = A_1 + \alpha I \tag{A.3}$$

where

$$A_1 = \begin{bmatrix} A & 0 \\ C_1 & 0 \end{bmatrix}. \tag{A.4}$$

It is well known (Anderson and Moore 1971) that the complete controllability of $(A_1 + \alpha I, B)$ follows from the complete controllability of (A_1, B). On the other hand (Power and Porter, 1970b) the necessary and sufficient conditions for complete controllability of (A_1, B) are given by (a) i.e. (b) Q.E.D.

References

Anderson, B. D. O., Moore, J. B. (1971), "Linear Optimal Control", Prentice Hall.

Davison, E. J., Smith, H. W. (1971), "A note on the design of industrial regulators: integral feedback and feedforward controllers", *Automatica*, Vol. **10,** pp. 329.

Lapidus, L., Amundson, N. R. (1950), "Stagewise absorption and extraction equipment", *Ind. Eng. Chem.*, Vol. **42,** pp. 1071.

Park, H., Seborg, D. E. (1974), "Eigenvalue assignment using proportional-integral feedback control", *Int. J. Control*, Vol. **28,** pp. 517.

Porter, B. (1971), "Optimal control of multivariable linear systems incorporating integral feedback", *Electronics Letters*, Vol. **7,** pp. 170.

Power, H. M., Porter, B. (1970a), "Eigenvalue assignment in multivariable systems incorporating integral feedback", *Electronics Letters*, Vol. **6,** pp. 795; (1970b) "Necessary and sufficient conditions for controllability of multivariable systems incorporating integral feedback", *Electronics Letters*, Vol. **6,** pp. 815.

Smith, H. W., Davison, E. J. (1972), "Design of industrial regulators-integral feedback and feedforward control", *Proc IEEE*, Vol. **119,** pp. 1210.

Компенсация недоступных измерению возмущений путем использования пропорционально-интегральных регуляторов с заданным уровнем стабильности

Д. ПЕТКОВСКИ, М. РАКИЦ

(Нови Сад)

Определение мультивариабельного, инвариантного во времени оптимального управления с заданным уровнем стабильности расширено таким образом, что решение проблемы включает мультивариабельную обратную связь. Интегральная обратная связь, получаемая путем расширения вектора состояния, ликвидирует отклонения отдельных векторных элементов, возникающих из-за непрерывных или медленно изменяющихся, неподдающихся измерению возмущений на входе. Разработан способ получения регуляторов с неполной обратной связью, достаточно простой для практической реализации; этот способ включает в себя как локальные, так и иерархические типы контрольных систем. Приведены примеры расчета таких систем.

D. Petkovski
Massachusetts Institute of Technology
Room No. 35—205
77 Massachusetts Avenue
Cambridge, Massachusetts 02139
USA

Problems of Control and Information Theory, Vol. 10 (3), pp. 213—218 (1981)

BOOK REVIEW

Pattern formation by dynamic systems and pattern recognition. (ed. **H. Haken**; Springer-Verlag, Berlin—Heidelberg—New York, 1979)

The volume contains the papers delivered at the International Symposium on *Synergetics* held at Elmau, Federal Republic of Germany in 1979. This new interdisciplinary sphere of science deals with the systematic study of the cooperative effects in physical systems far from equilibrium and in other disciplines. The basic concept is that there exist profound analogies in the behaviour of large classes of complex systems, though the systems themselves may be quite different.

The late Norbert Wiener said that "one of the most interesting aspects of the world is that it can be considered to be made up of patterns. A pattern is essentially an arrangement. It is characterized by the order of the elements of which it is made rather than by the intrinsic nature of these elements".

Although synergetics is mainly dealing with the question whether there are general principles by which the formation of macroscopic patterns can be understood, the most recent developments in this field include the using of pattern recognition methods, especially the syntactic algorithms.

A few of the most interesting topics are the following:

— pattern formation in fluids
— turbulence and chaos
— pattern recognition and pattern formation in biology
— pattern formation in ecology, sociology and history.

The latest advances in syntactic pattern recognition are presented by the most well-known experts of the field, including K. S. Fu and U. Grenander.

L. Molnár

Jeffrey M. Tobias (Ed.): Language Design and Programming Methodology. Proceedings of a Symposium held in Sydney, Australia, 10—11 September, 1979. In Lecture Notes in Computer Scienc; vol. **79**, Springer-Verlag, Berlin–Heidelberg–New York, 1980, 255 pp.

"If the seventies are to go down in the annals of computer science history as a decade of technological achievement in the manifacture of semiconductor devices, then one might predict that the eighties would see a commensurate expansion in the development of programming tools and methods."

The introduction of the book is started with these sentences explaining the interest in the future of programming and its methodology. Over two hundred members of the scientific community met at a Symposium in Sydney in September, 1979, to discuss language design and programming techniques. This volume is a collection of 17 lectures of the Symposium.

In the first paper Professor Niklaus Wirth of ETH Zurich renowned for his development of a number of superly engineered programming tools presents a programming

language concept called module as a means to partition system effectively. The module allows to encapsulate local details and to specify explicitly those elements that are part of the interface between modules.

In the second paper Dr. Dennis Ritchie of Bell Laboratories USA, the originator of the programming language C and co-author of the UNIX time-sharing system give a brief history of the early development of the UNIX operating system. The paper concentrates on the evolution of the file system, the process-control mechanism, and the idea of pipelined commands.

The third paper compares the two programming language PASCAL and C. While the two languages have comparable data and control structures, the program structure of C appears superior.

In the fourth paper the author deals with recursion as a programming technique. He considers the objection to it and the claims for it.

In the fifth paper the problem of findings prime numbers which has been much favoured by expositors of modern programming methodology is tackled once more, but from a broader view of the programming process than has previously been taken.

The sixth paper deals with the total effect of programs in order to produce correct software. Instead of using predicate transformers, a more powerful concept, the passback pair is introduced.

The seventh paper tries to embed the pattern matching facility of Snobol 4 in a more conventional language. It shows that the control structures that underlie pattern matching operations create difficulties, if the full power of Snobol 4 is to be retained.

The eighth paper introduces a language for describing concepts as programs, while in the ninth one another language is surwayed which may be of interest for computer animation of a human movement.

Paper 10 explains a simple connection between transition graphs and algebraic specifications of data types.

Paper 11 writes about the program DASIM1 which can construct and simulate data structures.

In paper 12 a new powerful programming language intended as a contribution to the development of PASCAL is discussed.

Paper 13 writes about Description-Oriented Programming Languages (DOPLs). Several examples are used to illustrate DOPLs, and the advantages of using DOPLs over Identifier-Oriented Programming Languages.

In paper 14 Professor Nikolaus Wirth shows that an inexpensive microcomputer, so-called work-station computer, holds great promise for the future as the development tool of the software engineer.

Paper 15 describes an experiment and its results, which was taken for testing the applicability of software science in the realm of student programming, and for obtaining quantitative inferences about the programming language PASCAL.

In paper 16 the author gives a critique of language MODULA. The emphasis is on the useability of MODULA, and whether it achieves its stated goals.

The last paper compares the notations for process communication given by Hoare and Brinch Hausen. The comparison is done by both qualitative and quantitative analyses. A significant variation in effectiveness with program class is shown.

For those who are interested in language design and programming methodology this book may be of great interest and is recommended by the rewiever.

B. Bago

F. Gaffal: Datenverarbeitung im Hochschulbereich der USA. Stand und Entwicklungstendenzen. (Data Processing in Higher Education of the USA. Status and tendencies of development). Springer-Verlag, Berlin–Heidelberg–New York, 1980, 199 pp.

This book is a volume of the series "Informatik-Fachberichte". The book is mainly the result of the author's 2-months visit in the USA at the end of 1978. During this trip he visited some central institutes (e.g. American Council of Education, Department of Health, Education and Welfare) and 22 universities and other organizations of higher

education, respectively (e.g. University of Illinois, University of California, Stanford University, MIT, EDUCOM, etc.).

The book contains six chapters and an appendix. The first chapter is an introduction in which the author describes the aim of the book and the highlights of his visit in the USA.

Chapter 2 deals with the structure of higher education in the USA.

In Chapter 3 the general organisation of data processing in higher education of the USA is discussed briefly.

Chapter 4 presents the data processing in education and research. Here are discussed some special (dedicated) systems with a centre of interest for education, such as "computer aided education" system PLATO, which is a really dedicated hardware-software system used in many states of the USA, in Canada (Quebec) and in Belgium (Bruxelles). Further systems described in this chapter are computer networks or organisations for data processing in higher education using computer networks, respectively. They are the following: Educatinal Network (EDUNET), New England Regional Computing Program, New Jersey Educational Computing Network, Merit — Network and Triangle Universities Computation Center. For each network the structure and hardware/software configuration, the organization, the actual usage and the plans for development are discussed in detail and with a plenty of actual data.

Chapter 5 deals with the data processing in the management and administration of higher education. The fields and methods of university management are discussed and shown by three examples (University of Michigan, University of Illinois and University of California, Berkeley). In each case the structure of data processing, the hardware/software configuration for administration and management, the history of EDP — application and tendencies of development are described.

In Chapter 6 the data processing in the libraries of universities is presented. The details are the following: aim of EDP-application in libraries, automation of libraries, implementation of the EDP-system, usage of the library, access methods of the literature. Three examples are discussed: Library Data Management System of the University of Chicago, Midwest Regional Library Network and the data processing in the libraries of the University of California.

The appendix contains the listing of the visited institutions of higher education and the listing of general and EDP-concerning data of the visited universities.

The main advantage of this book is its richness of actual and historical data concerning the data processing in higher education of the USA. The book is recommended for everybody dealing with organisation or management of EDP in any university.

Gy. Bikich

N. Ryska, S. Herda: Encryption Methods in Data Processing. Springer-Verlag, Berlin–Heidelberg–New York 1980, pp. 401 (Original title: Kryptographische Verfahren in der Datenverarbeitung; in German)

Security in commercial digital communication is becoming more and more important. Appropriate measures for physical security are well understood, if not always practiced. Fire, flood, power loss, disgruntled employees, etc. can, by and large, be held in check by convertional protection, prevention and recovery mechanisms. Data processing and communications systems, however, involve another dimension of security — namely information protection. This is one subject about which much is known and very little published, primarily because its principal area of application has been government communications security, where everything pertinent is highly classified.

The need for information security has two sources: privacy considerations and commercial intelligence protection. While the motivation, ethical considerations, and regulatory legislation will be radically different on these points, the basic principles employed to achieve the necessary security are essentially the same. Encryption has been the preferred solution from the time of Julius Ceasar to the present.

The aim of the authors in this book is to give a detailed description of all aspects of encryption methods in data processing starting from the formal description of crypto-functions to the practical implementation of hardware and software moduls. Great em-

phasis laid is upon systematizing and generalizing the subject discussed. However, it is based entirely on published material.

The book consists of seven chapters, a detailed list of references and a list of abbreviations. Chapter 1 is an introduction to cryptography. It comprises the functions of cryptography, the requirements for secure data communication, the development of algorithmic cryptography, selection of appropriate cryptographic processes, their limitations, etc. It is followed in chapter 2 with the considerations of risk of security in data processing systems.

Chapter 3 covers the subject of analysis of cryptographic processes. A detailed description is given here to such problems as structure of cryptographic processes, their mathematical description, properties of cryptographic functions and systems, analysis of cryptographic processes, etc.

Chapter 4 is devoted to the implementation of cryptographic processes in computer systems. The physical and logical interfaces, the implementation of cryptographic processes, the specification and supervision of keys, etc. are discussed in it.

Chapter 5 treats the problems of implementing cryptographic processes in computer networks. It is followed with the discussion of encryption methods in data base systems.

The last chapter deals with the authentication in the data processing systems. Among others it covers the problems related to accountability, authority, liability, responsibility, etc.

Since the presentation is very clear and adequate, the subject is very important, the reviewer believes that this book will be very valuable for everybody who is in some way involved in data processing. Management can understand the importance of implementing cryptographic processes in data processing systems, system engineers can get a complete survey of every field of the subject discussed, and even the specialists can find their interest in this book.

T. Gál

A. Bode, W. Händler: Computer Architecture. Springer-Verlag, Berlin–Heidelberg–New York 1980, pp. 278 (Original title: Rechenarchitektur; in German)

This book gives an introduction to computer architecture dealing with the principles and theoretical procedures which are necessary for understanding and designing computer structure. It contains the text for the first semester of a two-semester lecture on computer architecture.

In the first chapter the terms related to computer architecture are surveyed, the main phases of computer design are shown and some possible classification of computers are given.

It is followed in the second chapter with the discussion of performance evaluation. Giving a survey of the main methods of performance evaluation it discusses the problem from three different points of view: from the traffic theory, from the graph theory and from the probability theory. The examples given help the reader to understand the text more easily.

In chapter 3 the influence of technology on computer architecture is treated. The development in the field of computer hardware is summarised, the newest results of the technological development — the microprocessors and microcomputers are discussed and the technology of the future is outlined in this chapter.

Chapter 4 is devoted to some formal methods which are useful in describing, analysing and designing computer systems. Among others it deals with automata theory, Petri-nets and computer system description languages.

Chapter 5 presents the functional parts of computer systems and their interconnecting structure. A detailed discussion is given to the bus-concept and to the microprogramming.

References and an index are given at the end of the book. They can guide the reader in further study.

T. Gál

Stochastic Control Theory and Stochastic Differential Systems (Edited by **M. Kohlmann** and **W. Vogel**), Proceedings of a Workshop of the "Sonderforschungsbereich 72 der Deutschen Forschungsgemeinschaft an der Universität Bonn", Bad Honnef, January, 1979. Vol. **16** in "Lecture Notes in Control and Information Sciences", Springer-Verlag, Berlin–Heidelberg–New York

This book contains the contributions to a workshop on stochastic control theory and stochastic differential systems at the University of Bonn in January 1979. The main aim of this conference was to give a compact and possible rather complete presentation of the present state of the art and — in particular — of the scope of methods used in stochastic system theory, in order to stimulate the interest of scientists and students working in other areas of stochastic, in engineering sciences, in econometrics, etc.

This volume can be divided into two main parts. In the first part there are 12 survey lectures on different topics and methods of stochastic system theory. Thus this part can be considered as an introductory compendium for non-experts.

The survey lectures are:

- White Noise Models in Non-Linear Filtering and Control (by A. V. Balakrishnan),
- Optimal Impulsive Control Theory (by A. Bensoussan),
- An Introduction to Duality in Random Mechanics (by J. M. Bismut),
- Linear Stochastic Ito Equaton in Hilbert Spaces (by R. F. Curtain),
- Martingale Methods in Stochastic Control (by M. H. A. Davis),
- A Geometric Approach to Linear Control and Estimation (by T. E. Duncan),
- The Martingale Calculus and Applications (by R. J. Elliott),
- Interaction Between Stochastic Differential Equations and Partial Differential Equations (by A. Friedman),
- Approximation of Solutions to Differential Equations with Random Inputs by Diffusion Processes (by H. Kushner),
- Optimal Conditions and Sufficient Statistics for Controlled Jump Processes (by R. Rishel),
- Stochastic Filtering Theory: A Discussion of Concepts, Methods and Results (by J. van Schuppen),
- Introduction to the Theory of Optimal Stopping (by J. Zabczyk).

The second part of the book contains 35 research reports presented during the conference. These papers give a glimpse into the present state of research in the field of stochastic system theory. Not mentioning all of the papers we are going to give a short description of them.

E. O. Bertsch (Technical University, Berlin) in his paper presents some remarks concerning attainable sets of stochastic optimal control systems.

K. Borkar and P. Varaiya (University of California, Berkeley) deal with adaptive control of Markov chains. They considered a controlled Markov chain whose transition probabilities are parametrized by α known to be in a finite set A.

B. Doshi (Rutgers University, USA) gives a policy improvements algorithm for continuous time Markov decision processes with switching costs.

Th. Gasser and G. Dummermuth (University of Children' Clinic, Zurich) discuss a number of problems related to stochastic models and statistical methods used in electroencephalography. They present a variety of methods together with examples.

B. Grigelionis (Mokslu Akademija, Vilnius) deals with canonical form and local characteristics of semimartingales.

K. Helmes (Institute for Applied Mathematics, Bonn) gives a numerical comparison of non-linear with linear prediction for the transformed Ornstein–Uhlenbeck process. He investigates what advantages can be expected in passing from optimal linear prediction to non-linear prediction.

A. Kistner (University of Stuttgart, Stuttgart) makes some remarks on the solution and the moments of linear systems with randomly disturbed parameters.

A. J. Krener (University of California, USA) considers minimum convariance, minimax and minimum energy linear estimators. He shows that these estimators arise when one seeks to minimize the maximum error assuming that initial conditions and noises are bounded in norm in an appropriate Hilbert space.

E. Pardoux (IRIA Laboratory, France) presents a filtering method for a diffusion process with Poisson-type observation.

G. Sawitzki (Institute for Mathematics, Bochum) gives a recursive algorithm for exact filtering in exponential families; discrete case.

The papers listed above show the main topics of the Conference. Summarizing the 35 papers presented in the second part of this book cover a wide range of stochastic control theory and differential systems.

Consequently, the volume is recommended for experts and scientists working in this field.

Miklós Vajta, Jr.

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by guies, 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

L. Györfi

Technical University of Budapest

H–1111 Budapest XI., Stoczek u. 2, 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 (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 th paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) 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 статей в четырех экземплярах на русском языке или в двух на английском и в двух на русском языке следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50–100 слов и приложено резюме – реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в двух экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Pik, J.: Syntactic recognition of nonstylized handwritten numerals (*Пик Й.* Синтаксическое распознавание нестилизованных рукописных цифр) 149

Yemel'yanov, S. V., Korovin, S. K.: Development of feedback types and their application to design of closed-loop dynamic systems (*Емельянов С. В., Коровин С. К.* Пути развития типов обратных связей и их применение при построении замкнутых динамических систем) 161

Sinha, N. K., Rózsa, P.: Algorithm for transformation of a polynomial matrix to canonical form (*Синха Н. К., Рожа П.* Алгоритм для преобразования многочленных матриц к каноническому виду) 175

Kučera, V., Šebek, M.: Linear optimal dynamic regulators (*Кучера В., Шебек М.* Линейные оптимальные динамические регуляторы) 181

Bíla, J.: Modeling of morphological and functional structure in the symbolical system for the automatic synthesis of technical object (*Била И.* Моделирование морфологической и функциональной структуры в символической системе для автоматического синтеза технического объекта) 191

Petkovski, D., Rakić, M.: The compensation of unmeasurable disturbances via proportional-plus-integral regulators with a prescribed degree of stability (*Петковски Д., Ракиц М.* Компенсация недоступных измерению возмущений путем использования пропорционально-интегральных регуляторов с заданным уровнем стабильности) 199

Book review 213

HU ISSN 0370—2529

Index: 26 660

316.920

VOL. TOM 10 * NUMBER HOMEP 4

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1981

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Roumania, Viet-Nam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H–1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1981 Subscription Rate US Dollars 110.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)

G. BOGNÁR (Budapest)

COORDINATING EDITORS

S. V. EMELYANOV (Moscow)
N. N. KRASOVSKII (Sverdlovsk)
E. P. POPOV (Moscow)
V. S. PUGACHEV (Moscow
V. I. SIFOROV (Moscow)
E. D. TERYAEV (Moscow)
T. VÁMOS (Budapest)
L. VARGA (Budapest)
A. PRÉKOPA (Budapest)
S. CSIBI (Budapest)
I. CSISZÁR (Budapest)
J. KOCSIS (Budapest)
J. BENEŠ (Prague)
V. STREJC (Prague)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)

Г. БОГНАР (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

С. В. ЕМЕЛЬЯНОВ (Москва)
Н. Н. КРАСОВСКИЙ (Свердловск)
Е. П. ПОПОВ (Москва)
В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Е. Д. ТЕРЯЕВ (Москва)
Т. ВАМОШ (Будапешт)
Л. ВАРГА (Будапешт)
А. ПРЕКОПА (Будапешт)
Ш. ЧИБИ (Будапешт)
И. ЧИСАР (Будапешт)
Я. КОЧИШ (Будапешт)
И. БЕНЕШ (Прага)
В. СТРЕЙЦ (Прага)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 10 (4), pp. 219—235 (1981)

ПРИМЕНЕНИЕ МЕТОДОВ ОПТИМИЗАЦИИ К РЕЛЯТИВИСТСКОЙ ТЕОРИИ

В. Г. БОЛТЯНСКИЙ

(Москва)

(Поступила в редакцию 27 июня 1980 г.)

В статье исследуется управляемый объект, возникающий из релятивистских соображений при условии, что математическая модель пространства времени обладает свойствами анизотропии. Описываются световые (т. е. оптимальные по быстродействию) траектории этого объекта и возможные эффекты «гравитационной» фокусировки световых траекторий. В конце статьи даются примеры астрофизической интерпретации этих эффектов, причем эти примеры (кроме первого) связаны с явлениями, труднообъяснимыми в рамках эйпштейновской общей теории относительности.

Принято считать, что свет распространяется (в вакууме) с одной и той же скоростью c во всех направлениях. «Гравитационные» же частицы (обладающие ненулевой массой покоя) могут двигаться лишь со скоростями, м е н ь ш и м и c по величине. Таким образом, движения всевозможных частиц описываются неравенством

$$|\dot{x}| \leq c, \tag{1}$$

где $x = (x_1, x_2, x_3)^T$ — радиус-вектор движущейся частицы (здесь и далее верхний индекс T означает транспонирование, т. е. x является вектор-столбцом). Неравенство (1) можно также записать в виде

$$\dot{x} = u, \qquad |u| \leq c. \tag{2}$$

Если же свет распространяется не в вакууме, а в среде с переменным показателем преломления $n = n(x)$, то соотношения (2) видоизменяются:

$$\dot{x} = \frac{u}{n(x)}, \quad |u| \leq c. \tag{3}$$

Заметим, что одно только соотношение (2) (или (3)) еще не определяет световых траекторий. Ведь какова бы ни была гладкая кривая l, можно представить себе движение $x(t)$, осуществляющееся вдоль этой кривой со скоростью $|\dot{x}(t)| \equiv c$; однако это вовсе не будет означать, что l есть световая траектория. Возможные световые траектории определяются *принципом Ферма*, согласно которому линия l лишь в том случае является световой траекторией, если

каждый ее (достаточно малый) кусок AB является траекторией наибыстрейшего движения из A в B среди всех движений, удовлетворяющих условию (2) (или (3)). Иными словами, если соотношение (2) (или (3)) считать задающим *управляемый объект* (x — фазовая координата, u — управляющий параметр), то *световыми траекториями будут локально оптимальные по быстродействию траектории этого объекта*. Если, например, через $n(x)$ обозначить показатель преломления воздушной среды в точке x, то принцип Ферма, примененный к объекту (3), позволит рассчитать явление рефракции в земной атмосфере.

Сравним (3) с управляемым объектом более общего вида:

$$\dot{x} = f(x,\ u), \quad u \in U. \tag{4}$$

Если через $F(x)$ обозначить множество всех векторов фазовой скорости, которые может объект (4) иметь в точке x:

$$F(x) = \{f(x, u):\ u \in U\},$$

то (4) можно записать в виде дифференциального включения

$$\dot{x} \in F(x). \tag{5}$$

В многочисленных примерах, рассматриваемых в рамках математической теории оптимальных процессов, множества $F(x)$ не являются, как правило, центрально симметричными (а если и являются, то их центр симметрии не совпадает, как правило, с нулевой точкой фазового пространства), что резко отличает общий случай (5) от (2) или (3). Это обстоятельство и побудило автора рассмотреть (вначале — в чисто математическом плане) те следствия, к которым может привести предположение о наличии *световой анизотропии*, т. е. положения вещей, при котором скорости распространения света в двух противоположных направлениях могут быть не одинаковыми по величине.

Как это ни покажется на первый взгляд странным, вопрос о световой анизотропии не может быть решен опытным путем; ответ на этот вопрос зависит лишь от нашего желания. Чтобы пояснить эту мысль, рассмотрим следующий мысленный эксперимент. В двух далеко расположенных (и неподвижных друг относительно друга) космических лабораториях работают два космонавта A и B. Космонавт A производит локацию корабля второго космонавта (т. е. отправляет в его сторону световой сигнал, а затем ловит отраженный сигнал). Оказалось, что время распространения сигнала (туда и обратно) составляет 4 секунды. Заключив отсюда, что световой сигнал шел 2 секунды туда и 2 секунды обратно (и приняв скорость света примерно равной 300 000 км/сек), космонавт A находит, что расстояние между лабораториями составляет 600 000 км.

Затем, установив контакт и научившись понимать друг друга, космонавты намереваются синхронизировать часы. Космонавт A сообщает свою систему единиц времени: секунда — это четверть локационного времени (которое космонавт B может замерить самостоятельно), 60 секунд — минута, 60 минут — час и т. д. После этого космонавт A излагает свой план синхронизации часов: ровно в 12 часов он подает обусловленный сигнал, а космонавт B в момент получения сигнала поставит стрелки на $12^h\ 00'02''$ (т. е. 12 часов 00 минут 02 секунды); тем самым часы будут синхронизированы. Проведенное на следующий день измерение скорости света подтвердило наличие световой изотропии: световой сигнал, поданный космонавтом A в $12^h00'00''$, был принят космонавтом B в $12^h00'02''$, а отраженный сигнал пришел к A в $12^h00'00''$; таким образом, расстояние 600 000 км свет прошел (как туда, так и обратно) за 2 секунды, т. е. скорость света в обоих направлениях составляла 300 000 км/сек.

Представим себе, однако, что космонавт B при осуществлении синхронизации часов по ошибке поставил часы на $12^h00'01''$ (в момент получения обусловленного сигнала от A). Тогда при измерении скорости света окажется, что световой сигнал, поданный космонавтом A в $12^h00'00''$, был принят B в $12^h00'01''$, а отраженный сигнал вернулся к A в $12^h00'04''$. Иначе говоря, идя от A к B, свет прошел 600 000 км за одну секунду, а обратно — за три секунды, т. е. скорость света «туда» составляла 600 000 км/сек, а «обратно» 200 000 км/сек; отношение этих скоростей равно 3. Космонавт B при желании мог бы добиться того, чтобы отношение скоростей света «туда» и «обратно» было равно не 3, а, скажем, 1/3: достаточно в момент получения обусловленного сигнала поставить стрелки на $12^h00'03''$.

Таким образом, наличие или отсутствие световой анизотропии (так же как и величина этой анизотропии, т. е. отношение скоростей света «туда» и «обратно») зависит от способа синхронизации часов, т. е. в конечном счете от нашего желания. Какого-то «правильного» способа синхронизации часов, имеющего предпочтение перед другими способами и позволяющего «объективно» решить вопрос о наличии или отсутствии световой анизотропии, не существует. Ведь синхронизация осуществляется посылкой светового сигнала (скорость которого мы заранее не знаем), а для измерения скорости света нужно в свою очередь иметь каким-то образом синхронизированные часы. Выход из этого порочного круга не может быть найден с помощью какого-либо физического эксперимента: либо в эксперименте участвуют о д н и часы, локализованные в одной точке, и тогда мы замеряем у с р е д н е н н у ю скорость «туда-обратно», либо же имеется д в о е часов, расположенных в разных точках, и неизбежно всплывает проблема их синхронизации (даже если эти часы расположены в одной комнате).

Разумеется, мы можем п о т р е б о в а т ь, чтобы использовались только такие системы отсчета пространства-времени, в которых нет световой анизотропии (чем будет в каждой такой системе отсчета однозначно задан способ синхронизации часов). Однако этот волюнтаризм (общепринятый в теории относительности) вряд ли является оправданным, поскольку, как мы увидим ниже, имеются явления, которые проще и естественнее описываются в системах отсчета, обладающих световой анизотропией.

С этой точки зрения эйнштейновский принцип относительности целесообразно разбить на два отдельных предположения: *гипотезу релятивизма*, утверждающую, что любое отдельно взятое световое движение имеет *одну и ту же* скорость в двух системах отсчета, равномерно движущихся друг относительно друга, и *гипотезу изотропии*, утверждающую, что в одной системе отсчета скорость света во всех направлениях одинакова по величине (при надлежащем выборе масштабов по осям). Из сказанного выше понятно, что мы намереваемся о т к а з а т ь с я от гипотезы изотропии, сохранив лишь первую часть эйнштейновского принципа (т. е. гипотезу релятивизма).

Прежде всего рассмотрим, как в условиях анизотропии видоизменяются преобразования Лоренца. Для простоты рассмотрим лишь одномерный «мир», т. е. движения частиц вдоль некоторой прямой l. Каждое событие (вспышка, встреча двух частиц и т. п.) имеет в заданной на прямой l *системе отсчета* S однозначно определенные *время* и *координату*. Переход от одной системы отсчета S_α к другой системе S_β локально осуществляется линейными формулами:

$$\begin{cases} dt_\beta = A dt_\alpha + B dx_\alpha, \\ dx_\beta = C dt_\alpha + D dx_\alpha. \end{cases} \tag{6}$$

Движение описывается в системе S_α соотношением вида $x_\alpha = \varphi(t_\alpha)$; число $v_\alpha = \dfrac{dx_\alpha}{dt_\alpha}$ называется скоростью этого движения. Скорость этого же движения в системе S_β равна

$$v_\beta = \frac{dx_\beta}{dt_\beta} = \frac{C\,dt_\alpha + D\,dx_\alpha}{A\,dt_\alpha + B\,dx_\alpha} = \frac{C + Dv_\alpha}{A + Bv_\alpha}. \tag{7}$$

Движение, воспринимаемое в системе отсчета S_α как свет, идущий вправо (влево) по прямой l, будет и в системе S_β восприниматься как свет, идущий вправо (влево). При этом, согласно гипотезе релятивизма, скорость света, идущего вправо, о д и н а к о в а в системах S_α, S_β; обозначим эту скорость через σ. Точно так же, скорость света, идущего влево, о д и н а к о в а в

системах S_α, S_β; обозначим ее через ϱ. Таким образом, если $v_\alpha = \sigma$, то и $v_\beta = \sigma$ и аналогично для ϱ), и потому в силу (7) справедливы соотношения

$$\sigma = \frac{C + D\sigma}{A + B\sigma}, \qquad \varrho = \frac{C + D\varrho}{A + B\varrho}.$$

Далее, обозначим через v скорость системы S_α относительно системы S_β. Это означает, что при $v_\alpha = 0$ (частица «покоится» в системе S_α) мы имеем $v_\beta = v$. Таким образом, в силу (7) имеем

$$v = \frac{C}{A}.$$

Освободимся в последних трех соотношениях от знаменателей:

$$\begin{cases} B\sigma^2 - C - D\sigma = -A\sigma, \\ B\varrho^2 - C - D\varrho = -A\varrho, \\ C = Av. \end{cases}$$

Из этой системы линейных уравнений находим

$$B = -\frac{v}{\varrho\sigma}A, \quad C = vA, \quad D = \left(1 - v\left(\frac{1}{\varrho} + \frac{1}{\sigma}\right)\right)A.$$

Теперь для определителя $\Delta = AD - BC$ преобразования (6) имеем:

$$\Delta = A^2\left(1 - \frac{v}{\varrho}\right)\left(1 - \frac{v}{\sigma}\right).$$

Считая, что преобразование (1) является эквиаффинным (сохраняющим площади), т. е. $\Delta = 1$ (физически это равенство соответствует предположению о «совпадении» масштабов в системах S_α и S_β [1]), мы получаем отсюда

$$A = \frac{1}{\sqrt{\left(1 - \frac{v}{\varrho}\right)\left(1 - \frac{v}{\sigma}\right)}}. \tag{8}$$

Таким образом, преобразование (6) принимает следующий вид, где A определяется равенством (8):

$$\begin{cases} dt_\beta = \left(dt_\alpha - \frac{v}{\varrho\sigma}dx_\alpha\right)A, \\ dx_\beta = \left(vdt_\alpha + \left(1 - v\left(\frac{1}{\varrho} + \frac{1}{\sigma}\right)\right)dx_\alpha\right)A. \end{cases} \tag{9}$$

Это — *обобщенные преобразования Лоренца*. При выполнении гипотезы изотропии (т. е. $\sigma = c$, $\varrho = -c$, или, иначе, $\varrho + \sigma = 0$) соотношения (9) превращаются в обычные формулы преобразований Лоренца. Непосредственно проверяется, что квадратичная форма

$$dx^2 - (\varrho + \sigma)\,dx\,dt + \varrho\sigma\,dt^2 \tag{10}$$

инвариантна относительно преобразований (8), (9).

Для дальнейшего удобно ввести обозначения $c = \dfrac{|\varrho| + |\sigma|}{2} = \dfrac{\sigma - \varrho}{2}$, $f = \dfrac{\sigma + \varrho}{2}$. Тогда $\varrho = -c + f$, $\sigma = c + f$, $\varrho\sigma = f^2 - c^2$, и квадратичная форма (10) записывается следующим образом:

$$(dx - f\,dt)^2 - c^2\,dt^2. \tag{11}$$

Квадратичная форма (11) превращается в эйнштейновскую квадратичную форму $dx^2 - c^2dt^2$ при $f = 0$ (т. е. при выполнении гипотезы изотропии). Однако, более интересно то, что форма (11) приводится к эйнштейновской с помощью введения фиктивной координаты x^*, представляющей собой линейную комбинацию пространства и времени: $dx^* = dx - f\,dt$. Отсюда ясно, что формулы (8), (9) могли бы быть получены из преобразований Лоренца заменой $dx^* = dx - f\,dt$. Геометрически это означает, что формулы (8), (9) представляют собой запись преобразований Лоренца в неортогональной системе координат. Это, во-первых, позволяет предложить новый вывод формул (8), (9) — как линейных эквиаффинных преобразований, оставляющих инвариантной квадратичную форму (11). В частности, в трехмерном случае нужно рассмотреть преобразования, оставляющие инвариантной форму

$$(dx_1 - f_1\,dt)^2 + (dx_2 - f_2\,dt)^2 + (dx_3 - f_3\,dt)^2 - c^2\,dt^2.$$

Во-вторых, становится очевидным, что физические соотношения (например, уравнения Максвелла) остаются инвариантными и в анизотропном случае; отметим, что в [2] инвариантность уравнений Максвелла (в анизотропном случае) подтверждена непосредственными вычислениями.

Рассмотрим теперь вопрос о возможной области применения формул анизотропного релятивизма (8), (9). Опыт Майкельсона достаточно убедительно показал, что равномерное движение одной системы отсчета относительно другой не приводит к появлению световой анизотропии. Вполне закономерным является вопрос: не будет ли *ускорение* одной системы отсчета относительно другой (или наличие гравитационного поля, являющегося причиной появления ускорений) приводить к световой анизотропии. Так возникает гипотеза о том, что гравитация может создавать анизотропию. Непосред-

ственная проверка (экспериментальная) этой гипотезы не представляется возможной, так как гравитационные поля и ускорения, с которыми нам приходится иметь дело, ничтожны, а послать лабораторию в окрестность какой-либо черной дыры не представляется возможным. Однако это вовсе не означает, что гипотеза гравитационной анизотропии не может быть обсуждена в теоретическом плане. Ниже проведем такое обсуждение: как мы увидим, следствия, получаемые из принятия этой гипотезы, имеют ряд экспериментальных подтверждений, что можно считать косвенной проверкой гипотезы гравитационной анизотропии.

Простейшая модель, которая в этом плане может быть рассмотрена, заключается в том, что (1) заменяется уравнением

$$\dot{x} = f + u, \quad |u| \leq c, \tag{12}$$

где «вектор смещения» $f = f(t, x)$ зависит от времени t и точки $x = (x_1, x_2, x_3)^T$, причем характер этой зависимости определяется распределением масс и характером движения системы отсчета. Геометрически уравнение (12) можно истолковать следующим образом. Рассматривается управляемый объект (5), где множество $F(x)$ (т. е. множество всех возможных скоростей, с которыми частицы могут выходить из точки x) представляет собой, согласно (12), шар радиуса c, центр которого находится в точке f (рис. 1).

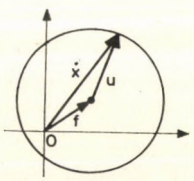


Рис. 1

Если мы по-прежнему будем считать, что (в соответствии с принципом Ферма) световыми траекториями являются локально оптимальные (в смысле быстродействия) траектории управляемого объекта (12), то для отыскания световых траекторий можно будет применить принцип максимума. Для этого нужно составить функцию

$$H = \psi f + \psi u, \tag{13}$$

представляющую собой скалярное произведение вспомогательного вектора $\psi = (\psi_1, \psi_2, \psi_3)$ (он представляет собой вектор-строку) на правую часть уравнения (12), и дифференциальные уравнения для вспомогательного вектора:

$$\dot{\psi} = -\frac{\partial H}{\partial x} = -\psi \frac{\partial f}{\partial x}. \tag{14}$$

Согласно принципу максимума, для оптимальности траектории $x(t)$ необходимо существование такого нетривиального решения уравнения (14), что управление $u(t)$ (соответствующее траектории $x(t)$) осуществляет в каждый момент t максимум функции H по переменному u. Из вида функции H (см. (13)) вытекает (в силу принципа максимума и ограничения $|u| \leq c$), что управление $u = u(t)$ определяется равенством

$$u = \frac{c\psi^T}{|\psi|}. \tag{15}$$

Таким образом, уравнение (12) принимает вид

$$\dot{x} = f + \frac{c\psi^T}{|\psi|}. \tag{16}$$

Уравнения (14), (16) и определяют световые траектории.

Рассмотрим, в частности, случай, когда имеется единственная масса m, неподвижно расположенная в начале координат. Наиболее простое и естественное предположение, которое в этом случае можно сделать, заключается в том, что вектор f направлен к началу координат и имеет величину $\frac{\gamma m}{|x|^2}$, где γ — коэффициент пропорциональности, т. е.

$$f = -\frac{\gamma m}{|x|^3} \cdot x. \tag{17}$$

Таким образом, управляемый объект (12) приобретает вид

$$\dot{x} = -\frac{\gamma m}{|x|^3} \cdot x + u, \quad |u| \leq c. \tag{18}$$

Из (17) видно, что число

$$r_0 = \sqrt{\frac{\gamma m}{c}} \tag{19}$$

обладает следующим свойством: при $|x| < r_0$ имеем $|f| > c$, и потому согласно (19)

$$x^T \dot{x} = x^T(f + u) = -|x| \cdot |f| + x^T u < -|x| \cdot c + |x| \cdot c = 0,$$

т. е. вектор $\dot{x}$ образует с радиусом-вектором x т у п о й угол. Это означает, что все световые траектории, начинающиеся в точках шара $|x| < r_0$, остаются целиком внутри этого шара. Иными словами, ни одна световая траектория, начинающаяся в шаре $|x| \leq r_0$, не выходит из этого шара наружу; условимся этот шар называть *черной дырой*. Формула (19) дает ее радиус.

Далее, несложно проверяется [3], что окружность радиуса $r_0 \cdot \sqrt[4]{3}$, расположенная в любой плоскости, проходящей через начало координат и имеющая центр в начале, является замкнутой световой траекторией. На рис. 2 показаны рассчитанные на ЭВМ (эту работу выполнил Л. К. Николаев [4]) световые траектории, расположенные в плоскости x_1, x_2 и пересекающие ось x_2 под пря-

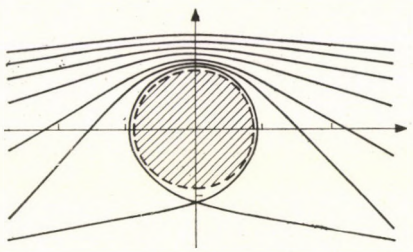


Рис. 2

мым углом. На рис. 3 изображен пучок световых траекторий, исходящих из удаленной точки S; этот рисунок показывает, что, какова бы ни была точка A, найдется световая траектория, исходящая из точки S, огибающая черную дыру O и приходящая в точку A. Рис. 4 (обсуждавшийся в [3], [5]) показывает, что если в точках $S_1, S_2, \ldots, S_k$ расположены «светящиеся объекты», т. е. рассматриваются пучки световых траекторий, исходящие из этих точек, то к на-

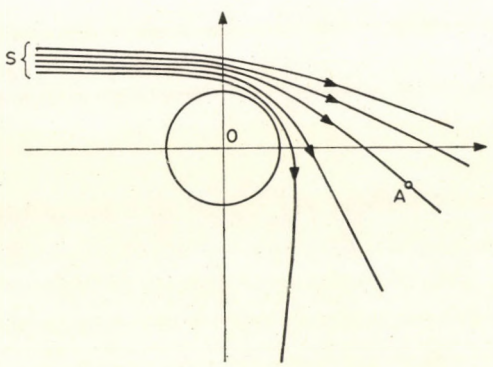


Рис. 3

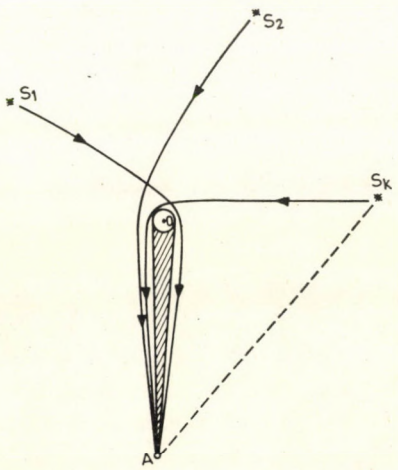


Рис. 4

блюдателю, расположенному в точке A, приходят световые траектории от всех объектов $S_1, S_2, \ldots, S_k$. Черная дыра O как бы фокусирует свет, идущий от всех источников $S_1, S_2, \ldots, S_k$, и направляет его узким пучком (как бы исходящим из окрестности точки O) к наблюдателю A. Внутри этого светового пучка наблюдается маленькое черное пятнышко угловых размеров $2r_0\sqrt[4]{3}/d$, где d — расстояние от наблюдателя A до черной дыры O.

Наконец, на рис. 5 изображено еще одно явление фокусировки (см. [5]), имеющее место, если светящийся объект S, наблюдатель A и черная дыра O расположены на одной прямой. В этом случае от S к A идет бесконечно много

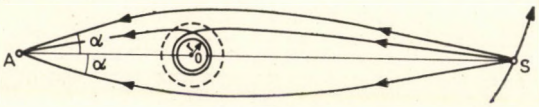


Рис. 5

световых траекторий. В каждой плоскости, проходящей через прямую AS, имеются две «крайние» из таких траекторий (лишь слегка искривляющиеся при прохождении мимо черной дыры O), а также бесконечное число траекторий, совершающих вокруг O различное число витков. В пространстве (ограничиваясь лишь крайними траекториями) наблюдатель A будет видеть объект S в виде светящегося кольца, имеющего угловой радиус α. Если, однако, (вследствие большой удаленности) угол α очень мал, так что наблюдатель не отличает это светящееся кольцо от точки, то A как бы видит объект S сквозь черную дыру O, причем видит его во много раз более ярким, чем в отсутствии черной дыры O (так как площадь кольца во много раз больше, чем площадь пятнышка, которое получилось бы при наблюдении объекта S в прямых лучах). Расчет, приведенный в [6], показывает, что коэффициент увеличения яркости за счет фокусировки черной дырой O (т. е. отношение площади кольца к площади пятнышка) примерно равен $\frac{8R\alpha}{q} \approx \frac{8h}{q}$, где h — расстояние крайней световой траектории от черной дыры O, а q — диаметр «зрачка» телескопа, с помощью которого A проводит свои наблюдения.

Попытаемся теперь применить эти математические соображения к объяснению некоторых астрофизических явлений. Именно, мы рассмотрим те выводы, которые получатся, если свойства пространства—времени описывать не уравнениями Эйнштейна, а поведением траекторий управляемого объекта (18). Иначе говоря, за световые траектории мы примем не изотропные (нулевые) геодезические того искривленного пространства, которое рассматривается в общей теории относительности (для случая одной точечной массы метрический тензор этого пространства дается так называемым решением

Шварцшильда), а локально оптимальные в смысле быстродействия траектории управляемого объекта (18).

Обратимся теперь к рассмотрению некоторых выводов, вытекающих из принятых предположений. При этом будем считать (см. [6]), что коэффициент γ в (18) зависит от длины волны λ электромагнитного излучения:

$$\gamma = q\sqrt{\lambda}. \tag{20}$$

Искривление световых траекторий вблизи Солнца. Наблюдениями во время полных солнечных затмений установлено, что световая траектория, идущая от далекой звезды и проходящая мимо края солнечного диска, немного искривляется. Угол между первоначальным направлением световой траектории и направлением, под которым она подходит к Земле, составляет примерно 1,7 угловой секунды. Это дает возможность вычислить константу γ, т. е. подобрать ее таким образом, чтобы при* $m = 2 \cdot 10^{33}$ (масса Солнца), $r = 7 \cdot 10^{10}$ (радиус Солнца), $c = 2{,}998 \cdot 10^{10}$ (скорость света) угол отклонения световой траектории, проходящей на расстоянии r от начала (ср. рис. 2) составлял для управляемого объекта (18) как раз 1,7″. Расчеты, проведенные Л. К. Николаевым на ЭВМ [4], позволили определить требуемое значение этой константы: $\gamma = 1{,}2 \cdot 10^{-4}$.

Итак, подбор нужного значения γ позволяет полностью согласовать экспериментальные данные об искривлении световых траекторий в поле тяготения Солнца (считающиеся экспериментальным подтверждением эйнштейновской общей теории относительности) с выводами анизотропной теории относительности. В дальнейшем же именно это найденное значение γ послужит объяснению ряда других экспериментальных фактов. Чтобы перейти к рассмотрению других экспериментов, примем, что найденное значение γ относится к длине волны $\lambda = 5{,}5 \cdot 10^{-5}$ (т. е. к «среднему» значению длины волны светового диапазона, простирающегося от $\lambda \approx 3 \cdot 10^{-5}$ до $\lambda \approx 7 \cdot 10^{-5}$). Подставляя это значение λ (и найденное γ) в формулу (20), находим $q = 0{,}016$. Теперь, согласно (19), получаем для радиуса черной дыры следующую формулу:

$$r_0 = \sqrt{\frac{\gamma m}{c}} = \sqrt{\frac{q\sqrt{\lambda}\, m}{c}} = 7{,}3 \cdot 10^{-7} \cdot \sqrt[4]{\lambda}\,\sqrt{m}, \tag{21}$$

а для величины $r_0 \cdot \sqrt[4]{3}$, которую (ср. рис. 4) условимся называть *фокусирующим радиусом* массы m, получаем

$$r_0\sqrt[4]{3} = 9{,}6 \cdot 10^{-7} \cdot \sqrt[4]{\lambda}\,\sqrt{m}. \tag{22}$$

* Если не указаны наименования, все данные приводятся в системе CGS.

Радиоизлучения «спокойного» Солнца. Так как, согласно (21), радиус черной дыры пропорционален $\sqrt[4]{\lambda}$, то одно и то же тело может (в зависимости от его размеров и плотности) быть черной дырой для одних частот и не быть ею для других. В частности, при $m = 2 \cdot 10^{33}$ для радиоизлучения свободного водорода ($\lambda = 21$) получаем из (21), (22): $r_0 = 6{,}99 \cdot 10^{10}$, $r_0 \cdot \sqrt[4]{3} = 9{,}2 \cdot 10^{10}$. Иными словами, *для радиочастот Солнце является черной дырой.* Радиоизлучение, возникающее в солнечной короне вне сферы радиуса $r_0 \cdot \sqrt[4]{3}$, фокусируется этой сферой (рис. 6), благодаря чему эта сфера воспринимается как источник радиоизлучения, более яркий на *крае* сферы, чем внутри нее.

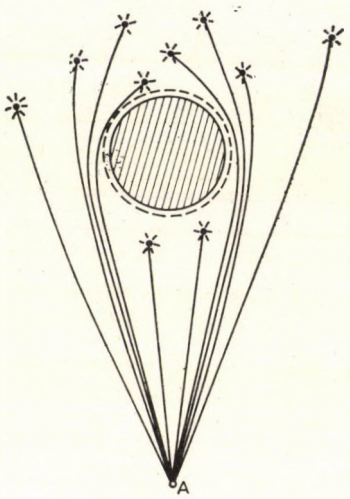


Рис. 6

Радиоизлучение как бы интенсивно льется с воображаемой сферы радиуса $9{,}2 \cdot 10^{10}$ (несколько большего, чем радиус Солнца $7 \cdot 10^{10}$). Сказанное хорошо согласуется с результатами экспериментальных наблюдений: на рис. 7 приведен взятый из книги [7] (стр. 96) график интенсивности излучения (на волне $\lambda = 21$), возникающего на различных расстояниях (за единицу принят радиус Солнца). В метровом диапазоне аналогичное явление наблюдается на сфере несколько большего радиуса (в соответствии с формулой (22)). Для видимого же света (для которого Солнце не является черной дырой) явление фокусировки и связанный с ним эффект повышенного свечения края отсутствует. Подчеркнем, что приведенное объяснение наблюдаемых в радиодиапазоне эффектов характерно для анизотропной теории относительности, тогда как в рамках эйнштейновской общей теории относительности (где размеры черных дыр на много порядков меньше) эти явления труднобъяснимы.

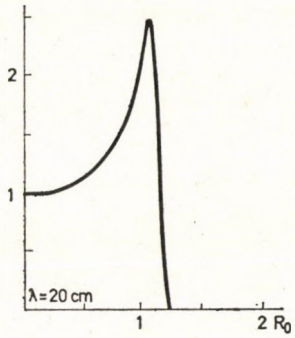


Рис. 7

Радиовспышки Солнца. Предположим, что при перемещении Солнца по небесной сфере его центр в некоторый момент совместится (для земного наблюдателя A) с очень удаленным источником радиоизлучения S. В этот момент, за счет фокусировки, интенсивность приходящего к A радиоизлучения увеличится, как отмечалось выше, примерно в $\frac{8h}{q}$ раз (рис. 5). Считая диаметр «зрачка» (т. е. антенны радиотелескопа равным 100 метров и учитывая, что h больше радиуса Солнца, находим, что коэффициент увеличения интенсивности радиоизлучения $\approx 10^8$. Ввиду быстрого движения Солнца по небесной сфере (за счет перемещения Земли по орбите и собственного движения Солнца) эта «радиовспышка» будет весьма кратковременной. Например, так как скорость движения Земли по орбите составляет $3 \cdot 10^6$, то (отвлекаясь от взаимного перемещения Солнца и удаленного радиоисточника) точка точной фокусировки «пробежит» по антенне радиотелескопа длиной 100 метров за $3{,}3 \cdot 10^{-3}$ секунды. Если же удаленный радиоисточник не совместится точно (для земного наблюдателя) с центром Солнца, а пройдет лишь близко от центра, то явление фокусировки будет более слабым (усиление интенсивности на 4—5 порядков), но и длиться оно будет дольше (десятые доли секунды). Для такого не очень точного совмещения радиоисточника с центром Солнца фокусировка будет производиться не всей сферой радиуса $r_0 \cdot \sqrt[4]{3}$, а лишь небольшой ее частью (наиболее близкой к радиоисточнику, а также наиболее удаленной от него). Ввиду быстрого движения Солнца, этот кажущийся источник радиоизлучения (т. е. точка сферы, ближайшая к заслоненному Солнцем радиоисточнику) будет перемещаться по поверхности Солнца со скоростями порядка сотен и даже тысяч километров в секунду. Эти теоретические выводы на основе анизотропной теории относительности подтверждаются экспериментальными наблюдениями: время от времени фиксируются

«радиовспышки» Солнца, которые длятся доли секунды и производятся «радиоисточниками», перемещающимися над поверхностью Солнца со скоростями 1000 км/сек и более.

Заметим еще, что при $m = 7{,}4 \cdot 10^{25}$ (масса Луны) и $\lambda = 1{,}95 \cdot 10^5$ получаем из (22) $r_0 \cdot \sqrt[4]{3} = 1{,}74 \cdot 10^8$, что совпадает с радиусом Луны. Таким образом, на волнах 1950 и более метров Луна должна обладать фокусирующими свойствами, аналогичными тем, о которых выше говорилось в связи с Солнцем. Установив на спутнике приемник, работающий в двухкилометровом диапазоне и имеющий постоянно направленную на Луну антенну, можно надеяться получить еще одно экспериментальное подтверждение описанного выше явления фокусировки радиоволн.

Вспышки красных карликов. Пусть A — земной наблюдатель, O — красный карлик и S — звезда голубого спектрального класса (рис. 5). Далее, AO примем равным 6 парсеков (расстояние до ближайших красных карликов). Принимая диаметр q «зрачка» телескопа равным 5 метрам и считая h большим, чем 300 000 км (средний радиус красного карлика), получаем для коэффициента усиления интенсивности излучения приближенное значение $\frac{8h}{q} > 4{,}8 \cdot 10^8$. Таким образом, звезда S, которая обычно ненаблюдаема вследствие ее большой удаленности, может вызвать явление внезапной интенсивной «голубой вспышки» красного карлика O. Если красный карлик O движется со скоростью 0,6″ в год (что соответствует средней скорости перемещения звезд этого класса в Галактике), то продолжительность этой «голубой вспышки» составит ≈ 5 сек. Такие вспышки красных карликов в несвойственной для них голубой части спектра, имеющие продолжительность несколько секунд, подтверждены многократными астрофизическими наблюдениями. Заметим, что для возникновения вспышки не обязательно, чтобы A, O и S точно находились на одной прямой; достаточно, чтобы звезда S находилась внутри узкого конуса с осью AO, раствор которого определяется диаметром q «зрачка», расстоянием AO и массой m красного карлика O. Рассчитанная на основе этих соображений вероятность вспышки очень хорошо согласуется с наблюдаемой частотой вспышек.

Радиоизлучения Крабовидной туманности. Крабовидная туманность является наиболее ярким радиоисточником, проходящим в окрестности Солнца. При подходе к Солнцу диск Крабовидной туманности искажается, вытягиваясь почти в 3 раза в направлении Солнца (на волне 7,9 м это происходит на расстоянии около 13 солнечных радиусов, на волне 3,7 м — около 10

радиусов). Это можно было бы объяснить наличием большой электронной концентрации в сверхкороне Солнца. Однако тогда, за счет рефракции должна была бы наблюдаться неравномерность видимого движения Крабовидной туманности (относительно Солнца). Как показывают наблюдения, нарушение равномерности движения ничтожно, откуда, напротив, можно заключить, что электронная концентрация очень невелика. Наконец, наблюдениями установлено, что вскоре после указанного выше искажения диска радиоизлучение Крабовидной туманности вовсе пропадает, вновь появляясь по другую сторону Солнца.

Как показал Л. К. Николаев, все эти особенности поведения крабовидной туманности (труднообъяснимые в рамках обычных физических представлений) хорошо объясняются — не только качественно, но и количественно — в рамках описанных выше представлений анизотропной теории относительности. Именно, управляемый объект (18) заменяется следующим (ср. (3)):

$$\dot{x} = -\frac{\gamma m}{|x|^3} \cdot x + \frac{u}{n(x)}, \quad |u| \leq c, \tag{23}$$

где γ определяется формулой (20) при $q = 0{,}016$, а $n(x)$ — показатель преломления среды в точке x для данной частоты излучения. Теперь с помощью принципа максимума можно находить оптимальные по быстродействию траектории объекта (23), т. е. (согласно принципу Ферма) траектории распространения радиоволн. Первоначальная идея состояла в следующем. При наличии электронной концентрации среда обладает в отношении распространения радиоволн показателем преломления $n < 1$. Поэтому в отсутствии гравитационного воздействия (т. е. при отбрасывании первого слагаемого в правой части (23)) траектории отклоняются от Солнца. Напротив, влияние только одной гравитации (т. е. при $n(x) \equiv 1$, что соответствует распространению волн в вакууме) вызывает искривление траекторий к Солнцу. При определенной электронной концентрации оба фактора могут компенсировать друг друга, в результате чего траектории оказываются малоискривленными (что соответствует равномерному наблюдаемому движению диска Крабовидной туманности).

Количественные расчеты на ЭВМ, осуществленные Л. К. Николаевым, были основаны на использовании формул для электронной концентрации (а через нее — для показателя преломления $n(x)$), взятых из [8]. Рис. 8 дает представление о результатах проведенного математического эксперимента. На нем изображены траектории, отходящие от Крабовидной туманности (из «бесконечности»). Поверхность, зачерчиваемая этими траекториями (вначале «параллельными») как бы образует складку, т. е. огибающую семейства траек-

торий. При движении Земли по орбите вокруг Солнца (пунктир на рис. 8) происходит «пропадание» Крабовидной туманности в течение времени, пока Земля пробегает дугу BC (куда радиоизлучения не проникают). В моменты же прохождения складок (т. е. точек B, C) за счет пересечения «бесконечно близких» траекторий в точках огибающей возникает легко объяснимое вытягивание диска Крабовидной туманности. Точные расчеты дали хорошее совпадение с результатами экспериментов. В частности, «пропадание» должно по расчетам происходить примерно на расстоянии 13 радиусов от Солнца на волне $\lambda = 7{,}9$ м и на расстоянии 10 солнечных радиусов на волне $\lambda = 3{,}7$ м.

Отметим еще, не входя в подробности, что в рамках анизотропной теории относительности удовлетворительное объяснение получают также и многие другие явления: эффект Мессбауэра, замедление светового сигнала при прохождении его мимо Солнца, гравитационное искривление пространства и смещение перигелия Меркурия, различие величины гравитационного красного смещения в центре солнечного диска и на его крае и др.

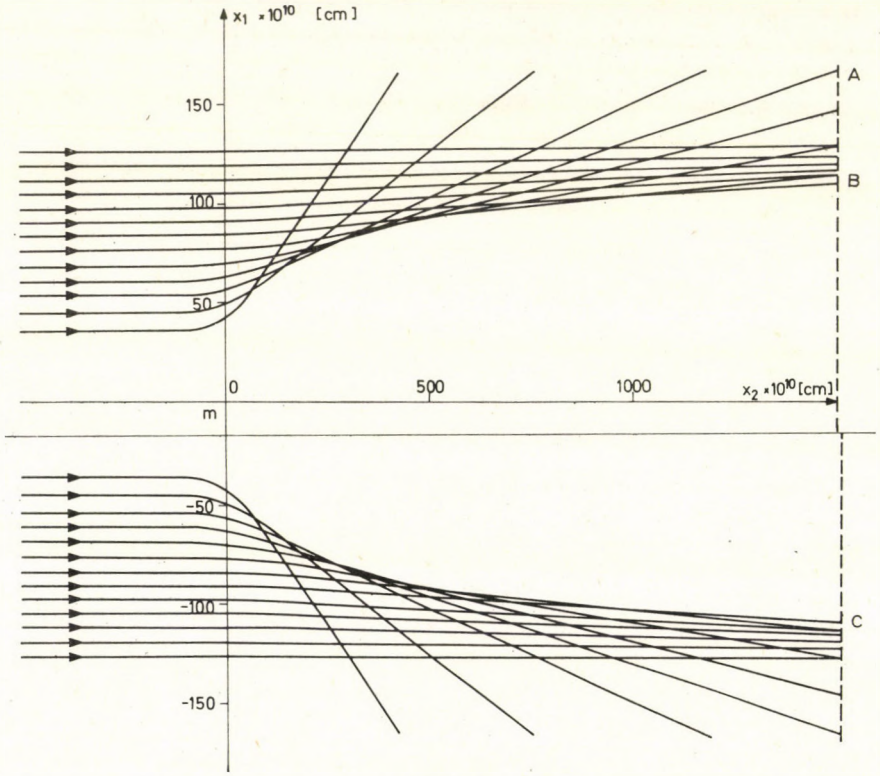


Рис. 8

Литература

1. *Болтянский В. Г.* Анизотропный релятивизм. Дифференциальные уравнения, **1** (№ *12*, 1974.
2. *Стрельцов В. Н.* Об инвариантности уравнений Максвелла. Сообщения ОИЯИ, Р2–7328. Дубна, 1973.
3. *Болтянский В. Г.* Принцип максимума, черные дыры и квазары. Дифференциальные уравнения **15,** № *10*, 1979.
4. *Болтянский В. Г. Николаев Л. К.* Машинный эксперимент по расчету световых траекторий. Дифференциальные уравнения, **16,** № *3*, 1980.
5. *Болтянский В. Г.* Анизотропная теория относительности и оптимизация. Дифференциальные уравнения, **15,** № *11*, 1979.
6. *Болтянский В. Г.* Является ли Солнце радиолинзой. Дифференциальные уравнения, **16,** № *11*, 1980.
7. Радиоастрономия. Парижский симпозиум, 1958 г. ИЛ, 1961.
8. *Аллен К. У.* Астрофизические величины. «Мир», 1977.

Methods of optimization applied to the relativistic theory

V. G. BOLTYANSKY

(Moscow)

In the article a control object is studied, connected with the relativistic theory.

A measuring of the light-velocity (if the light goes from a point A to a point B) is possible only in the case, when we have two synchronized clocks, situated in the points A, B. But the synchronization of the two clocks, situated in A, B, is possible only with the help of a light signal (under the condition, that we know the light-velocity). If we (with Einstein) allow to consider only such systems of reference, for which the extension of light is isotropic (i.e. the light-velocity in a direction is numerically equal to the light-velocity in the opposite direction), then we obtain a way to synchronize the clocks. But it is impossible to determine experimentally, if there exists a light-unisotropy in an inertial system.

From this point of view it is interesting to construct a relativistic theory in unisotropic systems. In the article a sketch of such theory is given. The formulas of time-space transformations for unisotropic systems are written. These transformations form an generalised Lorentz-group.

The global picture of light-trajectories is (in virtue of the Fermat-principle) the picture of the time-optimal trajectories of the control object

$$\dot{} = f + u, \quad |u| \leq c, \tag{1}$$

where $x = (x_1, x_2, x_3)$ etc. The "deviation-vector" $f = f(x)$ determines a light-unisotropy. This vector is defined by the mass-distribution. In the article some time-optimal trajectories of the control object (1) are described. In particular, in the case when the vector field $f(x)$ is spherically symmetrical and $|f| = \frac{\gamma m}{|x|^2}$, circle trajectories are found, curvature of the other trajectories is investigated and several focal properties of these trajectories are discussed.

The obtained description of the time-optimal trajectories of the control object (1) allows to give some astrophysical interpretations. Some of such interpretations are in the end of the article indicated. In particular, it is possible to explain the radioemission of the "quiet" Sun, the firings of the small red stars, the anomalies of the radioemission of star clusters.

В. Г. Болтянский
Всесоюзный НИИ системных исследований
СССР, Москва, ул. Рылеева, 29

Problems of Control and Information Theory, Vol. 10 (4), pp. 237–251 (1981)

USE OF COORDINATE-PARAMETRIC FEEDBACK IN DESIGN OF CONTROL SYSTEMS

S. V. YEMEL'YANOV, S. K. KOROVIN, V. I. SIZIKOV

(Moscow)

(Received October 4, 1980)

The paper is concerned with the design of closed-loop dynamic systems whose properties do not change much with an arbitrary wide range of variations of the process parameters. Parametric invariance is obtained by using coordinate-parametric feedback with finite gains and without resort of sliding modes. For various cases, relations are given which lead to systems that feature the desired properties.

1. Statement of the problem

It is often desirable to have a control system whose properties would not change much in face of process parameter variations. There is extensive literature on methods to solve this problem [2–7]. The most effective of these rely on searching and non-searching adaptation [3, 4, 6] and sliding modes [5, 7]. The existing adaptive approaches proceed from the assumption that the process parameters vary in a quasistationary way, an assumption which is not always valid in real life. In variable structure systems (VSS) [5, 7] the system motion is independent of the parameters only in a sliding mode which is in many cases difficult to obtain, especially if it is to feature some desired properties.

Consequently, methods are desired which would help solve this problem with arbitrary variations of the process parameters in a wide range without sliding modes and with finite gains in the main control loop. This paper is devoted to the description and justification of one such method.

The method embodies the ideas of [1] and employs the use of two kinds of feedback, namely coordinate and coordinate-parametric. The former is obtained by conventional methods of control engineering and the latter is intended to handle the variations of the former in order to make the properties of the closed-loop system independent of the process variables.

In formal terms the problem is stated as follows. The free process motion is described as

$$\begin{cases} \dot{x}_i = x_{i+1}, \quad i = 1, \ldots, n-1 \\ \dot{x}_n = -\sum_{i=1}^{n} a_i x_i + u \end{cases} \tag{1.1}$$

where the process parameters a_i $(i = 1, \ldots, n)$ may vary in an arbitrary way and take on any values form the ranges

$$a_i^{\min} \leq a_i(t) \leq a_i^{\max}; \quad i = 1, \ldots, n. \tag{1.2}$$

In (1.2) $a_i^{\min}$ and $a_i^{\max}$ are constants of known values. It is desired to design a control which would insure absolute asymptotic stability of solutions to (1.1) and constancy of their properties in face of any variations of $a_i(t)$ $(i = 1, \ldots, n)$ within (1.2).

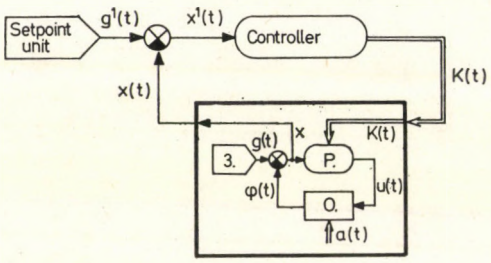


Fig. 1

The solution is sought in the framework of the structural diagram of Fig. 1, where the control u is assumed to be chosen in the form

$$u = \sum_{i=1}^{m+1} k_i x_i, \quad 1 \leq m \leq n - 2 \tag{1.3}$$

where the parameters k_i $(i = 1, \ldots, m + 1)$ can take on only finite values from within the range

$$-k_i^0 \leq k_i \leq k_i^0; \quad i = 1, \ldots, m + 1. \tag{1.4}$$

In (1.4) k_i^0 are constant and positive. With the ranges of $a_i(t)$ in (1.2) wide enough the quality of solutions to the closed-loop system (1.1)–(1.4) does not satisfy the above requirements with any fixed k_i from (1.4). It is desired to change the parameters of the law (1.3) so as to reduce the dependence of the solution on a_i. This change will be seen to be provided by coordinate-parametric feedback loops.

The basic features of the method are explained for the case of constant process parameters; then the findings are extended to variable parameters.

2. Equations of system motion

In the design of a coordinate-parametric loop the desired variations of the controlled variable are specified by the output of the loop setpoint unit (Fig. 1). With constraints (1.2) and (1.4) the rates of variation of $x(t)$ stay within certain limits. There are no constraints on the rate at which the parameters k_i in (1.3) vary. This indicates the essential possibility of making the rates of the coordinate and coordinate-parametric loops very different and permits, above all, the error x^1 to vanish. Since that time the coordinate $x(t)$ varies in compliance with the setpoint which is $g^1(t)$.

For instance, let the output from the setpoint unit of the coordinate-parametric loop be given by the expression

$$g^1(t) = \sum_{i=1}^{n} c_i x_i + x_1, \; c_n \triangleq 1 \tag{2.1}$$

where fixed parameters c_i, characterize the requirements for the dynamic properties of system (1.1), (1.3). Then the error x^1 is

$$x^1 = g^1(t) - x_1(t) = \sum_{i=1}^{n} c_i x_i \,. \tag{2.2}$$

Assume that the controller of the coordinate-parametric loop insures, starting from a certain time, the equality since that time the behaviour of system (1.1), (1.3) is described by the equation

$$\begin{cases} \dot{x}_i = x_{i+1}, \quad i = 1, \ldots, n-2 \\ \dot{x}_{n-1} = -\sum\limits_{i=1}^{n-1} c_i x_i \end{cases} \tag{2.3}$$

which depends only on the parameters c_i of the setpoint unit.

In this research the controller of the coordinate-parametric loop is chosen in the form

$$\dot{\varkappa} = \begin{cases} -\alpha_i \operatorname{sign} x^1 & \text{with} \quad \varkappa_i \in [-k_i^0, k_i^0] \\ 0 & \text{with} \quad \varkappa_i \notin [-k_i^0, k_i^0] \end{cases} \tag{2.4}$$

$$k_i = \varkappa_i \operatorname{sign} x_i, \quad i = 1, \ldots, m,$$

where α_i are loop gains.

The main effect of control (2.4) is that with finite α_i the equality $x^1(t) = 0$ cannot be maintained in strict terms, but this relation is valid in the limit as $\alpha_i \to \infty$. As for finite α_i, in the vicinity of $x^1 = 0$, oscillations arise whose amplitude reduces against α_i. What is important is that the presence of such oscil-

lations with their amplitude small enough does not have any significant effect on the shape of motions in system (1.1), (1.3) which is close to that in system (2.3). In particular, if system (2.3) is asymptotically stable, so is system (1.1), (1.3), (2.4). From these remarks it follows that parametric design of the system is completed if the values of laws (1.3) and (2.4) are specified such that the motions in the vicinity of $x^1 = 0$ stay within permissible limits and the describing point, with the initial conditions being arbitrary, is bound to reach that vicinity. The analysis which is to follow will result in appropriate estimates.

Let us introduce some formal notation and definitions. System (2.3) will be referred to as limit. In the neighborhood of the manifold $x^1 = 0$, let us determine the region $G(x)$ given by the relations

$$G(x) = \{x : S^+ \cdot S^- < 0\} \tag{2.5}$$

where the surfaces $S^+ = 0$ and $S^- = 0$ are boundaries for the region $G(x)$ described by the equations

$$\begin{cases} S^+(x) = \sum\limits_{i=1}^{m} c_i^+ x_i + \sum\limits_{i=m+1}^{n} c_i x_i \\ S^-(x) = \sum\limits_{i=1}^{m} c_i^- x_i + \sum\limits_{i=m+1}^{n} c_i x_i . \end{cases} \tag{2.6}$$

In (2.6) the parameters c_i^+ and c_i^- are obtained in the form

$$\begin{cases} c_i^+ = c_i + \Delta c_i \operatorname{sign} x_i \\ c_i^- = c_i - \Delta c_i \operatorname{sign} x_i, \quad i = 1, \ldots, m \end{cases} \tag{2.7}$$

where the positive constants Δc_i ($i = 1, \ldots m$) characterize the "size" of the region $G(x)$. In particular, with $\Delta c_i = 0$ ($i = 1, \ldots, m$) the surfaces $S^+ = 0$ and $S^- = 0$ coincide with the hyperplane $x^1 = 0$. The region $G(x)$ defines the vicinity in which all motions meet the desired requirements.

Let us consider in $G(x)$ a certain surface $S(t)$ defined by the equation

$$S(t) = l(t)\, S^+ + [1 - l(t)]\, S^- = 0, \quad 0 \leq l \leq 1 \tag{2.8}$$

in which the parameter $l(t)$ is chosen so that in all solutions to system (1.1), (1.3) in the region $G(x)$ it is true that $S(t) = 0$. Outside the region $G(x)$ the parameter $l(t)$ takes on the limit value which follows from the relations

$$l = \begin{cases} 0 & \text{with} \quad S^+S^- > 0 \ \& \ x^1 < 0 \\ 1 & \text{with} \quad S^+S^- > 0 \ \& \ x^1 > 0. \end{cases} \tag{2.9}$$

Relations (2.8) and (2.9) lead to a parameter $l(t)$ at each time and then its value is used for generating a coefficient k_{m+1} which plays an auxiliary role

in the law (1.3). We are concerned with the case where

$$k_{m+1} = -[c_n^- - a_i - (c_{n-1} - a_n)\, c_{m+1} + 2l \cdot \Delta c_m \operatorname{sign} x_m. \tag{2.10}$$

The parameter k_{m+1} in the law (1.3) and its representation in the form (2.10) are needed to keep the describing point in the region. In many cases k_{m+1} will be seen to be unnecessary. Equations (1.1) and (1.3)–(2.10) determine between them the behaviour and they should be used in analyzing the system properties when $x(t) \in G(x)$.

Now equations are needed to analyze the system behaviour outside $G(x)$. They are necessary to establish the conditions under which the describing point $x(t)$ from initial conditions outside $G(x)$ is bound to reach that region.

Equations for motion outside $G(x)$ are easily obtained from formula (1.1), (1.3), and (2.4) because, with any $\varkappa_i(0)$ and α_i large enough, the parameters k_i within a period of about $\frac{k_i^0}{\alpha_i}$ take on their limit values $\pm k_i^0$ $(i = 1, \ldots, m)$ and, consequently, the system behaviour outside $G(x)$ is described by the equation

$$\begin{cases} \dot{x}_i = x_{i+1}, \quad i = 1, \ldots, n-1 \\ \dot{x}_n = -\sum\limits_{i=1}^{n} a_i x_i + \sum\limits_{i=1}^{m} k_i^0 |x_i| \operatorname{sign} x^1 + k_{m+1} x_{m+1} \end{cases} \tag{2.11}$$

where k_{m+1} is given by equation (2.10) in which $l(t)$, computed from (2.8) should be substituted. Equation (2.10) describes a certain VSS [5] for which it is rather easy to find whether the describing point reaches $x^1 = 0$. Since $x^1 = (0)$ belongs to $G(x)$, to reach the region $G(x)$ is equivalent to reaching $x^1 = 0$ and this problem can be solved by using the VSS theory methods [7].

3. Basic results for the case of constant parameters

Let us first formulate the results for the case of constant parameters of the process. Assuming that a_i $(i = 1, \ldots, n)$ are constant and known, the following contention holds.

Theorem 1. If for system (1.1), (1.3)–(2.10) at certain time t_1 $x^1(t_1) = 0$, then with the relations

$$\frac{c_{i-1} - a_i}{c_i} = c_{n-1} - a_n, \quad i = m+2, \ldots, n-1; \tag{3.1}$$

$$k_i^0 > \max_{j=\pm} |c_{i-1}^j - a_i - (c_{n-1} - a_n)\, c_i^j| \quad i = 1, \ldots, m; \tag{3.2}$$

$$\alpha_i > \max\,[4\omega + k_i^0;\ -4\omega - k^0], \quad i = 1, \ldots, m \tag{3.3}$$

where ω^+ *and* ω^- *have the form*

$$\begin{cases} \omega^+ = \max\limits_{i,j=\pm} \dfrac{k_i^0 + |c_{i-1}^j - a_i - (c_{n-1} - a_n)\, c_i^j|}{2\Delta c_i} \\ \omega^- = \min\limits_{i,j=\pm} \dfrac{-k_i^0 - |c_{i-1}^j - a_i - (c_{n-1} - a_n)\, c_i^j|}{2\Delta c_i} \end{cases} \tag{3.4}$$

for all $t > t_1$ *the system paths belong to the region or* $x(t) \in G(x)$. The proof is given in Appendix 1. This theorem establishes the relations with which solution to the system which start in $G(x)$ stay within that region in subsequent motion. Note that, in compliance with (3.1)–(3.3) with the parameters c_i and Δc_i specified, the choice of α_i and k_i^0 can prove relatively simple. The explicit form of motion equations in $G(x)$ is rather complex for analysis and so let us transform them into an implicit but equivalent form. By adding up equations (2.6) with regard to (2.7), the following equation can be obtained which holds for all

$$x^1(t) = (1 - 2l(t)) \sum_{i=1}^{m} \Delta c_i\, |x_i|. \tag{3.5}$$

In motion in $G(x)$ the parameter $l(t)$ varies in the interval $(0, 1)$ and so during these motions the equation holds

$$|x^1(t)| < \sum_{i=1}^{m} \Delta c_i\, |x_i(t)|\,. \tag{3.6}$$

Replace in this system the variable x_n with x^1 consistently with the expression $x_n = x^1 - \sum\limits_{i=1}^{n-1} c_i x_i$. In new coordinates $(x_1, \ldots, x_{n-1}, x')$ the system is described by the equations

$$\begin{cases} \dot{x}_i = x_{i+1}, \quad i = 1, \ldots, n-2 \\ \dot{x}_{n-1} = -\sum\limits_{i=1}^{n-1} c_i x_i + x^1(t)\,. \end{cases} \tag{3.7}$$

In (3.7) there is no expression for $\dot{x}^1$ because this is not significant for further analysis and the only important thing is that for all $t > t_1$, (3.6) holds. Furthermore, in analyzing the system stability, investigation of equation (3.7) alone would do because with the condition $x_i(t) \to 0$ true as $t \to \infty$ for $i = 1, \ldots, n-1$, it follows from (2.17) that $x^1(t) \to 0$ and, consequently, $x_n(t) \to 0$.

Since (2.3) is limit for this system, it would be natural to require asymptotic stability from (2.3). According to [5] this is true if and only if all the roots of the equation

$$p^n + \sum_{i=m+2}^{n} a_i p^{i-1} + \sum_{i=1}^{m+1} (c_{i-1} + c_i p_0) p^{i-1} = 0 \tag{3.8}$$

except the obvious root $p = -p_0 = c_{n-1} - a_n$ have negative real parts. If with some m this condition is not valid, the number m should be increased; this would entail an increase in the number of derivatives used in (1.3). Let this condition hold with a certain m, or the system (2.3) be asymptotically stable, and its stability margin be known to be equal to $\delta_0 > 0$. With these assumptions the sufficient conditions for asymptotic stability of the system (3.6) and (3.7) and, consequently, of the system (1.1), (1.3)–(2.11) in the region $G(x)$ can be defined.

Theorem 2. If the stability margin of the system (2.3) is $-\delta_0$ then the truth of the inequality

$$\max_i \Delta c_i < \frac{\delta_0}{c_0} \tag{3.9}$$

is sufficient for asymptotic stability of the system (3.6), (3.7).

In (3.9) c_0 is a certain number related to the parameters c $(i = 1, \ldots, n-1)$.

The proof follows from the Bellman lemma.

Theorem 2 specifies the boundaries of the region $G(x)$ in which any motions are asymptotically stable if system (2.3) is so.

As for reaching the region $G(x)$ from arbitrary initial conditions, this is found by methods of the VSS theory for the system (2.11). The necessary and sufficient conditions are given in Theorem 2 of [8] but they are too complex to be ckecked and so easily verifiable sufficient conditions are of interest. In the VSS theory many such conditions have been obtained; for instance, the inequality

$$c_{n-1} - a^n < 0 \tag{3.10}$$

insures reaching the region $G(x)$. The inequality (3.10) cannot always be met; if it is not, then, according to [5], the reaching can be insured by increasing the parameter k_1^0 in (1.3). An accurate estimate of the parameter k_1^0 which insures the reaching is given in Theorem 3.

Theorem 3. In system (2.11) with $m = 1$, *in the case of truth of relations (2.12), and of the inequality*

$$k_1^0 \geq \max_{i, a_i<0} [|a_i| n]^{\frac{n}{n-i+1}} \tag{3.11}$$

the describing point is certain to reach $x^1 = 0$.

The Proof is given in Appendix 2.

Consequently, if for system (1.1), (1.3), (1.4) the conditions of Theorems 1, 2 and 3 hold, then the system is asymptotically stable on the whole. Taken together, the relations of the Theorems and inequality (3.10) permit designing a system of desired properties; for instance, this can be done in the following sequence: by choosing the parameters c_i, the dynamic properties of the limit system (2.3) are determined; then by specifying Δc_i and with due regard for (3.9), the permissible deflections of the system to be designed from the limit system are specified; k_i^0 is found through (3.2), (3.10) or (3.11); finally, α_i are chosen through (3.3) and (3.4). Note that with reduction of Δc_i to zero the parameters α_i grow to infinity, the magnitude of $|x^1(t)|$ tends to zero and this implies that the properties of the system approach those of system (2.3). What is important is that this effect is obtained without resort to sliding modes and with finite coefficients in the channels through which the process is influenced.

Now we treat some particular cases which are not covered by the above Theorems. Assume first that in law (1.3) $m = n - 1$*. In that case there is no need for using signals of x_n to keep the paths in the region and one can assume that $k_{m+1} = k_n = 0$. The relations for the design of such systems change somewhat in particular, the conditions (3.1) do not hold, there is no need for computing the parameter $l(t)$ and there is a wide lattitude in choosing the parameters c_i of the limit system. Then, in place of Theorem 1, we have

Theorem 4. If for system (1.1), (1.3)–(2.10) with $m = n - 1$ *and* $k_n \equiv 0$ *at time* t_1 $x^1(t_1) = 0$, *then, with the conditions*

$$k^0 > \max_{j=\pm} |c_{i-1}^j - a_i - (c_{n-1}^j - a_n) c_i^j|, \quad i = 1, \ldots, n-1 \tag{3.12}$$

$$\alpha_i > \max [4\omega^+ k_i^0 - 4\omega^- k_i^0], \quad i = 1, \ldots, n-1 \tag{3.13}$$

* Design of this kind may be needed, in particular, for insuring asymptotic stability of (2.3) or because the process parameters $a_i(t)$ are variable.

valid where ω^+ *and* ω^- *are given by the relations*

$$\begin{cases} \omega^+ = \max\limits_{i,j=\pm} \dfrac{k_i^0 + |c_{i-1}^j - a_i - (c_{n-1}^j - a_n)\, c_i^j|}{2\Delta c_i} \\ \omega^- = \min\limits_{i,j=\pm} \dfrac{-k_i^0 - |c_{i-1}^j - a_i - (c_{n-1}^j - a_n)\, c_i^j|}{2\Delta c_i} \end{cases} \tag{3.14}$$

for all $t \geq t_1$ $x(t) \in G(x)$.

The conditions for asymptotic stability on the whole for this system are defined by Theorem 2 and some sufficient condition for the describing point to reach the region $G(x)$.

The potential is still greater with $m = n$. For such systems all properties of systems with $m = n - 1$ hold and, furthermore, there is no need to bother with reaching the region $G(x)$. The following result is true:

Theorem 5. System (1.1), (1.3)–(2.10), with asymptotic stability of system (2.3) and truth of inequality (3.9), is asymptotically stable on the whole if the following relations hold

$$k_i^0 > \max_{j=\pm} |c_{i-1}^j - a_i|, \qquad i = 1, \ldots, n \tag{3.15}$$

$$\alpha_i > \max\,[4\omega^+ k_i^0 - 4\omega^- k_i^0], \qquad i = 1, \ldots, n \tag{3.16}$$

where the values of ω^+ *and* ω^- *are given by the expressions*

$$\begin{cases} \omega^+ = \max\limits_{1 \leq i \leq n-1, j=\pm} \dfrac{k_i^0 + |c_{i-1}^j - a_i|}{2\Delta c_i} \\ \omega^- = \min\limits_{1 \leq i \leq n, j=\pm} \dfrac{-k_i^0 - |c_{i-1}^j - a_i|}{2\Delta c_i} \end{cases} . \tag{3.17}$$

Theorems 4 and 5 can be proved with a reasoning as in the proof of Theorem 1.

4. Finding for the case of variable process parameters

Let us extend our findings to the case of time variable process parameters a_i $(i = 1, \ldots, n)$ which satisfy the constraints (1.2). With the variable $a_i(t)$, relations (3.1) cannot be equalities and in the control law (1.3), a signal of x_{n-1} should be used at least; another option is a signal of x_n. The procedure of design with variable at $a_i(t)$ $(i = 1, \ldots, n)$ is maintained in that by choosing the parameters k_i^0 and α_i the region $G(x)$ is bound to be reached and the subsequent motion will occur there. Both with $m = n - 1$ and $m = n$ there is no need for special generation of the parameter k_{m+1} in consistence with (2.10).

Now let us give formulations which contain the relations needed in design and note that there are no requirements for the nature of variations in $a_i(t)$. The proofs are analogous with those of Theorem 1.

Theorem 6. If for the system (1.1)–(2.10) with $m = n - 1$ *and* $k_n \equiv 0$ *at a certain time* t_1 $x^1(t_1) = 0$ *then the relations*

$$k_i^0 > \max_{a_j, j=\pm} |c_{i-1}^j - a_i - (c_{n-1}^j - a_n)\, c_i^j| \qquad i = 1, \ldots, n-1 \tag{4.1}$$

$$\alpha_i > \max\,[4\omega^+ k_i^0, \quad -4\omega^- k_i^0], \qquad i = 1, \ldots, n-1 \tag{4.2}$$

are true, where the magnitudes of ω^+ *and* ω^- *are given by the relations*

$$\begin{cases} \omega^+ = \max\limits_{a_i, i, j=\pm} \dfrac{k_i^0 + |c_{i-1}^j - a_i - (c_{n-1}^j - a_n)\, c_i^j|}{2\Delta c_i} \\ \omega^- = \min\limits_{a_i, i, j=\pm} \dfrac{-k_i^0 - |c_{i-1}^j - a_i - (c_{n-1}^j - a_n)\, c_i^j}{2\Delta s_i}. \end{cases} \tag{4.3}$$

for all $t > t_1$ $x_1(t) \in G(x)$.

Absolute stability on the whole for the system is obtained if the conditions of Theorem 2 and the inequality hold [13]

$$\sup_{t \geq 0} [c_{n-1} - a_n(t)] < 0. \tag{4.4}$$

For the other case where $m = n$ there is no need for conditions for reaching the region $G(x)$; the conditions for stability on the whole are given in

Theorem 7. System (1.1)–(2.10) with $m = n$, *asymptotic stability of system (2.3) and truth of the inequality (3.9) is asymptotically stable on the whole if the following relations hold*

$$k_i^0 > \max_{a_i, j=\pm} |c_{i-1}^j - a_i|, \qquad i = 1, \ldots, n \tag{4.5}$$

$$\alpha_i > \max\,[4\omega^+ k_i^0; \quad -4\omega^- k_i^0], \qquad i = 1, \ldots, n \tag{4.6}$$

where the magnitudes of ω^+ *and* ω^- *are given by the relations*

$$\begin{cases} \omega^+ = \max\limits_{a_i, 1 \leq i \leq n, j=\pm} \dfrac{k_i^0 + |c_{i-1}^j - a_i|}{2\Delta c_i} \\ \omega^- = \min\limits_{a_i, 1 \leq i \leq n, j=\pm} \dfrac{-k_i^0 - |c_{i-1}^j - a_i|}{2\Delta c_i}. \end{cases} \tag{4.7}$$

In relations (4.1)–(4.7), unlike (3.12)–(3.17), the operations max and min are computed additionally from the parameters a_i within constrains (1.2).

Appendix 1

To prove Theorem 1 it is sufficient to show that if its conditions hold, the paths in boundary points of the region $G(x)$ are directed into the region. According to (2.8), the boundary $s^+ = 0$ is associated with a value of l equal to 1 and the boundary $s^- = 0$, equal to 0. Consequently, the Theorem is true if for all $t > t_1$ the function $l(t)$ belongs to an interval $[0, 1]$ and for its derivative computed through motion equations, the following inequalities hold

$$\dot{l}_{l=1} \leq 0, \qquad \dot{l}_{l=0} \geq 0. \tag{A.1}$$

To verify (A.1), let us compute the derivative $\dot{l}$ using equations (1.1), (1.3)–(2.10) by using relation (2.8) for the purpose; we have

$$\dot{l} = -\frac{\dot{s} - l(\dot{s}^+ - \dot{s}^-)}{s^+ - s^-}. \tag{A.2}$$

From (2.6) we have $s^+ - s^- = \sum_{i=1}^{m} 2\Delta c\,|x_i|$ and, consequently,

$$\dot{s}^+ - \dot{s}^- = \sum_{i=1}^{m} 2\Delta c_{i-1}\, x_i \operatorname{sign} x_{i-1} + 2\Delta c_m\, x_{m+1} \operatorname{sign} x_{m1},\ \Delta c_0 \triangleq 0. \tag{A.3}$$

To compute the derivative of $s^-(t)$, let us make use of the second equation in (2.6); then

$$s^- = \sum_{i=1}^{m} c_i^-\, x_{i+1} + \sum_{i=m+1}^{n-1} c_i\, x_{i+1} + \dot{x}_n.$$

Substitute this into the expression for $\dot{x}_n$ with regard to (1.1) and (1.3); exclude from the resultant equation the coordinate x_n by using the relation (2.8)

$$ls^+ + (1 - l)\, s^- = 0$$

which is always true with $x \in G(x)$; then, performing collection of terms, we have ($c_0 \triangleq 0$)

$$\dot{s}^- = \sum_{i=1}^{m} [c_{i-1}^- - a_i - (c_{n-1} - a_n)\,[c_i^+ l + (1 - l)\, c_i^- \cdot k_i \operatorname{sign} x_i]\, x_i +$$

$$+ \sum_{i=m+2}^{n-1} [c_{i-1} - a_i - (c_{n-1} - a_n)]\, x_i + [c_m^+ - a_{m+1} - (c_{n-1} - a_n)\, c_{m+1} + k_{m+1}]\, x_{m+1}.$$

This expression is transformed by virtue of equation (2.10) for k_{m+1} and equalities of (3.1); ultimately, for $\dot{s}^-$ we have an equation

$$\dot{s}^- = \sum_{i=1}^{m} [\{c_{i-1}^- - a_i - (c_{n-1} - a_n)\,[c_i^+ \cdot l + (1 - l)\, c_i^-]\} \operatorname{sign} x_i + k_i]\,|x_i| -$$

$$- 2l\Delta c_m \operatorname{sign} x_m \cdot x_{m+1}. \tag{A.4}$$

Substitute now (A.3) and (A.4) into (A.2); performing elementary transformations we have an expression for $\dot{l}$

$$\dot{l} = \frac{\sum_{i=1}^{m} [\{c_{i-1}^- - a_i - (c_{n-1} - a_n)[c_i^+ l + (1 - l)\, c_i^-] + 2l\Delta c_{i-1} \operatorname{sign} x_{i-1}\} \operatorname{sign} x_i + k_i]\,|x_i|}{\sum_{i=1}^{m} 2\Delta c_i\,|x_i|}. \tag{A.5}$$

Compute the value of $\dot{l}$ in the boundary points of the region $G(x)$; in other words with $l = 0$ and $l = 1$ we have the following respective expressions

$$\text{(A.6)} \qquad \dot{l}|_{l=0} = -\frac{\sum_{i=1}^{m} [\{c_{i-1}^{-} - a_i - (c_{n-1} - a_n)\, c_i^{-}\} \operatorname{sign} x_i + k_i]|x_i|}{\sum_{i=1}^{m} 2\Delta c_i |x_i|}$$

$$\text{(A.7)} \qquad \dot{l}|_{l=1} = \frac{\sum_{i=1}^{m} [\{c_{i-1}^{-} - a_i - (c_{n-1} - a_n)\, c_i^{+} + 2\Delta c_{i-1} \operatorname{sign} x_{i-1}\} \operatorname{sign} x_i + k_i]|x_i|}{\sum_{i=1}^{m} 2\Delta c_i |x_i|}.$$

Assuming that in (A.6) $k_i = -k_i^0$ and in (A.7) $k_i = k_i^0$ for all $i = 1, \ldots, m$ and, furthermore, the equalities hold

$$\text{(A.8)} \qquad \begin{cases} k_i^0 > |c_{i-1}^{-} - a_i - (c_{n-1} - a_n)\, c_i^{-}| \\ k_i^0 > |c_{i-1}^{+} - a_i - (c_{n-1} - a_n)\, c_i^{+}|, \quad i = 1, \ldots, m \end{cases}$$

which are identical with the inequalities (3.2), then relations (A.1) are found to be true. Therefore to complete the proof one should see that the parameters $k_i(t)$ $(i = 1, \ldots, m)$ which change by virtue of law (2.4), and with the conditions of the Theorem being true, always take on their extreme values when the system paths approach the boundaries of $G(x)$.

To prove this fact, we will need a minorant and a majorant estimate of the rate of change of $l(t)$ in (A.5). For this purpose we will use easily verifiable inequalities

$$\text{(A.9)} \qquad \min_i \frac{\gamma_i}{\beta_i} \leq \frac{\sum_{i=1}^{m} \gamma_i |x_i|}{\sum_{i=1}^{m} \beta_i |x_i|} \leq \max_i \frac{\gamma_i}{\beta_i}; \quad \beta_i > 0, \quad i = 1, \ldots, m.$$

Let ω denote the relation of appropriate coefficients in (A.6); and making use of (A.9), we have

$$\text{(A.10)} \qquad \min_i \omega_i(t) \leq \dot{l}(t) \leq \max_i \omega_i(t).$$

The dependence of ω_i on time handicaps the use of estimates (A.10) and so let us further estimate the right- and left-hand sides of (A.10). Denoting by ω_i^{+} the minimum and by ω_i^{-} the maximum of $\omega_i(t)$ computed for all motion variables in the expression for $\omega_i(t)$. It is easily seen that

$$\text{(A.11)} \qquad \begin{cases} \omega_i^{+} = \max\limits_{j=\pm} \dfrac{k_i^0 + |c_{i-1}^{j} - a_i - (c_{n-1} - a_n)\, c_i^{j}|}{2\Delta c_i} \\ \omega_i^{-} = \min\limits_{j=\pm} \dfrac{-k_i^0 - |c_{i-1}^{j} - a_i - (c_{n-1} - a_n)\, c_i^{j}|}{2\Delta c_i}. \end{cases}$$

By virtue of (A.8) it is seen from (A.11) that $\omega_i^{+} > 0$ and $\omega^{-} < 0$ for all $i = 1, \ldots, n$ Now let $\omega^{+} = \max\limits_i \omega^{+}$, $\omega^{-} = \min \omega^{-}$; then for $\dot{l}$ we have finally in the region $G(x)$ the following estimates

$$\text{(A.12)} \qquad \omega^{-} \leq \dot{l}(t) \leq \omega^{+}$$

where $\omega^{+} > 0$ and $\omega^{-} < 0$.

Now let us take up again the variations of $k_i(t)$ under the initial condition $x^1(t_1) = 0$. The reference time is $t_1 = 0$; then by virtue of the initial condition $l(0) = \frac{1}{2}$ and the

estimates (A.12) the function varies within the limits

$$\frac{1}{2} + \omega^- t \leq l(t) \leq \frac{1}{2} + \omega^+ t. \tag{A.13}$$

Assume that under the above initial conditions the parameter $l(t)$ has gone beyond the segment [0, 1]; in particular, at time T the following inequality holds

$$\dot{l}(T)|_{l(T)=0} < 0. \tag{A.14}$$

For time T from (A.13) we have an estimate

$$T \geq \frac{1}{2\omega^-}. \tag{A.15}$$

In this motion the parameter k_i which varies in compliance with (2.4) from the initial condition $k_i(0)$ takes at time T values that satisfy the inequality

$$k_i(T) \leq k_i(0) + \frac{\alpha_i}{2\omega^-}, \quad i = 1, \ldots, m. \tag{A.16}$$

If α_i satisfies the inequalities $\alpha_i > -4k_i^0\,\omega^-$, $i = 1, \ldots, m$ which is identical with (3.3), then we have from (A.16)

$$k_i(T) < -k_i^0, \quad i = 1, \ldots, m. \tag{A.17}$$

Consequently, by virtue of (A.6) and (A.8) the inequality $\dot{l}(T)_{l(T)=0} \geq 0$ holds at time T which is inconsistent with the assumption (A.14). Assuming now that at a certain time the path leaves the region $G(x)$ with $l = 1$ and using the same reasoning the condition $\alpha_i > 4k_i^0\,\omega^+$ can be found to be inconsistent with this assumption. Consequently, if $\alpha_i > \max(4k_i^0\omega^+; -4k_i^0\omega^-)$ and (A.8) hold for all $t \geq t_1$ the paths of the system (1.1), (1.3)–(2.10) belong to the region.

Appendix 2

Truth of Theorem 3 follows from the reasoning of [5, p. 142] and a remark of [7, p. 163] whereby the plane $x^1 = 0$ is reached with stable motion by increasing the parameter k_1^0 for both constant and variable parameters of the process. These papers assert only the existence of $\tilde{k}_1^0$ such that for all $k_1^0 \geq \tilde{k}_1^0$ the region is reached. We have obtained an estimate of the quantity $\tilde{k}_1^0$ by using the findings of [8] that lack of nonnegative real roots of the characteristic equations of system (15) which with $m = 1$ has the form

$$\lambda^n + \sum_{i=1}^{n} a_i \lambda^{i-1} = 0 \tag{A.25}$$

is the necessary and sufficient condition for reaching. In (A.25) all a_i, $i \neq 1,2$ coincide with the values in (1.1) and a_2 allows for the increment caused by the presence of a coefficient k_{m+1} in the form (2.10) in (2.11) and $a_1 = a_1^* + k_1^0$ where a_1^* coincide with the associated a in (1.1).

In the differential equation associated with (A.25) the variables can be replaced in compliance with the formulae [7]

$$\tau = t/\varrho, \; y_i = \varrho^{i-1} x_i, \; \varrho = (k_1^0)^{-1/n}, \; i = 1, \ldots, n.$$

Then instead of (A.25) we have an equivalent characteristic equation of the form

$$f(\lambda) = \lambda^n + a_n \varrho \lambda^{n-1} + a_{n-1} \varrho^2 \lambda^{n-2} + \ldots a_1 \varrho^n + 1 = 0. \tag{A.26}$$

For (A.26) let us find the value of ϱ_0 under which it has no nonnegative real roots. For simplicity let us consider the case of constant a_i.

Represent $f(\lambda)$ in the form

$$f(\lambda) = \sum_{i=1}^{n} f_i(\lambda) \tag{A.27}$$

where

$$f_i(\lambda) = \frac{1}{n}\lambda^n + a_i \lambda^{i-1} \varrho^{n-i+1} + \frac{1}{n}. \tag{A.28}$$

Note now that if $a_i > 0$, then the function $f_i(\lambda)$ is positive with all $\lambda \geq 0$; this implies that (A.27) does not vanish with $\lambda \geq 0$ or has no roots on that ray.

Assume that with some $i = k$, $a_k \leq 0$ and consider the associated function $f_k(\lambda)$. It is positive if the inequality holds

$$\varrho < \frac{1}{[|a_k|\, n]^{\frac{1}{n-k+1}}}. \tag{A.29}$$

This follows from an inequality which results from (A.28) and (A.29)

$$f_k(\lambda) > \frac{1}{n}(\lambda^n - \lambda^k + 1)$$

whose right-hand side is positive on the ray $\lambda \geq 0$, which is obvious if the following inequality holds

$$(1 - \lambda^k) + \lambda^n > 0 \qquad \text{with} \qquad \lambda \in [0, 1]$$

$$\lambda^n - \lambda^k + 1 > 0 \qquad \text{with} \qquad \lambda > 1.$$

Consequently, if

$$\varrho \leq \frac{1}{\max [|a_k|\, n]^{\frac{1}{n-k+1}}} \tag{A.30}$$

where the max operation is performed for all k which are associated with negative a_k, then each addend in (A.27) is positive with $\lambda \geq 0$ and, consequently, $f(\lambda) = 0$ has no roots on that ray. By virtue of the relation between ϱ and k_1^0, we finally obtain an estimate (3.10)

$$\hat{k}_1^0 \geq \max_{k,\, a_k < 0} [|a_k|\, n]^{\frac{n}{n-k+}}.$$

In conclusion of the proof the resultant estimates are also applicable to the case of $m > 1$.

References

1. *Yemel'yanov, S. V., Korovin,* Development of Feedback Types and Their Application in Design of Closed-loop Dynamic Systems. This issue.
2. *Meerov, M. V.,* Design of High Precision Control Systems, 2nd ed., Nauka, Moscow, 1967, 423 pp. (in Russian)
3. Modern control systems theory. Edited by *C. T. Leonders.* McGrow-Hill Book Company, Inc. New York, 1966.
4. *Petrov, B. N., Rutkovskii, V. Yu., Krutova, I. N., Zemlyakov, S. D.,* Design of Principles of Self-Adjusting Control Systems. Mashinostroyenie. Moscow. 1972, 260 pp. (in Russian)
5. Theory of Variab e Structure Systems. Ed. by S. V. Yemel'yanov, Nauka, Moscow 1970, 592 pp. (in Russian)

6. *Krasovskii, A. A.*, Dynamics of Continuous Self-Adjusting Systems. Fizmatgiz, Moscow, 1963 (in Russian)
7. *Utkin, V. I.*, Sliding Modes and Their Application in Variable Structure Systems. Mir Publishers, Moscow, 1978.
8. *Bezvodinskaya, T. A., Sabaev, Ye. F.*, Conditions for Stability on the Whole of Variable Structure Systems. Automation and Remote Control Vol. **35**, No. *10*.

Применение координатно-параметрической обратной связи при синтезе систем автоматического управления

С. В. ЕМЕЛЬЯНОВ, С. К. КОРОВИН, В. И. СИЗИКОВ

(Москва)

Описывается и обосновывается метод синтеза замкнутых систем автоматического управления, основанный на использовании двух контуров регулирования: координатного и координатно-параметрического. Первый контур регулирования организуется в соответствии с обычными методами теории управления, а второй — по предлагаемым правилам, гарантирующим малую зависимость динамических и статических свойств системы управления по отношению к существенному изменению параметров объекта в широких диапазонах.

Отмечается, что этот эффект достигается при конечных коэффициентах усиления в контуре, замкнутом через объект управления, и без использования в нем скользящих режимов.

Приводятся соотношения, позволяющие осуществить синтез системы управления при различных предположениях о свойствах объекта управления. Установлено достаточное условие асимптотической устойчивости в целом синтезированных систем.

С. В. Емельянов
Всесоюзный НИИ системных исследований,
СССР, Москва, ул. Рылеева, 29

Problems of Control and Information Theory, Vol. 10 (4), pp. 253–266 (1981)

BOUNDS OF THE ASYMPTOTIC RATE OF CONVERGENCE OF THE ERROR PROBABILITIES OF A ROBUST TEST*

J. Á. VÍŠEK

(Prague)

(Received January 10, 1980)

Bounds of the asymptotic rate of convergence of the error probabilities of a robust version of the likelihood ratio test are found. At first this is done for the case when the true distribution is assumed to be one of the pair of the least favourable ones, then under the hypothesis that data have not been contaminated. Mutual consistence of the results are proved, too.

1. Introduction

In order to avoid the misleading influence of outliers, models of contamination were proposed (Example 1 illustrates the mentioned bad influence of the outliers). When minimax tests for these models had been found ([2], [3], [4], [5]) (asymptotic) properties of them have been studied ([6], [7]). In the present paper bounds of the rate of convergence of the error probabilities for a minimax test constructed by Khalfina and Khalfin are given. The bound derived for the case, when the true distribution is one from the least favourable pair of distributions, guarantees, that the rate of convergence of the error probabilities of the robust version of the test is usually only a little worse than the rate of the nonrobust one (Theorem 1, Corollary 1). On the other hand, having been sure that data were contaminated we should use the robust test, even in the case when the mentioned bound would admit a rather great deterioration. However, there are situations when the presence of the outliers is doubtful. Then the following interesting question arises: What is the price we must pay when we use the robust version of the test (in order to insure ourselves against the outliers) but the data have not been contaminated and therefore there exists a better likelihood ratio test. An answer is given in Theorem 2 in a form of the bound of the rate of convergence of the error pro-

* Paper presented at the Third Czechoslovak–Soviet–Hungarian Seminar on Information Theory held at Liblice, June 23–27, 1980.

AMS subject classification: 62F35, 62F03.

Key words and phrases: Asymptotic behaviour of robust tests.

babilities of robust test under the hypothesis that data have not been contaminated. To have any sense, this bound must not be greater than that in Theorem 1, because the bound in Theorem 1 was obtained for the least favourable pair of distributions. This consistency is proved in Lemma 1.

2. Notations

Let $\mathfrak{N}$ be the set of all positive integers, R the real line, $\emptyset$ — the empty set.

Let $(\mathfrak{X}, \mathfrak{A})$ be a measurable space, P_0 and P_1 probability measures defined on it. Let us denote by p_0 and p_1 the densities of P_0 and P_1 with respect to a measure μ, respectively. Let $H_\alpha(p_0, p_1)$ denote the α-entropy of the measures P_0 and P_1, i.e.

$$H_\alpha(p_0, p_1) = \int p_0^\alpha p_1^{1-\alpha} \, d\mu .$$

Finally, let $H(p_0, p_1)$ be the minimal α-entropy of the measures P_0 and P_1, i.e.

$$H(p_0, p_1) = \inf_{0<\alpha<1} H_\alpha(p_0, p_1).$$

For any $A \subset \mathfrak{X}$ let $A^c = \mathfrak{X} - A$ and let $\mathfrak{M}$ be the set of all probability measures on $(\mathfrak{X}, \mathfrak{A})$.

Now let us define two families of probability measures as follows:

$$\mathfrak{S}_i = \{Q; Q \in \mathfrak{M}, Q = (1 - \varepsilon_i) P_i + \varepsilon_i F, F \in \mathfrak{M}\} \quad (i = 0, 1)$$

where ε_i are fixed numbers, $0 \leq \varepsilon_i < 1$, $\varepsilon_0 \neq \varepsilon_1$. Let us suppose that ε_i $(i = 0, 1)$ are taken in such a way that the families $\mathfrak{S}_0$ and $\mathfrak{S}_1$ are disjoint. Let us put

$$r(x) = (1 - \varepsilon_1) p_1(x) - (1 - \varepsilon_0) p_0(x)$$

and assume hereafter that $\varepsilon_0 > \varepsilon_1$. (In the opposite case interchange the indexes 0 and 1.) Let, moreover, $X_1 = \{x \in \mathfrak{X}, r(x) > 0\} \neq \emptyset$.

Further, for any $c \in R$, let $A(c) = \left\{x \in \mathfrak{X}; \frac{p_1(x)}{p_0(x)} > c\right\}$ and find $\mathfrak{C} \in R$, $\mathfrak{C} \geq \frac{1 - \varepsilon_0}{1 - \varepsilon_1}$ so that

$$1 = \frac{1}{\varepsilon_0 - \varepsilon_1} \int_{A(\mathfrak{C})} r(x) \, d\mu(x) . \tag{1}$$

(The existence of such $\mathfrak{C}$ is guaranteed by the assumption $X_1 \neq \emptyset$.)

Then the minimax test of the hypothesis $H : Q \in \mathfrak{S}_0$ against the alternative $A : Q \in \mathfrak{S}_1$ is based on the likelihood ratio $\frac{q_0(x)}{q_1(x)}$, where q_0 and q_1 are

defined as follows (see [4]):

$$q_0(x) = \begin{cases} (1 - \varepsilon_0)\, p_0(x) + \dfrac{\varepsilon_0}{\varepsilon_0 - \varepsilon_1} r(x) & \text{for} \quad x \in A(\mathcal{C}), \\ (1 - \varepsilon_0)\, p_0(x) & \text{otherwise} \end{cases}$$

and

$$q_1(x) = \begin{cases} (1 - \varepsilon_1)\, p_1(x) + \dfrac{\varepsilon_1}{\varepsilon_0 - \varepsilon_1} r(x) & \text{for} \quad x \in A(\mathcal{C}), \\ (1 - \varepsilon_1)\, p_1(x) & \text{otherwise.} \end{cases}$$

Remark 1. A straightforward computation gives $q_0(x) = q_1(x)$ for $x \in A(\mathcal{C})$.

Let $n \in \mathfrak{N}$. Put $\mathfrak{X}_n = \overset{n}{\underset{i=1}{\times}} \mathfrak{X}^{(i)}$, where $\mathfrak{X}^{(i)} = \mathfrak{X}$ for all $i = 1, 2, \ldots, n$. Further for f, g being densities and $\gamma \in R$, denote

$$\mathfrak{K}_{n,\gamma}(f, g) = \{x \in \mathfrak{X}_n;\ \prod_{i=1}^{n} f(x_i) \leq \exp\{n\gamma\} \prod_{i=1}^{n} g(x_i)\}.$$

When γ is equal to zero, we use a simpler form $\mathfrak{K}_n(f, g)$ instead of $\mathfrak{K}_{n,0}(f, g)$. Finally, define for

$$e_n(f, A) = \int_A f d\mu$$

and

$$e_n(f, g, A) = e_n(f, A) + e_n(g, A^c).$$

3. Results

Let us start with the promised example.

Example 1. Let p_0 and p_1 be densities of the uniform distribution over $\left[\frac{1}{4}, \frac{5}{4}\right]$ and $[0, 1]$, respectively. Let $\varepsilon_0 = \frac{11}{32}$ and $\varepsilon_1 = \frac{1}{8}$. (Let us consider $\mathfrak{X} = \left[0, \frac{5}{4}\right]$). Then

$$r(x) = \begin{cases} \dfrac{7}{8} & \text{for} \quad x \in \left[0, \dfrac{1}{4}\right], \\ \dfrac{7}{32} & \text{for} \quad x \in \left[\dfrac{1}{4}, 1\right], \\ -\dfrac{21}{32} & \text{for} \quad x \in \left[1, \dfrac{5}{4}\right] \end{cases}$$

and

$$\int_0^{\frac{1}{4}} r(x)\,dx = \frac{7}{8}\cdot\frac{1}{4} = \frac{7}{32} = \frac{11}{32} - \frac{1}{8} = \varepsilon_0 - \varepsilon_1 .$$

So, let us take as $\mathcal{C}$ any real number from the interval $(1, \infty)$. Then, $\left(\text{assuming } \frac{1}{0} = \infty\right)$ $A\,(\mathcal{C}) = \left[0, \frac{1}{4}\right)$. (If you want to avoid the use of the convention $\frac{1}{0} = \infty$, you can construct Example 1 with interchanged indexes and use the results of Khalfina and Khalfin, which they have formulated for $\varepsilon_1 > \varepsilon_0$). Further,

$$q_0(x) = \begin{cases} \frac{11}{8} & \text{for } x \in \left[0, \frac{1}{4}\right), \\ \frac{21}{32} & \text{for } x \in \left[\frac{1}{4}, \frac{5}{4}\right] \end{cases}$$

and

$$q_1(x) = \begin{cases} \frac{11}{8} & \text{for } x \in \left[0, \frac{1}{4}\right), \\ \frac{7}{8} & \text{for } x \in \left[\frac{1}{4}, 1\right), \\ 0 & \text{for } x \in \left(1, \frac{5}{4}\right]. \end{cases}$$

Let us test in the i. i. d. model the hypothesis $H : q = q_0$ against the alternative $A : q = q_1$ by means of the test

$$\varphi_n^* = I_{\{\mathfrak{K}_n(p_0, p_1)\}}, \tag{2}$$

where $I_{\{\mathfrak{K}(p_0, p_1)\}}$ denotes the indicator of the critical region $\mathfrak{K}_{n(p_0, p_1)}$. Now we are going to evaluate a probability of type I, i.e. the quantity

$$\int_{\mathfrak{K}_n(p_0, p_1)} \prod_{i=1}^{n} q_0(x_i)\,dx_i .$$

Let $\mathfrak{T}_n = \left\{x \in \mathfrak{X}_n;\ x_i \in \left[\frac{1}{4}, \frac{5}{4}\right] \text{ for all } i = 1, 2, \ldots, n\right\}$ and $x^0 \in \mathfrak{X}_n - \mathfrak{T}_n$. Then there is $i_0 \in \{1, \ldots, n\}$ such that $x_{i_0}^0 < \frac{1}{4}$. This means that

$$\prod_{i=1}^{n} p_0(x_i^0) = 0 \leq \prod_{i=1}^{n} p_1(x_i^0),$$

i.e. $x^0 \in \mathcal{H}_n(p_0, p_1)$. So we have obtained $\mathcal{X}_n - \mathcal{S}_n \subset \mathcal{H}_n(p_0, p_1)$. On the other hand, we have

$$\int_{\mathcal{X}_n} \prod_{i=1}^{n} q_0(x_i)\, dx_i = 1,$$

$$\int_{\mathcal{S}_n} \prod_{i=1}^{n} q_0(x_i)\, dx_i = \left(\frac{21}{32}\right)^n.$$

From this we conclude that

$$\int_{\mathcal{H}_n(p_0, p_1)} \prod_{i=1}^{n} q_0(x_i)\, dx_i \geq \int_{\mathcal{X}_n - \mathcal{S}_n} \prod_{i=1}^{n} q_0(x_i)\, dx_i = 1 - \left(\frac{21}{32}\right)^n,$$

converging to 1 when $n \to \infty$.

Theorem 1. Let $B \in \mathfrak{A}$ be a set such that

$$\int_B r(x)\, d\mu(x) \geq \varepsilon_0 - \varepsilon_1. \tag{3}$$

Then

$$H(q_0, q_1) \leq (1 - \varepsilon_1)\, H(p_0, p_1) + \varepsilon_0 + (1 + \varepsilon_0) \int_B p_0(x)\, d\mu(x).$$

Proof. Let $\alpha \in (0, 1)$ be such a real number for which

$$H(p_0, p_1) = H_\alpha(p_0, p_1).$$

(If $H(p_0, p_1) = \lim_{\alpha \to 0+} H_\alpha(p_0, p_1)$ or $H(p_0, p_1) = \lim_{\alpha - 1-} H_\alpha(p_0, p_1)$, then the proof must be modified by inserting limits.) Then

$$H(q_0, q_1) \leq \int q_0^\alpha(x)\, q_1^{1-\alpha}(x)\, d\mu(x) = (1 - \varepsilon_0)^\alpha\, (1 - \varepsilon_1)^{1-\alpha} \int_{A^c(\mathcal{C})} p_0^\alpha(x)\, p_1^{1-\alpha}(x)\, d\mu(x) +$$

$$+ \int_{A(\mathcal{C})} q_0(x)\, d\mu(x) \leq (1 - \varepsilon_1)\, H(p_0, p_1) + (1 - \varepsilon_0) \int_{A(\mathcal{C})} p_0(x)\, d\mu(x) + \tag{4}$$

$$+ \frac{\varepsilon_0}{\varepsilon_0 - \varepsilon_1} \int_{A(\mathcal{C})} r(x)\, d\mu(x) = (1 - \varepsilon_1)\, H(p_0, p_1) + \varepsilon_0 + (1 - \varepsilon_0) \int_{A(\mathcal{C})} p_0(x)\, d\mu(x).$$

Let at first $\mathcal{C} > \dfrac{1 - \varepsilon_0}{1 - \varepsilon_1}$. From (3) we have

$$\int_B r(x)\, d\mu(x) \geq \int_{A(\mathcal{C})} r(x)\, d\mu(x) \tag{5}$$

because of

$$\int_{A(\mathcal{C})} r(x)\,d\mu(x) = \varepsilon_0 - \varepsilon_1.$$

Subtracting $\int_{B\cap A(\mathcal{C})} r(x)\,d\mu(x)$ from both sides of (5) we obtain

$$\int_{B-A(\mathcal{C})} r(x)\,d\mu(x) \geq \int_{A(\mathcal{C})-B} r(x)\,d\mu(x). \tag{6}$$

However, for $x \in B - A(\mathcal{C})$ we have $p_1(x)\,|\,p_0(x) \leq \mathcal{C}$ and for $x \in A(\mathcal{C}) - B$, on the contrary, $p_1(x)\,|\,p_0(x) > \mathcal{C}$. Assuming it, we may derive from (6)

$$\left(\mathcal{C}\frac{1-\varepsilon_0}{1-\varepsilon_1} - 1\right)(1-\varepsilon_0) \int_{B-A(\mathcal{C})} p_0(x)\,d\mu(x) \geq \tag{7}$$

$$\geq \left(\mathcal{C}\frac{1-\varepsilon_0}{1-\varepsilon_1} - 1\right)(1-\varepsilon_0) \int_{A(\mathcal{C})-B} p_0(x)\,d\mu(x).$$

We have assumed $\mathcal{C} > \dfrac{1-\varepsilon_0}{1-\varepsilon_1}$ and therefore from (7) we obtain

$$(1-\varepsilon_0) \int_{B-A(\mathcal{C})} p_0(x)\,d\mu(x) \geq (1-\varepsilon_0) \int_{A(\mathcal{C})-B} p_0(x)\,d\mu(x)$$

and finally

$$(1-\varepsilon_0) \int_{B} p_0(x)\,d\mu(x) \geq (1-\varepsilon_0) \int_{A(\mathcal{C})} p_0(x)\,d\mu(x). \tag{8}$$

Combination of (4) and (8) concludes the proof for $\mathcal{C} > \dfrac{1-\varepsilon_0}{1-\varepsilon_1}$. For the case $\mathcal{C} = \dfrac{1-\varepsilon_0}{1-\varepsilon_1}$ the only set B fulfilling (3) is $A(\mathcal{C})$, because of nonpositivity of $r(x)$ on $A^c(\mathcal{C})$. Q.E.D.

Corollary 1. Let f be a density of a probability measure defined on $(R, \mathfrak{B})$, where $\mathfrak{B}$ is the Borel σ-algebra. Let us denote by $\tilde{q}_\alpha$ the α-quantile of the density f and let for two positive numbers ε_0 and ε_1, $1 > \varepsilon_0 > \varepsilon_1$, a real number Δ is taken so that

$$\tilde{q}_{\varepsilon_1} \leq \tilde{q}_{\varepsilon_0} + \Delta. \tag{9}$$

Further let us suppose that there exists $K \in R$ so that $f(x-\Delta)\,|\,f(x)$ is a non-decreasing function on the set $\{x \in R, x \geq K\}$. Finally let us put $p_0(x) = f(x)$ and $p_1(x) = f(x-\Delta)$ and assume that the inequality

$$\int_{A\left(\frac{f(K-\Delta)}{f(K)}\right)} r(x)\,dx \geq \varepsilon_0 - \varepsilon_1 \tag{10}$$

holds. Then

$$H(q_0, q_1) \leq (1 - \varepsilon_1)\, H(p_0, p_1) + \varepsilon_0 + \varepsilon_1.$$

Proof. From the assumption that $\dfrac{f(x - \Delta)}{f(x)}$ is a nondecreasing function on $\{x \in R, x \geq K\}$ follows that for any $c \in R$, $c \geq \dfrac{f(K - \Delta)}{f(K)}$ the set $A(c)$ has a form of an interval (a, ∞) or $[a, \infty)$. Owing to (10) we can derive that

$$A(\mathcal{C}) = (a_1, \infty)$$

or

$$A(\mathcal{C}) = [a_1, \infty)$$

and $a_1 \geq K$.

Now let us compute $\dfrac{1}{\varepsilon_0 - \varepsilon_1} \int\limits_{\tilde{q}_{\varepsilon_1}}^{\infty} r(x)\, dx$

$$\frac{1}{\varepsilon_0 - \varepsilon_1} \int\limits_{\tilde{q}_{\varepsilon_1}}^{\infty} r(x)\, dx = \frac{1}{\varepsilon_0 - \varepsilon_1} \int\limits_{\tilde{q}_{\varepsilon_1}}^{\infty} \{(1 - \varepsilon_1)\, p_1(x) - (1 - \varepsilon_0)\, p_0(x)\}\, dx =$$

$$= \frac{1}{\varepsilon_0 - \varepsilon_1} \int\limits_{\tilde{q}_{\varepsilon_1}}^{\infty} \{(1 - \varepsilon_1)\, f(x - \Delta) - (1 - \varepsilon_0)\, f(x)\}\, dx =$$

$$= \frac{1}{\varepsilon_0 - \varepsilon_1} \left\{(1 - \varepsilon_1) \int\limits_{\tilde{q}_{\varepsilon_1}}^{\infty} f(y)\, dy - (1 - \varepsilon_0)\, \varepsilon_1\right\} \geq$$

$$\geq \frac{1}{\varepsilon_0 - \varepsilon_1} \left\{(1 - \varepsilon_2) \int\limits_{\tilde{q}_{\varepsilon_0}} f(y)\, dy - (1 - \varepsilon_0)\, \varepsilon_1\right\} =$$

$$= \frac{1}{\varepsilon_0 - \varepsilon_1} \{(1 - \varepsilon_1)\, \varepsilon_0 - (1 - \varepsilon_0)\, \varepsilon_1\} = 1,$$

where assumption (9) was used. From this we conclude that $a_1 \geq \tilde{q}_{\varepsilon_1}$. But this inequality yields that

$$\int\limits_{A(\mathcal{C})} p_0\, dx \leq \int\limits_{\tilde{q}_{\varepsilon_1}}^{\infty} p_0\, dx = \varepsilon_1. \tag{11}$$

Finally, put $B = A(\mathcal{C})$. Now the assertion of the Corollary follows from (1), (11) and Theorem 1.

Q.E.D.

Theorem 2. It holds that

$$\limsup_{n\to\infty} \frac{1}{n} \log e_n(p_0, p_1, \mathfrak{K}_n(q_0, q_1)) \leq$$

$$\leq \max \left\{ \log H(p_0, p_1) + \log (1 - \varepsilon_1) - \log (1 - \varepsilon_0), \log \left[H(p_0, p_1) + \int_{A(\mathcal{C})} p_1 d\mu \right] \right\}.$$

Proof. Let us recall that we consider only the case $\varepsilon_0 > \varepsilon_1$. From Lemma 1A of [8] we can write

$$\limsup_{n\to\infty} \frac{1}{n} \log e_n(p_0, p_1, \mathfrak{K}_n(q_0, q_1)) = \tag{12}$$

$$= \max \left\{ \limsup_{n\to\infty} \frac{1}{n} \log e_n(p_0, \mathfrak{K}_n(q_0, q_1)), \limsup_{n\to\infty} \frac{1}{n} \log e_n(p_1, \mathfrak{K}_n^c(q_0, q_1)) \right\}.$$

Let us find a bound for $\limsup_{n\to\infty} \frac{1}{n} \log e_n(p_0, \mathfrak{K}_n(q_0, q_1))$.

Denote $\mathbf{N}_n = \{1, 2, \ldots, n\}$ and $B(x) = \{i \in \mathbf{N}_n, q_0(x_i) = q_1(x_i)\}$ and let $B^c(x) = \mathbf{N}_n - B(x)$. Let $x^0 \in \mathfrak{K}_n(q_0, q_1)$. Then

$$\prod_{i \notin B(x^0)} q_0(x_i^0) \leq \prod_{i \notin B(x^0)} q_1(x_i^0). \tag{13}$$

Let us use a symbol $||A||$ for the cardinal number of a set A. From (13) we arrive at the relation

$$(1 - \varepsilon_0)^{||B^c(x^0)||} \prod_{i \notin B(x^0)} p_0(x_i^0) < (1 - \varepsilon_1)^{||B^c(x^0)||} \prod_{i \notin B(x^0)} p_1(x_i^0). \tag{14}$$

For $i \in B(x^0)$ we have $q_0(x_i^0) = q_1(x_i^0)$ and so $x_i^0 \in A(\mathcal{C})$ for $i \in B(x^0)$, i.e.

$$\frac{p_1(x_i^0)}{p_0(x_i^0)} > \mathcal{C} \geq \frac{1 - \varepsilon_0}{1 - \varepsilon_1} \qquad i \in B(x^0).$$

The last inequality together with (14) gives

$$(1 - \varepsilon_0)^n \prod_{i=1}^n p_0(x_i^0) < (1 - \varepsilon_1)^n \prod_{i=1}^n p_1(x_i^0)$$

and it implies that

$$\mathfrak{K}_n(q_0, q_1) \subset \mathfrak{K}_{n,\gamma}(p_0, p_1),$$

where $\gamma = \log \frac{1 - \varepsilon_1}{1 - \varepsilon_0}$. Using Lemma 1 of [8] we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log e_n(p_0, \mathfrak{K}_n(q_0, q_1)) \leq \limsup_{n\to\infty} \frac{1}{n} \log e_n(p_0, \mathfrak{K}_{n,\gamma}(p_0, p_1)) \leq$$

$$\leq \log H(p_0, p_1) + \log (1 - \varepsilon_1) - \log (1 - \varepsilon_0). \tag{15}$$

To be able to estimate $\limsup_{n\to\infty} \frac{1}{n} \log e_n(p_1, \mathcal{K}_n^c(q_0, q_1))$ let us introduce the following notation. Let $n \in \mathcal{N}$ and $C_n \subset \mathbf{N}_n$.

Then put (for two densities f and g)

$$\mathcal{K}_{C_n}(f, g) = \{x \in \mathcal{X}_n;\ \prod_{i \in C_n} f(x_i) \leq \prod_{i \in C_n} g(x_i)\}$$

and

$$A_n(\mathcal{C}, C_n) = \{x \in \mathcal{X}_n;\ x_i \in A(\mathcal{C}) \text{ for all } i \in C_n\}.$$

Let $x^0 \in \mathcal{K}_n^c(q_0, q_1)$. Then

$$\prod_{i \notin B_n(x_0)} q_0(x_i^0) > \prod_{i \notin B_n(x^0)} q_1(x_i^0).$$

Using the inequality $\frac{1 - \varepsilon_1}{1 - \varepsilon_0} > 1$

we obtain

$$\prod_{i \notin B_n(x^0)} p_0(x_i^0) > \prod_{i \notin B_n(x^0)} p_1(x_i^0).$$

From $i \in B_n(x^0)$ follows $x_i^0 \in A(\mathcal{C})$ and therefore we have found that

$$x^0 \in \mathcal{K}^C_{B_n^c(x^0)}(p_0, p_1) \cap A(\mathcal{C}, B_n(x^0))$$

and finally

$$\mathcal{K}_n^c(q_0, q_1) \subset \bigcup_{B \subset \mathbf{N}_n} \{\mathcal{K}_B^C(p_0, p_1) \cap A(\mathcal{C}, B^C)\}.$$

Denoting $D_k = \{1, 2, \ldots, k\}$ and recalling that we are studying an i.i.d. model we can write

$$\limsup_{n\to\infty} \frac{1}{n} \log e_n\left(p_1, \mathcal{K}_n^c(q_0, q_1)\right) \leq \tag{16}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \log \sum_{k=0}^{n} \binom{n}{k} \int_{\mathcal{K}^C_{D_k^C}(p_0, p_1) \cap A(\mathcal{C}, D_k^C)} \prod_{i=1}^{n} p_1(x_i)\, d\mu(x_i) =$$

$$= \limsup_{n\to\infty} \frac{1}{n} \log \sum_{k=0}^{n} \binom{n}{k} \int_{\mathcal{K}^C_k(p_0, p_1)} \prod_{i=1}^{k} p_1(x_i)\, d\mu(x_i) \left[\int_{A(\mathcal{C})} p_1(x_i)\, d\mu(x_i)\right]^{n-k}.$$

Now let $\{k_n\}_{n=1}^{\infty}$ be a nondecreasing sequence of integers such that

$$\lim_{n\to\infty} k_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{k_n}{n} = 0, \tag{17}$$

and $\Delta > 0$. Let us take $n_0 \in \mathfrak{N}$ so, that for any $n \in \mathfrak{N}$, $n \geq n_0$ for any $k \in \mathfrak{N}$, $k \geq k_n$

$$\left| \frac{1}{k} \log \int\limits_{\mathfrak{K}_k(p_0, p_1)} \prod_{i=1}^{k} p_1(x_i)\, d\mu(x_i) - \log H(p_0, p_1) \right| < \Delta$$

holds. (It is possible to find such an n_0, since

$$\lim_{n \to \infty} \frac{1}{n} \int\limits_{\mathfrak{K}_n(p_0, p_1)} \prod_{i=1}^{n} p_1(x_i)\, d\mu(x_i) = \log H(p_0, p_1),$$

see [1].) From this we obtain for any $k \in \mathfrak{N}$, $k \geq k_n$ for some $n \in \mathfrak{N}$, $n \geq n_0$

$$[e^{-\Delta} H(p_0, p_1)]^k < \int\limits_{\mathfrak{K}_k^c(p_0, p_1)} \prod_{i=1}^{k} p_1(x_i)\, d\mu(x_i) < [e^{\Delta} \cdot H(p_0, p_1)]^k. \tag{18}$$

Now let us rewrite (16) into the form (again using Lemma 1A of [8])

$$\limsup_{n \to \infty} \frac{1}{n} \log e_n\big(p_1, \mathfrak{K}_n^c(q_0, q_1)\big) \leq \tag{19}$$

$$\leq \max \left\{ \limsup_{n \to \infty} \frac{1}{n} \log \sum_{k=0}^{k_n} \binom{n}{k} \int\limits_{\mathfrak{K}_k^c(p_0, p_1)} \prod_{i=1}^{k} p_1\, d\mu \left[\int\limits_{A(\mathcal{C})} p_1\, d\mu \right]^{n-k}, \right.$$

$$\left. \limsup_{n \to \infty} \frac{1}{n} \log \sum_{k=k_n+1}^{n} \binom{n}{k} \int\limits_{\mathfrak{K}_k^c(p_0, p_1)} \prod_{i=1}^{k} p_1\, d\mu \left[\int\limits_{A(\mathcal{C})} p_1\, d\mu \right]^{n-k} \right\}.$$

From (17) follows that there exists $n_1 \in \mathfrak{N}$ such that for any $n \in \mathfrak{N}$, $n \geq n_1$ we have $k_n < \frac{n}{2}$. Let $n \in \mathfrak{N}$, $n \geq \max\{n_1, n_0\}$.

Then

$$\frac{1}{n} \log \sum_{k=0}^{k_n} \binom{n}{k} \int\limits_{\mathfrak{K}_k^c(p_0, p_1)} \prod_{i=1}^{k} p_1 d\mu \left[\int\limits_{A(\mathcal{C})} p_1\, d\mu \right]^{n-k} \leq \tag{20}$$

$$\leq \frac{1}{n} \log \left\{ k_n \cdot \binom{n}{k_n} \cdot \left[\int\limits_{A(\mathcal{C})} p_1\, d\mu \right]^{n-k_n} \right\}.$$

Using Lemma 1A (see the Appendix) we obtain

$$\lim_{n \to \infty} \frac{1}{n} \log \left\{ k_n \binom{n}{k_n} \cdot \left[\int\limits_{A(\mathcal{C})} p_1\, d\mu \right]^{n-k_n} \right\} = \log \int\limits_{A(\mathcal{C})} p_1\, d\mu. \tag{21}$$

Let us now turn our attention to the second member of the maximum in (19). Let $n \in \mathfrak{N}$, $n \geq n_0$. Then from (18) follows

$$\frac{1}{n}\log \sum_{k=k_n+1}^{n} \binom{n}{k} \int_{\mathfrak{K}_k^{\varepsilon}(p_0,p_1)} \prod_{i=1}^{k} p_1\, d\mu \left[\int_{A(\mathfrak{C})} p_1\, d\mu\right]^{n-k} \leq$$

$$\leq \frac{1}{n}\log \sum_{k=k_n+1}^{n} \binom{n}{k} [e^{\Delta}\cdot H(p_0,p_1)]^k \left[\int_{A(\mathfrak{C})} p_1\, d\mu\right]^{n-k} \leq$$

$$\leq \frac{1}{n}\log\left[e^{\Delta} H(p_0,p_1) + \int_{A(\mathfrak{C})} p_1\, d\mu\right]^n.$$

Having taken limits of $n \to \infty$ and $\Delta \to 0$, we obtain

$$\limsup_{n\to\infty} \frac{1}{n}\log \sum_{k=k_n+1}^{n} \binom{n}{k} \int_{\mathfrak{K}_k^{\varepsilon}(p_0,p_1)} \prod_{i=1}^{k} p_1\, d\mu \left[\int_{A(\mathfrak{C})} p_1\, d\mu\right]^{n-k} \leq \tag{22}$$

$$\leq \log\left\{H(p_0,p_1) + \int_{A(\mathfrak{C})} p_1\, d\mu\right\}.$$

Now the assertion of the Theorem follows from (12), (15) (19), (20), (21) and (22).

Q.E.D.

Remark 2. To have any sense the bounds given in Theorem 1 and in Corollary 1 have to be less than one. Similarly the bound in Theorem 2 has to be negative. In the opposite cases these bounds should be substituted by ones and zero, respectively. The bound in Theorem 1 is an upper one for an expression

$$\limsup_{n\to\infty} \frac{1}{n}\log \inf_{\varphi_n \in \Psi} \sup_{P' \in \mathfrak{P}_0, P'' \in \mathfrak{P}_1} \{E_{P'}\varphi_n + E_{P''}(1-\varphi_n)\},$$

where Ψ denotes the set of all tests and symbol E stands for the mean value. Therefore, for the test

$$\varphi_n^* = I_{\{\mathfrak{K}_n(q_0,q_1)\}}$$

(see (2)) and p_0 and p_1 we have

$$\limsup_{n\to\infty} (E_{P_0}\varphi_n^* + E_{P_1}\varphi_n^*) = \limsup_{n\to\infty} \frac{1}{n}\log e_n(p_0,p_1,\mathfrak{K}_n(q_0,q_1)) \leq$$

$$\leq \log\{(1-\varepsilon_1)\, H(p_0,p_1) + \varepsilon_0 + (1-\varepsilon_0) \int_{A(\mathfrak{C})} p_0\, d\mu\}.$$

Therefore, in order to show that Theorem 2 brings an improvement of the assertion of Theorem 1, we should prove that

$$\min\left\{\max\left\{\log H(p_0, p_1) + \log\frac{1-\varepsilon_1}{1-\varepsilon_0},\ \log\left[H(p_0, p_1) + \int\limits_{A(\mathcal{C})} p_1\, d\mu\right]\right\}, 0\right\} \leq$$

$$\leq \min\left\{\log\left[(1-\varepsilon_1)\, H(p_0, p_1) + \varepsilon_0 + (1-\varepsilon_0) \int\limits_{A(\mathcal{C})} p_0\, d\mu\right], 0\right\} \tag{23}$$

is true.

Lemma 1. The relation (23) holds.

Proof. Let us use in this proof H and A instead of $H(p_0, p_1)$ and $A(\mathcal{C})$, respectively. Let the left-hand side of inequality (23) equal zero. Then either

$$H\frac{1-\varepsilon_1}{1-\varepsilon_0} > 1 \tag{24}$$

or

$$H + \int\limits_A p_1\, d\mu > 1. \tag{25}$$

Let at first (24) be fulfilled. Then

$$\log\left[(1-\varepsilon_1)\, H + \varepsilon_0 + (1-\varepsilon_0) \int\limits_A p_0\, d\mu\right] \geq$$

$$\geq \log\left[(1-\varepsilon_0) + \varepsilon_0 + (1-\varepsilon_0) \int\limits_A p_0\, d\mu\right] \geq 0.$$

If (25) holds, then, using (1), we derive

$$(1-\varepsilon_1)\, H + \varepsilon_0 + (1-\varepsilon_0) \int\limits_A p_0\, d\mu =$$

$$= (1-\varepsilon_1)\left[H + \int\limits_A p_1\, d\mu\right] + \varepsilon_1 > (1-\varepsilon_1) + \varepsilon_1 = 1.$$

Therefore the right-hand side of (23) is equal to zero, too. To finish the proof it is necessary to verify that

$$\max\left\{\log\left(H \cdot \frac{1-\varepsilon_1}{1-\varepsilon_0}\right),\ \log\left(H + \int\limits_A p_1\, d\mu\right)\right\} \leq \tag{26}$$

$$\leq \log\left[(1-\varepsilon_1)\, H + \varepsilon_0 + (1-\varepsilon_0) \int\limits_A p_0\, d\mu\right]$$

under the condition that

$$\max\left\{\log\left(H \cdot \frac{1-\varepsilon_1}{1-\varepsilon_0}\right),\ \log\left(H + \int_A p_1\, d\mu\right)\right\} < 0. \tag{27}$$

From (1) we have

$$(1-\varepsilon_1)H + \varepsilon_0 + (1-\varepsilon_0)\int_A p_0\, d\mu = (1-\varepsilon_1)H + \varepsilon_1 + (1-\varepsilon_1)\int_A p_1\, d\mu. \tag{28}$$

Using (1) once again, we find

$$\frac{\varepsilon_0 - \varepsilon_1}{1-\varepsilon_1} \leq \int_A p_1\, d\mu$$

and from (27) $\varepsilon_0 H \cdot \dfrac{1-\varepsilon_1}{1-\varepsilon_0} < \varepsilon_0$.

Finally,

$$(1-\varepsilon_1)H + \varepsilon_1 + (1-\varepsilon_1)\int_A p_1\, d\mu \geq \frac{1-\varepsilon_1}{1-\varepsilon_0}[(1-\varepsilon_0)H + \varepsilon_0 H] = \frac{1-\varepsilon_1}{1-\varepsilon_0}H. \tag{29}$$

A similar straightforward computation gives

$$H + \int_A p_1 d\mu \leq (1-\varepsilon_1)H + \varepsilon_1 + (1-\varepsilon_1)\int_A p_1\, d\mu. \tag{30}$$

Now (28) together with (29) and (30) proves (26).

Q.E.D.

4. Appendix

Lemma 1A. Let $\{k_n\}_{n=1}^{\infty}$ be a nondecreasing sequence of positive integers such that

$$\lim_{n\to\infty} k_n = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{k_n}{n} = 0.$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \log \binom{n}{k_n} = 0.$$

Proof. By means of Stirling's formula we find

$$\lim_{n\to\infty} \frac{1}{n} \log \binom{n}{k_n} = \lim_{n\to\infty} \frac{1}{n}\left\{n \cdot \log \frac{1}{1-\frac{k_n}{n}} + k_n \log (n-k_n) - k_n \log k_n\right\} =$$

$$= \lim_{n\to\infty} \left\{\log \frac{1}{1-\frac{k_n}{n}} + \frac{k_n}{n} \log\left(1-\frac{k_n}{n}\right) - \frac{k_n}{n} \log \frac{k_n}{n}\right\} = 0.$$

References

1. *Chernoff, H.* (1952): A measure of asymptotic efficiency for tests of hyohesis based on the sum of observations. Ann. Math. Statist. **23**, 493–507.
2. *Huber, P. J.* (1952): A robust version of the probability ratio test. Ann. Math. Statist. **36**, 1753–1758.
3. *Huber, P. J., Strassen V.* (1973): Minimax tests and the Neyman–Pearson lemma for capacities. Ann. Statist. **1**, 251–263.
4. *Халфина, Н. М., Халфин, Л. А.* (1975): Об устойчивом варианте теста отношения правдоподобия. Теория вероятностей и ее применения, Том XX, выпуск *1*, 203–206.
5. *Rieder, H.* (1977): Least favorable pairs for special capacities. Ann. Statist. **5**, 909–922.
6. *Rieder, H.* (1978): A robust asymptotic testing model. Ann. Statist. **6**, 1080–1094.
7. *Víšek, J. Á.* (1980): An asymptotic rate of convergence of the error probabilities of a robust test. Proceedings of the Third Czechoslovak–Soviet–Hungarian seminar on information theory, 213–218.
8. *Víšek, J. Á.* (1981): Asymptotic behaviour of the likelihood ratio test under presence of deviations of the model. Submitted to the Kybernetika.

Границы предельной скорости сходимости вероятностей ошибок устойчивого теста

Й. А. ВИШЕК

(Прага)

В работе исследуется эффективность устойчивого варианта теста правдоподобия, который был приведен Н. М. Халфиной и Л. А. Халфиным. Верхние границы скорости сходимости вероятностей ошибок для этого теста были получены для случая, когда наблюдения загрязнены и для случая, когда наблюдения без загрязнения. В конце статьи показана состоятельность результатов.

J. Á. Víšek
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Pod vodárenskou věží 4
182 08 Praha 8 — Libeň
Czechoslovakia

Problems of Control and Information Theory, Vol. 10 (4), pp. 267—277 (1981)

SOME SYSTEM THEORY APPLICATIONS OF EGERVÁRY'S ALGORITHM FOR TRANSFORMATION OF A MATRIX TO THE HERMITE NORMAL FORM

N. K. SINHA, P. RÓZSA

(Hamilton), (Budapest)

(Received April 9, 1980)

Egerváry's outer product algorithm for the transformation of a matrix to the Hermite normal form is reviewed. Four applications of this algorithm to system theory are described. These are (1) minimum realization of a transfer function matrix in the state space form, (2) transformation of state equations to canonical forms, (3) reduction of high-order systems by partial realization, and (4) determination of controllability and observability of large scale composite systems.

1. Introduction

Egerváry [1] had proposed an algorithm utilizing outer products for linear transformation of matrices, and later Rózsa [2] proposed utilizing this algorithm for transformation of a matrix to the Hermite normal form. This algorithm is very efficient computationally, and has found several applications in system theory.

The object of this paper is to describe some applications of the algorithm. Before going into these applications, it will be desirable to first review this algorithm.

A square matrix is said to be in the Hermite normal form if the following conditions are satisfied [3].

(i) it is of the upper triangular form with the elements on the main diagonal being either 1 or 0,

(ii) if a certain diagonal element is equal to 1, then all other elements in the corresponding column are 0, and

(iii) if a certain diagonal element is equal to 0, then all elements in the corresponding row are 0.

It is evident that the rank of a matrix in the Hermite normal form is equal to the number of 1's on the main diagonal.

Now the procedure for tranformation of any rectangular matrix to this form will be described. Consider a matrix S with n rows and m columns, i.e.

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1m} \\ s_{21} & s_{22} & \cdots & s_{2m} \\ \vdots & \vdots & & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nm} \end{bmatrix}. \tag{1}$$

We first select any nonzero element, say $s_{i_1 1}$, in the first column (it is computationally advantageous to select the element with the largest absolute value), and then obtain the matrix

$$S^{(1)} = S - \frac{1}{s_{i_1 1}} (Se_1 - e_{i_1})(e^{i_1} S) \tag{2}$$

where e_{i_1} is the i_1th unit column vector of dimension n and e^{i_1} is the i_1th unit row vector of dimension m.

Note that we may also write

$$S^{(1)} = T_1 S \tag{3}$$

where

$$T_1 = I - \frac{1}{s_{i_1 1}} (Se_1 - e_{i_1}) e^{i_1} \tag{4}$$

and

$$\det T_1 = \frac{1}{s_{i_1 1}}. \tag{5}$$

Hence, $S^{(1)}$ is a nonsingular transformation of S. Also, the resulting matrix $S^{(1)}$ has the property that its first column is the unit vector e_{i_1}.

The next step in the transformation is to select a nonzero element in the second column of $S^{(1)}$, say $s^{(1)}_{i_2 2}$, with $i_2 \neq i_1$ and obtain

$$S^{(2)} = S^{(1)} - \frac{1}{s^{(1)}_{i_2 2}} (S^{(1)} e_2 - e_{i_2})(e^{i_2} S^{(1)}). \tag{6}$$

It follows that

$$S^{(2)} = T_2 S^{(1)} = T_2 T_1 S \tag{7}$$

where T_2 is also nonsingular, and the first two columns of $S^{(2)}$ are e_{i_1} and e_{i_2}, respectively.

Continuing in this manner, after a maximum of m steps, we get a matrix $S^{(m)}$ which can be brought to the Hermite normal form by an appropriate permutation of its rows, i.e., by premultiplication with a nonsingular permutation matrix P. Hence, we have

$$H = PS^{(m)} = PT_m T_{m-1} \ldots T_1 S. \tag{8}$$

It follows that the rank of S is equal to the number of independent unit column vectors in $S^{(m)}$.

2. Minimal realization of a transfer function matrix [4]

Let $G(s)$ be an $m \times r$ system transfer function matrix, each element of which is a strictly proper rational function of s. Then the minimal realization problem is to obtain the triple (A, B, C) where A, B and C are $n \times n$, $n \times r$ and $m \times n$ constant matrices, for minimum n, such that

$$G(s) = C(sI - A)^{-1} B. \tag{9}$$

Because of the strictly proper nature of $G(s)$, it can be expanded as

$$G(s) = J_0 s^{-1} + J_1 s^{-2} + J_2 s^{-3} + \ldots \tag{10}$$

where

$$J_i = CA^i B, \quad i = 0, 1, 2, \ldots \tag{11}$$

are called the Markov parameters of the system.

Following Ho and Kalman [5] define the block Hankel matrix S_{ij} as

$$S_{ij} = \begin{bmatrix} J_0 & J_1 & \ldots & J_{j-1} \\ J_1 & J_2 & \ldots & J_j \\ \vdots & \vdots & & \\ J_{i-1} & J_i & \ldots & J_{i+j-2} \end{bmatrix}. \tag{12}$$

Then the rank of S_{ij} is n, provided that $i \geq \beta$ and $j \geq \alpha$, where α is the controllability index of the system and β its observability index. From (11) and (12), it follows that

$$S_{ij} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{i-1} \end{bmatrix} [B \; AB \; \ldots \; A^{j-1}B] = V_i U \tag{13}$$

where U_j and V_i are called the controllability and observability matrices, respectively, each with rank equal to n, and n is the order of the minimal realization.

It is evident that the factorization indicated in equation (13) is not unique. It can, therefore, be simplified by requiring that the r columns of B be columns of the $n \times n$ identity matrix, I_n. This will be justified when the rank

of B is equal to r, as will be the case with all systems having r independent inputs

The algorithm for determining A, B and C in this canonical form will now be described. From the given $G(s)$, first the Markov parameters are calculated by long division of the numerator of each element of $G(s)$ by the corresponding denominator. Forming the Hankel matrix S_{ij} it is transformed into the Hermite normal form, H. Since the rank of H is n, any additional rows, consisting of only zero elements are deleted from H. The first r columns of the resulting matrix will be $e_1, e_2, \ldots, e_r$, and these constitute the matrix B, whereas the next n columns will give the matrix A. The columns of C are obtained from the first m rows of the matrix S_{ij} corresponding to the locations of the n unit column vectors of H. Further details of the procedure are given in reference [4].

It may be added that in a recent comparison [6] between different methods for obtaining the minimal realization of a transfer function matrix, it was shown that the method described above was computationally the most efficient.

3. Transformation of state equations to canonical forms

In the design of feedback control systems, it is often required to transform the state equations of multivariable systems to their canonical form. Such transformations were first described by Luenberger [7] who also proposed two special canonical forms known as (i) the column-companion form and (ii) the controllable canonical form. Luenberger's method requires the determination of a set of n linearly independent vectors from the controllability matrix of the system, selected in a particular order. The transformation matrix, consisting of these vectors, has then to be inverted to obtain A, B and C in the column-companion form. The transformation to the controllable canonical form requires the inversion of two $n \times n$ matrices in addition to determining a set of linearly independent vactors selected in a special order from the controllability matrix.

The application of the algorithm described in section 1 makes it possible to obtain the canonical form without any matrix inversion. For example, consider first the transformation to the column-companion form. In this case, one first starts with the columns of the controllability matrix rearranged in the following form

$$S = [b_1 \; Ab_1 \ldots A^n b_1 \; b_2 \; Ab_2 \ldots A^{n-1} b_2 \ldots b_r \; Ab_r \ldots A^{n-r-1} b_r] \qquad (14)$$

where b_i is the ith column of B.

Transformation of S to the Hermite normal form gives

$$H = [e_1 e_2 \ldots e_{n_1} \alpha_1 * * e_{n_1+1} \ldots \alpha_2 * * \ldots e_n \alpha_r * * \ldots] \tag{15}$$

where the columns marked with asterisk are of no particular significance.

The columns of the canonical form matrices are now readily obtained as follows

$$\overline{B} = [e_1 e_{n_1+1} e_{n_1+n_2+1} \ldots e_{n-n_r+1}] \tag{16}$$

$$\overline{A} = [e_2 e_3 \ldots e_{n_1} \alpha_1 e_{n_1+1} \ldots \alpha_2 \ldots e_n \alpha_r] \tag{17}$$

and

$$\overline{C} = C\,[b_1 \;\; Ab_1 \ldots A^{n_1-1} b_1 \ldots b_r \ldots A^{n_r-1} b_r] = CP. \tag{18}$$

It may be noted that the columns of the transformation matrix P are obtained as the columns of S in equation (14) corresponding to the independent unit column vectors in H.

The transformation to the controllable canonical form can be obtained in a similar manner, without requiring any matrix inversion. In this case, two reductions to the Hermite normal form are required, first on a column basis and then on a row basis. Details of the algorithm are given in [8].

4. Reduction of high-order systems by partial realization

Another important problem in the design of control system is the determination of a low-order model which approximates a given high-order system. Consider a system described by the equations

$$\dot{x} = Ax + Bu \tag{19}$$

$$y = Cx \tag{20}$$

where $x \in R^n$, $u \in R^m$ and $y \in R^p$, with $n \gg m, p$. The problem of reduction is the determination of the equations

$$\dot{z} = Fz + Gu \tag{21}$$

$$\hat{y} = Hz \tag{22}$$

where $z \in R^r$, and $r \ll n$, such that $\hat{y}(t)$ is a close approximation of $y(t)$ for all $u(t)$ and all t.

As in section 2, the transfer function matrix of the high-order system is given by

$$G(s) = C(sI - A)^{-1} B = J_0 s^{-1} + J_1 s^{-2} + J_2 s^{-3} + \dots \tag{23}$$

where

$$J_i = CA^i B, \; i = 0, 1, 2, \dots . \tag{24}$$

If the matrix A is nonsingular, one may also expand $G(s)$ in the following Laurent series

$$G(s) = -\sum_{i=0}^{\infty} T_i s^i \tag{25}$$

where

$$T_i = CA^{-(i+1)} B, \qquad i = 0, 1, 2, \dots . \tag{26}$$

It may be recognized that $\{-1/iT_i\}$ is the set of time moments of the inverse Laplace transform of $G(s)$. It may also be noted that

$$J_{-i} = T_{i-1} \quad \text{for} \quad i \geq 1. \tag{27}$$

Next, define a Hankel matrix of order (i, j) and index k as

$$S_{ij}^{(k)} = \begin{bmatrix} CA^k B & CA^{k+1} B & \dots & CA^{k+j-2} B \\ CA^{k+1} B & CA^{k+2} B & \dots & CA^{k+j} B \\ \vdots & \vdots & & \vdots \\ CA^{k+i-1} B & CA^{k+i} B & \dots & CA^{k+i+j-2} B \end{bmatrix}. \tag{28}$$

One may now factor the Hankel matrix as below

$$S_{ij}^{(k)} = V_i^{(\delta)} U_j^{(\varepsilon)} = \begin{bmatrix} CA^\delta \\ CA^{\delta+1} \\ \vdots \\ CA^{\delta+i-1} \end{bmatrix} [A^\varepsilon B A^{\varepsilon+1} B \dots A^{\varepsilon+j-1} B] \tag{29}$$

where δ and ε are integers, and $\delta + \varepsilon = k$.

It is easily shown that transformation of the Hankel matrix leads to a minimal realization of the system, as in section 2. However, if the process is stopped after r steps then a partial realization is obtained which matches some of the generalized Markov parameters instead of all, as would be the case with minimal realization. This may be regarded as the generalization of Pade approximation to the reduction of multivariable systems [9].

An advantage of the method is its iterative nature. If the reduced model of order r does not provide a satisfactory match with $G(s)$, we may try one of order $r + 1$ by transforming S one more time. It is conceptually easy and ideally suited for either hand calculation or implementation on a digital computer.

5. Determination of controllability and observability of large-scale composite systems

The investigation of the controllability and observability of large-scale composite linear time-invariant systems by direct application of the conventional criteria is generally very cumbersome since it requires the determination of the ranks of matrices containing terms of the form A^iB and CA^j where the matrix A is usually of a high order. In fact, there may often be problems due to truncation errors caused by repeated multiplications by a large matrix. An important contribution to the field was made by Davison and Wang [10] in a recent paper in which they avoid the need to obtain the product terms of the form A^iB and CA^j. However, their approach involves computation of both eigenvalues and rank. In a recent paper, Porter [11] has presented a method which also does not require the product terms, only the computation of the rank of a larger matrix is necessary. For a system of order n, with m inputs and p outputs, this method requires the determination of the rank of an $n^2 \times n(n + m - 1)$ matrix for controllability and the rank of an $n(n + p - 1) \times n^2$ matrix for observability. Porter has applied this method for determining the controllability and observability of feedback and series-parallel connections. Extension to the cascade interconnection is fairly straightforward [12].

Although Porter's approach is a great step forward from the approach of Davison and Wang, the main difficulty with this method has been summarized by Porter as follows, "the only computational constraint in the use of this method is the availability of an effective algorithm for the computation of the rank of a large matrix".

An effective algorithm of this type can be obtained easily by a simple modification of the algorithm for transformation to the Hermite normal form discussed in section 1. As was mentioned there, the rank of the matrix is simply the number of independent unit vectors obtained after transformation to the Hermite normal form. The total number of arithmetic operations (additions, multiplications and divisions) required for the general case of an $n \times m$ matrix of rank r is given by

$$N = r(2n + 1)(2m - r + 1). \tag{30}$$

In practice, the actual number of arithmetic operations will be much less if the matrix is sparse, with a large number of zero entries, and advantage is taken of this fact in programming.

The procedure can be further improved if it is recognized that to determine the rank of a given matrix, it is sufficient to make each column, starting with the first, have zero elements below the pivot element which is transformed into one. The resulting matrix, by a simple permutation of rows, takes the form of an upper triangular matrix, and hence its rank is determined immediately. It may be added that this permutation is not necessary for our purpose.

The proposed modification for determining the rank of an $n \times m$ matrix S will now be described in detail for the case when $m \geq n$ [12]. For the case when $m < n$, the same procedure can be carried out starting with the transpose of S. After selecting a suitable nonzero element s_{i-1} in the first column of S as the pivot, the following matrix is obtained

$$S^{(1)} = S - \frac{1}{s_{i_1 1}} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ s_{i_1 1} - 1 \\ s_{i_1+1} \quad 1 \\ \vdots \\ s_{n1} \end{bmatrix} [s_{i_1 1}\, s_{i_1 2} \cdots s_{i_1 m}]. \tag{31}$$

$S^{(1)}$ is seen to be a nonsingular transformation of S, since the determinant of the transformation matrix is again $1/s_{i_1 1}$.

It may be noted that the transformation in equation (31) differs from that in equation (2) in that all elements above $s_{i_1 1}$ are made zero in the outer product, with the result that the rows above i_1 are unchanged by the transformation.

Continuing this process, in r steps a matrix $S^{(r)}$ is obtained, where $r \leq n$, after a nonzero pivot has been utilized from each row. Then the rank of S is equal to r.

The total number of arithmetic operations required for this improved algorithm is given by

$$M = (2n + 1)mr + \frac{1}{3} r(r - 1)(2r - 1) - \frac{1}{2} r(r - 1)(2m + 2n + 1). \tag{32}$$

In general, M will be much smaller than N given by equation (30). For example, if $n = 121$ and $m = 132$, assuming that $r = 121$, we have

$$N = 2,117,016 \tag{33}$$

and

$$M = 1,350,844 \tag{34}$$

which indicates a reduction of about 36%.

Further reduction in the number of arithmetic operations can be obtained by taking advantage of the sparsity of the matrix. This simply requires testing each element in the row vector in the outer product and changing only those columns for which this element is nonzero.

In an example given in reference [12], the total time required for determining the rank on a CDC 6400 computer was 105 seconds using the algorithm described in section 1. With the modification suggested in this section, the computation time was reduced to 63.16 seconds without taking advantage of the sparsity of the matrix. When additional modification to utilize the advantage of sparsity was introduced, the computation time was reduced to only 2.87 seconds.

6. Conclusions

The algorithm for transformation of a matrix to the Hermite normal form, utilizing outer products, as proposed by Egerváry has been utilized to solve a number of problems in systems engineering. In particular, with some modifications, it can be used for determining the rank of a large matrix in a very efficient manner. A computer program, based on this modified algorithm has also been written [12].

By reviewing some applications of the algorithm originally proposed by Egerváry [1] and later formalized by Rózsa [2], it is hoped that further applications to problems in some other disciplines will be encouraged.

Acknowledgements

This work was made possible by the visit of Professor Rózsa to McMaster University under the cultural exchange program between McMaster University and the Institute of Cultural Relations. An earlier version of this paper was presented by Professor Sinha in Budapest at a talk sponsored by the Bolyai János Mathematical Society of Budapest.

References

1. *Egerváry, E.*, "Über die Faktorisation von Matrizen und ihre Anwendung auf die Lösung von linearen Gleichungssystemen", Zeitschrift für Angewandte Mathematik und Mechanik, vol. **35**, 1935, pp. 111–118.
2. *Rózsa, P.*, "Lineáris Algebra és Alkamazásai" (Budapest: Műszaki Könyvkiadó), 1974.
3. *MacDuffee, C. C.*, "Vectors and Matrices" (Buffalo, N. Y.: The Mathematical Association of America), 1943.
4. *Rózsa, P., Sinha, N. K.*, "Efficient algorithm for irreducible realization of a rational matrix", Int. J. Control, vol. **20**, 1974, pp. 739–751.
5. *Ho, B. L., Kalman, R. E.*, "Effective construction of linear state-variable models from input–output functions", Proc. 3rd Allerton Conf. on Circuit and System Theory, (Urbana, Ill.), 1965, pp. 449–459.
6. *Sinha, N. K.*, "Minimal realization of transfer function matrices: a comparative study of different methods", Int. J. Control, vol. **22**, 1975, pp. 627–640.
7. *Luenberger, D. G.*, "Canonical forms for linear multivariable systems", IEEE Trans., vol. **AC–12**, 1967, pp. 290–293.
8. *Hickin, J., Sinha, N. K.*, "An efficient algorithm for transformation of linear multivariable systems to canonical forms", Problems of Control and Information Theory, vol. **8** (3), 1979, pp. 193–211.
9. *Hickin, J., Sinha, N. K.*, "A new method for reducing multivariable systems", Proc. 7th Annual Conf. on Modeling and Simulation (Pittsburgh, Pa.), 1976, pp. 259–263.
10. *Davison, E. J., Wang, S. H.*, "New results on the controllability and observability of general composite systems", IEEE Trans., vol. **AC–20**, 1975, pp. 123–128.
11. *Porter, B.*, "Necessary and sufficient conditions for the controllability and observability of general composite systems", Proc. IFAC Symp. Large Scale Systems Theory and Applications (Udine, Italy), 1976, pp. 265–269.
12. *Sinha, N. K., ElSherief, H.*, "Efficient algorithm for determining controllability and observability of large-scale composite systems", Int. J. of Systems Science, vol. **9**, 1978, pp. 1295–1302.

Appendix

Some comments on the computational procedure

The computational procedure for the transformation of a matrix to the Hermite normal form can be further improved by following the procedure described below. It is assumed that the matrix has more columns than rows, i.e. $m \geq n$.

(i) Starting with the first column of the matrix, search for the element in this column which has the largest absolute value. This element will be used as the pivot.

(ii) Interchange the first row of the matrix with this row, and then set up the column vector as well as the row vector for the outer product as the first column and the first row of the matrix, with 1 subtracted from the pivot element, p, in the column vector. Let the elements of the column vector be denoted by b_i, $i = 1, 2, \ldots, n$ and those of the row vector be e_j, $j = 1, 2, \ldots, m$.

(iii) Replace each element s_{ij}, of the matrix S by $s_i^{(1)}$ of the matrix $S^{(1)}$ as below

$$s_{ij}^{(1)} = \frac{p s_{ij} - b_i c_i}{p}.$$

Note that it is better to calculate $s_i^{(1)}$ as above than as $s_{ij} - (b_i c_j)/p$ because the truncation error which may be introduced if the division is performed before subtraction.

It is convenient to perform this computation on one column at a time. In particular, if $c_j = 0$, then the jth column of $S^{(1)}$ is the same as the jth column of S. This fact can be utilized to avoid unnecessary computation.

(iv) After the above has been completed, examine the second column of $S^{(1)}$, starting from the second row to search for the next pivot in the same way as in step (i). Interchange rows if necessary, as in (ii) and proceed as in (iii).

(v) The process can be continued until all columns and rows are exhausted. The exception is the case when in a particular column a nonzero-pivot cannot be found. In this case, simple proceed to the next column.

(vi) A computer program for efficient transformation of an $n \times m$ matrix to the Hermite normal form has been written and is available.

Преобразование матриц к эрмитовой нормальной форме с помощью алгоритма Эгервари и некоторые приложения к теории систем

Н. К. СИНХА, П. РОЖА

(Будапешт)

Доказывается, что любая матрица может быть приведена к эрмитовой нормальной форме, используя метод Эгервари разложения матриц на сумму внешних произведений. Затем приводятся следующие четыре приложения алгорифма к теории систем: (1) минимальная реализация матрицы передаточных функций в форме пространства состояния, (2) преобразование уравнений состояния к канонической форме, (3) преобразование систем большого порядка с помощью частичных реализаций, (4) определение управляемости и наблюдаемости больших сложных систем.

N. K. Sinha
Group on Simulation, Optimization and Control
Faculty of Engineering
McMaster University
Hamilton, Ontario, Canada L85 4L7

P. Rózsa
Department of Mathematics
Faculty of Electrical Engineering
Technical University of Budapest
Budapest, Hungary

Problems of Control and Information Theory, Vol. 10 (4), pp. 279–285 (1981)

A REMARK ON THE ESTIMATION OF FUNCTIONALS OF A PROBABILITY DENSITY

W. WERTZ

(Vienna)

(Received September 4, 1980)

The functional $d(f) := \int (f''(x))^2\,dx$, where f is an unknown probability density, plays an important role in nonparametric estimation of the density. Based on an expansion of f'' by means of the Hermite orthonormal system, an estimator of $d(f)$ is constructed and its consistency in the quadratic mean is proved.

Introduction

In several simulation problems it is necessary to generate random numbers according to a distribution which is not exactly but only empirically known. One possible method is the following: first we recall a general rejection procedure, given by Văduva [15] (it is based on Lemma 4 below): let f and g be probability density functions and let an $\alpha > 0$ exist such that $f(x) \leq \alpha \cdot g(x)$ for every real x. Let u be a random number from the uniform distribution over $[0, 1]$ and y be a random number corresponding to g, then y is taken as a random number from the distribution with density f if and only if $0 \leq u \leq f(y)/[\alpha \cdot g(y)]$ (if $g(x) = 0$, also $f(y) = 0$ and the right-hand side equals zero by definition). In the case f is unknown, one can substitute it by an estimate $\hat{f}_n(x_1, \ldots, x_n; .)$ (such as described below), choose an appropriate density g with $\hat{f}_n(x_1, \ldots, x_n; x) \leq \alpha \cdot g(x)$ for every real x and accept y as a random number of the unknown distribution if and only if $0 \leq u \leq \hat{f}_n(x_1, \ldots, x_n; y)/[\alpha \cdot g(y)]$. If $\hat{f}_n(x_1, \ldots, x_n; .)$ is a good estimate of f, then the resulting random numbers can be considered as a good approximation of random numbers from the distribution with density f.

The well-known kernel-method for estimating unknown probability density functions has been investigated thoroughly (for a survey see Wertz [16]). Roughly speaking, the density f (with respect to one-dimensional Lebesgue-measure) is estimated by means of independent random variables $\xi_1, \ldots, \xi_n$, each one distributed with density f; the observed values will be denoted by $x_1, \ldots, x_n$; as an estimate of f, the function

$$x \to \hat{f}_n(x_1, \ldots, x_n; x) := (1/nb_n) \sum_{i=1}^{n} K((x - x_i)/b_n)$$

is used; here K is a probability density and (b_n) a positive sequence, tending to zero.

Let E_f denote expectations computed with density f. The mean integrated square error (MISE) of $\hat{f}_n$ at f is defined by $E_f \int [\hat{f}_n(\xi_1, \ldots, \xi_n; x) - f(x)]^2 dx$ and serves as the risk of $\hat{f}_n$. Under the condition that f is bounded and has two continuous derivatives, the asymptotically optimal choice for K and b_n is:

$$K(t) = \begin{cases} [3/(4.5^{1/2})] \cdot (1 - t^2/5) & \text{for } |t| \leq 5^{1/2} \\ 0 & \text{for } |t| > 5^{1/2}, \end{cases}$$

$b_n = (12.5^{-3/2})^{1/5} \cdot n^{-1/5} / \int (f''(x))^2 dx$. The order of the MISE is then $o(n^{-4/5})$ (see Epanehnikov [6], Rosenblatt [10]). In the expression for b_n, f enters only via the term $\int (f''(x))^2 dx$, and when applying the above result in practice, one has to know this quantity. Hence it is necessary to have first an estimate of $d(f) := \int (f''(x))^2 dx$.

The problem of estimating nonlinear functionals of the density has been treated by several authors. There are rather general results by Dmitriev and Tarasenko [4], [5], further methods for estimating $\int f^2(x) dx$ (Bhattacharya and Roussas [2]), the Fisher-information $\int (f'(x))^2/f(x) dx$ (Bhattacharya [1]), and so on. But most of these methods are based on the kernel technique, hence again the problem of choosing a suitable bandwidth b_n arises. On the other hand, there is an extensive literature on estimating probability density functions by orthogonal series (e.g. Čencov [20], [21], Schwartz [14], Bosq [3], Mirzahmedov and Hašimov [8], [9], Siraždinov, Mirzahmedov and Hosni [17], Hašimov [18], [19] — for further references see Schneider and Wertz [12]). Schüler and Wolff [13] use this method to estimate $\int f^2(x) dx$, in particular, they treat the case of the trigonometric system. In this note, we modify their results to prove consistency of an estimator of $\int (f''(x))^2 dx$, which is based on the Hermite orthonormal system, and moreover, consider the speed of convergence. In particular, for densities with unbounded support, Hermite functions are more convenient than other systems like the trigonometric system.

In an obvious way it is possible to formulate our results for other orthonormal systems, for which bounds like in Lemma 1 and Lemma 3 are valid.

Results

Let f be a *probability density*, which is *twice continuously differentiable* and let f'' be of *bounded variation*. We assume f *and* f'' *to belong to* $L_1 \cap L_2$.

Let $H_k(x) = (-1)^k \cdot e^{x^2} \cdot \dfrac{d}{dx^k} e^{-x^2}$ be the k-th Hermite polynomial $(k = 0, 1, 2, \ldots)$ and $\psi_k(x) = d_k \cdot e^{-x^2/2} \cdot H_k(x)$ be the k-th normed Hermite function, $d_k = (\pi^{1/2} \cdot 2^k \cdot k!)^{-1/2}$. $\{\psi_k; k = 0,1, 2, \ldots\}$ is a complete ortho-

normal system of L_2. (For details about Hermite functions see e.g. Erdély, Magnus, Oberhettinger and Tricomi [7], Sansone [11].)

We denote by $c_k = \hat{f''}(k) = \int f''(x) \cdot \psi_k(x)\, dx$ the *k-th Fourier coefficient of* f''. Under the above conditions, the Fourier series of f'', $\sum_{k=0}^{\infty} c_k \psi_k(x)$ converges in L_2-norm and also pointwise uniformly on compact intervals.

Since f is a uniformly continuous density and ψ_k'' bounded (Lemma 1)

$$\lim_{x \to \pm\infty} f(x) \cdot \psi_k''(x) = 0$$

and partial integration yields

$$c_k = \int f(x) \cdot \psi_k''(x)\, dx .$$

Obviously,

$$\hat{c}_k(x_1, \ldots, x_n) := [n \cdot (n-1)]^{-1} \cdot \sum_{\substack{i \neq j \\ 1 \leq i,j \leq n}} \psi_k''(x_i) \cdot \psi_k''(x_j)$$

defines an *unbiased estimator of* c_k^2; by the equation

$$\psi_k''(x) = (x^2 - 2k - 1) \cdot \psi_k(x)$$

it can also be written

$$\hat{c}_k(x_1, \ldots, x_n) = [n \cdot (n-1)]^{-1} \cdot \sum_{\substack{i \neq j \\ 1 \leq i,j \leq n}} (x_i^2 - 2k - 1) \cdot (x_j^2 - 2k - 1) \cdot \psi_k(x_i) \cdot \psi_k(x_j).$$

Let (q_n) be a sequence of natural numbers tending to infinity; then

$$\hat{d}_n(x_1, \ldots, x_n) := \sum_{k=0}^{q_n} \hat{c}_k(x_1, \ldots, x_n) \tag{1}$$

will be used as an estimator of $d(f)$.

By Parseval's formula,

$$d(f) = \sum_{k=0}^{\infty} c_k^2 = \sum_{k=0}^{q_n} c_k^2 + \sum_{k=q_n+1}^{\infty} c_k^2,$$

where the last term tends to zero with $n \to \infty$; but $\hat{d}_n$ is an unbiased estimator of $\sum_{k=0}^{q_n} c_k^2$, hence $\hat{d}_n$ is asymptotically unbiased.

Theorem. Let f be $s+2$ times differentiable ($s \geq 0$) and the function $x \to e^{x^2/2} \cdot \frac{d^s}{dx^s}[e^{-x^2/2} \cdot f''(x)]$ be square integrable over the real line. Let us

denote by

$$M_n(f) := E_f[\hat{d}_n(\xi_1, \ldots, \xi_n) - d(f)]^2$$

the mean square error of $\hat{d}_n$ *at* f. *In the case* $s = 0$ $\lim_{n\to\infty} M_n(f) = 0$ *for every choice of* $q_n = 0(n^{1/4})$, *with* $q_n \to \infty$, *for* $s \geq 1$ *this holds even for every* $q_n = 0(n^{1/3})$, $q_n \to \infty$.

More precisely, the appropriate choice of q_n *and the corresponding speed of convergence of* $M_n(f)$ *are given in the following table:*

s	q_n	$M_n(f)$
0	$o(n^{1/4})$	$o(1)$
1	$o(n^{1/3})$	$o(1)$
2	$n^{1/4}$	$O(n^{-1/2})$
3	$n^{1/5}$	$O(n^{-4/5})$
4	$n^{1/6}$	$O((\log n)^2/n)$
≥ 5	*between* $n^{1/(2s-2)}$ *and* $n^{1/6}$	$O(n^{-1})$

Lemma 1 (Bosq [3]). *There is a constant* $a_1 > 0$ *such that for each* k *and* $l \geq 0$

$$\sup_{x \in R} \left| \frac{d^l}{dx^l} \psi_k(x) \right| \leq a_1 \cdot [2 \cdot (k + l)]^{l/2}.$$

Lemma 2. Let $\varphi_i(x) := \psi_i''(x)$, *for short. Then for the variance of* d_n, *formed with* f, *the following expression is valid:*

$$V_f \hat{d}_n(\xi_1, \ldots, \xi_n) = \{2/[n \cdot (n-2)]\} \cdot \sum_{k,l=0}^{q_n} [E_f(\varphi_k(\xi) \cdot \varphi_l(\xi))]^2 +$$

$$+ \{4 \cdot (n-2)/[n \cdot (n-1)]\} \cdot \sum_{k,l=0}^{q_n} c_k \cdot c_l \cdot E_f[\varphi_k(\xi) \cdot \varphi_l(\xi)] -$$

$$- \{(4n-6)/[n \cdot (n-1)]\} \cdot \sum_{k,l=0}^{q_n} c_k^2 \cdot c_{l_i}^2.$$

This is essentially Lemma 1 of Schüler and Wolff [13].

Lemma 3 (Schwartz [14]). *Let* f'' *be* s*-times differentiable and the function*

$$x \mapsto e^{x^2/2} \cdot \frac{d^s}{dx^s}[e^{-x^2/2} \cdot f''(x)]$$

be in L_2. Then $|c_k| \leq b \cdot k^{-s/2}$ for every $k \geq 1$, where $b > 0$ is a constant depending on f'' and s.

Proof of the theorem. First we note that

$$M_n(f) = V_f \hat{d}_n(\xi_1, \ldots, \xi_n) + \left[\sum_{k=q_n+1}^{\infty} c_k^2\right]^2.$$

Substituting the formula of Lemma 2 into this expression, we have four sums to be estimated. a_i are positive constants.

First term: by Lemma 1,

$$\sum_{k,l=0}^{q_n} E_f[\varphi_k(\xi) \cdot \varphi_l(\xi)]^2 \leq 4 \cdot a_1^2 \cdot \sum_{k,l=0}^{q_n} [(k+2) \cdot (l+2)]^2 =$$

$$= 4 \cdot a^2 \cdot \left[\sum_{k=0}^{q_n} (k+2)^2\right]^2 \leq a_2 \cdot q_n^6,$$

hence the first term is $O(q_n^6/n^2)$.

Second term: we recall the formula

$$\sum_{i=1}^{m} x^p \leq \begin{cases} [(m+1)^{p+1} - 1]/(p+1) & \text{for } p \geq 0 \\ [m^{p+1} + p]/(p+1) & \text{for } -1 < p \leq 0 \\ \log m + 1 & \text{for } p = -1 \\ a_3(p) & \text{for } p < -1 \end{cases}$$

which will be applied in the following. By Lemma 1 and Lemma 3,

$$\left| \sum_{k,l=0}^{q_n} c_k \cdot c_l \cdot E_f[\varphi_k(\xi) \cdot \varphi_l(\xi)] \right| \leq a_4 \cdot \left[\sum_{k=0}^{q_n} k^{-s/2} \cdot (k+2)\right]^2.$$

Now

$$\sum_{k=0}^{q_n} k^{-s/2} \cdot (k+2) = \begin{cases} O(q_n^{2-s/2}) & \text{for } s = 0, 1, 2, 3 \\ O((\log q_n)^2) & \text{for } s = 4 \\ O(1) & \text{for } s \geq 5. \end{cases}$$

This yields the order of the second term:

$$\begin{cases} O(q_n^{4-s}/n) & \text{for } s = 0, 1, 2, 3; \\ O((\log q_n)^2/n) & \text{for } s = 4; \\ O(1/n) & \text{for } s \geq 5. \end{cases}$$

Third term:

$$\{(4n-6)/[n\cdot(n-1)]\}\cdot\sum_{k,l=0}^{q_n} c_k^2\cdot c_l^2 = \{(4n-6)/[n\cdot(n-1)]\}\cdot\left[\sum_{k=0}^{q_n} c_k^2\right]^2 \leq$$

$$\leq \{(4n-6)/[n\cdot(n-1)]\}\cdot\left[\sum_{k=0}^{q_n} k^{-s}\right]^2 = \begin{cases} O(q_n^2/n) & \text{for } s=0 \\ O((\log q_n)^2/n) & \text{for } s=1 \\ O(1/n) & \text{for } s\geq 2. \end{cases}$$

Since $f''\in L_2$, the *fourth term* $\left[\sum_{k=q_n+1}^{\infty} c_k^2\right]^2$ converges to zero with $q_n\to\infty$ for every $s\geq 0$. For $s\geq 2$ a nontrivial bound can be established:

$$\sum_{k=q_n+1}^{\infty} c_k^2 \leq \sum_{k=q_n+1}^{\infty} k^{-s} \leq \int_{q_n}^{\infty} x^{-s}\,dx = \frac{q_n^{1-s}}{s-1},$$

hence $\left(\sum_{k=q_n}^{\infty} c_k^2\right)^2 = O(q_n^{2-2s})$.

Compiling these four bounds, the result follows.

q.e.d.

The following lemma is the basis of the rejection procedure for the generation of random numbers as described in the introduction.

Lemma 4 (Văduva [15], *Lemma* 3.2). *Let ξ and η be one-dimensional random variables with probability densities f and g respectively; we assume that $f(x)>0$ implies $g(x)>0$ and define $f(x)/g(x):=0$ if $f(x)=0$. Let an $\alpha>0$ exist such that $f(x)/g(x)\leq\alpha$ and let ζ be uniformly distributed in $[0,1]$ and independent of η. Then the conditional distribution of η given $0\leq\zeta\leq f(\eta)/[\alpha\cdot g(\eta)]$ has density f.*

References

1. *Bhattacharya, P.*, Estimation of a probability density function and its derivatives. Sankhyā Ser. A. **29** (1967), 373–382.
2. *Bhattacharya, P., Roussas, G. R.*, Estimation of a certain functional of a probability density function. Skand. Aktuarietidskr. **52** (1969), 201–206.
3. *Bosq, D.*, Contribution à la théorie de l'estimation fonctionelle. Publ. Inst. Statist. Univ. Paris **19** (1970), fasc. 2, 1–96.
4. *Dmitriev, Yu. G., Tarasenko, F. P.*, On the estimation of functionals of the probability density and its derivatives. Theor. Probab. Appl. **18** (1973), 628–633.

5. *Dmitriev, Yu. G., Tarasenko, F. P.*, On a class of nonparametric estimates of nonlinear functionals of the density. Theor. Probab. Appl. **19** (1974) 390–394.
6. *Epanechnikov, V. A.*, Nonparametric estimation of a multidimensional probability density. Theor. Probab. Appl. **14** (1969), 153–158.
7. *Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F. G.*, Higher Transcendental Functions, II. Bateman Manuscript Project, McGraw-Hill, New York, 1953.
8. *Mirzahmedov, M. A., Hašimov, Š. A.*, On the estimate of a probability density. Theor. Probab. Math. Statist. **9** (1976), 145–153.
9. *Mirzahmedov, M. A., Hašimov, Š. A.*, On some properties of density estimation. Colloquia Mathematica Societatis János Bolyai, Budapest, 1974, 535–546.
10. *Rosenblatt, M.*, Curve estimates. Ann. Math. Statist. **42** (1971), 1815–1842.
11. *Sansone, G.*, Orthogonal Functions. Interscience, New York, 1959.
12. *Schneider, B., Wertz, W.*, Statistical density estimation: a bibliography. Internat. Statist. Rev. **47** (1979), 155–175.
13. *Schüler, L., Wolff, H.*, Zur Schätzung eines Dichtefunktionals. Metrika **23** (1976), 149–154.
14. *Schwartz, S.*, Estimation of a probability density by an orthogonal series. Ann. Math. Statist. **38** (1967), 1261–1265.
15. *Văduva, I.*, Modele de Simulare cu Calculatorul. Editura Tehnica, Bucureşti, 1977.
16. *Wertz, W.*, Statistical Density Estimation — A Survey. Vandenhoeck & Ruprecht, Göttingen, 1978.
17. *Сираждинов, С. Х., Мирзахмедов, М. А., Носни, А. К.*, Об оценке плотности вероятности по зависимым выборкам. ДАН УзССР **1977**, *2* (1977), 3–6.
18. *Хашимов, Ш. А.*, О некоторых свойствах непараметрических оценок плотности вероятности. ДАН УзССР **1977**, *1* (1977), 3–5.
19. *Хашимов, Ш. А.*, О некоторых свойствах гауссовских процессов, связанных с непараметрической оценкой плотности распределения. Trans. Eighth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Academia, Prague, 1978, 319–323.
20. *Ченцов, Н. Н.*, Оценка неизвестной плотности распределения по наблюдениям. ДАН СССР **147**, *1* (1962), 45–48.
21. *Ченцов, Н. Н.*, Статистические решающие правила и оптимальные выводы. Теория вероятностей и математическая статистика. «Наука» Москва, 1972.

Заметка об оценке функционала плотности вероятности

В. ВЕРТЦ

(Вена)

Функционал $d(f) = \int (f''(x)^2\, dx$, где f — неизвестная плотность вероятности, играет важную роль в непараметрическом оценивании плотности. Употребляя разложение Хермита f'', строится оценка для $d(f)$ и доказывается, что она состоятельна в среднем квадратическом.

W. Wertz

Institut für Statistik und Wahrscheinlichkeitstheorie

Technische Universität Wien

Karlsplatz 13

A-1040 Wien

Österreich

Problems of Control and Information Theory, Vol. 10 (4), pp. 287—291 (1981)

ARTIFICIAL INTELLIGENCE

(Second International Workshop in Repino)

Second International Artificial Intelligence Workshop was held in Repino (near Leningrad) from 12 to 19 October, 1980. 150 Soviet and 40 foreign researchers from Italy, France, Finland, India, England, Canada, Sweden, Japan, Czechoslovakia, Bulgaria, GDR, Yugoslavia, Hungary and Poland took part in the workshop activity.

The workshop had the following sessions:

- artificial intelligence methodology,
- knowledge representation and data bases,
- interactive and dialogue systems,
- robots and walking machines,
- game problems and planning,
- artificial intelligence applications.

59 papers were presented at the workshop by 32 Soviet and 27 foreign researchers.

Methodology of artificial intelligence was considered in papers of G. S. Pospelov (USSR), V. V. Martynov (USSR), A. Andrew (England) M. Colombetti (Italy), V. Masnikos (Yugoslavia), M. N. Vaintsvaig (USSR) and V. V. Chavchanidze (USSR).

G. S. Pospelov stressed strong relation between system analysis and artificial intelligence which manifests itself, in particular, in the case of frame knowledge representation. As knowledge representation is closely related to the mathodological problems of artificial intelligence, it is difficult to draw a boundary line between them.

Paper of V. V. Martynov (USSR) also may be classified as methodological. It is dedicated to profound semantic representations of texts. Universal semantic code (USC) proposed by the author enables coding of any text segment (SAO).

The paper of A. Andrew (England) discusses the "everlasting problem" of relation between discrete and continuous as applied to artificial intelligence problems. A hypothesis about structure and organization of knowledge in human brain was presented by Colombetti (Italy).

The gist of V. Masnikos' paper was an attempt to use multilayer perceptron supplied with more complicated set of operations for design of artificial intelligence systems.

Very interesting paper of M. N. Vaintsvaig (USSR) may be also regarded as methodological because it tries to construct an artificial intelligence system built around modelling of associative subconcious processes taking place in human brain and their interactions with concious level. Artificial intelligence systems of this type seem to be very promising.

Direct relation to methodological problems has the paper of V. V. Chavchanidze (USSR) "Four-dimensional space-time conceptualized modelling of system evolution".

Paper of I. Kellemen (Czechoslovakia) belongs to the group devoted to knowledge representation. It presents a classification according to some attributes of various knowledge representations (semantic networks, scenaries, frame-like structures). Extension of the knowledge representation system was the topic of the papers presented by A. A. Koz'min (USSR) and D. Ribbens and his co-authors (Belgium).

A. A. Koz'min proposes a semantic network language with operators in the form of pragmatic aim factors. D. Ribbens discusses ways to representation of hierarchical terminological systems in associative storage structures enabling more effective organization of descriptive data fetch systems.

A program for classification of object frames with respect to their structures was discussed by F. Sobik and E. Zommerfeld (GDR).

Finally, Helbig (GDR) presented a paper on "Semantic representation of data given in natural language" where a system is discussed which approaches data representation in natural language as its structure becomes more involved.

Session "Interactive dialogue systems" had several interesting papers devoted to

- communication with various systems (planning of control, design and learning) in natural language,
- solution of problems by their description,
- generation of models by their description in natural language.

V. Sembugamurthy (India) considered RLAS system which was used for realization of language education models for pre-school children.

Paper of F. Chine and J. Guyo also was devoted to a system able to teach synthesis of functions and predicates. A dialogue system for programming teaching was described in the paper by E. Papon (France) and A. M. Kondratov (USSR).

Paper of N. Cordie (France) was devoted to a design technique for pragmatically oriented dialogue system for narrow application areas such as special-purpose robot control system.

Separate aspects of dialogue systems were considered by Polish researchers M. Perkovsky (MULTCOMP program) and A. Salvicky.

L. I. Mikulich (USSR) presented DISPUT system used for communication with end users in natural language.

Finally, V. F. Khoroshevskii (USSR) discussed instrumental set for design of natural-language processors for dialogue systems.

Papers dedicated to robot engineering aroused great interest. Prof. T. Vámos (Hungary) reviewed studies on use of technical sight for object identification, localization, trajectory planning, putting parts into contact and fine adjustment. Laboratory studies make use of a 6-degree orthogonal manipulator with control system based on one powerful computer and four microcomputers.

Paper of D. E. Okhotsimskii presented some new results on mathematical and physical modelling of integral moving (walking and wheeled) robot. Major problems encountered in design of walking robots were formulated, and new results were discussed in detail which had been obtained in using technical sight systems in the sensor unit of a moving integral robot.

E. Paghello and G. Trainito (Italy) considered a three-level control system of assembling robot. Major attention was paid to the development of intermediate level software which is translator of a sort between upper and lower levels. It is realized on computer coordinating operation of lower level microprocessors.

Paper of D. E. Devyanin and V. S. Gurfinkel presented results of studies of robot manipulator having three-component force cells. Signals of the cells were demonstrated to enrich robot's functional capabilities.

Paper of Prof. J. Giralt and R. Chatille (France) described moving robot HILARE equipped with TV camera, ultrasonic near location transducers and infrared navigation transducer. Topological properties of the environment are given to the upper level of control while routing is done through connectivity graph and data coming from sensor systems as the robot moves.

Prof. J. Giralt demonstrated also a film showing effective use of manipulator mechanical compliance in assemblying.

The paper of V. Schoenborn (GDR) described a system for visual determination of position and orientation of single part moving along conveyor. The

latter stops when a part gets into the field of TV camera vision, and the part is taken from the conveyor by manipulator according to signals from the camera.

S. K. Dulin describes an approach to development of algorithms for visual perception of robot-assembler. The problem of data capture for assembling was posed at the stage of assembly planning when even assembly sequence was defined. The problem solved by the author is classified as belonging to the class of computer experiments.

G. Jappinen (Finland) discussed trade-offs between fast, but rigid programs for industrial robots and flexible, but slow artificial intelligence systems. The compromise as seen by the author is program forming in robot a set of habits. An important place in robot programming is given to learning with the aid of an instructor.

Results obtained by A. V. Timofeev and his collaborators witness to effective application of adaptive manipulator control algorithms in cases of obstacle by-passing under indeterminate conditions of loading and interaction with soil.

In the paper of M. Vukobratovic "Dynamic robot and manipulator control algorithms and their application" consideration is given to a two-level system of control with decentralized distributed suboptimal control at lower level based on exact mechanism model. The paper is illustrated by a physical device with six degrees of freedom.

From the papers treating theorem proving and programming automation, those by V. M. Gluskov, Yu. V. Kapitonova and A. A. Lyutichevskii ("Mathematical informational environment and artificial intelligence system design") were devoted to software design automation attracts attention. Into the same category belongs the paper of N. N. Nepeivoda devoted to problems involved into logic programming regarded as search of proofs in appropriate constructive theory and program generation on the basis of these proofs.

Programming automation was also considered in the paper of N. I. Il'in and V. S. Strizhevskii "Intellectual programming systems" devoted to REGENT system integrating synthesis of program synthesis automation system understanding tasks in natural language with use of intellectual data base.

An interesting information given by M. V. Donskoi about the last chess program tournament should be classified as paper devoted to combinatorial and chess problems. Papers of M. M. Botvinnik and his collaborators working out PIONER system, and G. M. Adel'son-Vel'skii and A. L. Futer witness to an important trend to use chess players' experience in selection programs which dramatically reduces selection especially in the case of PIONER program.

It should be noted that new selection reduction methods obtained in chess programs are used, as it may be seen from the papers, in combinatorial

planning algorithms and scheduling theory. These methods have found use, in particular, in planning power station equipment maintenance repair.

A series of papers was devoted to applied issues of artificial intelligence. Among them R. Seregy's one (Hungary) on realization and popularity in his country of PROLOG language may be mentioned.

Recognition problems were discussed by G. V. Avaliani in his paper "Human experience as applied to recognition system design".

The problem of fuzzy sets as posed and developed by Zadeh is traditional to artificial intelligence meetings. It was discussed by D. I. Shapiro and I. V. Ezkhova who proposed a technique enabling one to use jointly objective data and subjective estimates expressed in fuzzy set terms.

N. G. Zagoruiko (USSR) treated an applied program package for processing of experimental data appropriately tabulated. Programs isolate informative attributes of gained data and identify by their analysis existing regularities. They also enable process forecasting.

Along with basic programs, the package also involves service programs including those which aid users in formulation of their problems. The package is widely used in various areas such as medicine, geology, process control systems, etc.

The papers presented at the workshop featured high-theoretical level in combination with orientation to applications, thus attesting that the ideas and methods of the artificial intelligence penetrate into design of physical computerized systems used in various industries.

F. M. Kulakov

Problems of Control and Information Theory, Vol. 10 (4), pp. 293—298 (1981)

VIIIth ALL-UNION CONFERENCE ON CONTROL SCIENCES (USSR, Tallinn, October 1980)

The VIIIth All-Union conference on control sciences was held in Tallinn (USSR) from 6 to 10 October, 1980. It was sponsored by the USSR and Estonian Academies of Sciences, the USSR National IFAC Committee, the Institute of Cybernetics of Estonian Academy of Sciences and the Tallinn Polytechnical Institute.

The conference summarized the development of the control theory and its applications during three years passed since the VIIth All-Union Conference. Wide exchange of data, useful discussions on new basic theoretical and applied results enabled definition of major directions for control science development. Researchers, experts from industrial institutes, scientific industrial combines and enterprises, scientists from academic and scientific institutes participated at the conference.

The conference was opened by the paper of Academician V. A. Trapeznikov "On some trends in the development of control systems". He emphasized the importance of contacts between science and industry. First-day plenary papers were delivered by Academician B. G. Tamm ("Role of computer engineering in automatic control systems"), Academician of Latvian Academy of Sciences, E. Ya. Yakubatitis ("Computer network architechture"), Academician O. M. Belotserkovsky ("Computer experiment and the problem of personnel training for control applications"), and Dr. M. A. Aizerman ("Problems of voting theory").

The conference had the following sessions: 1. Dynamics of control systems, 2. Methods for functional and structural optimization, 3. Data processing and control in stochastic systems, 4. Computer simulation of control systems, 5. Identification of complex systems, 6. Adaptive control, 7. Methodology of design of control systems for processes, enterprises, industries, process control system cost-effectiveness, 8. Methods for processing and representation of data arrays in control systems (picture processing, identification of informational parameters, preprocessing and data compression), 9. Control of economic and management systems, 10. Control in biology and medicine, 11. Control of moving objects, 12. Design of man-machine systems, 13. Control

of robots and robot-engineering systems, 14. Design automation for control systems and hardware, 15. Automation of scientific researches, 16. Design of systems and networks for data processing and control, 17. Design philosophy for control hardware and software, 18. Control of economic systems, 19. Theoretical problems of reliability, technical diagnostics of control systems, 20. Systems of logic control.

For the convenience of reviewing, all the topics were categorized into four groups: control theory (Sessions 1 to 3, 5, 6 and 20), simulation and computerized automation (Sessions 4, 7, 8, 12, 14, 15), control for particular application classes (Sessions 9 to 11, 13, 18), and development of control means (Sessions 16, 17, 19).

Session 1 from the first group had four surveys: M. A. Krasnosel'ski ("On some problems of the nonlinear system theory"), V. I. Utkin ("Sliding models in control systems"), A. A. Voronov ("State-of-art and problems in the stability theory"), and N. A. Kuznetsov ("Controllability of dynamic systems").

Papers at Session 2 were presented by Yu. N. Andreev ("Difference-geometrical methods in the control theory") and R. P. Fedorenko ("Methods for approximate solution of optimal control problems").

Papers of V. S. Pugachev ("Application of stochastic differential equations") and E. S. Kochetkov ("Controllable queueing systems") were presented at Session 3.

At Session 5, papers were presented by I. E. Kazakov ("Classification of identification methods"), V. P. Zhivoglyadov ("Identifier control systems"), and N. S. Raibman and V. M. Chadeev ("Current trends in identification studies").

Papers of I. B. Yadykin ("Searchless methods") and B.T. Polyak ("Search methods") were delivered at Session 6.

A. A. Tal' gave at Session 20 presentation of his "Study of logic control systems".

Results of studies witness intensive efforts applied to these areas of the control theory. Adaptive approach to control is at present widely accepted. Studies were carried out on dynamic plant adaptive control concerning estimation of control quality and use of adaptive state observers. Promising studies concerning use of a priori information for optimization of adaptive algorithms were made for statistical control plants.

Control system design methods were proposed based on application of the MIS philosophy to the domain of identification. Systems of direct digital control of complex plants for which there is no mathematical model were designed and implemented on the basis of these methods. New computational techniques and optimization algorithms were developed. Studies has been

started on the control theory for quantum mechanical systems and mobile control systems, on optimization of thermodynamic processes and stabilization of plasma configuration.

New results were obtained in the theory of optimal control of statistical and dynamic systems of various classes, on optimization and synthesis of closed systems (multivariate, distributed, variable-parameter, linear, nonlinear). New methods were developed for optimal control synthesis, some generalizations to the maximum principle were found, theory of degenerate problems and special modes of optimal control was developed. For stochastic systems some optimal control and optimal observation problems were solved, new methods for study of queueing systems were worked out, theory was developed for processing multidimensional information in complex systems. Study has been started on program realization of logic control systems.

The conference demonstrated the need for further development of the general system theory and, on its basis, of the classical areas of the control theory.

It was noted that the scope of economic problems solved by means of computers has widened essentially. The developed optimization software is used in management information systems (MIS) for enterprises, industries. Papers of Kalashnikov ("Problems of simulational modelling of control systems") and B. Ya. Kogan et al. ("General problems of computer simulation") have demonstrated wide study of analogue, digital and hybrid means for computer simulation, of simulation languages, applied software enabling one to pass to problem-oriented simulation systems.

Session 7 had papers of O. I. Aven ("Methodology of MIS design") E. L. Itskovich ("MIS in industry"), A. G. Mamikonov ("Nonindustrial MIS"), and K. B. Norkin ("Topical issues of agricultural production control").

Papers of I. B. Muchnin ("Structural methods for analysis of empirical image curves") and S. M. Borodkin ("Structural methods for analysis of large arrays of empirical data") were presented at Session 8.

New results were obtained in the situational control theory, man-computer communication, especially in systems oriented to understanding of limited natural language. Paper "Problems of man-computer system design" was presented by A. G. Chachko at Session 12. Progress is observed in studies of problem on "human operator in control system", methods of engineering and psychological design. Ways were proposed to reduce number of errors in operator's actions under normal operating conditions, and also were proposed effective procedures for interaction in emergency situations.

The range of economic problems solved by modern computers, has dramatically extended. Developed optimization software is used for enterprise and industrial MIS.

Application of computers to industrial problems is continuously growing. Papers of V. L. Epstein ("Computer-aided design of data processing systems") and V. V. Solodovnikov ("Computer-aided design of process control systems") presented at Session 14 have demonstrated that currently there is no industry without MIS and process control systems at enterprises and industrial combines.

Further progress was made in the area of study and development of computer-aided design for control systems. Session 15 had papers of G. K. Krug on "Current problems of research automation" and G. A. Fomin on "Mathematical support of research automation systems".

It was noted that studies of languages, development of formalized methods and optimization procedures, development of user-oriented design methodologies for cases with no data supporting design process, improvement of various applied program packages, etc. are most topical problems in the area of theory, methodology and program realization of computer-aided design systems.

Among most important problems requiring further study are development of unique methodology of computer-aided design of MIS satisfying requirements of simplicity and convenience, able to be modernized, cost-effective, development of methods of substantiated formulation of guidelines for design of large MIS, development of realization methods for large MIS projects and estimation of their effectiveness.

Five papers on various problems were presented at Session 9: "Control in management systems" byV. N. Burkov, "Mathematical description of socio-economic systems" by A. A. Petrov, "Multilevel planning schemes. Algorithms for economic control in the presence of disturbances" by V. V. Tokarev, "Numerical methods for economic planning problems" by A. P. Uzdemir and "Decision making theory" by V. A. Irikov.

All efforts of researchers and engineers should be concentrated on attaining reasonable trade-off between studies on fundamental control theory problems, new highly effective control hardware, new advanced technology and studies of control in involved management and industrial-economic systems. Studies should be oriented to stepwise growth.

At Session 10, A. M. Petrovskii delivered a paper on "System approach to control in biology and medicine", a paper on "Control of flying vehicles" was presented at Session 11, and at Session 13 Yu. M. Svirezhev read his paper dealing with "Problems in economic system control". Studies of these scientists have born out extensive application of control-theoretical results to health protection problems and wide application of system approach to health protection and simulational modelling of medical and biological processes.

Significant increase of the level of studies in robot engineering area was demonstrated at Session 13 by papers of E. P. Popov and K. S. Yushchenko ("Manipulational and industrial robots") and D. A. Pospelov ("Artificial intelligence problems"). Industrial robots are continuously improving and introduced in various industries. Work is going on theory and methods of control for sensitized manipulators and mobile robots. Progress has been made in design of algorithms and firmware for basic subsystems of integrated robots: perception and analysis of environmental data, actuator control action planning. New results were obtained by the situational control theory, in design of interactive systems especially those oriented to understanding of limited natural language. Studies of human operator in control systems proved to be successful. Methods were developed for structural analysis and synthesis of man-machine control systems, for engineering-psychological design. Ways were proposed to reduction of operator's errors under normal operating conditions, and effective procedures were developed for interactions in emergency situations.

Development of methods for optimization and decision making in economic problems involved into large national programmes, and continuation of work on software for optimal economic planning and integration of the software into operating and designed control systems are major study areas in control of industrial-economic and management systems. These issues were treated at Session 16 by V. A. Pirogov ("Problems of hard- and software for computer networks") and V. M. Vishnevskii ("Problems of flow control and topological optimization of computer networks"). At Session 17 papers were presented by I. V. Prangishvili ("Design philosophy of advanced highly-effective computer systems") and V. V. Lipaev ("Major problems and current state of work on programming technology for control systems"). Papers on "Major tasks of technical diagnostics "were delivered by P. P. Parkhomenko and V. P. Grabovetskii.

The conference noted significant progress of computer engineering. Progress has been made by the theory of uniform rearrangeable-structure systems which are now commercially available; multiprocessor computer systems are designed; theory of microprocessor systems is developed; effective hybrid computer systems and automatic equipment for computation control are created; progress was made by the theory of system programming enabling generation of effective translation systems, improvement of operating systems and optimization of complex programs. New sets of hardware for process control were designed. New methods were developed for determination of statistical estimates of reliability of characteristics obtained by tests; reliability of computer system software was considered; system synthesis methods based on reliability criteria were developed; synthesis methods were created for built-

in self-checking circuits and for organization of fault-tolerant structures of computer systems; work was continued on computer-aided design of diagnostic support.

Further studies should be oriented to identification of new ideas for design of highly effective hardware, design and practicalization of highly effective computer systems, primarily microprocessor rearrangeable-structure ones; to development of data base computer networks and distributed systems; development and use of computer systems for analysis and estimation of reliability data obtained through tests, methods for estimation and improvement of MIS reliability by means of built-in self-checking testing equipment and equipment for self-repair and test diagnostics; development and practicalization of methods for system availablility analysis methods.

Design of unified set of devices of the State instrument system is one of most topical tasks in hardware design.

Animated roundtables devoted to "Objectives, problems and perspectives of control sciences" (lead by B. G. Tamm, I. M. Makarov and I. V. Prangishvili) demonstrated that wide application of experience gained by theory and practice of control would enable design of better control systems for various applications, enhance scientific and technological progress and significantly improve labour effectiveness and quality.

N. I. Rodionova

BOOK REVIEW

Optimization Techniques. Parts 1–2. Edited by **K. Iracki, K. Malanowski** and **Walukiewicz.** Lecture Notes in Control and Information Sciences Series. Vols. **22–23.** Ed.: A. V. Balakrishnan and M. Thoma. Springer-Verlag.

These Proceedings contain most of the papers presented at the 9th IFIP Conference on Optimization Techniques held in Warsaw, Poland on September 4–8, 1979.

The Conference was sponsored by the IFIP Technical Committee on System Modelling and Optimization. It was organized by the System Research Institute of the Polish Academy of Sciences (PAS) with cooperation of the Mathematical Institute of the PAS and the Institute of Automatic Control of the Technical University of Warsaw.

The Conference was attended by 284 scientists from 29 countries. The Conference Program comprised 8 plenary lectures, 131 contributed papers classified into 15 sections, and a Round Table Session on Systems Techniques in Economics.

The program offered a broad view of recent developments in theory and computational methods of optimization and their applications in various fields of science and technology. The emphasis was on advances in optimal control and mathematical programming techniques as well as their applications to modelling and control in particular in economics, and environmental and energy systems.

The Proceedings are composed of two volumes.

The first volume contains plenary lectures:

— Optimal Control Problems in Aeroelasticity by A. V. Balakrishnan,
— Optimization of Engineering Structures, Theory and Applications by W. Gutkowski,
— Mathematic-Economic Modelling of Scientifical and Technical Progress by L. Kantorovich,
— Game-Theoretical Optimization of Differential Systems by N. N. Krasovskii,
— Optimization of Regional Development — Integrated Models for Socio-Economic and Environmental Planning by R. Kulikowski,
— On the Foundations of the Optimal Control of Distributed Systems by J. L. Lions,
— Optimization Algorithms in 1979 by M. J. D. Powell,
— A Methodological Guide to Multiobjective Optimization by A. P. Wierzbicki,

panel addresses of the Round Table Session on Systems Techniques in Economics:

— Interactions Between Economics and Systems Theory by M. D. Intriligator,
— Modelling Man in Economics and System Theory by R. F. Drenick,
— The Importance of Mathematical Optimization in Economics by L. Kantorovich,
— The Global 2000 Study: An Attempt to Increase Consistency in Government Forecasting by J. M. Robinson,

and the contributed papers dealing with the following main topics (in brackets the number of papers)

— Stochatics Control (7)
— Differential Games (8)

— Optimal Control: Ordinary and Delay Differential Equations (12)
— Optimal Control: Partial Differential Equations (12)
— Multiobjective Problems (6)
— Applications: Biomedical Systems (5).

In the second volume there are collected papers devoted to the topics:

— Mathematical Programming: Theory (3)
— Mathematical Programming: Algorithms (14)
— Integer Programming (9)
— Software Problems (5)
— Graph and Networks (9)
— Scheduling and Allocation Problems (8)
— Applications: Economics and Econometrics (8)
— Applications: Environmental and Energy Systems (8)
— Applications: Industrial Processes (6).

Extensions of Linear-Quadratic Control Theory by **D. H. Jacobson, D. H. Martin, M. Pachter** and **T. Geveci.** Lecture Notes in Control and Information Sciences Series, Vol. **27**. Ed.: A. V. Balakrishnan and M. Thoma. Springer-Verlag.

These Lecture Notes give a concise survey of the theory of multivariable linear-quadratic optimal control and estimation and some details of various extensions to the theory developed over the past five years at National Research Institute for Mathematics and Sciences (NRIMS) of the Council for Scientific and Industrial Research, Pretoria.

The following lectures are included:

1. Review of Linear Dynamic Systems
2. Canonical Forms, Pole Assignment and State Observers
3. Lyapunov Stability Theory
4. Linear-Quadratic Optimal Control
5. Introduction to Kalman Filtering
6. The Maximum Principle and the Hamilton–Jacobi–Bellman Equation
7. The Non-convex Case
8. Controllability Subject to Controller Constraints
9. Linear-Quadratic Problems with Canonical Control Set
10. Special Non-Linear-Quadratic Formulations
11. Hybrid Criteria and Parameter Sensitivity
12. Conditional Definiteness of Quadratic Functionals
13. Exponential Performance Criteria and State-Dependent Noise
14. Differential Games
15. Optimal Control of Partial Differential Equations.

While each lecture naturally contains some basic subject matter, the first six lectures and the last two are in the nature of comprehensive surveys, covering material that is somewhat scattered in textbooks, research monographs. On the other hand, Lectures 7 to 13 present the results of more recent research done in the Mathematics Division of NRIMS. The style presupposes some familiarity with real vectors and matrices, with systems of ordinary differential equations, and in Lectures 12 and 15, with a little basic Hilbert space theory. However, the flavour is decidedly applied mathematical, and these notes should be of use principally to those concerned with process control in industry, dynamic modelling and decision theory in economics and management science, and to members of departments of applied mathematics and engineering at universities and technical colleges.

Cs. Bányász

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

ENGLISH TRANSLATION OF THE PAPER IN RUSSIAN

Problems of Control and Information Theory. Vol. **10**, No. *4* (1981)

APPLICATION OF OPTIMIZATION METHODS TO RELATIVISTIC THEORY

V. G. BOLTYANSKIY

The paper investigates a control system which is a conceptual system obtained by relativistic reasoning whose mathematical space — time model has anisotropic properties. The light, or time-optimal, paths of that system and possible effects of "gravitational" focusing of light paths are described. Examples of astrophysical interpretation of these effects are given all of which, except the first one, are phenomena that are difficult to explain in the framework of Einstein's general relativity theory

Light is generally believed to propagate (in vacuo) at the same speed in all directions while "gravitational" particles, whose rest mass is nonzero, can only move at speeds lower than c in magnitude. Consequently, the motions of various particles are described by the inequality

$$|\dot{x}| \leq c, \tag{1}$$

where $x = (x_1, x_2, x_3)^T$ is the radius vector of the moving particle (the superscript T denotes hereafter transposition, or x is a column vector). Inequality (1) can also be written in the form

$$x = u, \quad |u| \leq c. \tag{2}$$

If, however, light propagates in a medium with a variable index of refraction rather than in vacuo, relations (2) change somewhat

$$x = \frac{u}{n(x)}, \quad |u| \leq c. \tag{3}$$

Note that relation (2) (or (3)) alone does not determine light paths. For however smooth the curve l is, a motion $x(t)$ can be imagined along that curve at a speed $|\dot{x}(t)| \equiv c$; this does not imply, however, that l is a light path. Possible light paths are defined by the Fermat principle whereby a line l is not a light path unless each its (small enough) segment AB is the path of the fastest movement from A to B from among all motions which satisfy condition

(2) (or (3)). To put it differently, if the system is described by relation (2) (or (3)) (x being the phase coordinate and u the control variable), then the light paths are locally time-optimal paths of the systems. If, for instance, $n(x)$ denotes the air index of refraction in the point x, the Fermat principle applied to plant (3) will permit computing refraction in the earth atmosphere.

Compare (3) with a plant of a general form

$$\dot{x} = f(x, u), \qquad u \in U. \tag{4}$$

If $F(x)$ denotes the set of all vectors of phase velocity which plant (4) can have in the point x

$$F(x) = \{f(x, u): u \in U\},$$

then (4) can be written as a differential inclusion

$$\dot{x} \in F(x). \tag{5}$$

In numerous examples considered in the framework of theory of optimal processes the sets $F(x)$ are not, as a rule, centrally symmetrical (if they are, their center of symmetry does not coincide, as a rule, with the origin of coordinates in the phase space), which makes all the difference between (5) and (2) or (3). This fact has induced the author to consider (initially, in purely mathematical terms) the implications of assumed light anisotropy, or possible difference in the speed of light propagation in two opposite directions.

Strange as it may seem, the truth of the anisotropy hypothesis cannot be tested experimentally; the existence or otherwise of anisotropy depends only on our wishes. To explain this, let us run a conceptual experiment. In two space laboratories which are far from, and immobile with respect to, one another, there are two spacemen, A and B. A locates the other spaceman, or sends a light signal to him and then receives the reflected signal. The propagation time there and back amounts to four seconds. Believing that the light signal took two seconds to travel to B and two seconds to travel back and assuming that the speed of light is about 300,000 km/sec, spaceman A finds that the distance between the laboratories is 600,000 km.

Once in contact, the spacemen intend to synchronize their clocks. Spaceman A tells the other of his system of time units: a second is a quarter of the location time (which spaceman B can measure himself); 60 seconds are a minute; 60 minutes are an hour, etc. Following this, spaceman A describes his plan of synchronizing the clocks. At 12 hours sharp he gives the agreed signal and at the time spaceman B receives the signal he sets his clock to $12^{\text{h}}00'\ 02''$ (i.e. 12 hours 00 minutes 02 seconds); so the clocks are synchronized in this way. Measurement of the speed of light the next day confirms the

existence of light isotropy. A light signal sent at $12^h\,00'00''$ was received at $12^h\,00'02''$ and the reflected signal arrived back at $12^h\,00'04''$; consequently, light covered the 600,000 km in either direction in two seconds, which implies that the speed of light in both cases was 300,000 km/sec.

Let us imagine, however, that spaceman B erroneously set the clock to $12^h\,00'01''$ when he received the signal from A. Then in measuring the speed of light the signal from A appears to be received at $12^h\,00'01''$ by B and the reflected signal, to return to A at $12^h\,00'04''$. In other words, from A the light covered the 600,000 km in one second and back, in three seconds, or the speed of light on the way out was 600,000 km/sec and on the way back, 200,000 km/sec, the ratio of these speeds being 3 : 1. Spaceman B could, if he wished, make that ratio equal to 1 : 3 by setting the clock to $12^h\,00'03''$ when he received the signal.

Consequently, the existence or nonexistence of light anisotropy (and the value of anisotropy, or the ratio of the speeds out and back) depend on the way the clocks are synchronized, or in the final analysis, on our own wish. There is no "correct" way to synchronize clocks to be preferred over others that would help to ascertain "objectively" the existence or otherwise of light anisotropy. For synchronization performed by sending a light signal whose speed is unknown in advance while for measuring the speed of light clocks synchronized in some way should be available. There is no way to break this vicious circle by any kind of physical experimentation because either a single clock is involved and the averaged speed of the emitted and returned light is measured or two clocks are placed in different points and need synchronization even if they are in the same room.

We can certainly require that only those space-time frames of reference be used which do not feature light anisotropy (in this way the method of clock synchronization would be univalently specified). But this arbitrariness (generally accepted in the relativity theory) is hardly justified because, as we will see, there are phenomena which are simpler and more naturally described in frames of reference which feature light anisotropy.

Einstein's relativity principle can be divided into two parts, namely a hypothesis of relativity, whereby any separate movement of light has the same speed in two frames of reference which move uniformly with respect to each other, and isotropy, whereby in one frame of reference the speed of light has the same magnitude in all directions (with proper choice of scales along the axes). From our reasoning above, it is obvious that we intend to reject isotropy and retain only the first part of the principle, relativity.

Before all, let us see how the Lorentz transforms change under anisotropy. For simplicity let us consider only a single-dimensional "universe", or

motion of particles along a certain straight line l. Each event (eruption, collision of two particles, etc.) has in a frame of reference l, specified on the straight line, a time and a coordinate which are univalently determined. A change from one frame of reference system S_α to another, S_β, is locally performed in compliance with linear formulae

$$\begin{cases} dt_\beta = A dt_\alpha + B dx_\alpha, \\ dx_\beta = C dt_\alpha + D dx_\alpha. \end{cases} \tag{6}$$

The motion is described in the system S_α by a relation of the form $x_\alpha = \varphi(t_\alpha)$; the number $v_\alpha = \dfrac{dx_\alpha}{dt_\alpha}$ is referred to as the speed of that motion. The speed of the same motion in the system S_β is

$$v_\beta = \frac{dx_\beta}{dt_\beta} = \frac{C dt_\alpha + D dx_\alpha}{A dt_\alpha + B dx_\alpha} = \frac{C + D v_\alpha}{A + B v_\alpha}. \tag{7}$$

A motion perceived in a frame of reference S_α as light moving rightwards (leftwards) along the straight line l is perceived as such in a frame of reference S_β as well. Also, consistently with relativity, the speed is the same in the frames of references S_α and S_β. Denote that speed by σ. In the same way the speed of light moving leftwards is the same in both systems S_α, S_β; it will be denoted as ϱ. Consequently, if $v_\alpha = \sigma$, then $v_\beta = \sigma$ (the same holds for ϱ) and so, by virtue of (7), the relations hold

$$\sigma = \frac{C + D\sigma}{A + B\sigma}, \qquad \varrho = \frac{C + D\varrho}{A + B\varrho}.$$

Further, let v denote the speed of the system S_α with respect to the system S_β. This implies that with $v_\alpha = 0$ (the particle is "at rest" in S_α) we have $v_\beta = v$. Consequently, by virtue of (7)

$$v = \frac{C}{A}.$$

Let us eliminate the denominators in the last three relations

$$\begin{cases} B\sigma^2 - C - D\sigma = -A\sigma, \\ B\varrho^2 - C - D\varrho = -A\varrho, \\ C = Av. \end{cases}$$

From this set of linear equations we have

$$B = -\frac{v}{\varrho\sigma} A, \quad C = vA, \quad D = \left(1 - v\left(\frac{1}{\varrho} + \frac{1}{\sigma}\right)\right) A.$$

Then for the determinant $\Delta = AD - BC$ of transformation (6) we have

$$\Delta = A^2\left(1 - \frac{v}{\varrho}\right)\left(1 - \frac{v}{\sigma}\right).$$

Assuming that the transformation (6) is equiaffine (retaining the areas), i.e. $\Delta = 1$ (physically this is equivalent to "coincidence" of scales in the frames of reference S_α and S_β [1]), we have

$$A = \frac{1}{\sqrt{\left(1 - \frac{v}{\varrho}\right)\left(1 - \frac{v}{\sigma}\right)}}. \tag{8}$$

Consequently, transformation (6) takes a form where A is given by equality (6)

$$\begin{cases} dt_\beta = \left(dt_\alpha - \dfrac{v}{\varrho\sigma}\, dx_\alpha\right) A. \\ dx_\beta = \left(v dt_\alpha + \left(1 - v\left(\dfrac{1}{\varrho} + \dfrac{1}{\sigma}\right)\right) dx_\alpha\right) A. \end{cases} \tag{9}$$

These are generalized Lorentz transformations. If isotropy is true ($\sigma = c$, $\varrho = -c$ or $\varrho + \sigma = 0$), relations (9) become usual formulae of Lorentz transformations. The quadratic form

$$dx^2 - (\varrho + \sigma)\, dx\, dt + \varrho\sigma dt^2 \tag{10}$$

is readily seen to be invariant with the transformations (8) and (9).

For further discussion a new notation should be introduced $c = \dfrac{|\varrho| + |\sigma|}{2} = \dfrac{\sigma - \varrho}{2}$, $f = \dfrac{\sigma + \varrho}{2}$. Then $\varrho = -c + f$, $\sigma = c + f$, $\varrho\sigma = f^2 - c^2$ and the quadratic form (10) is represented as

$$(dx - f dt)^2 - c^2 dt^2. \tag{11}$$

The quadratic form (11) becomes an Einstein quadratic form $dx^2 - c^2 dt^2$ with $f = 0$ (i.e. for the case of isotropy). What is more interesting, however, is that form (11) is reduced to the Einstein form by introduction of a fictitious coordinate x^* which is a linear combination of space and time $dx^* = dx - f dt$. Thence it is clear that formulae (8) and (9) could be obtained from Lorentz transformations by subtraction $dx^* = dx - f dt$. Geometrically this signifies that formulae (8) and (9) represent the Lorentz transformations in a non-orthogonal frame of reference. As a result, formulae (8) and (9) can be derived as linear equiaffine transformations which retain the invariance of the quadratic form (11). In particular, if an analogous formula is to be obtained a

three-dimensional case transformations which retain the invariant form should be taken up

$$(dx_1 - f_1\,dt)^2 + (dx_2 - f_2\,dt)^2 + (dx_3 - f_3\,dt)^2 - c^2\,dt^2.$$

Furthermore, the physical relations (such as the Maxwell equations) remain true in the anisotropic case; this has been confirmed by direct computation [2].

As for possible application of anisotropic relativity formulated in equations (8) and (9), the Michaelson experiment is sufficient to prove that uniform motion of one frame of reference with respect to another does not lead to light anisotropy. Would not acceleration of one frame of reference with respect to another (or the presence of a gravity field which is the cause of accelerations) lead to light anisotropy? Direct experimental testing of this possibility is out of the question because gravity fields and accelerations which are within our reach are negligible; nor a laboratory can be sent to the vicinity of some black hole. This does not imply, however, that gravitational anisotropy cannot be discussed in theoretical terms. The corollaries of this assumption will be seen to be confirmed experimentally, and this fact can be viewed as indirect test of the gravitation anisotropy assumption.

The simplest model is represented by an equation which is a substitute of (1)

$$\dot{x} = f + u, \quad |u| \leq c, \tag{12}$$

where the "transfer vector" $f = f(t, x)$ is a function of time t and the point $x = (x_1, x_2, x_3)^T$, the shape of the function being dependent on the distribution of masses and the frame of reference. A geometrical interpretation of equation (12) is that it describes a system (5) where the set $F(x)$ (or the set of all possible speeds at which particles can issue from the point x) is a sphere of radius c whose center is in the point f (Fig. 1).

If, in compliance with the Fermat principle, the light paths are still believed to be locally (time-) optimal paths of system (12), then to find the light paths, the maximum principle should be used. A function is built

$$H = \psi f + \psi u, \tag{13}$$

which is a scalar product of the auxiliary vector $\psi = (\psi_1, \psi_2, \psi_3)$ (which is a row vector) by the right-hand side of equation (12) and differential equations for the auxiliary vector

$$\dot{\psi} = -\frac{\partial H}{\partial x} = -\psi \frac{\partial f}{\partial x}. \tag{14}$$

According to the maximum principle, for optimality of the path $x(t)$ a non-trivial solution to equation (14) should exist such that the control $u(t)$ as-

sociated with the path $x(t)$ maximizes at each time t the function H for the variable u. From the shape of the function H (see (13)) it follows (by virtue of the maximum principle and the constraint $|u| \leq c$) that the control $u = u(t)$ is given by the equality

$$u = \frac{c\psi^T}{|\psi|}. \tag{15}$$

Consequently, equation (12) takes the form

$$\dot{x} = f + \frac{c\psi^T}{|\psi|}. \tag{16}$$

Equations (14) and (16) describe the light paths.

Let us consider a case of a single mass m fixed in the origin of coordinates. A simplest and most natural assumption is that the vector f is directed towards the origin of coordinates and its magnitude is $\frac{\gamma m}{|x|^2}$ where γ is a factor, or

$$f = -\frac{\gamma m}{|x|^3} \cdot x. \tag{17}$$

Consequently, system (12) takes the form

$$\dot{x} = -\frac{\gamma m}{|x|^3} x + u, \quad |u| \leq c. \tag{18}$$

From (17) the number

$$r_0 = \sqrt{\frac{\gamma m}{c}} \tag{19}$$

is seen to feature a property that with $|x| < r_0$ we have $|f| > c$, and so, in compliance with (19),

$$x^T \dot{x} = x^T(f + u) = -|x| \cdot |f| + x^T u < -|x| \cdot c + |x| \cdot c = 0,$$

or the vector $\dot{x}$ makes on obtuse angle with the radius-vector x. This implies that all light paths which start from any point of the sphere stay within the sphere $|x| < r_0$. In other words, no light path which starts inside the sphere can leave the sphere; this sphere will be referred to as a black hole. Its radius is given by formula (19).

A circle of radius $r_0 \sqrt[4]{3}$ in any plane which goes through the origin of coordinates and has its center there is easily seen to be a closed light path. Figure 2 shows computed (by L. K. Nikolaev [4]) light paths in the plane (x_1, x_2) which intersect the axis x_2 at a right angle. Figure 3 shows a bundle

of light paths from a remote point S; wherever a point A is, there is a light path which originates in the point S, envelopes the black hole O, and arrives in the point A. Figure 4 (which has been discussed in [3], [5]) reveals that if in the points $S_1, S_2, \ldots, S_k$ there are light emitting objects, or bundles of light paths originate from these points, then the observer in the point A receives light paths from all objects, $S_1, S_2, \ldots, S_k$. The black hole O focuses, as it were, the light from all sources $S_1, S_2, \ldots, S_k$, and directs it as a narrow beam (as if issuing from the vicinity of the point O) to the observer A. Inside that light beam a small black spot is observed whose angular dimension is $2r_0\sqrt[4]{3}/d$ where d is the distance from the observer A to the black hole O.

Finally, Fig. 5 shows one more phenomenon of focusing (see [5]) which occurs if the light emitting object S, the observer A and the black hole O are on the same straight line. In this case there is an infinite number of light paths from S to A. In each plane which intersects the straight line AS there are two "extreme" paths (which curve only slightly in passing by the black hole O) and an infinite number of paths which make different numbers of turns around O. In space (we limit ourselves to the extreme paths) the observer A sees the object S as a radiant ring of an angular radius α. If, however, the angle α is very small because of the distance, then A sees the object, as it were, through the black hole and many times brighter that he would if the black hole were not there because the area of the ring is many times that of the spot which would be seen if the object S were observed in direct beams. The computation of Fig. 6 shows that the factor by which the brightness increases thanks to focusing by the black hole O, or the ratio of the ring area to the spot area, is approximately $\frac{8R\alpha}{q} \approx \frac{8h}{q}$, where h is the distance from the extreme light path to the black hole O and q is the diameter of the telescope "pupil".

Now let us try to employ this mathematical reasoning in interpretation of certain astrophysical phenomena. More specifically, let us see what conclusions can be made if the properties of space-time are described in terms of paths of plant (18) rather than of Einstein's equations. To put it differently, the light paths of our reasoning will not be isotropic (zero) geodetic curves of the curved space which is assumed in the general relativity theory (for the case of a point-like mass the metric tensor of that space is specified by the so-called Swartzschield solution); they will be locally time-optimal paths of the plant (18).

Following [5], the factor γ in (18) is assumed to be a function of the electromagnetic radiation

$$\gamma = q\sqrt{\lambda}. \tag{20}$$

Curvature of light paths in the vicinity of the Sun

Observations of full eclipses reveal that the light path from a distant star is somewhat curved in passing by the Sun disc. The angle made by the initial light path and the one at which it reaches the earth amounts to about 1.7 angular seconds. This permits computing a constant γ, or choosing it so that with* $m = 2\times10^{33}$ (the Sun mass), $r = 7\times10^{10}$ (the Sun radius) and $c = 2.998\times10^{10}$ (the speed of light) the deflection angle of a light path at a distance r from the origin of coordinates (cf. Fig. 2) would amount to exactly $1.7''$ for plant (18). L. K. Nikolaev's computations [4] have yielded that value, $\gamma = 1.2\times10^{-4}$.

Consequently, choice of the desired value of γ makes the experimental data on curvature of light paths in the Sun gravitation field (which are regarded as experimental confirmation of the Einstein theory of relativity) consistent with the conclusions of the anisotropic relativity theory. In further discussion this value of γ will help in explaining many other experimental facts. To do this, let us assume that the computed value of γ is associated with a wavelength $\lambda = 5.5\times10^{-5}$ (which is the average value of the light's wavelength from $\lambda \approx 3\times10^{-5}$ to $\lambda \approx 7\times10^{-5}$). Substituting this value of λ (and the above value of γ) into formula (20) we have $q = 0.016$. Now, from (19) the formula for the black hole radius is

$$r_0 = \sqrt{\frac{\gamma m}{c}} = \sqrt{\frac{q\sqrt{\lambda}\, m}{c}} = 7.3\cdot10^{-7}\sqrt[4]{\lambda}\,\sqrt{m}\,, \tag{21}$$

and for $r_0\sqrt[4]{3}$ which (cf. Fig. 4) will be referred to as the focusing radius of the mass m, we have

$$r_0\sqrt[4]{3} = 9.6\cdot10^{-7}\sqrt[4]{\lambda}\,\sqrt{m}\,. \tag{22}$$

Radio emission of the "quiet" Sun

Since, according to (21), the black hole radius is proportional to $\sqrt[4]{\lambda}$, the same body is, depending on the dimensions and density, a black hole for some frequencies and is not for others. In particular, with $m = 2\times10^{33}$ from (21) and (22) we have for radio emission of free hydrogen ($\lambda = 21$) $r_0 = 6.99\times10^{10}$ and $r_0\sqrt[4]{3} = 9.2\times10^{10}$. In other words, the Sun is now already a black hole for radio frequencies. The radio emission which is generated in the

* Unless the unit is specified, all data are given in the CGS system.

Sun corona outside a sphere of radius $r_0\sqrt[4]{3}$ is focused by that sphere; thanks to this the sphere is perceived as a radio emission source which is brighter at the edge of the sphere than inside it. The radio emission seems to pour intensively from an imaginary sphere of radius 9.2×10^{10} (somewhat larger than the Sun radius, 7×10^{10}). This result is in good agreement with experimental observations. Figure 7 shows a plot (borrowed from [7], p. 96), of emission intensities (at a wavelength $\lambda = 21$) for different distances, the Sun radius being the unit. In the meter band a similar phenomenon is observed on a sphere of a somewhat larger radius (which is consistent with formula (22)). For visible light (for which the Sun is not a black hole) no focusing or associated increased light emission of the edge occur. What is important is that these effects in the radio range are easily interpreted only in the framework of the anisotropic relativity theory whereas in the framework of Einstein's general relativity theory (where the black hole radii are many orders of magnitude lower) these phenomena are difficult to explain.

Radio eruptions of the Sun

Assume that in its motion on the celestial sphere the Sun center aligns (for an observer A on the Earth) with a very distant radio emission source S. As noted above, the radio intensity increases for A approximately $\frac{8h}{q}$ times because of focusing (Fig. 5). Assuming that the diameter of the pupil (i.e. the radio telescope aerial) is 100 m and since h is larger than the Sun radius, the factor by which the radio intensity increases is $\approx 10^8$. Because the Sun motion along the celestial sphere (because of the Earth's orbital motion and the proper motion of the Sun) that eruption is very short-lived. Thus, since the speed of the Earth's orbital motion is 3×10^6, then (disregarding displacement of the Sun and the distant radio source) the point of accurate focusing "runs" the 100 m wide aerial within 3.3×10^{-3} seconds. Its alignment is inaccurate for the observer on the Earth and the radio source travel near the center, the focusing effect is weaker (brightness intensifies by merely four or five orders of magnitude) but the phenomenon lasts longer (tenths of a second). For this, not very accurate, alignment of the radio-emission source and the Sun center the focusing is done by a small fraction (the closest to and the most distant from the radio emission source) rather than the entire sphere of radius $r_0\sqrt[4]{3}$. Because of the fast motion of the Sun this apparent source of radio emission (i.e. the point on the sphere closest to the radio wave source shaded by the Sun) moves along the Sun surface at speeds of hundreds or thousands kilometers

per second. These theoretical conclusions from the anisotropic relativity theory are confirmed by experimental observations, for from time to time Sun radio eruptions are noted which last fractions of a second and are caused by "radio emission sources" which cross the Sun surface at speeds at least 1000 km/sec.

Note also that with $m = 7.4 \times 10^{25}$ (the mass of the Moon) and $\lambda = 1.95 \times$ $\times 10^5$ we have from (22) $r_0 \sqrt[4]{3} = 1.74 \times 10^8$, which is equal to the Moon's radius. Consequently, at wavelengths of 1950 m and more the Moon must have focusing properties similar to the above properties of the Sun. A satellite receiver tuned to 2000 m and an aerial pointing to the Moon may be expected to furnish an experimental confirmation of radio wave focusing.

Eruptions of red dwarfs

Let A be an observer on Earth, O be a red dwarf, and S be a star of the blue spectral class, Fig. 5. Furthermore, assume that AO is equal to six parsecs (a distance to the nearest red dwarfs). Assume that the telescope "pupil" diameter q is five meters and h is larger than 300 000 km (average radius of a red dwarf); then we have that the brigtness gain is approximately $\frac{8h}{q} >$ $> 4.8 \times 10^8$. Consequently a star S which is usually unobservable, so far it is, can cause a sudden intensive "blue eruption" of the red dwarf O. If O moves at a rate of 0.6″ a year (the medium shift of stars of this class in the Galaxy), a "blue eruption" should last approximately 5 sec. Such eruptions of red dwarfs which last several seconds in the blue range of the spectrum, which is outside their own range, have been many times confirmed by astrophysical observations. Note that for an eruption to occur there is no need for A, O, and S to be accurately aligned; it is sufficient that S stays within a narrow cone whose axis is AO and the spread angle depends on the diameter q of the "pupil", distance AO, and mass m of the red dwarf O. The probability of an eruption computed with this reasoning in mind fits well the observed frequency of eruptions.

Radiations of the Crab nebula

The Crab nebula is the brightest radio wave source in the vicinity of the Sun. In approaching the Sun the disk of the nebula is distorted, elongating by a factor of about three towards the Sun (at a wavelength of 7.9 m this occurs at a distance of about 13 solar radii and at a wavelength of 3.7 m, of about 10 radii). This could be attributed to heavy electron concentration in

the Sun supercorona. But then refraction would cause nonuniformity in the visible motion of the nebula relative to the Sun. The observed nonuniformity is, however, negligible and so the electron concentration is very low. Finally, observations reveal that, following the disc distortion, the radio emission of the Crab nebula completely disappears to reappear on the other side of the Sun.

L. K. Nikolaev has shown that this behaviour (which can hardly be explained in the conventional framework) fits well, quantitatively as well as qualitatively, the anisotropic relativity theory. The system (18) is replaced by (cf. (3))

$$\dot{x} = -\frac{\gamma m}{|x|^3}x + \frac{u}{n(x)}, \quad |u| \leq c, \tag{23}$$

where γ is given by formula (20) with $q = 0.016$ and $n(x)$ is the medium index of refraction in the point for the actual radiation frequency. Now the maximum principle can help find time optimal paths of plant (23), i.e. (consistently with the Fermat principle) the paths of radio wave propagation. The initial idea was as follows. With an electron concentration the medium has an index of refraction $n < 1$ for radio waves. Therefore, in the absence of gravitation (or if the first addend in the right-hand side of (23) is eliminated) the paths deflect from the Sun. On the other hand, the effect of gravitation alone (or with $n(x) \equiv 1$, which is the case of wave propagation in vacuo) makes the paths curve towards the Sun. At a certain value of electron concentration both factors can compensate each other to make the curvature of the paths very little (which results in the uniform observed motion of the nebula disc).

In his quantative computations L. K. Nikolaev borrowed formula for electron concentration (and consequently, for the index of refraction $n(x)$) from [8]. Figure 8 represents the findings of a mathematical experiment. It shows paths outgoing from the Crab nebula (from "infinity"). The surface shaded by those paths (initially "parallel") makes a sort of fold, i.e. a family of paths has an envelope. As the Earth orbits the Sun (the dotted line in Fig. 8) the nebula disappears while the Earth moves along the BC arc (where the radio waves do not reach). When the folds (in the points B and C) are crossed easily explainable elongation of the nebula disc occurs through intersection of "infinitely close" paths in the points of the envelope. Accurate computation results fit well the experimental findings. In particular, the "disappearance" should occur, according to computation, at an approximate distance of 13 radii from the Sun at a wavelength of $\lambda = 7.9$ m and at a distance of 10 solar radii at a wavelength of $\lambda = 3.7$ m.

Without going into detail let us also note that the anisotropic relativity theory provides a satisfactory explanation for many other phenomena such

as the Mössbauer effect, the deceleration of the light signal in passing by the Sun, the gravitational curvature of space and shift of the Mercury perihelion, the difference of gravitational red shift in the center and at the edge of the Sun disc, etc.

References

1. *Boltyanskiy, V. G.*, Anisotropic relativity. Differentsial'nye uravneniya, **10,** No. *12,* 1974 (in Russian).
2. *Strel'tsov, V. N.*, On invariance of Maxwell equations. Soobschcheniya OIYaI, R2-7328. Dubna, 1973 (in Russian).
3. *Boltyanskiy, V. G.*, The maximum principle, black holes, and quasars. Differentsial'nye uravneniya, **15,** No. *10,* 1979 (in Russian).
4. *Boltyanskiy, V. G., Nikolaev, L. K.*, A computer experiment with light paths. Differentsial'nye uravneniya, **16,** No. *3,* 1980 (in Russian).
5. *Boltyanskiy, V. G.*, The anisotropic relativity theory and optimization. Differentsial'nye uravneniya, **15,** No. *11,* 1979 (in Russian).
6. *Boltyanskiy, V. G.*, Is the Sun a radio lens? Differentsial'nye uravneniya, **16,** No. *11,* 1980 (in Russian).
7. *Lang, K. P.*, Astrophysical formulae (a compendum for the physicist and astrophysicist). Springer-Verlag, Berlin–Heidelberg–New York, 1974.
8. Paris Symposium on Radioastronomy. IAU Symposium No. 9 and URSA Symposium No. 1 held from 30 July to 6 August 1958. Edited by R. N. Bracewell. Stanford University Press, Stanford, Calif. 1959.

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

L. Györfi

Technical University of Budapest

H – 1111 Budapest XI., Stoczek u. 2, 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 (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 and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) 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 wil lbe returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в четырех экземплярах на русском языке или в двух на английском и в двух на русском языке следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50 – 100 слов и приложено резюме – реферат объемом не менее 10 – 15% объема статьи на русском и на английском языке (в двух экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS ● СОДЕРЖАНИЕ

Болтянский В. Г. Применение методов оптимизации к релятивистской теории (*Boltyansky, V. G.* Methods of optimization applied to the relativistic theory) 219

Yemel'yanov, S. V., Korovin, S. K., Sizikov, V. I. Use of coordinate-parametric feedback in design of control systems (*Емельянов С. В., Коровин С. К., Сизиков В. И.* Применение координатно-параметрической обратной связи при синтезе систем автоматического управления) 237

Víšek, J. Á. Bounds of the asymptotic rate of convergence of the error probabilities of a robust test (*Вишек Й. А.* Границы предельной скорости сходимости вероятностей ошибок устойчивого теста) 253

Sinha, N. K., Rózsa, P. Some system theory applications of Egerváry's algorithm for transformation of a matrix to the Hermite normal form (*Синха Н. К., Рожа П.* Преобразование матриц к эрмитовой нормальной форме с помощью алгоритма Эгервари и некоторые приложения к теории систем) 267

Wertz, W. A remark on the estimation of functionals of a probability density (*Вертц В.* Заметка об оценке функционала плотности вероятности) 279

Artificial Intelligence (Second International Workshop in Repino) 287

VIIIth All-Union Conference on Control Sciences 293

Book Review 299

HU ISSN 0370—2529

Index: 26 660

316.920

VOL. ТОМ 10 * NUMBER НОМЕР 5

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1981

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

9

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Roumania, Viet-Nam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H–1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1981 Subscription Rate US Dollars 110.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

B. N. PETROV (Moscow)

G. BOGNÁR (Budapest)

COORDINATING EDITORS

S. V. EMELYANOV (Moscow)
N. N. KRASOVSKII (Sverdlovsk)
E. P. POPOV (Moscow)
V. S. PUGACHEV (Moscow
V. I. SIFOROV (Moscow)
E. D. TERYAEV (Moscow)
T. VÁMOS (Budapest)
L. VARGA (Budapest)
A. PRÉKOPA (Budapest)
S. CSIBI (Budapest)
I. CSISZÁR (Budapest)
J. KOCSIS (Budapest)
J. BENEŠ (Prague)
V. STREJC (Prague)

РЕДАКТОРЫ ЖУРНАЛА

Б. Н. ПЕТРОВ (Москва)

Г. БОГНАР (Будапешт)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

С. В. ЕМЕЛЬЯНОВ (Москва)
Н. Н. КРАСОВСКИЙ (Свердловск)
Е. П. ПОПОВ (Москва)
В. С. ПУГАЧЕВ (Москва)
В. И. СИФОРОВ (Москва)
Е. Д. ТЕРЯЕВ (Москва)
Т. ВАМОШ (Будапешт)
Л. ВАРГА (Будапешт)
А. ПРЕКОПА (Будапешт)
Ш. ЧИБИ (Будапешт)
И. ЧИСАР (Будапешт)
Я. КОЧИШ (Будапешт)
И. БЕНЕШ (Прага)
В. СТРЕЙЦ (Прага)

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

Problems of Control and Information Theory, Vol. 10 (5), pp. 301–307 (1981)

О НИЖНЕЙ ГРАНИЦЕ ДЛИНЫ B_4-КОДА

А. Г. ДЬЯЧКОВ, В. В. РЫКОВ

(Москва)

(Поступила в редакцию 12 августа 1980 г.)

В [1] введено понятие B_S-кода, возникающее при рассмотрении задачи кодирования для разрешения конфликта в некоторых системах связи со случайным множественным доступом [2]. В [1] получены нижние и верхние границы оптимальной длины B_S-кода. В данной заметке дается улучшение нижней границы из [1] для частного случая B_4-кода.

Нашей задачей будет получение нижней границы оптимальной длины кодов, предназначенных для кодирования адресной части сообщений в системах связи со случайным множественным доступом. Подробно анализ применимости таких кодов, в частности в системах спутниковой связи, дан в работе [1].

Обозначим через $\underline{x} = ||\underline{x}_i(j)||$, $i = \overline{1, N}$, $j = \overline{1, t}$, — двоичную (из 0 и 1) матрицу размера $N \times t$. Пусть $\underline{x}(j)$ — j-й столбец, а $\underline{x}_i$ — i-я строка этой матрицы.

Определение. [1] Матрица $\underline{x}$ называется B_S-кодом, если все C_{t+S-1}^{S} сумм ее столбцов вида $\underline{x}(j_1) + \underline{x}(j_2) + \ldots + \underline{x}(j_S)$, где $1 \leq j_1 \leq j_2 \leq \ldots \leq j_S \leq t$, отличны друг от друга. Пусть $N(S, t)$ — минимально возможное число строк (длина) B_S-кода.

Обозначим через $\underline{q} = (q_0, q_1, q_2)$ произвольный вектор с компонентами $0 \leq q_i \leq 1$, для которого $q_0 + 2q_1 + 2q_2 = 1$. Пусть

$$\underline{H}(\underline{q}) \triangleq -q_0 \log q_0 - 2q_1 \log q_1 - 2q_2 \log q_2 \tag{1}$$

— энтропия Шеннона соответствующего распределения вероятностей, причем здесь и далее используются логарифмы по основанию 2, а символ $\triangleq$ означает равенство по определению. Зададим вектор $\underline{q}^* = (q_0^*, q_1^*, q_2^*)$, где

$$q_0^* = 3 \cdot 6^{1/4}/2(3 + 2 \cdot 6^{1/4}) = 0{,}38296,$$

$$q_1^* = 1/2 - 2q_0^*/3 = 0{,}24469,$$

$$q_2^* = q_0^*/6 = 0{,}06383.$$

Теорема. При $t \to \infty$ справедливо асимптотическое неравенство

$$N(4, t) \geq [(\log 3)^{-1} + \underline{H}(\underline{q}^*)^{-1}]\, 2 \log t\, (1 + {}_0(1)). \tag{2}$$

В [1] было показано, что при $t \to \infty$

$$N(4,t) \geq \left[\frac{2}{3} + \left(\frac{2}{3} + \log 3\right)^{-1}\right] 2 \log t\,(1 + o(1)). \tag{3}$$

Для удобства сравнения границ (2) и (3) приведем результаты вычислений:

$$2\,[(\log 3)^{-1} + \underline{H}(\underline{q}^*)^{-1}] = 2{,}24661,$$

$$2\left[\frac{2}{3} + \left(\frac{2}{3} + \log 3\right)^{-1}\right] = 2{,}22158.$$

Доказательство теоремы. Рассмотрим произвольный B_4-код x размера $N \times t$. Зафиксируем число $n = \overline{0, N}$. Через $\underline{v} = (v_1, v_2, \ldots v_n)$, $\underline{w} = (w_1, w_2, \ldots, w_{N-n})$ будем обозначать столбцы с целочисленными компонентами $v_i = 0, \pm 1, \pm 2$, $w_i = 0, \pm 1, \pm 2$ длины n и $N - n$, соответственно. Как и в [1], введём $\mu_1(\underline{v}, \underline{w})$ — число столбцов $\underline{x}$, которые имеют вид $(\underline{v}, \underline{w})$. Пусть

$$\mu_2(\underline{v}, \underline{w}) \stackrel{\triangle}{=} \sum_{\underline{v}'} \sum_{\underline{w}'} \mu_1(\underline{v}', \underline{w}')\, \mu_1(\underline{v} - \underline{v}', \underline{w} - \underline{w}').$$

Положим

$$\mu_S(\underline{v}, \cdot) \stackrel{\triangle}{=} \sum_{\underline{w}} \mu_S(\underline{v}, \underline{w}), \qquad s = \overline{1,2}$$

Далее, следуя [1], определим

$$l_n(s) \stackrel{\triangle}{=} \sum_{\underline{v}} \mu_S^2(\underline{v}, \cdot),$$

$$m_S(\underline{v}, \underline{w}) \stackrel{\triangle}{=} \sum_{\underline{w}'} \mu_S(,\underline{v}, \underline{w}')\, \mu_S(\underline{v}, \underline{w}' + \underline{w}). \tag{4}$$

$$m_S(\cdot, \underline{w}) \stackrel{\triangle}{=} \sum_{\underline{v}} m_S(\underline{v}, \underline{w}),$$

$$Z_S = \{\underline{w} : m_S(\cdot, \underline{w}) > 0\}, \quad s = \overline{1,2},$$

$$m_2(\underline{v}, Z_1) = \sum_{\underline{w} \in Z_1} m_2(\underline{v}, \underline{w}) \leq \mu_2^2(\underline{v}, \cdot), \tag{5}$$

$$m_2(\cdot, Z_1) = \sum_{\underline{v}} m_2(\underline{v}, Z_1).$$

В [1] была доказана

Лемма 1. Для любого B_4-кода x справедливы следующие утверждения.

1) Число

$$m_2(\cdot, Z_1) \leq 4t l_n\,(1).$$

2) Множество $Z_1 \subset Z_2$. При этом, если $\underline{w} \in Z_2 \setminus Z_1$, то $m_2(\cdot, \underline{w}) \leq 4$.

Введем тройку векторных случайных величин $(\underline{\eta}, \underline{\xi}, \underline{\tilde{\eta}})$, $\underline{\eta} = (\eta_1, \ldots, \eta_{N-n})$, $\underline{\tilde{\eta}} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_{N-n})$, $\underline{\xi} = (\xi_1, \ldots \xi_n)$ с распределением, задаваемым соотношениями

$$P_r\{\underline{\xi} = \underline{\nu}\} = \mu_2(\underline{\nu}, \cdot)t^{-2},$$

$$P_r\{\underline{\eta} = \underline{\tilde{w}}, \underline{\eta} = \underline{\tilde{w}} \mid \underline{\xi} = \underline{\nu}\} = \frac{\mu_1(\underline{\nu}, \underline{w}) \cdot \mu_2(\underline{\nu}, \underline{\tilde{w}})}{\mu_2^2(\underline{\nu}, \cdot)}. \tag{6}$$

Из (6), учитывая определение (4), имеем

$$P_r\{\underline{\eta} - \underline{\tilde{\eta}} = \underline{w}\} = t^{-2} \sum_{\underline{\nu}} \mu_2(\underline{\nu}, \cdot)^{-1} m_2(\underline{\nu}, \underline{w}).$$

Следовательно,

$$P_r\{\underline{\eta} - \underline{\tilde{\eta}} \in Z_1\} = t^{-2} \sum_{\underline{\nu}} \mu_2(\underline{\nu}, \cdot)^{-1} m_2(\underline{\nu}, Z_1) \leq t^{-2} \sum_{\underline{\nu}} \sqrt{m_2(\underline{\nu}, Z_1)} \leq$$

$$\leq t^{-2} \sqrt{3^n m_2(\cdot, Z_1)} \leq 2t^{-2} \sqrt{3^n t l_n(1)}, \tag{7}$$

где описанные выше оценки основаны на неравенстве (5), выпуклости вверх функции $y = \sqrt{x}$ и утверждении 1) леммы 1.

Пусть

$$H(\underline{\eta} - \underline{\tilde{\eta}}) \triangleq - \sum_{\underline{w} \in Z_2} P_r\{\underline{\eta} - \underline{\tilde{\eta}} = \underline{w}\} \log P_r\{\underline{\eta} - \underline{\tilde{\eta}} = \underline{w}\}$$

— энтропия Шеннона [3] векторной случайной величины $\underline{\eta} - \underline{\tilde{\eta}}$. Используя (7), утверждение 2) леммы 1 и очевидное неравенство

$$H(\underline{\eta} - \underline{\tilde{\eta}}) \geq - P_r\{\underline{\eta} - \underline{\tilde{\eta}} \in Z_2 \setminus Z_1\} \quad \log \left[\max_{\underline{w} \in Z_2 \setminus Z_1} P_r \{\underline{\eta} - \underline{\tilde{\eta}} = \underline{w}\} \right],$$

приходим к оценке

$$H(\underline{\eta} - \underline{\tilde{\eta}}) \geq [1 - 2t^{-2} \sqrt{3^n t l_n(1)}] \log (t^2/4). \tag{8}$$

Ниже будет доказана

Лемма 2. Энтропия разности

$$H(\underline{\eta} - \underline{\tilde{\eta}}) \leq (N - n) \underline{H}(q^*), \tag{9}$$

где число $\underline{H}(q^*)$ определено при формулировке теоремы.

Неравенства (8) и (9) означают, что нами получено

Предложение 1. Если для B_4-кода размера $N \times t$ и целого числа n, $0 \leq n \leq N$, величина $l_n(1) < t \log t$, то

$$N \geq n + 2\underline{H}(q^*)^{-1} [1 - 2t^{-2} \sqrt{3^n t^2 \log t}] \log (t/2). \tag{10}$$

Следующим шагом доказательства теоремы, к которому сейчас приступим, является вывод неравенства, аналогичного (10), для случая, когда $l_n(1) \geq t \log t$. Введем

$$\tilde{m}(\underline{w}) \stackrel{\triangle}{=} \sum_{\underline{w}'} m_1(\cdot, \underline{w}')\, m_1(\cdot, \underline{w} - \underline{w}'),$$

$$\tilde{Z}_2 = \{\underline{w} : \tilde{m}(\underline{w}) > 0\},$$

$$\tilde{m}(Z_1) = \sum_{\underline{w} \in Z_1} \tilde{m}(\underline{w}).$$

Несложно проверить, что для любого $\underline{w}$ число $m(\underline{w}) \leq \tilde{m}_2(\cdot, \underline{w})$. Поэтому из леммы 1 вытекает

Лемма 3. Для любого B_4-кода имеют место следующие два утверждения. 1) Число $\tilde{m}(Z_1) \leq 4t\, l_n(1)$. 2) Множество $Z_1 \subset \tilde{Z}_2$, причем, если $\underline{w} \in \tilde{Z}_2 \setminus Z_1$, то $\tilde{m}(\underline{w}) \leq 4$.

Рассмотрим пару $(\underline{\xi}_1, \underline{\xi}_2)$, где $\underline{\xi}_i = (\xi_1^{(i)}, \ldots, \xi_{N-n}^{(i)})$, независимых векторных случайных величин с одним и тем же распределением вероятностей

$$P_r\{\underline{\xi}_i = \underline{w}\} = m_1(\cdot, \underline{w})/l_n(1), \quad i = \overline{1, 2}.$$

Очевидно,

$$P_r(\underline{\xi}_1 + \underline{\xi}_2 = \underline{w}\} = m(\underline{w})/l_n^2(1).$$

Следовательно, утверждение 1) леммы 3 означает

$$P_r\{\underline{\xi}_1 + \underline{\xi}_2 \in Z_1\} \leq 4t/l_n(1). \tag{11}$$

С помощью утверждения 2) леммы 3 и (11) аналогично выводу (8) доказывается

$$H(\underline{\xi}_1 + \underline{\xi}_2) \geq [1 - 4t/l_n(1)] \log (l_n^2(1)/4). \tag{12}$$

Непосредственной аналитической проверкой нетрудно установить, что справедлива

Лемма 4. Пусть $(\underline{\xi}_k^{(1)}, \underline{\xi}_k^{(2)})$ — независимые, одинаково распределенные случайные величины, принимающие значения 0, 1, -1 с вероятностями

$$P_r\{\xi_k^{(i)} = 0\} = p_k,\ P_r\{\xi_k^{(i)} = 1\} = P_r\{\xi_k^{(i)} = -1\} = (1 - p_k)/2,$$

где $0 \leq p_k \leq 1$, $k = \overline{1, N - n}$. Тогда энтропия суммы $H(\xi_k^{(1)} + \xi_k^{(2)})$ как функция параметра p_k, $1/2 \leq p_k \leq 1$, достигает своего максимального значения

$$H_4 \stackrel{\triangle}{=} -\sum_{i=0}^{4} C_4^i 2^{-4} \log (C_4^i 2^{-4}) = 2{,}03064 < 2{,}03096 = H(q^*), \tag{13}$$

когда $p_k = 1/2$.

Из леммы 4 и свойства субаддитивности энтропии [3] получаем неравенство

$$H(\underline{\xi}_1 + \underline{\xi}_2) \leq \sum_{k=1}^{N-n} H(\xi_k^{(1)} + \xi_k^{(2)}) \leq (N-n) H_4,$$

из которого с учетом (12) следует

Предложение 2. Если для B_4-кода размера $N \times t$ и целого числа n, $0 \leq n \leq N$, значение $l_n(1) \geq t \log t$, то

$$N \geq n + 2H_4^{-1}(1 - 4/\log t) \log (t/2). \tag{14}$$

Полагая в (10) и (14)

$$n \triangleq \lfloor 2 \log (t/\log t)/\log 3 \rfloor$$

(здесь $\lfloor a \rfloor$ — наибольшее целое, меньшее или равное a), приходим в силу (13) к утверждению теоремы. Для завершения вывода теоремы осталось провести

Доказательство леммы 2. Введем

$$\mu_2^{(j)}(\underline{v}, w) \triangleq \sum_{\underline{w} \in W_j(w)} \mu_2(\underline{v}, \underline{w}),$$

где $j = \overline{1, N-n}$, $w = \overline{0, 2}$, а множество

$$W_j(w) \triangleq \{\underline{w} = (w_1, w_2, \ldots, w_{N-n}) : w_j = w\}.$$

Из (6) легко понять, что совместное распределение пары $(\eta_j, \tilde{\eta}_j)$ задается соотношением

$$P_r\{\eta_j = w, \tilde{\eta}_j = \tilde{w}\} = \sum_{\underline{v}} \frac{\mu_2^{(j)}(\underline{v}, w)\, \mu_2^{(j)}(\underline{v}, \tilde{w})}{t^2\, \mu_2(\underline{v}, \cdot)}, \tag{15}$$

а маргинальные распределения являются биномиальными

$$P_r\{\eta_j = w\} = P_r\{\tilde{\eta}_j = w\} = \sum_{\underline{v}} t^{-2} \mu_2^{(j)}(\underline{v}, w) =$$

$$= C_2^w \left(\frac{|\underline{x}_{n+j}|}{t}\right)^w \left(1 - \frac{|\underline{x}_{n+j}|}{t}\right)^{2-w}, \quad j = \overline{1, N-n}, \tag{16}$$

где $|\underline{x}_{n+j}| \triangleq \sum_{k=1}^{t} x_{n+j}(k)$ — число единиц в $(n+j)$-ой строке B_4-кода. Рассмотрим

$$M(\eta_j \cdot \tilde{\eta}_j) = \sum_{\underline{v}} \frac{\mu_2(\underline{v}, \cdot)}{t^2} \left[\sum_{w=0}^{2} w \frac{\mu_2^{(j)}(\underline{v}, w)}{\mu_2(\underline{v}, \cdot)}\right]^2 \geq \left[\sum_{\omega=0}^{2} w P_r\{\eta_j = w\}\right]^2 = (M\eta_j)^2,$$

где использованы (15), (16) и свойство выпуклости вниз функции $y = x^2$. Следовательно, для любого $j = \overline{1, N-n}$

$$M(\eta_j - \tilde{\eta}_j^2) = 2[M\eta_j^2 - M(\eta_j \cdot \tilde{\eta}_j)] \leq 2D\eta_j = 4|\underline{x}_{n+j}|(t - |x_{n+j}|)/t^2 \leq 1. \quad (17)$$

Заметим, что $\eta_j - \tilde{\eta}_j$ имеет распределение следующего вида

$$(\eta_j - \tilde{\eta}_j) \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ q_2 & q_1 & q_0 & q_1 & q_2 \end{pmatrix},$$

где $q_0 + 2q_1 + 2q_2 = 1$, $0 \leq q_i \leq 1$, $i = \overline{0,2}$. Неравенство (17) означает, что вектор $\underline{q} = (q_0, q_1, q_2)$ удовлетворяет условию $8q_2 + 2q_1 \leq 1$, которое равносильно

$$q_1 \geq 1/2 - 2q_0/3. \quad (18)$$

Поэтому энтропия Шеннона

$$H(\eta_j - \tilde{\eta}_j) \leq \max_{(18)} \underline{H}(\underline{q}), \qquad j = \overline{1, N-n}, \quad (19)$$

где $\underline{H}(\underline{q})$ определена (1), а максимум берется при условии (18). Простая аналитическая проверка (которую опускаем) показывает, что максимум в (19) достигается на векторе q^*, заданном при формулировке теоремы. Применяя свойство субаддитивности энтропии

$$H(\underline{\eta} - \underline{\tilde{\eta}}) \leq \sum_{i=1}^{N-n} H(\eta_i - \tilde{\eta}_i),$$

из (19) получаем неравенство (9).

Лемма 2 доказана.

Выводы

В статье исследована минимальная длина кода, позволяющего различить слагаемые в сумме, если число слагаемых не выше четырех. Слагаемыми являются кодовые слова. При этом получена асимптотическая нижняя оценка для этой минимальной длины при стремлении к бесконечности числа слов в коде. Основной результат статьи представлен неравенством (2). Проведено сравнение результата с ранее полученной границей из работы [1]. Приведенный метод оценки длины B_S-кода для случая $S = 4$ позволяет надеяться на возможность уточнения полученной в [1] границы и в общем случае.

Литература

1. *Дьячков А. Г., Рыков В. В.* Об одной модели кодирования для суммирующего канала с множественным доступом. Проблемы передачи информации, 1981, **17**, *2*, 26–38.
2. *Цыбаков Б. С., Михайлов В. А.* Свободный синхронный доступ пакетов в широковещательный канал с обратной связью. Проблемы передачи информации, 1978, **14**, *4*, 32–59.
3. *Галлагер Р. Г.* Теория информации и надежная связь, М. «Сов. радио», 1974.

On lower bound of B_4-code

A. G. DYACHKOV, V. V. RYKOV

(Moscow)

B_S-code definition has introduced in [1]. It appears in considering the coding problem to decide a conflict in some communication systems with random multiple access [2]. Some lower and upper bounds of optimal length of B_S-code have founded in [1]. Improvement of lower bound from [1] is given in this paper for particular case of B_4-code.

А. Г. Дьячков
Московский Государственный университет им. М. В. Ломоносова, механико-математический факультет
СССР, Москва В-234,
Ленинские горы

Problems of Control and Information Theory, Vol. 10 (5), pp. 309—316 (1981)

НИЖНЯЯ ГРАНИЦА ОБЛАСТИ ДОПУСТИМЫХ СКОРОСТЕЙ КОДОВ ДЛЯ СУММИРУЮЩЕГО КАНАЛА С КОРРЕЛИРОВАННЫМИ ИСТОЧНИКАМИ

В. Н. ДЫНЬКИН, Г. Л. КАЦМАН, А. П. КУРДЮКОВ

(Москва)

(Поступила в редакцию 12 августа 1980 г.)

Выводится нижняя граница мощности кодов, предназначенных для передачи информации по суммирующему каналу с коррелированными источниками.

В последние годы в процессе развития информационно-управляющих систем усложнилась их структура, что привело к усложнению структуры информационных потоков и к необходимости применения принципиально различных процедур обращения пользователей к системе.

Общепринятой моделью системы связи считается классическая модель, в которой один передатчик посылает информацию одному приемнику по зашумленному каналу. Эта модель достаточно строго описывает много типов действующих систем связи.

Поэтому возникла необходимость исследовать наряду с классической и другие более общие модели систем связи. Эти исследования составляют бурно развивающуюся ветвь теории информации, называемую теорией связи с многими пользователями. Одно из направлений последней составляют исследования каналов с множественным доступом, то есть каналов с несколькими входами и одним выходом.

С другой стороны большинство существующих в настоящее время систем спутниковой связи используют мультиплексирование с разделением по частотам, когда каждый канал соответствует одному источнику и одному приемнику для соответствующего момента времени. Однако систему связи со спутником можно рассматривать как систему связи, включающую канал с множественным доступом, что позволяет улучшить ее некоторые характеристики.

В связи с вышеизложенным становятся актуальными вопросы исследования области допустимых скоростей передачи по каналам с множественным доступом.

В данной работе строятся нижние границы области допустимых скоростей для суммирующего канала с коррелированными источниками.

1. Постановка задачи

Будем рассматривать задачу кодирования для канала следующей структуры. Пусть имеются три источника I_0, I_1 и I_2, выдающие сообщения j, i и k соответственно. Кодер 1 строит двоичный вектор u длины n по сообщениям i и j, а кодер 2 соответственно вектор v по сообщениям k и j. Эти вектора поступают на входы суммирующего канала, на выходе которого появляется вектор $u + v$. Сложение векторов здесь и далее понимается как покомпонентное сложение, причем компоненты складываются как обычные действительные числа. Декодер по вектору $u + v$ на выходе канала должен определить сообщения j, i и k. Такой канал называется суммирующим каналом с коррелированными источниками [1].

Нашей задачей будет оценка мощностей источников M_0, M_1 и M_2, при которых возможно определение сообщений j, i и k с нулевой вероятностью ошибки.

Мощностям M_0, M_1 и M_2 поставим в соответствие тройку скоростей R_0, R_1 и R_2 по формуле $R_i = (\log_2 M_i)/n$, $i = 0, 1, 2$. Тройки скоростей, при которых возможно определение сообщений j, i и k с нулевой вероятностью ошибки, будем называть допустимыми.

Пусть задано M_0 пар двоичных кодов длины n (U_1, V_1), (U_2, V_2), ..., (U_{M_0}, V_{M_0}), $|U_j| = M_1$, $|V_j| = M_2$, $j = 1, \ldots, M_0$ таких, что для любых j_1 и j_2 и любых четырех векторов u, u', v, v', $u \in U_{j_1}$, $v \in V_{j_1}$, $u' \in U_{j_2}$, $v' \in V_{j_2}$ равенство $U + V = U' + V'$ возможно тогда и только тогда, когда $j_1 = j_2$ и $u = u'$, $V = V'$. Такую последовательность пар кодов будем называть разделимой.

Легко видеть, что если существует разделимая последовательность пар кодов с мощностями M_0, M_1 и M_2, то тройка скоростей R_0, R_1, R_2 является допустимой ($R_i = (\log_2 M_i)/n$, $i = 0, 1, 2$).

2. Выпуклость области допустимых скоростей

Лемма 1. Если существует разделимая последовательность пар кодов длины n с тройкой скоростей R_0, R_1, R_2, то для любого натурального m существует разделимая последовательность кодов длины mn с теми же скоростями.

Из существования допустимой тройки скоростей R_0, R_1, R_2 следует существование бесконечного множества разделимых последовательностей пар кодов со скоростями R_0, R_1, R_2.

Лемма 2. Если существует две разделимые последовательности пар кодов со скоростями R_0, R_1, R_2 и R_0', R_1', R_2', то существует разделимая последова-

тельность пар кодов со скоростями

$$\frac{l}{m}R_0 + \left(1 - \frac{l}{m}\right)R'_0,$$

$$\frac{l}{m}R_1 + \left(1 - \frac{l}{m}\right)R'_1,$$

$$\frac{l}{m}R_2 + \left(1 - \frac{l}{m}\right)R'_2.$$

Лемма 2 доказывает возможность так называемого режима «разделения времени» между двумя разделимыми последовательностями пар кодов.

Следствие. Множество троек допустимых скоростей является выпуклым.

Доказательство следствия очевидно следует из леммы 2 и определения выпуклого множества.

Лемма 3. Если существует допустимая тройка скоростей (R_0, R_1, R_2), то для любых $0 \leq r_1 \leq R_1$ и $0 \leq r_2 \leq R_2$ существует допустимая тройка скоростей $(r_1 + r_2 + R_0, R_1 - r_1, R_2 - r_2)$.

Доказательство. Пусть $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ — разделимая последовательность пар кодов с мощностями M_1, M_2. Разобьем каждый код U_j на l_1 равномощных непересекающихся подкодов мощности $\left[\frac{M_1}{l_1}\right]$ — $U_{j_1}, \ldots, U_{jl_1}$, $j = 1, 2, \ldots, M_0$. Аналогично разобьем каждый код V_j на l_2 равномощных непересекающихся подкодов мощности $\left[\frac{M_2}{l_2}\right]$ — $V_{j_1}, \ldots, V_{jl_2}$ $j = 1, 2, \ldots M_0$.

Рассмотрим последовательность пар кодов (U_{11}, V_{11}), $(U_{12}, V_{12}), \ldots,$ $\ldots, (U_{11}, V_{1l_2}), (U_{12}, V_{11}), \ldots, (U_{M_0 l_1}, V_{M_0 l_2})$. Легко видеть, что эта последовательность является разделимой с мощностями $M_0 l_1 l_2, \left[\frac{M_1}{l_1}\right], \left[\frac{M_2}{l_2}\right]$. Отсюда, очевидно, следует утверждение леммы 3.

3. Нижняя граница области допустимых скоростей

Теорема 1. Любая тройка скоростей, удовлетворяющая неравенствам $0 \leq R_0 < \log_2 3 - \frac{2}{3}$, $R_1 + R_2 \leq \frac{2}{3}$, $R_1 \geq 0$, $R_2 \geq 0$ является допустимой.

Доказательство. Из лемм 2 и 3 и из очевидного соображения, что если допустима тройка скоростей R_0, R_1, R_2, то допустима и тройка скоростей

R_0, R_2, R_1 следует, что достаточно доказать допустимость тройки скоростей $R_0 = \log_2 3 - \frac{2}{3} - \varepsilon, R_1 = \frac{2}{3}$, $R_2 = 0$ для любого $\varepsilon > 0$.

Для доказательства этого построим разделимую последовательность пар кодов длины n с мощностями $M_0 = C_n^{n/3}$, $M_1 = 2^{2n/3}$, $M_2 = 1$.

Занумеруем произвольным образом числами от 1 до $C_n^{n/3}$ все сочетания из n по $n/3$. В качестве кода V_i рассмотрим код, состоящий из единственного вектора длины n и веса $n/3$, в котором единицы стоят лишь на позициях, входящих в i-тое сочетание. Таких кодов может быть построено $C_n^{n/3}$. Далее, в качестве кода U_i рассмотрим код длины n и мощности $2^{2n/3}$, состоящий из всех векторов, у которых на позициях, принадлежащих i-тому сочетанию, стоят единицы, а остальные позиции произвольны. Пусть $u \in U_i$, $v \in V_i$, $u' \in U_j$, $v' \in V_j$. Если $i \neq j$, то $u + v \neq u' + v'$, поскольку в векторе $u + v$ двойки стоят на позициях, принадлежащих i-тому сочетанию, а в векторе $u' + v'$ — принадлежащих j-тому сочетанию. Если же $i = j$, то $v = v'$, и из того, что $u + v = u' + v'$, следует, что $u = u'$.

Таким образом, получена разделимая последовательность пар кодов с мощностями $M_0 = C_n^{n/3}$, $M_1 = 2^{2n/3}$, $M_2 = 1$. Поскольку $\lim\limits_{n\to\infty} \frac{1}{n} \log_2 C_n^{n/3} = \log_2 3 - \frac{2}{3}$, то отсюда следует утверждение теоремы.

Теорема 2. Допустима любая тройка скоростей (R_0, R_1, R_2), удовлетворяющих неравенствам $0 \leq R_1 < 1$, $0 \leq R_2 < 1$, $0 \leq R_0$, $R_0 + R_1 + 2R_2 < 3 - \log_2 3$.

Для доказательства этой теоремы докажем предварительно следующую лемму.

Лемма 4. Существует не более $2 \cdot 6^n$ упорядоченных четверок двоичных векторов длины n, u, v, u', v' таких, что $u + v = u' + v'$.

Доказательство. Пусть $u + v = u' + v' = w$ и число единиц в векторе w равно l. Тогда упорядоченная пара векторов (u, v) такая, что $u + v = w$ может быть выбрана 2^l способами. Таким же числом способов может быть выбрана пара векторов (u', v'), такая, что $u' + v' = w$. Таким образом, упорядоченных четверок векторов u, v, u', v' таких, что $u' + v = u' + v' = w$ существует не более $2 \cdot 4^l$. Но число векторов w, таких, что число единиц в них равно l есть $C_n^l \, 2^{n-l}$. Поэтому число упорядоченных четверок u, v, u', v' двоичных векторов длины n таких, что $u + v = u' + v'$ не превосходит $2 \sum\limits_{l=0}^{n} C_n^l \, 2^{n-l} 4^l = 2 \cdot 6^n$. Что и требовалось доказать.

Перейдем теперь к доказательству теоремы 2. Рассмотрим всевозможные последовательности $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ пар кодов длины n, $|U_i| = M_1$, $|V_i| = M_2$, $i = 1, 2, \ldots, M_0$. Различных таких последовательностей существует $2^{(M_1+M_2)M_0 n} = A$. При этом каждый код U_i (V_i) есть упорядоченная совокупность M_1 (M_2) векторов длины n, среди которых возможно есть совпадающие. Разобьем каждый из кодов U_i на $s = M_1/l$ подмножеств $U_{i1}, \ldots, U_{is}$ мощности l (без ограничения общности можно считать что M_1 делится на l). Подмножество U_{ij} $(i = 1, 2, \ldots, M_0; j = 1, 2, \ldots, s)$ будем называть плохим, если для любого вектора $u \in U_{ij}$ найдутся такие три вектора u', v, v', что $u' \in U_p, v \in V_i, v' \in V_p$, $u + v = u' + v'$. Последовательность пар кодов $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ будем называть плохой, если найдется хотя бы одно плохое подмножество U_{ij} $1 \leq i \leq M_0$; $1 \leq j \leq s$. Легко видеть, что число плохих последовательностей оценивается сверху величиной

$$\frac{M_0 M_1}{l} [2 \cdot 6^n M_0 M_1 l M_2^2]^l \cdot 2^{[(M_1+M_2)M_0 - 4l]n} <$$

$$< 2^l M_0^{l+1} M_1^{l+1} M_2^{2l} 6^{nl} \cdot 2^{[(M_1+M_2)(M_0-4l)]n} = B.$$

Следовательно, если $B \leq A$, т. е.

$$M_2^2 (M_0 M_1)^{1/l} M_0 M_1 \leq 8^n/2 \cdot 3^n, \tag{1}$$

то найдется по меньшей мере одна последовательность пар кодов, не являющаяся плохой.

Положим $l = n$, $R_0 = \frac{1}{n} \log_2 M_0$, $R_1 = \frac{1}{n} \log_2 M_1$, $R_2 = \frac{1}{n} \log_2 M_2$. Тогда неравенство (1) можно переписать в виде

$$R_0 + R_1 + 2R_2 + \frac{1}{n}(R_0 + R_1) \leq 3 - \log_2 3 - \frac{1}{n}.$$

Таким образом при достаточно большом n и $R_0 + R_1 + 2R_2 < 3 - \log_2 3$ существует последовательность пар кодов, не являющаяся плохой, со скоростями R_0, R_1, R_2.

Пусть $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ — последовательность пар кодов, не являющаяся плохой. Обозначим $U_i' = \{u_{ij} \mid u_{ij} \in U_{ij}$ такой, что $\neg \exists v, u', v', v \in V_i, u \in U_p, v' \in V_p$, что $u_{ij} + v = u' + v', p = 1, 2, \ldots, M_0\}$. Заметим, что такой вектор u_{ij} в каждом множестве U_{ij} обязательно найдется, поскольку последовательность пар кодов $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ не является плохой. Кроме того, без ограничения общности можно считать, что из каждого из подмножеств U_{ij} берется только один вектор, даже если их можно взять несколько.

Легко видеть, что полученные таким образом коды U_i' $i = 1, \ldots, M_0$ имеют мощность $\frac{M_1}{n}$, а последовательность $(U_1', V_1), \ldots, (U_{M_0}', V_{M_0})$ является разделимой последовательностью пар кодов. Тройка скоростей для этой последовательности есть $R_0, R_1 - \frac{\log n}{n}, R_2$. Следовательно, для любой тройки скоростей R_0, R_1, R_2, удовлетворяющих неравенствам $0 \leq R_1 < 1, 0 \leq R_2 < 1$, $0 < R_0$, $R_0 + R_1 + 2R_2 < 3 - \log 3$, при достаточно большом n существует разделимая последовательность пар кодов с тройкой скоростей R_0, R_1, R_2. Что и требовалось доказать.

На основе теорем 1 и 2, следствия из леммы 2 и очевидного соображения, что в условиях теоремы 2 скорости R_1 и R_2 могут меняться местами, может быть построена нижняя оценка области допустимых троек скоростей, которая приведена на рис. 1.

Координаты точки A есть $R_0(A) = \log_2 3$, $R_1(A) = R_2(A) = 0$,

точки B — $R_0(B) = \log_2 3 - \frac{2}{3}$, $R_1(B) = \frac{2}{3}$, $R_2(B) = 0$,

точки C — $R_0(C) = 2 - \log_2 3$, $R_1(C) = 1$, $R_2(C) = 0$,

точки D — $R_0(D) = 0$, $R_1(D) = 1$, $R_2(D) = 0$,

точки E — $R_0(E) = 0$, $R_1(E) = 1$, $R_2(E) = 1 - \frac{1}{2}\log_2 3$.

Координаты точек B', C', D' и E' могут быть получены соответственно из координат точек B, C, D и E заменой R_1 на R_2 и R_2 на R_1.

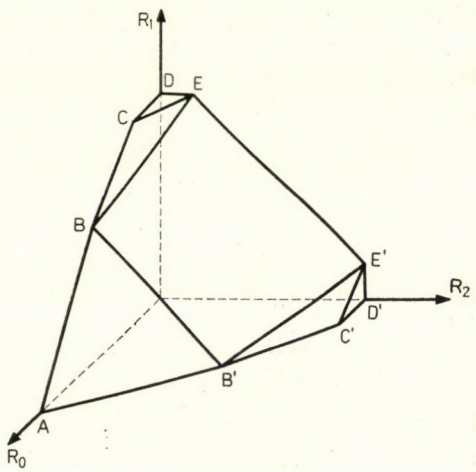


Рис. 1

4. Заключение

Остановимся теперь на возможности улучшения полученной нижней оценки области допустимых троек скоростей. Очевидно, что в области $R_0 \geq \log_2 3 - \frac{2}{3}$ полученная оценка не может быть улучшена, поскольку она совпадает с очевидной верхней оценкой $R_0 + R_1 + R_2 \leq \log_2 3$. При других значениях R_0 оценка может быть улучшена, например, за счет рассмотрения в плоскости (R_1, R_2), $(R_0 = 0)$ допустимых пар скоростей для обычного суммирующего канала [2], [3]. При $R_1 < 1$, $R_2 \geq 1 - \frac{1}{2}\log_2 3$ в [2] были получены пары кодов со скоростями, лежащими выше прямой EE'. Разделение времени этих пар кодов с последовательностями пар кодов, скорости которых представлены на рис. 1, улучшает полученную нижнюю границу на участке $B\,E\,E'\,B'$.

Таким образом, в работе предложена процедура построения кодов, позволяющая либо полностью использовать пропускную способность канала, либо достаточно близко к ней приблизиться.

Предлагаемый способ может быть распространен на случай, когда количество испочников равно $2^k - 1$, а число кодеров равно k.

Литература

1. *Slepian, D., Wolf, J. K.*, A Coding Theorem for Multiple Access Channels with Correlated Sources. BSTJ, 1973, **52**, No. 7, 1037–1076.
2. *Kasami, T., Lin, S.*, Bounds on the Achievable Rates of Block Coding for a Memoryless Multiple-Access Channel. IEEE Trans. Inform. Theory, 1978, **24**, 2, 187–197.
3. *Weldon, E. J.*, Coding for a Multiple-Access Channel. *I&C*, **36**, 256–274 (1978).

Lower bound of the region of acceptable code rates for a summary channel with correlated sources

V. N. DYNKIN, G. L. KATSMAN, A. P. KURDYUKOV

(Moscow)

Consideration is given to coding of messages arriving from three dependent sources via two coding units into the summary channel. The following lower bound for the region of acceptable rates for binary channel inputs is proposed.

Let M_0, M_1, M_2 be capacities of the sources and

$$R_i = \frac{1}{n}\log_2 M_i, \qquad i = 0, 1, 2.$$

Hence, if

$$(1) \qquad 0 \le R_0 \le \log_2 3 - \frac{2}{3}, \quad R_1 + R_2 \le \frac{2}{3}, \quad R_1 \ge 0, \quad R_2 \ge 0$$

or

$$(2) \quad 0 \le R_1 < 1, \quad 0 \le R_2 < 1, \quad R_0 \ge 0, \quad R_0 + R_1 + 2R_2 < 3 - \log_2 3,$$

the triple rate (R_0, R_1, R_2) is acceptable.

В. Н. Дынькин
Институт проблем управления
СССР, 117342, Москва ГСП-7,
Профсоюзная ул., 65

Problems of Control and Information Theory, Vol. 10 (5), pp. 317–327 (1981)

ON CONSISTENT PARAMETER ESTIMATION IN ADAPTIVE FILTERING

N. A. KUZNETSOV, A. I. YASHIN

(Moscow)

(Received July 11, 1980)

The paper investigates the necessary and sufficient conditions for consistency of parameter estimates in adaptive Kalman filtering. The unknown parameter is assumed to dictate the evolution of both the observed and unobserved processes and to take on values from a certain numerical set. In deriving the necessary and sufficient conditions use is made of the fact that these conditons are equivalent to singularity of probabilistic measures associated with the observable process with fixed values of the parameter.

1. Introduction

In many cases the coordinates of partially observable plants subjected to random noise can now be successfully estimated (filtered) [1,2], by using good a priori data on the structure and numerical values of the parameters.

In real life, however, some characteristics or parameters of the plant may prove impossible to determine before the experiment starts. In this case adaptive filtration methods are relevant to apply [3–6]. One common property is very important in all these adaptive algorithms, that the parameter estimates should in some sense tend to their true values as the number of observation grows. Such estimates are referred to as consistent [7–10].

This paper obtains the necessary and sufficient conditions for consistency of these parameter estimates in adaptive Kalman filtering.

2. Statement of the Problem

The problem of parameter estimation in Kalman filtering can be investigated in the framework of the following formal description.

On a probabilistic space $(\Omega, \mathcal{F}, P)$ a random quantity $\beta(\omega)$ is specified which takes on values from a certain numerical set $\{\beta_i\}$, $i \in \mathbf{N}$ with a priori probabilities $p_i = P(\beta = \beta_i)$ $p_i > 0$, $\sum_i p_i = 1$. Let $\mathcal{F}^\beta$ be a σ-algebra in the space Ω that is generated by values of the parameter β. On the same space a random process $(\theta, \xi) = (\theta_n, \xi_n)_{n \geq 0}$ is specified. Denote as $\mathcal{F}_n^{\theta\xi}$ and $\mathcal{F}_n^{\xi}$ the σ-algebras that are generated by values of the process (θ, ξ) and (ξ) until

time n, $n \geq 0$; $\mathcal{F}_n^{\beta\theta\xi} = \mathcal{F}^{\beta} \vee \mathcal{F}_n^{\theta\xi}$, $n \geq 0$ are σ-algebras generated by union of σ-algebras $\mathcal{F}^{\beta}$ and $\mathcal{F}_n^{\theta\xi}$ while $\mathcal{F}_n^{\beta\xi} = \mathcal{F}^{\beta} \vee \mathcal{F}_n^{\xi}$, $n \geq 0$ are σ-algebras generated by union of σ-algebras $\mathcal{F}^{\beta}$ and $\mathcal{F}_n^{\xi}$, $n \geq \epsilon$. Let $\mathcal{F}^{\theta\xi} = V_n \mathcal{F}_n^{\theta\xi}$, $\mathcal{F}^{\beta\theta\xi} = V_n \mathcal{F}_n^{\beta\theta\xi} \equiv \mathcal{F}$, $\mathcal{F}^{\beta\xi} = V_n \mathcal{F}_n^{\beta\xi}$.

Assume that on the probabilistic space $(\Omega, \mathcal{F}^{\beta\theta\xi}, P)$ the process $(\theta\xi)$ can be represented as a set of stochastic equations in discrete time

$$\theta_{n+1} = a(\beta, n)\,\theta_n + b_1(\beta, n)\,\varepsilon_{1n+1} + b_2(\beta, n)\,\varepsilon_{2n+1}, \theta_0$$

$$\xi_{n+1} = A(\beta, n)\,\theta_n + B_1(\beta, n)\,\varepsilon_{1n+1} + B_2(\beta, n)\,\varepsilon_{2n+1}, \xi_0 \qquad (1)$$

where $(\theta_n)_{n\geq 0}$ is a sequence of k-dimensional vectors; $(\xi_n)_{n\geq 0}$ is a sequence of l-dimensional vectors and $(\varepsilon_{1n}, \xi_{2n})_{n\geq 0}$ is a sequence of l-independent Gaussian random quantities of dimension k_1 and k_2, respectively, with zero means and identity covariance matrices; these quantities are independent of the initial values of θ_0 ξ_0 and the random value of β. The matrix $a(\beta, n)$ is $(k\times k)$, $A(\beta, n)$ is $l\times k$ and the matrices $b_1(\beta, n)$, $b_2(\beta, n)$, $B_1(\beta, n)$, and $B_2(\beta, n)$ are $(k\times k_1)$, $(k\times k_2)$, $(l\times k_1)$ and $(l\times k_2)$ respectively, and are constrained functions of time and the random parameter β. The process θ describes the time variation of the unobservable coordinate of a certain dynamic plant while the process ξ models the measurement of the coordinate θ against the background of random noises.

Introduce on $(\Omega, \mathcal{F}^{\beta\theta\xi})$ a numerical family of probabilistic measures P^i, $i \in \mathbf{N}$ using the relations

$$P^i(A) = \frac{p(A \cap \{\beta = \beta_i\})}{p_i}, \quad A \in \mathcal{F}^{\beta\theta\xi}, \quad i \in \mathbf{N}.$$

Denote $\pi_j(n) = P(\beta = \beta_j/\mathcal{F}_n^{\xi})$, $B \circ B_i = B_1(\beta_i, n)\,B_1^*(\beta_i, n) + B_2(\beta_i, n) \cdot B_2^*(\beta_i, n)$ and assume that $B \circ B_i$ is nonsingular with any $i \in \mathbf{N}$, $n \geq 0$. Let P_n, $\bar{P}$ and $\bar{P}_n$ denote restrictions of the measure P to σ-algebras $\mathcal{F}_n^{\beta\theta\xi}$, $\mathcal{F}^{\xi}$ and $\mathcal{F}_n^{\xi}$, respectively, while P_n^i, $\bar{P}^i$, and $\bar{P}_n^i$, restrictions of the measures P^i to $\mathcal{F}^{\beta\theta\xi}$, $\mathcal{F}^{\xi}$, and $\mathcal{F}_n^{\xi}$. Let Z_n^{ij} denote the Radon–Nikodim derivative of the measure $\bar{P}_n^i$ with respect to the measure $\bar{P}_n^j$ in the case of absolute continuity of $\bar{P}_n^i$ in relation to $\bar{P}_n^j$. Let $\hat{\beta}_{1n} = E(\beta/\mathcal{F}_n^{\xi})$ and $\hat{\beta}_{2n} = \arg\max \pi_j(n)$.

Definition. $\mathcal{F}_n^{\xi}$ coordinated process $(\beta_n)_{n\geq 0}$, is referred to as a strongly consistent estimate of the parameter β if $\lim_n \hat{\beta}_n = \beta$ with a probability of 1.

It is required to investigate the conditions for consistency of the estimates $\hat{\beta}_{in}$, $i = 1, 2$; $n \geq 0$.

3. Result

Assume that the initial conditions θ_0 and ξ_0 are such that

$$\cdot E_i(||\theta_0||^2 + ||\xi_0||^2) < \infty, \quad i \in \mathbf{N} \tag{2}$$

where E_i denotes the mathematical expectation with respect to the measure P_i. From the fact that the coefficients are constrained and from condition (2) it follows that with any $n < \infty$ and $i \in \mathbf{N}$

$$E_i(||\theta_n||^2 + ||\xi_n||^2) < \infty. \tag{3}$$

Assume also that the simultaneous distribution of the random quantities θ_0 and ξ_0 are Gaussian with respect to each measure P^i, $i \in \mathbf{N}$. Hereinafter $b_1(i)$, $b_2(i)$, $B_1(i)$, $B_2(i)$, $A(i)$, and $a(i)$ will replace $b_1(\beta_i, n)$, $b_2(\beta_i, n)$, $B_1(\beta_i, n)$, $B_2(\beta_i, n)$, and $A(\beta_i, n)$, $a(\beta_i, n)$ for convenience of notation.

Introduce the quantities $D_i^{\dagger}(n)$ by using the relations

$$D_i(n) = (B \circ B_i + A(i)\,\gamma_i(n)\,A(i))^{-1}, \quad i \in \mathbf{N}, \quad n \geq 0 \tag{4}$$

where $\gamma_i(n)$ satisfies the following equations

$$\gamma_i(n+1) = a(i)\,\gamma_i(n)\,a^*(i) + b \circ b_i - K_i(n)\,(b \circ B_i + a(i)\,\gamma_i(n)\,A(i))^*, \quad \gamma(0)$$

while

$$K_i(n) = (b \circ B_i + a(i)\,\gamma_i(n)\,A^*(i))\,(B \circ B_i + A(i)\,\gamma_i(n)\,A^*(i))^+ \tag{5}$$

$$\begin{aligned} b \circ B_i &= b_1(i)\,B_1^*(i) + b_2(i)\,B_2^*(i) \\ b \circ b_i &= b_1(i)\,b_1^*(i) + b_2(i)\,B_2^*(i). \end{aligned} \tag{6}$$

The primary finding of the paper is in the following

Theorem 1. For strong consistency of the estimates β_{in} $i = 1, 2$; $n \geq 0$ it is necessary and sufficient that for any $i, j \in \mathbf{N}$, $i \neq j$ the series diverge

$$\sum_{n=1}^{\infty} f_{ij}(n) \tag{7}$$

where

$$f_{ij}(n) = \ln \frac{|I + D_i(n)\,D_j^{-1}(n)|^{1/2}\,|D_j(n)|^{1/4}}{\sqrt{2}\,|D_i(n)|^{1/4}} + \frac{1}{4} \operatorname{Sp} S_{ij}(n)\,\Phi_{ij}(n)\,S_{ij}(n) \times$$

$$\times D_i(n)\,(D_i(n) + D_j(n))^{-1}\,D_i^*(n) \tag{8}$$

$$S_{ij}(n) = (0, -A_i(n), A_j(n)) \tag{9}$$

and $\Phi_{ij}(n)$ are solutions of the equations

$$\Phi_{ij}(n+1) = R_{ij}(n)\,\Phi_{ij}(n)\,R_{ij}^{*}(n) + G_{ij}(n)\,\theta\,G_{ij}^{*}(n) \tag{10}$$

with

$$R_{ij}(n) = \begin{pmatrix} a(i) & 0 & 0 \\ K_i(n)\,A(i) & a(i) - K_i(n)\,A(i) & 0 \\ K_j(n)\,A(j) & 0 & a(j) - K_j(n)\,A(j) \end{pmatrix} \tag{11}$$

$$G_{ij}(n) = \begin{pmatrix} b_1(j) & b_2(j) & 0 \\ K_i(n)\,B_1(j) & K_i(n)\,B_2(j) & 0 \\ K_j(n)\,A(j) & K_j(n)\,B_2(j) & 0 \end{pmatrix} \tag{12}$$

$$Q_{ij} = \begin{pmatrix} I_1 & 0 & 0 \\ 0 & I_2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{13}$$

and initial conditions

$$\Phi_{ij}(0) = \begin{pmatrix} E_j\,\theta_0\theta_0^{*} & E_j\,m_0^{j}\,m_0^{i*} & E_j\,m_0^{j}\,m_0^{j*} \\ E_j\,m_0^{j}\,m_0^{j*} & E_j\,m_0^{j}\,m_0^{j*} & E_j\,m_0^{i}\,m_0^{j*} \\ E_j\,m_0^{j}\,m_0^{j*} & E_j\,m_0^{j}\,m_0^{i*} & E_j\,m_0^{i}\,m_0^{j*} \end{pmatrix} \tag{14}$$

I_1 is a $k_1 \times k_1$ identity matrix;

I_2 is a $k_2 \times k_2$ identity matrix;

$|D_i(n)|$ is the determinant of the matrix $D_i(n)$, $m_0^{i} = F_i(\theta_0/\mathfrak{F}_0^{\xi})$.

Corollary 1. Assume that the coefficients of equations (1) do not depend on time n and with any $i \in \mathbf{N}$ there exist stationary solutions of equations (6). Then for consistent estimates $\hat{\beta}_k$, $k = 1, 2$ it is sufficient that with any $i, j \in \mathbf{N}$, $i \neq j$ the inequalities hold

$$B \circ B_i + A_i\,\gamma_i\,A_i^{*} \neq B \circ B_j + A_j\,\gamma_j\,A_j^{*}. \tag{15}$$

Truth of Corollary 1 follows from the first addend of the expression for $f_{ij}(n)$ and the stationarity assumption.

Corollary 2. In addition to the statement of Corollary 1 let a stationary solution of equation (10) exist. Then for the estimates $\hat{\beta}_k$, $k = 1, 2$ to be consistent it is sufficient that for any $i, j \in \mathbf{N}$, $i \neq j$

$$\operatorname{Sp} S_{ij} \Phi_{ij} S_{ij} D_i (D_i + D_j)^{-1} D_j^* \neq 0 \tag{16}$$

where Φ_{ij} is a stationary solution of equation (10).

The proof of Corollary 2 follows from the second addend in the expression for $f_{ij}(n)$ and from the stationarity assumption.

Remark. In the case of a finite number of possible values of the parameters β and constant coefficients in equations (4), conditions (15) and (16) can be verified before the observations are made.

4. Proof of Theorem 1

The proof will be preceded by several auxiliary Lemmas.

Lemma 1. On probabilistic spaces $(\Omega, \mathfrak{F}^{\beta\xi}, P^i)$, $i \in \mathbf{N}$ the process ξ can be represented as

$$\xi_{n+1} = A(i)\, m_n^i + D_i^{-\frac{1}{2}}(n)\, \tilde{\varepsilon}_{n+1}, \quad \xi_0, P^i - \text{a.s.} \tag{17}$$

where $D_i(n)$ satisfy (5), $\{\bar{\varepsilon}_n\}_{n \geq 0}$ is a sequence of independent Gaussian vectors with zero means and identity covariance matrices, $m_n^i = E_i(\theta_n / \mathfrak{F}_n^{\xi})$.

The proof of this lemma can be found in [14].

The properties of the process ξ are found to insure equivalence of the probabilistic measures $\bar{P}_n^i$ and $\bar{P}_n^j$, $i \neq j$.

Lemma 2. The measures $\bar{P}_n^i$ and $\bar{P}_n^j$ are equivalent and the derivative $z_{n+1}^{ij} = \dfrac{d\bar{P}_n^i}{d\bar{P}_n^j}$ has the form

$$z_{n+1}^{ij} = \prod_{k=0}^{n} \frac{|D_i(k)|^{1/2}}{|D_j(k)|^{1/2}} \exp\left\{ -\frac{1}{2} (\xi_{k+1} - A(i)\, m_k^i)^* D_i(k) (\xi_{k+1} - A(i)\, m_k^i) + \right.$$

$$\left. + \frac{1}{2} (\xi_{k+1} - A(j)\, m_k^j)^* D_j(k) (\xi_{k+1} - A(j)\, m_k^j) \right\}. \tag{18}$$

The proof of this lemma follows from the results of [11, 12]. Equivalence of the probabilistic measures $\bar{P}_n^j$ and $\bar{P}_n^i$, $i \neq j$, $i, j \in \mathbf{N}$ permits formulating the necessary and sufficient condition for consistency of an estimate such as conditional mathematical expectation.

Lemma 3. For the estimate $(\hat{\beta}_{1n} = E(\beta/\mathcal{F}_n^{\xi}))_{n\geq 0}$ to be consistent it is necessary and sufficient for the measures $\bar{P}^i$ and $\bar{P}^j$, $i \neq j$, $i, j \in \mathbf{N}$ to be singular.

The proof is to be found in [13].

Following the findings of Kabanov et al. [11] the necessary and sufficient condition for singularity of the probabilistic measures $\bar{P}^i$ and $\bar{P}^j$ with equivalence of the measures $\bar{P}_n^i$ and $\bar{P}_n^j$, $i \neq j, i, j \in \mathbf{N}$, $n \geq 0$ can be represented as

$$\bar{P}^i\left\{\sum_{n=1}^{\infty}\left(1 - E_j((\alpha_n^{ij})^{1/2}/\mathcal{F}_{n+1}^{\xi})\right) = \infty\right\} = 1 \quad i, j \in \mathbf{N}, \quad i = j \tag{19}$$

where $\alpha_n^{ij} = z_n^{ij}(z_{n-1}^{ij})^{-1}$ and z_n^{ij} are determined by using relations (18). Note that the series

$$\sum_{n=1}^{\infty}[1 - E_j((\alpha_n^{ij})^{1/2}/\mathcal{F}_{n-1}^{\xi})]$$

consists of nonnegative $\bar{P}^j$ a.s. terms (and, consequently, by virtue of equivalence of the measures $\bar{P}_n^i$ and $\bar{P}_n^j$, of non-negative $\bar{p}^i$ a.s. terms).

Indeed, in agreement with the Jensen inequality

$$E_j((\alpha_n^{ij})^{1/2}/\mathcal{F}_{n-1}^{\xi}) \leq (E_j(\alpha_n^{ij}/\mathcal{F}_{n-1}^{\xi}))^{1/2} = 1 \qquad \bar{P}^j - \text{a.s.} \tag{20}$$

Since the series

$$\sum_{n=1}^{\infty}|\ln E_j((\alpha_n^{ij})^{1/2}/\mathcal{F}_{n-1}^{\xi})|$$

$$\sum_{n=1}^{\infty}\left(1 - E_j((\alpha_n^{ij})^{1/2}/\mathcal{F}_{n-1}^{\xi})\right)$$

under condition (19) converge and diverge simultaneously with $\bar{P}^i$ and $\bar{P}$ a.s., in the case of Gaussian measures it is easier to use singularity criteria in terms of divergence of the series

$$\sum_{n=1}^{\infty}|\ln E_j((\alpha_n^{ij})^{1/2}/\mathcal{F}_{n-1}^{\xi})|. \tag{21}$$

Using the fact that the process ξ is Gaussian on probabilistic spaces $(\Omega, \mathcal{F}^{\xi}, \bar{P}^i)$, $i \in \mathbf{N}$ and the representation (18) for z_n^{ij} the following expressions for α_{n-1}^{ij} are readily obtained

$$\alpha_n^{ij} = \frac{|D_i(n)|^{1/2}}{|D_j(n)|^{1/2}}\exp\left\{-\frac{1}{2}(\xi_n - A(i)\,m_{n-1}^i)^*\,D_i(n)\,(\xi_n - A(i)\,m_{n-1}^i) + \right.$$

$$\left. + \frac{1}{2}(\xi_n - A(j)\,m_{n-1}^j)^*\,D_j(n)\,(\xi_n - A(j)\,m_{n-1}^j)^*\,D_j(n)\,(\xi_n - A(j)\,m_{n-1}^j)\right\}, \tag{22}$$

$$i, j \in \mathbf{N}, \; i \neq j, \; n \geq 0.$$

Making simple transforms of this expression we receive the singularity condition in the form:

$$\bar{P}^j\left\{\sum_{n=0}^{\infty}\ln\frac{\sqrt{2}\,|D_i(n)|^{1/4}\,|D_j(n)|^{1/4}}{|D_i(n)+D_j(n)|^{1/2}}-\right. \tag{23}$$

$$\left.-\frac{1}{4}\,(A(i)\,m_n^i-A(j)\,m_n^j)^*\,D_i(n)\,(D_i(n)+D_j(n))^{-1}\,(A(i)\,m_n^i-A(j)\,m_n^j)=\infty\right\}=1.$$

Condition (23) is not very convenient for checking because convergence or divergence $\bar{P}^j$ — a.s. for the series

$$\sum_{n=1}^{\infty}(A(i)\,m_n^i-A(j)\,m_n^j)^*\,D_i(n)\,(D_i(n)+D_j(n))^{-1}\,D_j(n)\,(A(i)\,m_n^i-A(j)\,m_n^j)$$

$$i,j\in\mathbf{N},\ i\neq j \tag{24}$$

is difficult to ascertain directly. If the random vectors $(A(i)\,m_n^i-A(j)\,m_n^j)$ are Gaussian, investigation of convergence or divergence of the series (24) can be reduced to that of the deterministic series [11]:

$$\sum_{n=1}^{\infty}E_j(A(i)\,m_n^i-A(j)\,m_n^j)\,D_i(n)\,(D_i(n)+D_j(n))^{-1}\,(A(i)\,m_n^i-A(j)\,m_n^j). \tag{25}$$

The fact that these vectors are Gaussian follows from

Lemma 4. The processes $(m_n^i)_{n\geq 0}$, $i\in\mathbf{N}$ are Gaussian on the probabilistic spaces $(\Omega,\mathfrak{F}_t^{\xi},\bar{P}^k)$, $k\in\mathbf{N}$; k is not necessarily equal to i.

Proof. That the processes $(m_n^i)_{n\geq 0}$ $i\in\mathbf{N}$ investigated on probabilistic spaces $(\Omega,\mathfrak{F}^{\xi},\bar{P}^i)$, $i\in\mathbf{N}$ are Gaussian follows from the findings of Ch. 13 in Ref. [14]; let us prove that the processes $(m_n^i)_{n\geq 0}$, $i\in\mathbf{N}$ are Gaussian on the probabilistic spaces $(\Omega,\mathfrak{F}^{\xi},P^k)$, $k\neq i$, $k\in\mathbf{N}$. Because the simultaneous initial distribution of the quantities θ_0, ξ_0 on any of the probabilistic spaces $(\Omega,\mathfrak{F}^{\theta\xi},P^i)$, $i\in\mathbf{N}$ is Gaussian, the quantities $m_0^j=E_j(\theta_0/\mathfrak{F}_0^{\xi})$ are linear functions of ξ_0 with any $j\in\mathbf{N}$ and, consequently, simultaneous distributions of the quantities m_0^j, θ_0, $j\in\mathbf{N}$ are also Gaussian but also on the probabilistic spaces $(\Omega,\mathfrak{F}^{\theta\xi},\bar{P}^k)$ where $k\in\mathbf{N}$, $k\neq j$ as well as on the probabilistic spaces $(\Omega,\mathfrak{F}^{\theta\xi},\bar{P}^k)$. Substituting into the expression for the values of m_n^j ξ_n from equation (1) we have

$$m_{n+1}^j=[a(j)-K_j(n)\,A(j)]\,m_n^i+K_j(n)\,A(\beta)\,\theta_n+$$

$$+K_j(n)\,B_1(\beta,n)\,\varepsilon_{1n+1}+K_j(n)\,B_2(\beta,n)\,\varepsilon_{2n+1},\ m_0^j. \tag{26}$$

That the simultaneous distribution of (m_0^j,θ_0,ξ_0) is Gaussian and the expressions for θ_n and m_n^j, $j\in\mathbf{N}$, $n\geq 0$ are linear makes the processes $(m_n^j)_{n\geq 0}$,

$j \in \mathbf{N}$ and $(\xi_n)_{n\geq 0}$ Gaussian on the probabilistic spaces $(\Omega, \mathcal{F}^{\xi}, \bar{P}^k)$, $k \in \mathbf{N}$, $k \neq j$, which was required. Consequently, the expression $A(i)\, m_n^i - A(j)\, m_n^j$ is a Gaussian vector for any of the measures $\bar{P}^i$ and $\bar{P}^j$, $i \neq j$, $i, j \in \mathbf{N}$. Note now that for any two symmetrical nonnegatively definite matrices $D_i(n)$ and $D_j(n)$ it is true that

$$|D_i(n) + D_j(n)| \geq |D_i(n)| + |D_j(n)|.$$

Therefore

$$\frac{2\,|D_i(n)|^{1/2}\,|D_j(n)|^{1/2}}{|D_i(n) + D_j(n)|} \leq \frac{2\,|D_i(n)|^{1/2}\,|D_j(n)|^{1/2}}{|D_i(n)| + |D_j(n)|} \tag{27}$$

and, consequently,

$$\ln \frac{\sqrt{2}\,|D_i(n)|^{1/4}\,|D_j(n)|^{1/4}}{|D_i(n) + D_j(n)|^{1/2}} \leq 0\,. \tag{28}$$

Using expression (25), condition (28), and the fact that expression (27) is deterministic, the condition for singularity of the measures $\bar{P}^i$ and $\bar{P}^j$, $i \neq j$, $i, j \in \mathbf{N}$ can be rearranged into

$$\sum_{n=0}^{\infty} \ln \frac{|D_i(n) + D_j(n)|}{\sqrt{2}\,|D_i(n)|^{1/4}\,|D_j(n)|^{1/4}} + \frac{1}{4}\,[E_j(A(i)\, m_n^i - A(j)\, m_n^j)^*\, D_i(n) \times$$
$$\times\, (D_i(n) + D_j(n))^{-1}\, D_j(n)\, (A(i)\, m_n^i - A(j)\, m_n^j)] = \infty\,.$$

For calculation of the mathematical expectation incorporated into this condition perform certain additional constructions.

Let $\varphi_{ij}(n)$ denote the block matrix

$$\begin{pmatrix} E_j(\theta_n \theta_n^*) & E_j(\theta_n\, m_n^i{}^*) & E_j(\theta_n\, m_n^j{}^*) \\ E_j(m_n^i\, \theta_n^*) & E_j(m_n^i\, m_n^i{}^*) & E_j(m_n^i\, m_n^j{}^*) \\ E_j(m_n^j\, \theta_n^*) & E_j(m_n^j\, m_n^i{}^*) & E_j(m_n^i\, m_n^j{}^*) \end{pmatrix}$$

and S_{ij}, a block matrix

$$(0, -A(i), A(j))\mathbf{i}$$

Using the well-known property of matrix multiplication [15], one can write

$$E_j(A(i)\, m_n^i - A(j)\, m_n^j)^*\, D_i(n)\, (D_i(n) + D_j(n))^{-1}\, D_j(n)\, (A(i)\, m_n^i - A(j)\, m_n^j) =$$
$$= \mathrm{Sp}\, E_j(A(i)\, m_n^i - A(j)\, m_n^j)\, (A(i)\, m_n^i - A(j)\, m_n^j)^*\, D_i(n)\, (D_i(n) + D_j(n))^{-1}\, D_j(n).$$

Direct verification shows that the matrix $E_j(A(i)\, m_n^i - A(j)\, m_n^j) \times (A(i)\, m_n^i - A(j)\, m_n^j)^*$ can be represented in terms of the matrix $\varphi_{ij}(n)$

$$E_j(A(i)\, m_n^i - A(j)\, m_n^j)\, (A(i)\, m_n^i - A(j)\, m_n^j)^* = S_{ij}\, \varphi_{ij}(n)\, S_{ij}^*.$$

Consequently

$$E_j(A(i)\,m_n^i - A(j)\,m_n^j)^*\,D_i(n)\,(D_i(n) + D_j(n))^{-1}D_j(n)\,(A(i)\,m_n^i - A(j)\,m_n^j) =$$
$$= \operatorname{Sp} S_{ij}^{\blacksquare}\varphi_{ij}(n)\,S_{ij}^*\,D_i^{\blacksquare}(n)\,(D_i(n) + D_j(n))^{-1}D_j(n).$$

Find now the recurrence equation for the matrix $\varphi_{ij}^{\blacksquare}(n)$. Take up a block matrix $F_{ij}(n)$ which has the form

$$F_{ij}(n) = \begin{pmatrix} \theta_n \\ m_n^i \\ m_n^j \end{pmatrix}(\theta_n^*\, m_n^{i*}, m_n^{j*}) = \begin{pmatrix} \theta_n\,\theta_n^* & \theta_n\,m_n^{i*} & \theta_n\,m_n^{j*} \\ m_n^i\,\theta_n^* & m_n^i\,m_n^{i*} & m_n^i\,m_n^{j*} \\ m_n^j\,\theta_n^* & m_n^j\,m_n^{i*} & m_n^j\,m_n^{j*} \end{pmatrix}.$$

Using the expression for m_{n+1}^i, m_{n+1}^j, θ_{n+1} and the notation for the matrix $R_{ij}(n)$, $Q_i(n)$, $G_i(n)$, we shall obtain the equation for $\varphi_{ij}(n) = E_i(F_{ij}(n))$ which have the form of (10).

Consequently, the condition for singularity of the measures $\bar{P}^i$ and $\bar{P}^j$, $i \neq j$, $i, j \in \mathbf{N}$ is equivalent to condition (7) of the Theorem; thence, using the necessary and sufficient condition for consistency of the estimate in the form of conditional mean is obtained. Let us prove that condition (7) is sufficient for convergence of the estimate $\hat{\beta}_2$ that maximizes the a posteriori probability $\pi_j(n)$, $j \in \mathbf{N}$.

From the findings of [13] it follows that (7) is equivalent to convergence of $\pi_j(n) \to I(\beta = \beta_j)\ \bar{P}^j$ a.s. Let $\varepsilon > 0$ be an arbitrary small number. By virtue of the normalizing condition $\sum_{i \in \mathbf{N}} p_i = 1$ there is a number M such that

$$P\left(\bigcup_{i=M}^{\infty}\{\beta = \beta_i\}\right) \leq \sum_{i=M}^{\infty} p_i < \frac{\varepsilon}{2}. \tag{29}$$

Let us show that $I\left(\bigcup_{i=1}^{M}\{\beta = \beta_i\}\right)\hat{\beta}_{2n} \to T\left(\bigcup_{i=1}^{M}\{\beta = \beta_i\}\right)\beta\bar{P}$ — a.s. Indeed, from convergence of $\pi_j(n) \to I(\beta = \beta_j)\,\bar{P}$ — a.s. it follows that for the chosen ε an an arbitrary $\delta > 0$ there is an ordinal number m_j such that

$$P\Big(\bigcup_{n \geq m_j}\{\pi_j(n) < 1 - \delta\} \cap \{\beta = \beta_j\}\Big) < \frac{\varepsilon}{2^j}.$$

From the definition of $\hat{\beta}_{2n}$ with reasonably small δ and $n \geq m_j$ it follows that

$$\{\beta = \beta_j\} \cap \{\hat{\beta}_{2n} \neq \beta_j\} \subset \{\pi_j(n) < 1 - \delta\} \cap \{\beta = \beta_j\}.$$

Uniting the sets in both parts of the embedding, we have

$$\bigcup_{n \geq m_j}\{\beta = \beta_j\} \cap \{\hat{\beta}_{2n} \neq \beta_j\} \subset \bigcup_{n \geq m_j}\{\pi_j(n) < 1 - \delta\} \cap \{\beta = \beta_j\}.$$

Hence

$$P\Big(\bigcup_{n\geq m_j}\{\beta=\beta_j\}\cap\{\hat{\beta}_{2n}\neq\beta_j\}\Big)<\frac{\varepsilon}{2^j}.$$

Denoting $m=\max\limits_{1\leq j\leq M} m_j$ we have

$$P\Big(\bigcup_{n\geq m}\bigcup_{j=1}^{M}\{\beta=\beta_{jj}\}\cap\{\hat{\beta}_{2n}\neq\beta_j\}\Big)<\frac{\varepsilon}{2}. \tag{30}$$

Noticing that

$$P\Big(\bigcup_{n\geq m}\{\hat{\beta}_{2n}\neq\beta\}\Big)\leq P\Big(\bigcup_{n\geq m}\bigcup_{j=1}^{M}\{\beta=\beta_j\}\cap\{\beta_{2n}\neq\beta_j\}\Big)+P\Big(\bigcup_{j=M}^{\infty}\{\beta=\beta_j\}\Big)$$

we have from the estimates (29) and (30) that for an arbitrary $\varepsilon>0$ there is an ordinary number m such that

$$P\Big(\bigcup_{n\geq m}\{\hat{\beta}_{2n}=\beta\}\Big)<\varepsilon$$

which proves convergence of $\hat{\beta}_{2n}$ to $\beta\bar{P}$ — a.s.

Consistency of an estimate such as conditional mean is easily seen to follow from that of an arbitrary F^{ξ}-consistent estimate. Indeed, let $\hat{\beta}_n$ be an arbitrary F^{ξ}-consistent estimate of the random quantity β. Consistency of the estimate $\hat{\beta}_n$ entails $\mathfrak{F}^{\xi}$-measurability of the random quantity β. By virtue of the Levy theorem, the Bayesian estimate $\hat{\beta}_{1n}=E(\beta/\mathfrak{F}_n^{\xi})\to E(\beta/\mathfrak{F}^{\xi})$ a.s. and thence by virtue of $\mathfrak{F}^{\xi}$-measurability of β follows consistency of $\hat{\beta}_{1n}$, Q.E.D

References

1. *Pugachev, V. S.*, Theory of Random Functions and Its Application to Problems of Automatic Control. Fizmatgiz., Moscow, 1962 (in Russian).
2. *Fang, B.*, Kalman–Bucy Filter for Optimum Radioinertial Navigation. IEEE Trans. Automat. Contr. Vol. **12**, No. *4*, 1967.
3. *Novosel'tseva, Zh. A.*, A priori Information in Optimal Filtering Problems. Avtomatika i telemekhanika, Vol. **39**, No. *6*, 1968 (in Russian).
4. *Kitsul, P. I., Lubkov, A. V.*, On Design of Pseudogradient Algorithms for Adjustment of an Adaptive Kalman Filter. Avtomatika i telemekhanika, Vol. **37**, No. *4*, 1976 (in Russian).
5. *Mehra, R.*, "On-line Identification of Linear Dynamic Systems with Application to Kalman Filtering", IEEE Trans. Automat. Contr. Vol. **AC–16**, No. *1*, 1971.
6. *Leondes, C. T., Person, J. O.*, Kalman Filtering of Systems with Parameters Uncertainties, a Survey. Int. J. Control, Vol. **17**, No. *4*, 1973.
7. *Ljung, L.*, Convergence Analysis of Parametric Identification Methods. IEEE Trans. Automat. Contr. Vol. **23**, No. *4*, August 1978.
8. *Baram, Y., Sandell, N. R.*, An Information Theoretic Approach to Dynamical System Modeling and Identification. IEEE Trans. Automat. Contr. Vol. **AC–23**, No. *1*, 1978.
9. *Baram, Y., Sandell, N. R.*, Consistent Estimation on Finite Parameter Sets with Application to Linear Systems Identification. IEEE Trans. Automat. Contr. Vol. **AC–23**, No. *3*, 1978.

10. *Baram, Y.*, A Sufficient Condition for Consistent Discrimination Between Stationary Gaussian Models. IEEE Trans. Automat. Contr. Vol. **AC–23**, No. *5*, 1978.
11. *Kabanov, Yu. M., Liptser, R. Sh., Shiryaev, A. N.*, On Absolute Continuity and Singularity of Probabilistic Measures. Matematischeskiy sbornik. Vol. **104** (106), No. *2* (10), 1977 (in Russian).
12. *Kabanov, Yu. M., Liptser, R. Sh., Shiryaev, A. I.*, Absolute Continuity and Singularity of Locally Absolutely Continuous Probabilistic Distributions. Matematicheskiy sbornik, Vol. **107** (149), No. *3* (11), 1978 (in Russian).
13. *Yashin, A. I.*, On Consistency of Bayesian Parameter Estimates. Problem peredachi informatsii, No. *1*, 1981 (in Russian).
14. *Liptser, R. Sh., Shiryaev, A. I.*, Statistics of Random Processes. Nauka, Moscow, 1974 (in Russian).
15. *Gantmakher, F. R.*, Theory of Matrices. Nauka, Moscow, 1966 (in Russian).

О состоятельности оценивания параметров в схеме адаптивной фильтрации

Н. А. КУЗНЕЦОВ, А. И. ЯЩИН

(Москва)

В работе исследуются необходимые и достаточные условия состоятельности оценок параметров в схеме адаптивной Калмановской фильтрации. Считается, что неизвестный параметр определяет эволюцию как наблюдаемого, так и ненаблюдаемого процессов и принимает значения из некоторого счетного множества. Вывод необходимых и достаточных условий состоятельности использует эквивалентность этих условий свойству сингулярности вероятностных мер, соответствующих наблюдаемому процессу при фиксированных значениях параметра.

Условия сингулярности мер выражаются в терминах характеристик отношений правдоподобия мер на суженных σ-алгебрах. В случае адаптивной калмановской фильтрации эти условия удается выразить в терминах коэффициентов исходных уравнений и характеристик фильтра, что удобно для приложений, поскольку позволяет проверить свойство состоятельности оценок до проведения эксперимента или наблюдений.

Н. А. Кузнецов
Институт проблем управления
СССР, 117342, Москва, ГСП-7
Профсоюзная, 65

Problems of Control and Information Theory, Vol. 10 (5), pp. 329—339 (1981)

ПРИМЕНЕНИЕ АДАПТИВНЫХ ФИЛЬТРОВ КАЛМАНА ДЛЯ ПРЕДСКАЗАНИЯ НАГРУЗКИ В ЭЛЕКТРОЭНЕРГЕТИЧЕСКИХ СИСТЕМАХ НА КРАТКОВРЕМЕННЫЕ ИНТЕРВАЛЫ

Х. ХИНКЕЛЬ

(Берлин)

(Поступила в редакцию 7 ноября 1980 г.)

Рассматривается проблема предсказания нагрузки в электроэнергетических системах на интервал длины 10 мин — 1 час. Для предсказания применяется специальный метод адаптивной фильтрации. Практическое испытание было проведено как над моделированными, так и над реальными данными.

1. Введение

Рациональное производство и распределение электроэнергии представляет собой задачу большого народнохозяйственного значения. Для быстрого, целенаправленного и по возможности автоматизированного воздействия на производство и распределение электроэнергии особенно важно определение будущей нагрузки в общей системе или в подсистеме. Эта проблема называется проблемой предсказания нагрузки на кратковременные интервалы.

Цель проведенных в данной работе исследований состояла в изучении возможностей использования адаптивных фильтров Калмана для решения проблем предсказания.

Из-за отсутствия данных, описывающих непосредственно нагрузку системы, предсказание сначала было проведено для реальной последовательности данных для сетевой частоты, причем отклонение сетевой частоты от номинальной частоты $f = 50$ гц является мерой разности нагрузки и производства электроэнергии.

2. Оценка состояния при неизвестных ковариациях возмущений

Как известно, применение дискретных фильтров Калмана требует знания дискретной модели состоянии вида

$$\begin{aligned} x_{k+1} &= Fx_k + u_k, \\ y_k &= Hx_k + v_k, \end{aligned} \tag{2.1}$$

которая описывает предсказываемый процесс, где

$\{x_k\}$ — состояние системы

$\{y_k\}$ — наблюдаемый выходной процесс.

Возмущения $\{u_k\}$ и $\{v_k\}$ — дискретные случайные процессы белого шума. Предположим, что эти процессы являются центрированными, независимыми и стационарными.

Для построения фильтра Калмана, кроме знания матриц F и H, требуется и знание ковариационных матриц $Q = E(u_k u_k^T)$ и $R = E(v_k v_k^T)$ процессов возмущений.

В случае кратковременного предсказания нагрузки можно получить несложную модель предсказуемого процесса (см. разд. 4), но матрицы Q и R в этом случае априорно неизвестны. Поэтому сначала невозможно построить фильтр Калмана по известному соотношению

$$\hat{x}_{k+1} = F\hat{x}_k + L_k[y_k - H\hat{x}_k] \tag{2.2}$$

при

$$L_k = FM_k H^T(HM_k H^T + R)^{-1}, \quad \hat{x}(0) = Ex_0, \tag{2.3}$$

где ковариация $M_k = Ee_k e_k^T$ ошибки оценки $e_k = x_k - \hat{x}_k$ является решением дискретного матричного уравнения Риккати вида

$$M_{k+1} = FM_k F^T - FM_k H^T(HM_k H^T + R)^{-1} HM_k F^T + Q, \quad M_0 = Ex_0 x_0^T. \tag{2.4}$$

Проблемы этого рода и возможности их решения известны из литературы. Существует ряд работ, в которых рассматриваются так называемые адаптивные фильтры [2]. По нашему мнению для предсказания нагрузки на кратковременные интервалы самым пригодным методом из методов адаптивной фильтрации является метод корреляции [2], так как в данном случае приближенно выполняются требования этого метода.

3. Метод корреляции невязок

Основная идея метода корреляции состоит в том, чтобы провести корреляцию выходной величины системы или другой характерной для поведения системы величины непосредственно или после определенной линейной операции. В результате корреляции получим систему уравнений, в которой параметры системы соотнесены с оцененными автокорреляционными функциями. Сравнительные исследования показали, что так называемый метод корреляции невязок, где невязки (Innovations)

$$v_i = y_i - \hat{y}_i = y_i - H\hat{x}_i$$

выходного процесса $\{y_i\}$ взаимно коррелируются, особенно пригоден для предсказания нагрузки на кратковременные интервалы.

Условие для применимости этого метода состоит в том, что рассматриваемая система (2.1) является полностью управляемой и полностью наблюдаемой.

Вместо характеристики L_k фильтра, которая представляет собой функцию времени, используется постоянная характеристика L фильтра ($\lim_{k\to\infty} L_k = L$) и вместо формулы (2.2) в дальнейшем будет рассматриваться соответствующий стационарный фильтр (steady-state filter)

$$\hat{x}_{k+1} = F\hat{x}_k + L[y_k - H\hat{x}_k]. \tag{3.1}$$

Кроме того предположим, что фильтр рассматривается в установившемся режиме. Отсюда следует, что $M_k \equiv M$ const.

В том случае, если для ковариаций M и R знаем некоторые приближенные значения, можно определить характеристику L^* субоптимального фильтра и построить субоптимальный фильтр

$$x^*_{k+1} = Fx^*_k + L^*[y_k - Hx^*_k].$$

Существо метода корреляции невязок заключается в получении оценки характеристики L оптимального фильтра, зная оценки ковариаций последовательности невязок заданного субоптимального фильтра.

В литературе описаны два различных подхода к решению [2, 3]. В данной работе проведенные исследования имели сначала целью сравнения эффективности обоих методов корреляции невязок, а в результате этого использования более пригодного метода для предсказания нагрузки на кратковременные интервалы. При внимательном теоретическом рассмотрении удалось показать, что эти два подхода к решению в действительности несущественно отличаются друг от друга. Поэтому можно ограничиться использованием и реализацией одного из обоих методов.

В следующих разделах 3.1 и 3.2 изложены основные идеи обоих методов, а в разделе 3.3 показано соотношение между этими методами.

3.1 Метод корреляции невязок (варианта I) [2]

Будем исходить из субоптимального фильтра Калмана для системы (2.1) в виде

$$x^*_{i+1} = Fx^*_i + L^*[y_i - Hx^*_i]. \tag{3.2}$$

Последовательность невязок $\{v^*_i\}$ субоптимального фильтра при

$$v^*_i = y_i - y^*_i = y_i - Hx^*_i = He^*_i + v_i$$

не является белым шумом.

Для ковариаций этой последовательности невязок при условии, что субоптимальный фильтр работает в установившемся режиме имеем

$$T_k^* = E\{\nu_i^* \nu_{i-k}^{*T}\} = \begin{cases} HM^* H^T + R, & \text{если} \quad K = 0 \\ HEe_i^* e_{i-k}^{*T} H^T + HEe_i^* \nu_{i-k}^T, & \text{если} \quad K > 0. \end{cases}$$

Здесь M^* — ковариация ошибки $e_i^* = x_i - x_i^*$ субоптимального фильтра, т. е. $M^* = Ee_i^* e_i^{*T}$.

Система уравнений, полученная из соотношения действительного для T_k^* $(k = 1, \ldots n)$, преобразуется следующим образом

$$AM^* H^T = \begin{bmatrix} T_1^* + HL^* T_0^* \\ T_2^* + HL^* T_1^* \quad + HFL^* T_0^* \\ \vdots \\ T_n^* + HL^* T_{n-1}^* + \quad \ldots \quad + HF^{n-1} L^* T_0^* \end{bmatrix} = E, \quad (3.3)$$

где

$$A^T = [(HF)^T, (HF^2)^T, \ldots, (HF^n)^T].$$

Обозначая псевдоинверсию матрицы A знаком A^+, получим

$$M^* H^T = (A^T A)^{-1} A^T E = A^+ E.$$

В качестве асимптотически несмещенной и состоятельной оценки для T_k^* используется оценка

$$\hat{T}_k^* = \frac{1}{N} \sum_{i=k}^{N} \nu_i^* \nu_{i-k}^{*T} \qquad k = 1, 2, \ldots . \quad (3.4)$$

Зная $\widehat{M^* H^T}$, непосредственно можно определить характеристику L оптимального фильтра. Если для разности между ковариацией M ошибки оценки оптимального фильтра и ковариацией M^* введем обозначение $\delta M = M - M^*$, то оценку характеристики L оптимального фильтра можно представить в виде.

$$\hat{L} = F(\widehat{M^* H^T} + \delta M H^T)(\hat{T}_0^* + H\delta M H^T)^{-1}.$$

Можно показать, что величина δM удовлетворяет матричному разностному дифференциальному уравнению вида

$$\delta M = F[\delta M - (\widehat{M^* H^T} + \delta M H^T)(\hat{T}_0^* + H\delta M H^T)^{-1}(\widehat{HM^*} + H\delta M)] F^T + C, \quad (3.6)$$

где

$$C = L^* \widehat{HM^*} F^T + F\widehat{M^* H^T} L^{*T} - L^* \hat{T}_0^* L^{*T}.$$

Подставив определенную из этого уравнения величину δM в соотношение для $\hat{L}$, получим асимптотически несмещенные и состоятельные оценки для характеристики L оптимального фильтра.

3.2. *Метод корреляции невязок (варианта* II) [3]

Модель состояния вида (2.1) может быть известным образом формально преобразована в другую эквивалентную модель состояния (3.7), в которой оптимальные оценки $\hat{x}_i$ выступают в качестве переменных состояния

$$\begin{aligned} \hat{x}_{i+1} &= F\hat{x}_i + L\nu_i \\ y_i &= Hx_i + \nu_i. \end{aligned} \tag{3.7}$$

Последовательность $\{\nu_i\}$ возмущений здесь равна последовательности невязок оптимального фильтра.

Для системы (3.7) можно определить субоптимальный фильтр Калмана (3.2), с помощью которого получим субоптимальные оценки x^*_{i+1} для $\hat{x}_{i+1}$.

Ковариация P^* ошибки оценки $\tilde{e}^*_i = \hat{x}_i - x^*_i$ в установившемся режиме является существенной величиной в этом методе. Для указанной ковариации имеем

$$\begin{aligned} P^* &= E\{(\hat{x}_{i+1} - x^*_{i+1})(\hat{x}_{i+1} - x^*_{i+1})^T\} = \\ &= (F - L^*H)P^*(F - L^*H)^T + (L^* - L)T_0(L^* - L)^T, \end{aligned} \tag{3.8}$$

где

$$T_0 = E\nu_i\nu_i^T = HMH^T + R.$$

Аналогично методу корреляции невязок варианта I определяются ковариации T^*_k последовательности невязок субоптимального фильтра (3.2), работающего в установившемся режиме. Ковариации T^*_k здесь представлены в качестве функций от P^*.

Применяя для T^*_k ($k = 1, \ldots n$) оценки по формуле (3.4), можно найти соотношение, в котором матрица L^* представляет собой функцию от T_0 и P^*. Учитывая соотношение $T_0 = T^*_0 - HP^*H^T$, будем иметь три одновременных уравнения для P^*, T_0 и $\hat{L}$, которые могут быть решены посредством специального сходящегося алгоритма.

3.3. *Сравнение методов корреляции невязок* I *и* II

В результате наблюдений, проведенных в работе, было установлено, что оба метода должны вести к одному и тому же результату, так как величина P^* лишь незначительно отличается от δM. Это можно продемонстрировать следующим образом.

В обоих методах используется тот же самый субоптимальный фильтр (3.2). Между ковариацией M^* ошибки оценки $e_i^* = x_i - x_i^*$ и ковариацией P^* ошибки оценки $\tilde{e}_i^* = \hat{x}_i - x_i^*$ существует простая связь

$$\begin{aligned} M^* &= E e_i^* e_i^{*T} = E(x_i - x_i^*)(x_i - x_i^*)^T = \\ &= E[(x_i - \hat{x}_i) + (\hat{x}_i - x_i^*)][(x_i - \hat{x}_i) + (\hat{x}_i - x_i^*)]^T = \\ &= E(x_i - \hat{x}_i)(x_i - \hat{x}_i)^T + E(\hat{x}_i - x_i^*)(\hat{x}_i - x_i^*)^T. \end{aligned} \tag{3.9}$$

Из этого следует, что

$$M^* = M + P^*,$$

соответственно

$$P^* = M^* = -\delta M. \tag{3.10}$$

Итак, оба известных метода корреляции невязок отличаются только в подходе к определению величин P^* и L и соответственно величин δM и $\hat{L}$. Систему одновременных уравнений для определения величины P^*, T_0, $\hat{L}$ можно непосредственно преобразовать в уравнение, тождественное уравнению (3.6).

В результате выше сказанного можно ограничиться реализацией на ЭВМ только одного из этих методов. Из-за более простой структуры был реализован на ЭВМ метод корреляции невязок варианта I.

4. Результаты, полученные в эксперименте над моделированными и над реальными данными

В частности из-за практических потребностей в электрохозяйстве ГДР в этом специальном случае рассмотрена только проблема предсказания нагрузки на интервалы длины 10 мин — 1 час.

Для этого в литературе [4] задана следующая простая модель второго порядка системы (4.1), причем в качестве переменных состояния (составляющих от x_k) выбираются

— нагрузка системы ξ_k,
— приращение Δ_k нагрузки системы.

Модель системы имеет вид

$$\begin{bmatrix} \xi_{k+1} \\ \Delta_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \xi_k \\ \Delta_k \end{bmatrix} + \begin{bmatrix} u_k^1 \\ u_k^2 \end{bmatrix}, \tag{4.1}$$

$$y_k = [1 \quad 0] \begin{bmatrix} \xi_k \\ \Delta_k \end{bmatrix} + v_k.$$

Возмущения $\{u_k^1\}$ и $\{u_k^2\}$ являются дискретными стохастическими процессами, которые взаимно независимы и независимы от возмущения $\{v_k\}$.

Далее предполагается, что

$$E(u_k^1) = E(u_k^2) = E(v_k) = 0.$$

Для ковариаций возмущений имеем

$$E(u_i^1 u_j^1) = q_i^{11}\delta_{ij} \qquad E(u_i^2 u_j^2) = q_i^{22}\delta_{ij} \qquad E(v_i v_j) = R_i \delta_{ij},$$

где δ_{ij} — символ Кронекера. Величины q_i^{11}, q_i^{22} и R_i априорно неизвестны.

В дальнейшем эта простая модель используется и при предсказании сетевой частоты, причем сетевая частота и ее приращение являются переменными состояния. Предполагается, что в этом случае возмущения приближенно стационарные, и поэтому $Q_k = Q = \text{const.}$ и $R_k = R$ const.

4.1. *Эксперименты над моделированными данными*

Условием применения методов корреляции невязок является знание последовательности невязок субоптимального фильтра, работающего в установившемся режиме. Для построения такой последовательности должны быть выполнены два требования;

а) должна быть в наличии последовательность измеренных значений $\{y_n\}$ для рассматриваемой системы (2.1);
б) должна быть известна характеристика L^* субоптимального фильтра.

Такую последовательность $\{y_n\}$ можно получить при помощи моделирования соответствующей системы (2.1). Очевидно, что модель (4.1) неудобна для моделирования (корни характеристического уравнения $|\lambda I - F| = 0$ расположены на единичном круге). Поэтому с целью моделирования модель (4.1) была изменена таким образом, чтобы полученная система была как устойчивой, так и полностью управляемой и полностью наблюдаемой.

В качестве матриц было выбрано

$$F^M = \begin{bmatrix} 0.8 & 1 \\ 0 & 0.8 \end{bmatrix} \qquad H^M = [1\ 0]. \tag{4.2}$$

Для ковариаций возмущений стационарных процессов белого шума было задано

$$Q^M = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad R^M = 0.5. \tag{4.3}$$

В результате моделирования этой системы на ЭВМ была получена последовательность данных $\{y^M\}$, необходимая для выполнения требования

а). Характеристика L оптимального фильтра для системы, описанной соотношениями (4.2) и (4.3) может быть определена с помощью формул (2.3) и (2.4).

В том случае, если используются ковариации возмущений Q^{M*} и R^{M*} мало измененные по сравнению с матрицами Q^M и R^M, то с помощью формул (4.2) и (4.3) можно определить необходимые для выполнения требования б) характеристики L^* субоптимальных фильтров.

Используя последовательность моделированных данных $\{y^M\}$ при помощи программы, составленной для метода I корреляции невязок, были вычислены, исходя из соответствующих характеристик L^* субоптимальных фильтров, характеристики $\hat{L}$.

При сравнении характеристик $\hat{L}$ с характеристикой L оптимального фильтра оказалось, что качество полученных результатов незначительно зависит от выбора характеристик субоптимальных фильтров.

При выбранной для экспериментальной цели характеристике субоптимального фильтра вида $L^* = 0$ выяснилось, что в данном случае качество полученных результатов, хуже чем в случае теоретически определенных с помощью матриц Q^{M*} и R^{M*} характеристик субоптимальных фильтров, но, несмотря на него полученные в этом специальном случае моделирования результаты являются достаточно хорошими. Этот факт мотивирует использование пробной характеристики $L^* = 0$ при предсказании нагрузки на кратковременных интервалах, для которой не существует априорной информации о возмущениях.

Было показано, что фильтр, построенный на основе оцененной характеристики $\hat{L}$ оптимального фильтра, позволяет образовать последовательности невязок, свойства которых очень близки свойствам белого шума. Таким образом подтверждено, что действительно возможно построение почти оптимального фильтра на основе характеристики $\hat{L}$ фильтра, и согласно соотношениям

$$\hat{y}(K+\delta) = H\hat{x}(K+\delta) = HF^{\delta-1}\hat{x}(K), \quad (\delta \geq 1) \tag{4.4}$$

построение соответствующего упредителя в смысле Калмана.

На рис. 1 для установившегося режима показаны последовательность моделированных данных и полученные при помощи составленной здесь программы предсказания результаты одношагового предсказания (отрезок).

4.2. *Эксперименты над данными для сетевой частоты*

Тест программы для предсказания нагрузки на кратковременные интервалы посредством адаптивной фильтрации был проведен сначала над данными для сетевой частоты. Была задана последовательность с 5000 значениями, измеренными каждые 10 минут (продолжительность с 5000 значения-

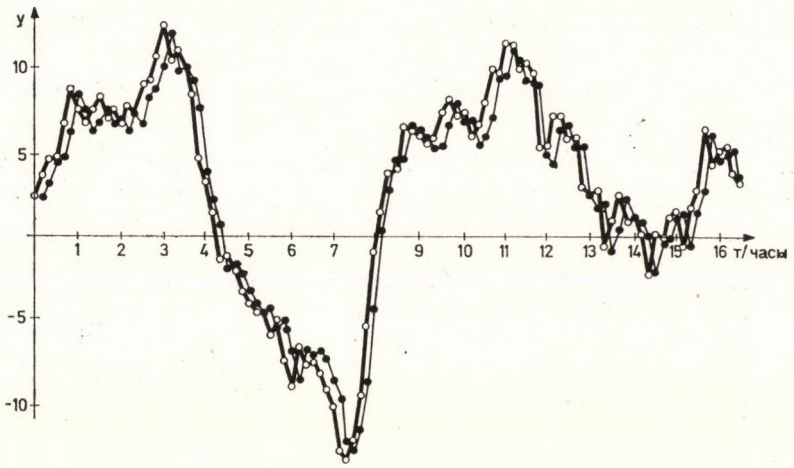


Рис. 1. Одношаговое предсказание для последовательности моделированных данных

ми, измеренными каждые 10 минут (продолжительность измерения: около 35 дней). В основу определения характеристики L фильтра была положена простая модель системы (4.1). При этом возмущения приближенно предполагались стационарными.

Чтобы оценить, каким образом следует выбрать число K рабочих шагов субоптимального фильтра до установившегося режима и число N необходимых для оценки T_K^* значений невязок, были использованы различные комбинации параметров ($K = 20$; $N = 1000$, $N = 1500$; $K = 50$; $N = 1000$, $N = 1500$, $N = 2000$).

$L^* = 0$ было выбрано в качестве характеристики субоптимального фильтра.

Но уже простые статистические критерии показали, что следует отвергнуть гипотезу, которая заключается в том, что последовательность невязок полученных фильтров является белым шумом. Поэтому невозможно построить оптимальный фильтр Калмана с помощью определенных здесь характеристик $\hat{L}$.

Причина этого может заключаться в невыполнении требования стационарности положенной в основу последовательности данных сетевой частоты. Другая причина может состоять в использовании слишком простой модели (4.1).

Несмотря на это оказалось, что можно получить достаточно хорошее одношаговое предсказание, применяя определенную, несмотря на неоптимальность, наилучшую характеристику фильтра ($K = 50$, $N = 1500$). Ис-

пользуя рис. 2, можно провести сравнение кривой измеренных данных с последовательностью данных $\{\tilde{y}^1\}$, полученных при одношаговом предсказании. При многошаговом предсказании были получены менее удовлетворительные результаты.

Чтобы оценить полученные результаты на возможность использования при технических условиях, исходили из следующих соображений:

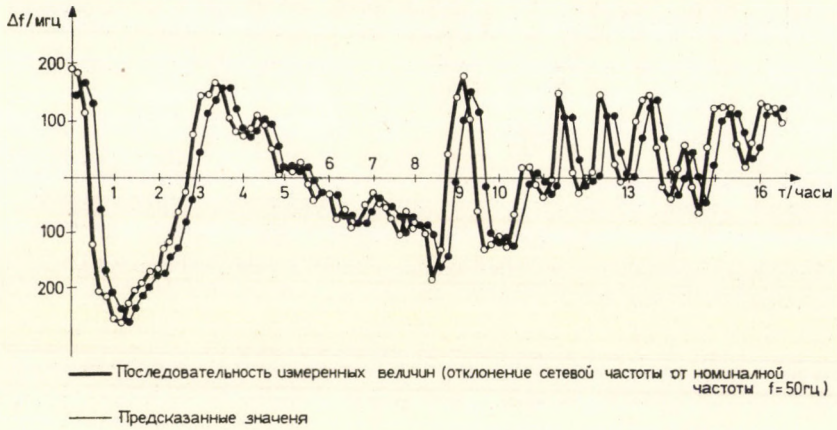


Рис. 2. Одношаговое предсказание для сетевой частоты

В том случае, если было бы возможно использование предсказания сетевой частоты в технике таким образом, что предсказываемое отклонение частоты от номинальной было бы устранено при соответственном воздействии на производство энергии, то было бы получено отклонение частоты на величине ошибки предсказания. Поэтому для оценивания результатов использована средняя ошибка σ_δ, где

$$|\sigma_\delta^2 = \frac{1}{s-\delta} \sum_{i=\delta+1}^{s} (y_i - \tilde{y}_i^\delta)^2, \qquad \delta = 1, 2, 3, 4.$$

Символом $\{\tilde{y}^\delta\}$ обозначается последовательность данных, полученных при δ-шаговом предсказании.

Для общей последовательности данных среднее отклонение измеренной частоты от номинальной (50 гц) составляет $\sigma_{\Delta f} = 162$ мгц.

Полученные результаты иллюстрируются на рис. 3. Например, результат предсказания нагрузки на интервале длины 10 минут (одношагового предсказания) приводит к улучшению среднего баланса активной мощности (отклонение частоты) на 55%.

В дальнейшем, как только соответствующие данные будут нам представлены, предполагается провести испытание составленных программ на последовательностях, описывающих нагрузку системы.



Рис. 3. Результат предсказания по методу фильтрации Калмана

Важным условием для применения методов корреляции невязок является стационарность представляемого процесса. Этому требованию не отвечает процесс, описывающий реальное развитие нагрузки. Но мы считаем, что в практическом случае возможно получить дополнительную информацию, которую можно использовать для преобразования процесса развития нагрузки таким образом, чтобы найденный процесс явился приближенно стационарным.

Литература

1. *Mehra, R. K.*, On the identification of variances and adaptive Kalman-filtering. IEEE Trans. **AC–15**, 1970, 175–184.
2. *Mehra, R. K.*, Approaches to adaptive filtering. IEEE Trans. **AC–17**, 1972, 693.
3. *Carew, B., Bélanger, P. R.*, Identification of optimum filtering steady-state gain for systems with unknown noise covariances. IEEE Trans. **AC–18**, 1973, 582–587.
4. *Toyoda, J., Chen, M., Inoue, Y.*, An application of state estimation to short-time load forecasting Part I: Forecasting modeling. IEEE Trans **PAS–89**, 1970, 1678–1682.

An application of adaptive Kalman filtering to short-term load forecasting for power systems

H. HINKEL

(Berlin)

The problem c short-term load forecasting in power systems for forecasting periods from 10 minutes to 1 hour is considered. Special adaptive Kalman filters are used for prediction. Computer programs were tested by using both simulated and real data for the frequency of the electrical power system.

H. Hinkel
Zentralinstitut für Kybernetik und Informationsprozess der AdW der DDR
Berlin

Problems of Control and Information Theory, Vol. 10 (5), pp. 341–351 (1981)

ON THE APPLICATION OF CONICAL APPROXIMATIONS TO AN OPTIMAL CONTROL PROBLEM FOR SYSTEMS DESCRIBED BY PARTIAL DIFFERENTIAL EQUATIONS OF PARABOLIC TYPE WITH TIME DELAY

A. KOWALEWSKI, W. KOTARSKI

(Kraków, Katowice)

(Received 7 August, 1980)

In this paper an optimal control problem for distributed parameter system (with time delay in the state) with quadratic performance functional and constrained control is considered.

Dynamics of the controlled systems is described by a linear partial differential equation of parabolic type, which solution is an element of certain Sobolev space.

The control time is fixed in our problem.

Making use of the Milutin–Dubovicki theorem [5], necessary and sufficient conditions of optimality are derived for the Dirichlet problem.

The flow chart of the algorithm which can be used in the numerical solving of certain optimization problems for distributed parameter system is also presented.

1. The Milutin–Dubovicki Method [5]

Milutin–Dubovicki's method arises from the geometric form of the Hahn–Banach theorem (a theorem about separating convex sets).

We shall show it on the example.

Let us assume that

E — a linear topological space, locally convex

$J(x)$ — a functional defined on E

A_i, $i = 1, 2, \ldots, n$ — sets in E with inner points which represent inequality constraints

B — a set in E without inner points representing equality constraint.

We must find some conditions necessary for a local minimum of the functional $J(x)$ on the set $Q = \bigcap_{i=1}^{n} A_i \cap B$, or find a point $x_0 \in E$, so that $J(x_0) = \min_{Q \cap U} J(x)$, where U means a certain environment of the point x_0. We define the set $A_0 = \{x : J(x) < J(x_0)\}$.

Then, we formulate the necessary condition of optimality as follows: in the environment of the local minimum point, the intersection of the system

of sets (the set on which the functional attains smaller values than $J(x_0)$, and the sets representing constraints) is empty, or $\bigcap_{i=0}^{n} A_i \cap B = \emptyset$.

The condition $\bigcap_{i=0}^{n} A_i \cap B = \emptyset$ is also equivalent to the one in which there are approximations of the sets A_i $(i = 0, 1, \ldots, n)$ and B instead of A_i or B ones. These approximations are cones with the vertex in point $\{0\}$.

We shall approximate the inequality constraints by the regular admissible cones RAC (A_i, x_0) $(i = 1, 2, \ldots, n)$, the equality constraint by the regular tangent cone RTC (B, x_0) and for the performance functional we shall construct the regular improvement cone RFC (J, x_0).

Then we have the necessary condition of the optimality $J(x)$ on the set $Q = \bigcap_{i=1}^{n} A_i \cap B$ in the form of Euler–Lagrange's equation

$$\sum_{i=0}^{n+1} f_i = 0 \tag{1}$$

where

f_i $(i = 0, 1, \ldots, n+1)$ — are linear, continuous functionals, all of them are not equal to zero at the same time and they belong to the adjoint cones

$$f_i \in [\mathrm{RAC}\,(A_i, x_0)]^*, \quad i = 1, 2, \ldots, n$$

$$f_{n+1} \in [\mathrm{RTC}\,(B, x_0)]^*, \quad f_0 \in [\mathrm{RFC}(J, x_0)]^*$$

$$\{f_i \in [\mathrm{RAC}(A_i, x_0)]^* \Leftrightarrow f_i(x) \geq 0 \ \ \forall\, x \in \mathrm{RAC}\,(A_i, x_0)\}.$$

For convex problems, i.e. problems in which the constraints are convex sets and the performance functional is convex, the Euler–Lagrange equation is the necessary and sufficient condition of optimality, if only certain additional assumptions are fulfilled (so-called Slater's condition).

Using Milutin–Dubovicki's theorem we shall derive the necessary and sufficient conditions of optimality for parabolic systems with time delay in the state by the quadratic performance functional and the constrained control.

2. Parabolic Equations with Time Delay without Control

At the beginning we consider an evalutional equation with time delay without control. Let $\tau_0 > 0$ be a given number representing time delay. We have to find a function $y(x, t)$ which fulfils the following conditions

$$y \in L^2(-\tau_0, T; H_0^1(\Omega)), \quad \frac{\partial y}{\partial t} \in L^2(-\tau_0, T; H^{-1}(\Omega)) \tag{2}$$

$$\text{i.e. } y \in W(-\tau_0, T)$$

and the solution of the equation is

$$\frac{\partial y}{\partial t} + A(t)\, y + y(t - \tau_0) = f_1(t) \qquad x \in \Omega,\ t > 0 \tag{3}$$

$$y(x, t) = g(x, t) \qquad x \in \Omega,\ t \in (-\tau_0, 0) \tag{4}$$

$$y(x, 0) = y_p(x) \qquad x \in \Omega \tag{5}$$

where $f_1 \in L^2(-\tau_0, T; H^{-1}(\Omega))$, $g \in W(-\tau_0, 0)$, $y_p(x) \in L^2(\Omega)$ are given functions, $\Omega \subset R^n$ — a bounded, open set with boundary Γ, which is a C^∞-manifold of dimension $(n - 1)$. Locally, Ω is totally on one side of Γ.

The operator $A(t)$ has the form

$$A(t)\, y = - \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} \left(a_{ij}(x, t) \frac{\partial y(x, t)}{\partial x_j} \right)$$

and the functions $a_{ij}(x, t)$ satisfy the conditions

$$\sum_{i,j=1}^{n} a_{ij}(x, t)\, \xi_i\, \xi_j \geq \alpha \sum_{i=1}^{n} \xi_i^2 \qquad \alpha > 0\ \forall\, (x, t) \in Q\ \ \forall\, \xi_i \in R$$

$$a_{ij}(x, t) \in C^1(Q) \qquad Q = \Omega \times (0, T).$$

Equations (3)÷(5) can be presented in a more convenient form, after defining the operator $M \in \mathcal{L}\big(L^2(-\tau_0, T; H_0^1(\Omega));\ L^2(-\tau_0, T; H_0^1(\Omega))\big)$ by the formula

$$My(t) = \begin{cases} y(t - \tau_0) & t > 0 \\ 0 & t < 0 \end{cases} \tag{6}$$

and the function

$$f(t) = \begin{cases} f_1(t) & t > 0 \\ \dfrac{\partial g}{\partial t} + Ag & t < 0 \end{cases} \tag{7}$$

$$y(x, 0) = g(x, 0). \tag{8}$$

Then equations (3)÷(5) has the form

$$\frac{\partial y}{\partial t} + Ay + My = f \qquad x \in \Omega,\ t \in (-\tau_0, T) \tag{9}$$

$$y(x, 0) = y_p(x) \qquad x \in \Omega. \tag{10}$$

It is easy to notice that in the interval $(0, T)$ equations (9)÷(10) is equivalent to (3). But in the interval $(-\tau_0, 0)$ from (9)÷(10), we get

$$\frac{\partial y}{\partial t} + Ay = \frac{\partial g}{\partial t} + Ag \tag{11}$$

$$y(x, 0) = g(x, 0). \tag{12}$$

On the basis of the theorem about the existence and the uniqueness of the solution of the parabolic equation (Theorem 1.2 [6]) we see that $y \equiv g$ in $(-\tau_0, 0)$.

Existence and uniqueness of the solution of equations (3)÷(5) prove using a constructing method, i.e. solving at first equations (3)÷(5) in the subinterval $(0, \tau_0)$ and in turn in $(\tau_0, 2\tau_0)$ one etc., until to covering the whole interval $(0, T)$; by this the solution in the previous step is the initial condition for the next one.

Really, in every interval the solution of equations (3)÷(5) exists and is unique (Theorem 1.2 [6]).

More general cases, e.g. when τ_0 is a function of time, are considered by Artola in [1] and Baiocchi in [2].

3. Dirichlet's Problem with the Quadratic Performance Functional and the Constrained Control. Distributed Control and Observation

Now, we formulate the control problem for the system described by the following equation

$$\frac{\partial y}{\partial t} + Ay + My = f + Bu \qquad x \in \Omega,\ t \in (-\tau_0, T) \tag{13}$$

$$y(x, 0) = y_p(x) \qquad x \in \Omega, \tag{14}$$

$$y(x, t) = 0 \qquad x \in \Gamma,\ t \in (-\tau_0, T) \tag{15}$$

where $y_p \in L^2(\Omega)$, $f \in L^2(-\tau_0, T; H^{-1}(\Omega))$

$$u \equiv u(x, t), \quad y \equiv y(x, t), \quad f \equiv f(x, t)$$

and the operator $B \in \mathcal{L}(U; H^{-1}(\Omega))$.

Let us denote by $Y = L^2(0, T; H_0^1(\Omega))$ the space of states and by $U = L^2(Q)$ the space of controls.

The control time T is fixed in our problem.

The performance functional is given by

$$I(y, u) = ||Cy - z_d||^2_{L^2(Q)} + \langle Nu, u\rangle_{L^2(Q)} \tag{16}$$

where $C \in \mathfrak{L}(L^2(0, T; H_0^1(\Omega)); L^2(Q))$ — is a linear operator representing the observation in a region Q.

z_d — is a given element in $L^2(Q)$

$N \in \mathfrak{L}(U, U)$ — is a hermitian operator which performed the coercive condition

$$\langle Nu, u\rangle_U \geq \gamma ||u||^2_U \quad \forall u \in U, \quad \gamma > 0.$$

We assume the following constraint on controls $u \in U_{ad}$ is a closed, convex set with non-empty interior, a subset of U. (17)

The solving of the stated optimal control problem is equivalent to a seeking a couple $(y_0, u_0) \in E = Y \times U$ which satisfies equations (13)÷(15) and minimizing performance functional (16) with the constraint on controls (17).

We formulate the necessary and sufficient conditions of the optimality in the form of Theorem 1.

Theorem 1. The solution of the optimization problem (13)÷(17) exists and is unique by the assumptions mentioned above and the necessary and sufficient conditions of the optimality are characterized by the following system of partial differential equations and inequalities:

equation of the system control

$$\frac{\partial y_0}{\partial t} + Ay_0 + y_0(t - \tau_0) = f + Bu_0 \qquad x \in \Omega,\ t \in (0, T) \tag{18}$$

$$\frac{\partial y_0}{\partial t} + Ay_0 = f \qquad x \in \Omega,\ t \in (-\tau_0, 0) \tag{19}$$

$$y_0(x, 0) = y_p(x) \qquad x \in \Omega \tag{20}$$

$$y_0(x, t) = 0 \qquad x \in \Gamma,\ t \in (-\tau_0, T) \tag{21}$$

adjoint equation

$$-\frac{\partial p}{\partial t} + A^* p + p(t + \tau_0) = C^*(Cy_0 - z_d) \qquad x \in \Omega,\ t \in (-\tau_0, T - \tau_0) \tag{22}$$

$$-\frac{\partial p}{\partial t} + A^* p = C^*(Cy_0 - z_d) \qquad x \in \Omega,\ t \in (T - \tau_0, T) \tag{23}$$

$$p(x, T) = 0 \qquad x \in \Omega \tag{24}$$

$$p(x, t) = 0 \qquad x \in \Gamma,\ t \in (-\tau_0, T) \tag{25}$$

maximum condition

$$\langle B^* p + Nu_0,\ u - u_0 \rangle_{L^2(Q)} \geq 0 \qquad \forall\ u \in U_{ad}. \tag{26}$$

Proof. According to the Milutin–Dubovicki theorem, we approximate the set representing the inequality constraint by the regular admissible cone, the equality constraint by the regular tangent cone and the performance functional by the regular improvement cone.

a) *Analysis of the Equality Constraint*

The set Q_1 representing the equality constraint has the form

$$Q_1 = \left\{ \begin{array}{ll} \dfrac{\partial y}{\partial t} + Ay + My = f + Bu & x \in \Omega,\ t \in (-\tau_0, T) \\ (y, u) \in E;\ y(x, 0) = y_p(x) & x \in \Omega \\ y(x, t) = 0 & x \in \Gamma,\ t \in (-\tau_0, T) \end{array} \right\}. \tag{27}$$

We construct the regular tangent cone of the set Q_1 using the Lusternik theorem (Theorem 9.1 [5]).

For this purpose we define the operator $P(y, u)$ in the form

$$P(y, u) = \left(\frac{\partial y}{\partial t} + Ay + My - f - Bu, y(x, t), y(x, 0) - y_p(x) \right). \tag{28}$$

The operator $P(y, u)$ is the mapping from the space $L^2(-\tau_0, T;\ H_0^1(\Omega)) \times$ $\times L^2(Q)$ into the space $L^2(-\tau_0, T;\ H^{-1}(\Omega)) \times L^2(-\tau_0, T;\ H_0^1(\Omega)) \times L^2(\Omega)$.

The Fréchet differential of the operator $P(y, u)$ can be written in the following form

$$P'(y_0, u_0)(\bar{y}, \bar{u}) = \left(\frac{\partial \bar{y}}{\partial t} + A\bar{y} + M\bar{y} - B\bar{u}, \bar{y}(x, t), \bar{y}(x, 0) \right). \tag{29}$$

Really, $\dfrac{\partial}{\partial t}$ (Theorem 2.8 [7]), $A(t)$ (Theorem 2.1 [6]) and M are linear and bounded mappings.

Using Theorem 1.2 [6] we can prove that $P'(y_0, u_0)$ is the operator "one to one" from the space $L^2(-\tau_0, T;\ H_0^1(\Omega)) \times L^2(Q)$ onto $L^2(-\tau_0, T;\ H^{-1}(\Omega)) \times$ $\times L^2(-\tau_0, T;\ H_0^1(\Omega)) \times L^2(\Omega)$.

Considering that there are fulfilled the assumptions of Lusternik's theorem we can write down the regular tangent cone for the set Q_1 in a point (y_0, u_0) in the form

$$\mathrm{RTC}\,(Q_1, (y_0, u_0)) = \{(\bar{y}, \bar{u}) \in E;\ P'(y_0, u_0)(\bar{y}, \bar{u}) = 0\}. \tag{30}$$

It is easy to notice that it is a subspace. Therefore using Theorem 10.1 [5] we know the form of the functional belonging to the adjoint cone

$$f_1(\bar{y}, \bar{u}) = 0 \quad \forall\, (\bar{y}, \bar{u}) \in \mathrm{RTC}\,(Q_1, (y_0, u_0))\,.$$

b) *Analysis of the Constraint on Controls*

The set $Q_2 = Y \times U_{ad}$ represented the inequality constraint is a closed and a convex one with non-empty interior in the set E.

Using Theorem 10.5 [5] we find the functional belonging to the adjoint cone $[\mathrm{RAC}\,(Q_2, (y_0, u_0))]^*$.

We can note if E_1, E_2 are two linear topological spaces, then the adjont space to $E = E_1 \times E_2$ has the form

$$E^* = \{f = (f_1, f_2);\ f_1 \in E_1^*, f_2 \in E_2^*\}$$

and

$$f(x) = f_1(x_1) + f_2(x_2).$$

So we note the functional $f_2(\bar{y}, \bar{u})$ as follows:

$$f_2(\bar{y}, \bar{u}) = (0, f_2')$$

where $f_1'(\bar{y}) = 0 \quad \forall\, \bar{y} \in Y$ (Theorem 10.1 [5])

$f_2'(\bar{u})$ is the support functional to the set U_{ad} in a point u_0 (Theorem 10.5 [5]).

c) *Analysis of the Performance Functional*

On the basis of Theorem 7.5 [5] we construct the regular improvement cone $\mathrm{RFC}\,(I, (y_0, u_0))$ for the performance functional.

The performance functional we shall write down in the following way

$$I(y, u) = \langle Cy - z_d, Cy - z_d\rangle_{L^2(Q)} + \langle Nu, u\rangle_{L^2(Q)}.$$

The Fréchet differential $I'(y_0, u_0)\,(\bar{y}, \bar{u})$ has the form

$$I'(y_0, u_0)\,(\bar{y}, \bar{u}) = 2\,\langle Cy_0 - z_d, C\bar{y}\rangle_{L^2(Q)} + 2\,\langle Nu_0, \bar{u}\rangle_{L^2(Q)}.$$

By virtue of Theorem 7.5 [5], we get

$$\mathrm{RFC}\,(I, (y_0, u_0)) = \{(\bar{y}, \bar{u}) \in E;\ I'\,(y_0, u_0)\,(\bar{y}, \bar{u}) < 0\} \tag{31}$$

and $f_3(\bar{y}, \bar{u}) \in [\mathrm{RFC}\,(I, (y_0, u_0))]^*$ is

$$f_3(\bar{y}, \bar{u}) = -\lambda_0\langle C^*(Cy_0 - z_d), \bar{y}\rangle_{L^2(Q)} - \lambda_0\langle Nu_0, \bar{u}\rangle_{L^2(Q)} \tag{32}$$

where $\lambda_0 \geq 0$.

It is easy to verify that $\lambda_0 \neq 0$. Because of $\lambda_0 = 0$ all the functionals f_1, f_2, f_3 in Euler's equation are equal zero. But according to the Milutin–Dubovicki theorem it is impossible. Then, $\lambda_0 > 0$.

1) *Analysis of the Euler–Lagrange's equation*

The Euler–Lagrange equation for our optimization problem has the form

$$f_2'(\bar{u}) = \lambda_0 \int_Q C^*(Cy_0 - z_d)\,\bar{y}\,dx\,dt + \lambda_0 \int_Q Nu_0\,\bar{u}\,dx\,dt \tag{33}$$

$$\forall\,(\bar{y}, \bar{u}) \in \mathrm{RTC}\,(Q_1, (y_0, u_0)).$$

Introducing the adjoint variable by the equation

$$-\frac{\partial p}{\partial t} + A^* p + M^* p = C^*(Cy_0 - z_d) \qquad x \in \Omega,\ t \in (-\tau_0, T) \tag{34}$$

$$p(x, T) = 0 \qquad x \in \Omega \tag{35}$$

$$p(x, t) = 0 \qquad x \in \Gamma,\ t \in (-\tau_0, T) \tag{36}$$

where

$$M^* p(t) = \begin{cases} p(t + \tau_0) & t \in (-\tau_0, T - \tau_0) \\ 0 & t \in (T - \tau_0, T) \end{cases}$$

and taking into account that $\bar{y}(x, 0) = 0$ (because it belongs to RTC $(Q_1, (y_0, u_0))$) in turn, we get

$$\int_{-\tau_0}^{T} \left\langle -\frac{\partial p}{\partial t} + A^* p + M^* p, \bar{y} \right\rangle dt = \int_{-\tau_0}^{T} \left\langle -\frac{\partial p}{\partial t}, \bar{y} \right\rangle dt + \int_{-\tau_0}^{T} \langle A^* p, \bar{y} \rangle\, dt +$$

$$+ \int_{-\tau_0}^{T} \langle M^* p, \bar{y} \rangle\, dt = \langle -p, \bar{y} \rangle \Big|_{-\tau_0}^{T} + \int_{-\tau_0}^{T} \left\langle p, \frac{\partial \bar{y}}{\partial t} \right\rangle dt + \int_{-\tau_0}^{T} \langle p, A\bar{y} \rangle\, dt +$$

$$+ \int_{-\tau_0}^{T} \langle p, My \rangle\, dt = \langle p(-\tau_0), \bar{y}(-\tau_0) \rangle - \langle p(T), \bar{y}(T) \rangle +$$

$$+ \int_{-\tau_0}^{T} \left\langle p, \frac{\partial \bar{y}}{\partial t} + A\bar{y} + M\bar{y} \right\rangle dt = \int_{-\tau_0}^{T} \langle p, B\bar{u} \rangle\, dt = \langle B^* p, \bar{u} \rangle_{L^2(Q)}. \tag{37}$$

Substituting (37) for the first component of the right-hand side of (33), we have

$$f_2'(\bar{u}) = \lambda_0 \langle B^* p + N u_0, u \rangle_{L^2(Q)}. \tag{38}$$

Using the definition of the support functional [5] and dividing both sides by λ_0, we finally get the maximum condition

$$\langle B^* p + N u_0, u - u_0 \rangle_{L^2(Q)} \geq 0 \quad \forall u \in U_{ad}. \tag{39}$$

Formulas (18)÷(21) and (22)÷(25) are obtained after substituting the operators M and M^* to equations (13)÷(15) and (34)÷(36).

In order to prove the sufficiency of the derived conditions of the optimality we use the fact that constraints and the performance functional are convex and Slater's condition satisfies Theorem 15.2 in [5]. Really, there exists a point $(\tilde{y}, \tilde{u}) \in \operatorname{int} Q_2$ such that $(\tilde{y}, \tilde{u}) \in Q_1$. This fact follows immediately from the existence of non-empty interior in the set Q_2 and from the existence of the solution of equations (13)÷(15), as well.

The uniqueness of the optimal control follows from the strict convexity of the performance functional (16).

This last remark finishes the proof of Theorem 1.

4. Final Remarks

We must remark that the optimal conditions derived above allow to obtain an analytical formula for the optimal control in particular cases only (e.g. there no constraints on controls).

It results from the following; the determining of the function $p(x, t)$ in the maximum condition is possible from the adjoint equation, if and only if we known $y_0(x, t)$ will suit the control $u_0(x, t)$.

These mutual connections make the practical usage of the derived optimization formulas difficult. So that we resign from the exact determining of the optimal control and we use approximation methods. In the case of non-coercive performance functional (in formula (16), $N = 0$) the optimal control problem (13)÷(17) will reduce to the minimizing of a functional on a closed and convex subset in a Hilbert space. Then the optimal control problem is equivalent to a quadratic programming one which can be solved by the use of the well-known algorithms, e.g. Gilbert's [4], Barr's [3] or Nahi–Wheeler's [8] ones. In this case we can show the course of calculations on the flow chart (Fig. 1).

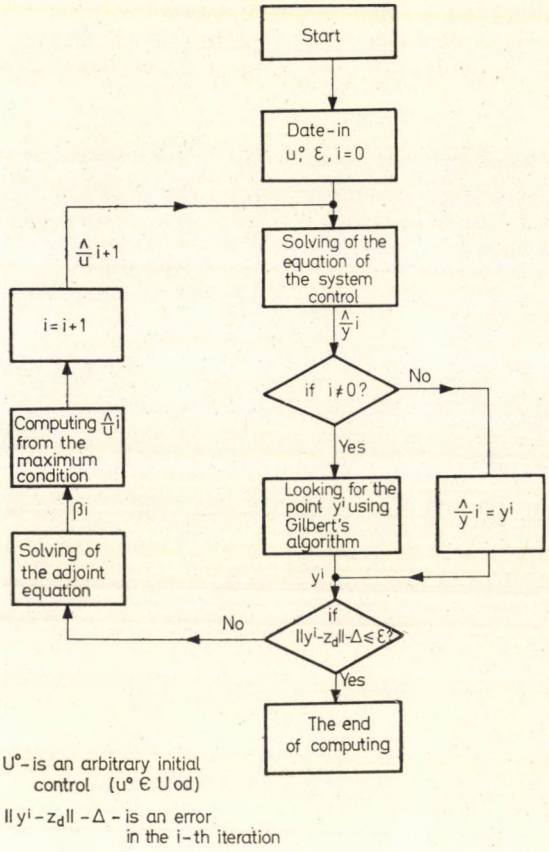


Fig. 1

References

1. *Artola, M.*, Equations paraboliques à retardement. C. R. Acad. Sci. Paris **264,** 668–671 (1967).
2. *Baiocchi, C.*, Sulle equazioni differenziale astratte lineari del primo e del secondo ordine negli spazi di Hilbert. Ann. Mat. Pura ed. Appl. **76,** 233–304 (1967).
3. *Barr, R. O.*, On efficient computational procedure for a generalized quadratic programming problem. SIAM Journal on Control, vol. **7,** 415–429 (1969).
4. *Gilbert, E. S.*, An iterative procedure for computing the minimum of a quadratic form on a convex set. SIAM Journal on Control, vol. **4,** 61–80 (1966).
5. *Girsanov, I. V.*, Lectures on the Mathematical Theory of Extremal Problems. Publication University of Moscow, 1970 (in Russian).
6. *Lions, J. L.*, Optimal Control of Systems Governed by Partial Differential Equations. Springer-Verlag, Berlin, Heidelberg, New York, 1971.
7. *Maslov, V. P.*, Operators Methods. Publication "Science", Moscow, 1973 (in Russian).
8. *Nahi, N. E., Wheeler, L. A.*, Optimal terminal control of continuous systems via successive approximation of the reachable set. IEEE Trans. Aut. Control, **AC-12,** 515–521 (1967).

Применение конических аппроксимаций в задачах оптимального управления системами, описываемыми уравнениями с частными производными параболического типа с запаздыванием

А. КОВАЛЕВСКИ (Краков) В. КОТАРСКИ (Катовице)

В работе рассматривается задача оптимального управления системами с распределенными параметрами с запаздыванием при квадратном функционале качества и ограничении управления.

Поведение системы описывается линейным дифференциальным уравнением в частных производных параболического типа, решение которого принадлежит некоторому пространству Соболева.

В нашей задаче время управления фиксировано.

Пользуясь теоремой Милютина–Дубовицкого [5] для проблемы Дирихлета, получены необходимые и достаточные условия оптимума.

Полученные результаты сходятся с результатами Лионса.

Представлена также блок-схема алгоритма, который можно использовать при численном решении некоторых оптимальных задач для распределенных систем.

A. Kowalewski
Institute of Computer Science and Control Engineering,
Stanislaw Staszic University of Mining and Metallurgy,
Al. Mickiewicza 30, 30–059 Kraków, Poland

W. Kotarski
Institute of Mathematics, Silesian University,
ul. Bankowa 14, 40–077 Katowice, Poland

Problems of Control and Information Theory, Vol. 10 (5), pp. 353–361 (1981)

ON A NEW HIERARCHICAL APPROACH TO THE LARGE-SCALE SYSTEM SOLUTION

L. BAKULE

(Prague)

(Received 1 September, 1980)

The paper deals with an algorithm derivation for effective structural large-scale systems design. The sparsity of matrix is employed to decompose a system model. New hierarchical algorithm for matrix inversion is developed. The application of results for the algebraic systems equations is presented. The numerical examples for the developed procedure are supplied to demonstrate its effectiveness.

Introduction

The aim of this contribution is further development and presentation of new results reached in the area of finite methods for matrix inversion (resp. algebraic system equations solution) using the specific properties of such matrices. It is supposed that a model of large-scale system is described by a positive definitive matrix (resp. nonsingular structural symmetric matrix).

The paper continues the results published in [1, 2, 3, 9, 13, 16]. The basic approach in large-scale system treatment consists in system tearing or structural system decomposition. It generally means that the sequence of three steps is applied:

1) The decomposition of a given system into isolated or free subsystems.
2) The solution of isolated subsystems.
3) The solution of the global system using the reached solutions of isolated subsystems and including the system interconnections.

The advantages of this approach are in the simplification of large-scale complex system solution and further in the possibility of obtaining new information on the system, which could not be reached by single-step design.

This design philosophy has been formulated by G. Kron for the structural physical systems design [1, 2, 6, 8, 10, 11, 16]. However, its importance and general validity in the large-scale system methodology has been appraised sufficiently till recently [8, 10–16].

The contribution continues the idea of "permutation" method of decomposition and the evaluation of the corresponding computational effort during computation of matrix inversion. New hierarchical procedure, solution of sparse algebraic system equations and numerical examples are presented in this contribution to illustrate the effectiveness of proposed procedure.

Problem formulation

Let us deal with the real numbers range.

It is supposed that a system model has the form

$$MX = Y, \tag{1}$$

where M is a sparse positive definitive matrix of order n. X, Y are n-component column vectors, at least one $Y_i \neq 0$, Y is a known vector.

The generalization to a square nonsingular structural symmetric sparse matrix M is possible. It is supposed in such a case that all necessary matrix inversions exist.

A sparse matrix is a matrix having only a small percentage of nonzero elements [16].

A structural symmetric matrix is a square matrix having a symmetric zero-nonzero structure and having nonzero diagonal elements.

Definition. A square matrix M has a double bordered block diagonal form (further on abbreviated as DBBDF) if it is structurally symmetric and has the form

$$M = \begin{Vmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{Vmatrix}, \tag{2}$$

where M_{11} is a quasidiagonal matrix. The orders of M_{11} (resp. M_{22}) are further denoted $n_{11}, n_{12}, \ldots, n_{1k}$ (resp. n_2).

The aim is:

1) The presentation of matrix inversion method for the matrix M in hierarchical approach using DBBDF of the matrix.
2) The derivation of operation criterion corresponding to the hierarchical DBBDF approach.
3) To apply the obtained results to the algebraic system equations solution.
4) To supply the presented results by examples.

Problem solution

Let us deal with the matrix inversion problem solved by DBBDF approach. Let us present the selected preliminaries.

The matrix inversion method using 2×2 partitioned matrices results in the following formulae [4, 5, 7, 9]:

$$M^{-1} = N = \begin{Vmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{Vmatrix}, \tag{3}$$

where

$$\begin{aligned} N_{22} &= (M_{22} - M_{21} M_{11}^{-1} M_{12})^{-1}, \\ N_{12} &= -M_{11}^{-1} M_{12} N_{22}, \\ N_{21} &= -N_{22} M_{21} M_{11}, \\ N_{11} &= M_{11}^{-1} - M_{11}^{-1} M_{12} N_{21}. \end{aligned}$$

The effectiveness of computation by finite methods is evaluated by the amount of arithmetical operations. The most important are multiplications and divisons. Concentrating on the performing of these operations we obtain for DBBDF (2) using (3)

$$J = n_2 \sum_i n_{1i}^2 + 2 n_1 n_2 + \sum_i f(n_{1i}) + f(n_2) + n_1^2 n_2, \tag{4}$$

where $f(n) = \dfrac{n^3}{2} + \dfrac{n^2}{2}$, $n_1 = \sum\limits_i n_{1i}$ [1]. The derivation in (4) is based on the operation evaluation in (3) for the iterated bordered method used for the necessary matrix inversions [1, 4, 5]. The number of operations using the iterated bordered method for a $n\times n$ matrix is equal [5]

$$J = \frac{n^3}{2} + \frac{n^2}{2}. \tag{5}$$

This method is the most effective one in the comparison with the other finite methods. These results are known and we shall not deal with them here in detail [1, 5].

A given matrix M and a DBBDF of this matrix M are similar matrices [1, 9, 13, 14, 16].

Definition. A feasible permutation P corresponding to a given $n\times n$ matrix M means that for

$$\hat{M} = P^{-1} M P, \tag{6}$$

where $\hat{M}$ is DBBDF of matrix M, $\bar{n}_2$ is such that

$$1 \leq n_2 \leq \bar{n}_2. \tag{7}$$

$\bar{n}_2$ is an upper bound such that for a $n \times n$ DBBDF matrix $\overline{M}$ (resp. $\overline{\overline{M}}$) with $n - \bar{n}_2$ (resp. $n - \bar{n}_2 - 1$) subsystems and $n_2 = \bar{n}_2$ (resp. $n_2 = \bar{n}_2 + 1$) for the corresponding $\underline{J} = \overline{J}$ (resp. $\underline{J} = \overline{\overline{J}}$) is

$$\overline{J} < J \tag{8}$$

and

$$\overline{\overline{J}} > J. \tag{9}$$

The feasible permutation corresponding to the minimal value of criterion (4) $\underline{J}^0$ forms an optimal decomposition. But only suboptimal algorithms are practically applicable during this optimisation for large n [1, 12, 13, 16].

The effectiveness of evaluation is given by the relation between $\underline{J}$ and J. If

$$\underline{J} < J, \tag{10}$$

then the decomposition is effective. $\underline{J}$ is meant as a value reached by application of a suboptimal algorithm [1].

Definition. A matrix M has a S property if it is permutable to DBBDF and there is $\underline{J}$ such that

$$\underline{J} < J. \tag{11}$$

Let us continue the consideration in the direction of construction of hierarchy using DBBDF approach. Let us concentrate on a 2-level solution.

Definition. The 2-level double block bordered form (2-level DBBDF) of matrix M

$$M = \begin{Vmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{Vmatrix} \tag{12}$$

means that M has DBBDF and that at least one M_i has DBBDF, where

$$M_{11} = \text{diag}\,(M_1, \ldots, M_k). \tag{13}$$

Full 2-level DBBDF is 2-level DBBDF for all M_i in DBBDF.

2-level partitioned method means that procedure (3) is applied but instead of "full" matrices M_i, computation procedure (3) is applied on such M_i, which are in DBBDF.

Let us formluate a theorem determining operation criterion for 2-level DBBDF using the 2-level partitioned method.

Theorem. Let us suppose that a positive definitive matrix M of order n has full 2-level DBBDF. Then the operation criterion J (2) using the 2-level partitioned method has the form

$$J(2) = n_2 \sum_i n_{1i}^2 + 2n_1 n_2 + n_1^2 n_2 + f(n_2) + \\ + \sum_i \left[m_{i0} \sum_{j=1}^{k_i} m_{ij}^2 + 2m_i m_{i0} + \sum_j f(m_{ij}) + f(m_{i0}) \right], \tag{14}$$

where n_1 (resp. n_2, resp. m_{ij}; resp. m_{i0}) is the order of matrix M_{11} (resp. M_{22}; resp. M_j^i; resp. M_{22}^i);

$$m_i = \sum_j m_{ij}, \; M_i = \begin{Vmatrix} M_{11}^i & M_{12}^i \\ M_{21}^i & M_{22}^i \end{Vmatrix},$$

$$M_{11}^i = \operatorname{diag}(M_1^i, \ldots, M_k^i). \tag{15}$$

Proof. Using the iterated bordered method the number of multiplications (resp. divisions) by $n \times n$ matrix inversion is $f_1(n)$ (resp. $f_2(n)$) [1, 5]

$$f_1(n) = \frac{n^3}{2} + \frac{n^2}{2} - n, \; f_2(n) = n. \tag{16}$$

We obtain for a nonhierarchical structure the number of multiplications $\underline{J}_m$ (resp. divisions $\underline{J}_d$)

$$\underline{J}_m = n_2 \sum_i n_{1i}^2 + 2n_1 n_2 + \sum_i f_1(n_{1i}) + f_1(n_2) + n_1^2 n_2, \tag{17}$$

$$\underline{J}_d = \sum_i f_2(n_{1i}) + f_2(n_2),$$

then

$$\underline{J} = \underline{J}_m + \underline{J}_d \tag{18}$$

as derived in detail in [1].

Applying now the 2×2 partitioned method DBBDF matrices M_i, $\forall i$ and substituting in (18) we obtain (14).

Comment. The theorem includes 2-level approach only. It is also the most important reduction reached during computation.

Definition. A matrix M has a S (2) property if there is corresopnding 2-level DBBDF for which holds

$$\underline{J}(2) < \underline{J}. \tag{19}$$

Comment. The relation between S (2) property and S property [1] cannot be generally formulated to compare these properties because of the operations reduction possibility in the left upper corner of the matrix when we use 2-level procedure. Of course, if a matrix has S property and further hierarchical reduction is possible then it has also S (2) property. Because of the same feature of 2-level DBBDF a feasible permutation for 2-level procedure can be defined more generally so that nonhierarchical approach definitin can be understood as its special case.

Definition. A feasible permutation P (2) for a matrix M using 2-level procedure according to the Theorem means that for

$$M\,(2) = P\,(2)^{-1} MP\,(2) \tag{20}$$

the following inequality for the corresponding $J(2)$ is valid

$$J\,(2) < J. \tag{21}$$

Comment. A feasible permutation for 2-level procedure is each permutation ensuring S (2) property for a given matrix M.

Let us now deal with an algebraic large sparse system equations solution using DBBDF approach now. We suppose that Gauss elimination is the most advantageous method from the point of view of computational effort [5]. We obtain for this method the number of multiplications and divisions Q equal to

$$Q = \frac{n^3}{6} + \frac{3}{2} n^2 + \frac{n}{3}. \tag{22}$$

Lemma. Let a matrix M has DBBDF (resp. 2-level DBBDF). Then using the 2×2 partitioned method (resp. 2-level partitioned method), the operation criterion for algebraic system equations, $\underline{Q}$ (resp. Q (2), is

$$\underline{Q} = \underline{J} + n^2$$

$$Q\,(2) = J\,(2) + n^2. \tag{23}$$

Proof. Simply follows from the addition of matrix inversion computation effort and the additional term n^2 given by a $n \times n$ matrix multiplied by an n-component column vector.

Definition. A matrix M has a $\overline{S}$ property (resp. a $\overline{S(2)}$ property) if there is such permutation that for the corresponding criterion $\underline{Q}$ (resp. $Q(2)$ it is valid

$$\underline{Q} < Q,$$
$$Q(2) < Q. \tag{24}$$

Comment. Feasible permutations for algebraic systems equations are defined quite analogously as by the matrix inversion.

Examples

Formulation. Let us deal with the positive definite matrices M having the following structure:

1) $n = 430$, $m_{11} = m_{12} = m_{21} = m_{22} = 10^2$, $m_{10} = m_{20} = 10$;
$n_2 = 10$, $k = 2$.

2) $n = 412$, $m_{11} = m_{12} = m_{21} = m_{22} = 10^2$, $m_{10} = m_{20} = 1$;
$n_2 = 10$, $k = 2$.

3) $n = 403$, $m_{11} = m_{12} = m_{21} = m_{22} = 10^2$, $m_{10} = m_{20} = 1$,
$n_2 = 1$, $k = 2$.

4) $n = 35$, $m_{11} = m_{12} = 10$, $n_{12} = 10$, $n_2 = 3$, $m_{10} = 2$.

Let us demonstrate the presented results on these examples.

Solution

We compute all criteria to evaluate the decomposition effectiveness.

1) $J = 39845950$, $Q = 13528660$,
$\underline{J} = 9931450$, $\underline{Q} = 10116350$,
$J(2) = 2488900$, $Q(2) = 2673800$,
$J : \underline{J} : J(2) = 1 : 0.24925 : 0.06246$,
$Q : \underline{Q} : Q(2) = 1 : 0.74778 : 0.19764$.

2) $J = 36600267$, $Q = 11910508$,
$\underline{J} = 8415606$, $\underline{Q} = 8585350$,
$J(2) = 1791644$, $Q(2) = 1931388$,
$J : \underline{J} : J(2) = 1 : 0.22993 : 0.048951$,
$Q : \underline{Q} : Q(2) = 1 : 0.72082 : 0.162158$.

3) $J = 32806618$, $Q = 11152219$,

$\underline{J} = 8400595$, $\underline{Q} = 8563004$,

$J(2) = 282295$, $Q(2) = 444704$,

$J : \underline{J} : J(2) = 1 : 0.2560 : 0.00860$,

$Q : \underline{Q} : Q(2) = 1 : 0.7678 : 0.03988$.

4) $J = 22050$, $Q = 8995$,

$\underline{J} = 8462$, $\underline{Q} = 9687$,

$J(2) = 3700$, $\underline{Q}(2) = 4925$,

$J : \underline{J} : J(2) = 1 : 0.383 : 0.1678$,

$Q : \underline{Q} : Q(2) = 1 : 1.076 : 0.5475$.

All required properties: S property, $S(2)$ property, $\bar{S}$ property and $\overline{S(2)}$ property are tested. The effectiveness of the proposed procedure is obvious.

Conclusion

A new hierarchical approach for large sparse matrix inversion and large sparse algebraic system equations is presented. The effectiveness of this approach is demonstrated on examples.

References

1. *Bakule, L.*, On a new large-scale system decomposition method. Acta technica ČSAV No. *6*, 1976, pp. 707–725.
2. *Bakule, L.*, On a new decomposition solution of the sparse systems. Acta technica ČSAV, No. *4*, 1974, pp. 688–711.
3. *Brayton, R. K., Gustavson, F. G., Willoughby, R. A.*, Some Results on Sparse Matrices. Mathematics of Computation, vol. **24**, No. *112*, October 1970, pp. 937–954.
4. *Faddějev, D. K., Fadějevová, V. N.*, Numerické metody lineární algebry. SNTL, Praha, 1964.
5. *Fox, L.*, An Introduction to Numerical linear Algebra. Clerandon Press, Oxford, 1964.
6. *Frank, H.* (ed.), A Publication of the IEE Circuits and System Society. Special Issue on Large-scale Networks. IEEE Transactions on Circuit theory, May 1973.
7. *Gantmacher, F. R.*, Teorija matric. Izd. Nauka, Moskva, 1966.
8. *Ghausi, M.* (ed.), Special Issue on Large-Scale Networks and Systems. IEEE Transactions on Circuits and Systems, December 1976.
9. *Harary, F.*, A graph theoretic approach to matrix inversion by partitioning. Numerische Mathematik, Band, 1962, pp. 128–135.
10. *Himmelblau, D. M.* (ed.), Decomposition of Large-scale Problems. North Holland–Elsevier, 1973.
11. *Kron, G.*, Diakoptics. MacDonald, London, 1963.
12. *Read, R. C.* (ed.), Graph Theory and Computing. Academic Press, 1972.
13. *Reid, J. K.* (ed.), Large sparse sets of linear equations. Academic Press, 1971.
14. *Rose, D. J., Willoughby, R. A.* (ed.), Sparse matrices and their applications. Plenum Press, New York—London, 1972.
15. *Šiljak, D. D.*, Large-scale Dynamic Systems. North Holland, 1978.
16. *Tewarson, R. D.*, Sparse matrices. Academic Press, 1973.

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

О новом иерархическом подходе к решению больших систем

[Л. БАКУЛЕ

(Прага)

В статье дана формулировка иерархического структурного подхода к эффективному решению больших систем.

Большая система описывается положительно определенной редкой матрицей большого ранга. Структурная декомпозиция модели большой системы основана на использовании нулевых элементов матрицы к преобразованию подобия данной матрицы на двойственно окаймленную квазидиагональную форму. Эта форма используется с применением метода обращения матрицы, разделенной на блоки 2×2, к редукции числа арифметических операций. Двухуровневый метод основан на преобразовании подобия данной матрицы и современно диагональных субблоков левого верхнего блока на двойственно окаймленную квазидиагональную форму. Метод обращения матрицы, разделенной на блоки 2×2, применяется тоже к решению субматричных блоков левого верхнего блока данной формы матрицы. Для двухуровневого метода формулируются критерии эффективности. Предпочтительные качества иерархического подхода иллюстрируются на нескольких нетривиальных примерах.

L. Bakule
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Prague

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

ENGLISH TRANSLATION OF THE PAPERS IN RUSSIAN

Problems of Control and Information Theory, Vol. **10**, *No. 5.*

ON LOWER BOUND OF THE LENGTH OF B_4-CODE

A. G. D'YACHKOV, V. V. RYKOV

Consideration of coding of conflicts in some communication systems with random multiple access [2] has led to the notion of B_S-code [1]. Lower and upper bounds for the optimal length of B_S-code was obtained in [1]. The present publication improves the lower bound of [1] for the special case of B_4-code.

Our task is to establish the lower bound of optimal length of codes for address part of messages in communication systems with random multiple access. Detailed analysis of applicability of these codes (in particular, for satellite communication systems) may be found in [1].

Denote by $x = ||x_i(j)||$, $i = \overline{1, N}$, $j = \overline{1, t}$ a binary $N \times t$ matrix consisting of 0's and 1's. Let $x(j)$ and x_i be, respectively, the j-th column and the i-th row of the matrix.

Definition. [1] Matrix x is referred to as B_S-code if all the C^S_{t+s-1} sums of its columns of the form $x(j_1) + x(j_2) + \ldots + x(j_s)$ where $1 \leq j_1 \leq j_2 \leq \leq \ldots \leq j_s \leq t$ are different from each other. Let $N(s, t)$ be the minimal possible number of rows (length) of B_S-code.

Denote by $q = (q_0, q_1, q_2)$ an arbitrary vector with components $0 \leq q_i \leq 1$ for which $q_0 + 2q_1 + 2q_2 = 1$. Let

$$\underline{H}(q) \stackrel{\triangle}{=} -q_0 \log q_0 - 2q_1 \log q_1 - 2q_2 \log q_2 \qquad (1)$$

be Shannon entropy of an appropriate probability distribution, here and below logarithms with respect to base 2 are used, and $\stackrel{\triangle}{=}$ denotes equality by definition. Define a vector $q^* = (q_0^*, q_1^*, q_2^*)$ where

$$q_0^* = 3 \cdot 6^{1/4}/2(3 + 2 \cdot 6^{1/4}) = 0.38296,$$

$$q_1^* = 1/2 - 2q_0^*/3 = 0.24469,$$

$$q_2^* = q_0^*/6 = 0.06383;$$

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

Theorem. Under $t \to \infty$ the following asymptotic inequality

$$N(4, t) \geq [(\log 3)^{-1} + \underline{H}(q^*)^{-1}] 2 \log t (1 + o(1)) \tag{2}$$

holds. As it was demonstrated in [1], for $t \to \infty$

$$N(4, t) \geq \left[\frac{2}{3} + \left(\frac{2}{3} + \log 3\right)^{-1}\right] 2 \log t (1 + o(1)). \tag{3}$$

For more convenient comparison of boundaries (2) and (3), let us present the results of computations:

$$2\,[(\log 3)^{-1} + H(q^*)^{-1}] = 2.24661,$$

$$2\left[\frac{2}{3} + \left(\frac{2}{3} + \log 3\right)^{-1}\right] = 2.22158.$$

Proof. Consider an arbitrary B_4-code x of size $N \times t$. Fix the number $n = \overline{0, N}$. Denote by $v = (v_1, v_2, \ldots, v_n)$, $w = (w_1, w_2, \ldots, w_{N-n})$ columns with integer components $v_i = 0, \pm 1, \pm 2$, $w_i = 0, \pm 1, \pm 2$ of length n and $N - n$, respectively. Like in [1], introduce $\mu_1(v, w)$ which denotes the number of columns x of the form (v, w). Let

$$\mu_2(v, w) \stackrel{\triangle}{=} \sum_{v'} \sum_{w'} \mu_1(v' w') \mu_1(v - v', w - w').$$

Assume that

$$\mu_s(v, \cdot) \stackrel{\triangle}{=} \sum_{w} \mu_s(v, w), \qquad s = 1, 2.$$

Define then according to [1]

$$\begin{aligned}
&\ln(s) \stackrel{\triangle}{=} \sum_{v} \mu_s^2(v, \cdot), \\
&m_s(v, w) \stackrel{\triangle}{=} \sum_{w'} \mu_s(v, w') \mu_s(v, w' + w), \qquad (4) \\
&m_s(\cdot, w) \stackrel{\triangle}{=} \sum_{v} m_s(v, w), \\
&Z_s = \{w : m_s(\cdot, w) > 0\}, \quad s = \overline{1, 2}. \\
&m_2(v, Z_1) = \sum_{w \in z_1} m_2(v, w) \leq \mu_2^2(v, \cdot), \qquad (5) \\
&m_2(\cdot, Z_1) = \sum_{v} m_2(v, z_1).
\end{aligned}$$

The following lemma was proved in [1].

Lemma 1. The following statements hold for any B_4-code x.

1) The number

$$m_2(\cdot, Z_1) \leq 4\, tl_n(1).$$

2) The set $Z_1 \subset Z_2$, and if $w \in Z_2 \setminus Z_1$, $m_2(\cdot, w) \leq 4$.

Introduce a triple of vector random variables $(\eta, \xi, \tilde{\eta})$

$$\eta = (\eta_1, \ldots, \eta_{N-n}),\ \tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_{N-n}),\ \xi = (\xi_1, \ldots, \xi_n)$$

with distribution given by the following relations

$$P_r\{\xi = v\} = \mu_2(v, \cdot)\, t^{-2},$$

$$P_r\{\eta = w, \tilde{\eta} = \tilde{w} \mid \xi = v\} = \frac{\mu_2(v, w) \cdot \mu_2(v, \tilde{w})}{\mu_2^2(v, \cdot)}. \tag{6}$$

Allowing for (4), obtain from (6)

$$P_r\{\eta - \tilde{\eta} = w\} = t^{-2} \sum_v \mu_2(v, \cdot)^{-1}\, m_2(v, w).$$

Consequently,

$$P_r\{\eta - \tilde{\eta} \in Z_1\} = t^{-2} \sum_v \mu_2(v, \cdot)^{-1}\, m_2(v, Z_1) \leq$$

$$\leq t^{-2} \sum_v \sqrt{m_2(v, Z_1)} \leq t^{-2} \sqrt{3^n m_2(\cdot, Z_1)} \leq 2t^{-2} \sqrt{3^n\, tl_n(1)}, \tag{7}$$

the above estimates being based on inequality (5), upper convexity of the function $y = \sqrt{x}$, and statement 1) of Lemma 1.

Let

$$H(\eta - \tilde{\eta}) \stackrel{\triangle}{=} - \sum_{w \in Z_2} P_r\{\eta - \tilde{\eta} = w\} \log P_r\{\eta - \tilde{\eta} = w\}$$

be Shannon entropy [3] of the vector random variable $\eta - \tilde{\eta}$. Then (7), statement 2) of Lemma 1 and obvious inequality

$$H(\eta - \tilde{\eta}) \geq - P_r\{\eta - \tilde{\eta} \in Z_2 \setminus Z_1\} \log \left[\max_{w \in Z_2 \setminus Z_1} P_r\{\eta - \tilde{\eta} = w\}\right],$$

result in estimate

$$H(\eta - \tilde{\eta}) \geq \left[1 - 2t^{-2} \sqrt{3^n\, tl_n(1)}\right] \log (t^2/4). \tag{8}$$

Below, consideration will be given to the proof of

Lemma 2. The entropy of the difference

$$H(\eta - \tilde{\eta}) \leq (N - n)\, \underline{H}(q^*) \tag{9}$$

where $\underline{H}(q^*)$ is defined in theorem formulation.

Inequalities (8) and (9) mean that we have obtained

Statement 1. If for B_4-code of size $N \times t$ and integer n, $0 \leq n \leq N$, $l_n(1) < t \log t$ then

$$N \geq n + 2\underline{H}(q^*)^{-1}\left[1 - 2t^{-2}\sqrt{3^n t^2 \log t}\right] \log (t/2). \tag{10}$$

Now we proceed to the next step of the proof which involves derivation of an inequality similar to (10) for the case where $l_n(1) \geq t \log t$. Introduce

$$\tilde{m}(w) \stackrel{\triangle}{=} \sum_{w'} m_1(\cdot, w')\, m_1(\cdot, w - w'),$$

$$\tilde{Z}_2 = \{w : \tilde{m}(w) > 0\},$$

$$\tilde{m}(Z_2) = \sum_{w \in z_1} \tilde{m}(w).$$

One may readily see that for any w, $\tilde{m}(w) \leq m_2(\cdot, w)$; therefore, Lemma 1 leads to

Lemma 3. The following two statements hold for any B_4-code:

1) The number $m(Z_1) \leq 4tl_n(1)$.

2) The set $Z_1 \subset \tilde{Z}_2$, and if $w \in \tilde{Z}_2 \setminus Z_1$, $\tilde{m}(w) \leq 4$.

Consider a pair (ξ_1, ξ_2) (where $\xi_i = (\xi_1^{(i)}, \ldots, \xi_{N-n}^{(i)})$) of independent vector random variables with the same probability distribution

$$P_r\{\xi_i = w\} = m_1(\cdot, w)/l_n(1), \quad i = \overline{1, 2}.$$

Obviously,

$$P_r\{\xi_1 + \xi_2 = w\} = \tilde{m}(w)/l_n^2(1).$$

Consequently, Statement 1 of Lemma 3 implies that

$$P_r\{\xi_1 + \xi_2 \in Z_1\} \leq 4t/l_n(1). \tag{11}$$

Similarly to (8), one may prove by Statement 2) of Lemma 3 and (11) that

$$H(\xi_1 + \xi_2) \geq [1 - 4t/l_n(1)] \log (l_n^2(1)/4). \tag{12}$$

Direct analytical verification (we are not going to make it here) easily demonstrates that the following lemma holds.

Lemma 4. Let $(\xi_k^{(1)}, \xi_k^{(2)})$ be independent identically distributed random variables taking values $0, 1, -1$ with probabilities

$$P_r\{\xi_k^{(i)} = 0\} = p_k, \; P_r\{\xi_k^{(i)} = 1\} = P_r\{\xi_k^{(i)} = -1\} = (1 - p_k)/2$$

where $0 \leq p_k \leq 1$, $k = \overline{1, N-n}$. Then the entropy of the sum $H(\xi_k^{(1)} + \xi_k^{(2)})$ vs. the parameter p_k, $1/2 \leq p_k \leq 1$, reaches its maximum

$$H_4 \triangleq -\sum_{i=0}^{4} C_4^i 2^{-4} \log (C_4^i 2^{-4}) = 2.03064 < 2.03096 = \underline{H}(q^*) \qquad (13)$$

when $p_k = 1/2$.

The inequality

$$H(\xi_1 + \xi_2) \leq \sum_{k=1}^{N-n} H(\xi_k^{(1)} + \xi_k^{(2)}) \leq (N - n) H_4$$

follows from Lemma 4 and the sub-additivity of the entropy [3]. It results after allowance for (12) in

Statement 2. If for B_4-code of size $N \times t$ and integer $n, 0 \leq n \leq N$, the value of $l_n(1) \geq t \log t$, then

$$N \geq n + 2H_4^{-1}(1 - 4/\log t) \log (t/2). \qquad (14)$$

Taking into (10) and (14) that

$$n \triangleq \lfloor 2 \log (t/\log t)/\log 3 \rfloor$$

(here $\lfloor Q \rfloor$ is greatest integer less or equal to Q). come in virtue of (13) to the statement of the theorem. To complete the proof, one has to prove Lemma 2. Introduce

$$\mu_2^{(j)}(v, w) \triangleq \sum_{w \in W_j(w)} \mu_2(v, w)$$

where $j = \overline{1, N-n}$, $w = \overline{0, 2}$, and the set

$$W_j(w) \triangleq \{w = (w_1, w_2, \ldots, w_{N-n}) : w_j = w\}.$$

One may easily see from (6) that the joint distribution of pair $(\eta_j, \tilde{\eta}_j)$ is given by the relation

$$P_r\{\eta_j = w, \tilde{\eta}_j = \tilde{w}\} = \sum_v \frac{\mu_2^{(j)}(v, w)\, \mu_2^{(j)}(v, \tilde{w})}{t^2 \cdot \mu_2(v, \cdot)} \qquad (15)$$

and that marginal distributions are binomial

$$P_r\{\eta_j = w\} = P_r\{\tilde{\eta}_j = w\} = \sum_{\nu} t^{-2}\,\mu_2^{(j)}(v, w) =$$

$$= C_2^w \left(\frac{|x_{n+j}|}{t}\right)^w \left(1 - \frac{|x_{n+j}|}{t}\right)^{2-w}, \quad j = \overline{1, N-n} \tag{16}$$

where $|x_{n+j}| \stackrel{\triangle}{=} \sum_{k-1}^{t} x_{n+j}(k)$ is the number of 1's in $(n+j)$-th row of B_4-code. Consider

$$M(\eta_j \cdot \tilde{\eta}_j) = \sum_{\nu} \frac{\mu_2(v, \cdot)}{t^2} \left[\sum_{w=0}^{2} w \frac{\mu_2^{(j)}(v, w)}{\mu_2(v, \cdot)}\right]^2 \geq \left[\sum_{w=0}^{2} w P_r\{\eta_j = w\}\right]^2 = (M_{\eta_j})^2$$

where (15), (16) and downward convexity of the function $y = x^2$ are used. Therefore, for any $j = \overline{1, N-n}$

$$M(\eta_j - \tilde{\eta}_j)^2 = 2\,[M\eta_j^2 - M(\eta_j \tilde{\eta}_j)] \leq 2D\eta_j = 4\,|x_{n+j}|\,(t - |x_{n+j}|)\,|t^2 \leq 1. \tag{17}$$

Note that $\eta_j - \eta$ has the following distribution

$$(\eta_j - \tilde{\eta}_j) \sim \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ q_2 & q_1 & q_0 & q_1 & q_2 \end{pmatrix}$$

where $q_0 + 2q_1 + 2q_2 = 1$, $0 \leq q_i \leq 1$, $i = \overline{0, 2}$. Inequality (17) means that the vector $q = (q_0, q_1, q_2)$ meets the condition $8q_2 + 2q_1 \leq 1$ which is equivalent to

$$q_1 \geq 1/2 - 2q_0/3. \tag{18}$$

Therefore, the Shannon entropy

$$H(\eta_j - \tilde{\eta}_j) \leq \max_{(18)} H(q), \quad j = \overline{1, N-n} \tag{19}$$

where $H(q)$ is defined by (1), and maximum is taken under condition (18). Simple analytical verification (omitted here) demonstrates that maximum in (19) is reached over the vector q^* given in the theorem formulation. Application of the entropy sub-additivity

$$H(\eta - \tilde{\eta}) \leq \sum_{i=1}^{N-n} H(\eta_i - \tilde{\eta}_i)$$

enables one to obtain inequality (9) from (19).

Q.E.D.

Conclusions

The paper is dedicated to the study of minimal code-length for codes enabling discrimination of addends in the sum provided that their number is at most four. For this case, an asymptotic lower bound of the minimal length has been derived under the assumption that the number of words in the code tends to infinity. The major result of the paper is represented by inequality (2). The result is compared with a previous estimate from [1]. This method of B_S-code estimation for $S = 4$ gives rise to hopes for further improvement of the bound obtained in [1] also in the general case.

References

1. *D'yachkov, A. G., Rykov, V. V.*, On a coding model for multiple-access channel. Problemy peredachi informatsii, 1981, **17**, *2*, pp. 26–38 (in Russian).
2. *Tsybakov, B. S., Mikhailov, V. A.*, Free synchronous access of batches into broadcast feedback channel. Problemy peredachi informatsii, 1978, **14**, *4*, pp. 32–59 (in Russian).
3. *Gallager, R. G.*, Information theory and reliable communication, John Wiley & Sons, Inc., 1968.

LOWER BOUND OF THE ALLOWABLE RATE DOMAIN FOR ADDER-CHANNEL WITH CORRELATED SOURCES

V. N. DYN'KIN, G. L. KATSMAN, A. P. KURDYUKOV

Lower bound of code power is derived for codes intended for data communication through adder-channel with correlated sources.

Recent years have witnessed further development of information and control systems and sophistication of their structures which resulted in more involved structure of data flows and in need for basically different procedures of user access to the system.

The classical model where one source sends information to one receiver through a noisy channel is generally accepted because it describes rather well many kinds of communication systems.

However, a necessity arose for studies of more general communication systems. These studies comprise a rapidly developing area of the information theory called multiple-access channels, i.e. channels with multiple inputs and single output.

On the other hand, the majority of current satellite communication systems makes use of multiplexing with frequency separation where each channel corresponds at every instant to one source and one receiver. However, the satellite communication system may be regarded as a system with multiple access channel which enables one to improve some of its characteristics.

Owing to the above-said, studies of the domain of allowable transmission rates through multiple-access channels become currently topical.

The present paper derives lower bounds for allowable rates of an adder-channel with correlated sources.

1. Problem formulation

Consider coding for the following channel. Let there be three sources I_0, I_1 and I_2 sending messages j, i and k, respectively. Coder 1 generates a binary vector u of length n through mesages i and j, while coder 2 makes use of the messages k and j for the generation of the vector v. The vectors are fed into inputs of the adding channel which outputs the vector $u + v$. Here

and below vector addition is understood in the component-wise sense, components being added as conventional real numbers. The decoder has to discriminate messages j, i and k through channel output vector $u + v$. Such a channel is referred to as adding channel with correlated sources [1].

Our task is to estimate powers of sources (M_0, M_1 and M_2) under which the messages j, i and k may be discriminated with zero error probability.

Through the formula $R_i = (\log_2 M_i)/n$, $i = 0, 1, 2$, assign to the powers M_0, M_1 and M_2 a triple of rates R_0, R_1 and R_2. The triples of rates under which the messages j, i and k may be determined with zero error probability will be called allowable.

Let M_0 pairs of binary codes of length n $(U_1, V_1), (U_2, V_2), \ldots, (U_{M_0}, V_{M_0})$, $|U_j| = M_1$, $|V_j| = M_2$, $j = 1, \ldots, M_0$, be given such that for any j_1 and j_2 and any four vectors u, u', v, v', $u \in U_{j1}$, $v \in V_{j1}$, $u' \in U_{j2}$, $v' \in V_{j2}$ the equality $u + v = U' + V'$ holds if and only if $j_1 = j_2$ and $u = u'$, $v = v'$. Such a sequence of code pairs will be called separable.

It may be readily seen that if there is a separable sequence of pairs of codes with powers M_0, M_1 and M_2, then the triple of rates R_0, R_1, R_2 is allowable $(R_i = (\log_2 M_i)/n,\ i = 0, 1, 2)$.

2. Convexity of the domain of allowable rates

Lemma 1. If there is a separable sequence of code pairs of length n with rate triple R_0, R_1, R_2, then for any natural m there is a separable code sequence of length mn with the same rates.

Existence of an infinite set of separable code pair sequences with rates R_0, R_1, R_2 follows from the existence of the allowable triple of rates R_0, R_1, R_2.

Lemma 2. If there are two separable code pair sequences, R_0, R_1, R_2 and R_0', R_1', R_2', there is a separable code pair sequence with rates

$$\frac{l}{m} R_0 + \left(1 - \frac{l}{m}\right) R_0',$$

$$\frac{l}{m} R_1 + \left(1 - \frac{l}{m}\right) R_1',$$

$$\frac{l}{m} R_2 + \left(1 - \frac{l}{m}\right) R_2'.$$

Lemma 2 proves the possibility of time-sharing between two separable code pair sequences.

Corollary. The set of the allowable rate triples is convex.

The proof of the corollary is evident from Lemma 2 and the definition of convex set.

Lemma 3. If there is an allowable triple of rate (R_0, R_1, R_2), then for any $0 \leq r_1 \leq R_1$ and $0 \leq r_2 \leq R_2$ there is an allowable triple of rates $(r_1 + r_2 + R_0, R_1 - r_1, R_2 - r_2)$.

Proof. Let $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ be separable sequence of pairs of codes with powers M_1, M_2. Decompose each code U_j into l_1 disjoint subcodes of equal power $\left[\frac{M_1}{l_1}\right]$, $U_{j_1}, \ldots, U_{jl_2}$, $j = 1, 2, \ldots, M_0$. Similarly, decompose each code V_j into l_2 disjoint subcodes of equal power $\left[\frac{M_2}{l_2}\right]$, $V_{j_1}, \ldots, V_{jl_2}$, $j = 1, 2, \ldots, .$

Consider the sequence of code pairs (U_{11}, V_{11}), $(U_{12}, V_{12}), \ldots, (U_{11}, V_{1l_2})$, $(U_{12}, V_{11}), \ldots, (U_{M_0 l_1}, V_{M_0 l_2})$. It may be esaily seen that the sequence is separable with powers $M_0 l_1 l_2, \left[\frac{M_1}{l_1}\right], \left[\frac{M_2}{l_2}\right]$. Hence, the statement of Lemma 3 follows.

3. Lower bound of the allowable rate domain

Theorem 1. Any triple of rates satisfying the inequalities $0 \leq R_0 \leq \leq \log_2 3 - \frac{2}{3}$, $R_1 + R_2 \leq \frac{2}{3}$, $R_1 \geq 0$, $R_2 \geq 0$ is allowable.

Proof. It follows from Lemmas 2 and 3 and from evident reason that the rate triple R_0, R_1, R_2 is allowable if the rate triple R_0, R_1, R_2 is allowable, that it is sufficient to prove admissibility of the triple of rates $R_0 = \log_2 3 - - \frac{2}{3} - \varepsilon$, $R_1 = \frac{2}{3}$, $R_2 = 0$ for any $\varepsilon = > 0$.

To prove this statement, let us construct a separable code pair sequence of length n and powers $M_0 = C_n^{n/3}$, $M_1 = 2^{2n/3}$, $M_2 = 1$.

Enumerate in arbitrary fashion by numbers from 1 to $C_n^{n/3}$ all possible combinations from n by $n/3$. A code consisting of a single vector of length n and weight $n/3$ where 1's stand only in positions involved into the i-th combination is considered as code V_i. One can construct $C_n^{n/3}$ such codes. Further, a code of length n and power $2^{2n/3}$ consisting of vectors where positions of i-th combination have 1's while the rest of them is arbitrary is considered as code U_i. Let $u \in U_i$, $v \in V_j$, $u' \in U_j$, $v' \in V_j$. If $i \neq j$, $u + v = u' + v'$ because in vector $u + v$ two's stand in positions belonging to the i-th combination, and in $u' + v'$ they stand in positions of the j-th one. If $i = j$, then $V = V'$ and $u = u'$ follows from the fact that $u + v = u' + v'$.

Thus, we have obtained a separable sequence of pairs of codes with powers $M_0 = C_n^{n/3}$, $M_1 = 2^{2n/3}$, $M_2 = 1$. Since $\lim_{n\to\infty} \frac{1}{n} \log_2 C_n^{n/3} = \log_2 3 - \frac{2}{3}$ the theorem follows.

Theorem 2. Any triple of rates (R_0, R_1, R_2) satisfying inequalities $0 \leq R_1 < 1$, $0 \leq R_2 < 1$, $0 \leq R_0$, $R_0 + R_1 + 2R_2 < 3 - \log_2 3$ is allowable. To prove this theorem, let us at first prove the following lemma.

Lemma 4. There are at most $2 \cdot 6^n$ ordered quadruples of binary vectors of length n, u, v, u', v' such that $u + v = u' + v'$.

Proof. Let $u + v = u' + v' = w$, and the number of one's in the vector w be equal to l. Then there are 2^l ways to select an ordered pair of vectors (u, v) such that $u + v = w$. The same number of ways exists for the vector pair (u', v') such that $u' + v' = w$. Thus, there are at most $2 \cdot 4^l$ ordered quadruples of vectors u, v, u', v' such that $u + v = u' + v' = w$. And the number of vectors w such that the number of one's in them is l is equal to $C_n^l 2^{n-l}$. Therefore, the number of ordered quadruples u, v, u', v' of binary vectors of length n such that $u + v = u' + v'$ does not exceed $2 \sum_{l=0}^{n} C_n^l 2^{n-l} 4^l = 2 \cdot 6^n$.

Q.E.D.

Proceed now to the proof of Theorem 2. Consider all the possible sequences $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ of pairs of codes of length n, $|U_i| = M_1$, $|V_i| = M_2$, $i = 1, 2, \ldots, M_0$. There are $2^{(M_1+M_2)M_0 n} = A$ different sequences, each code $U_i(V_i)$ being an ordered set $M_1(M_2)$ of vectors of length n among which may be identical ones. Decompose each code into $s = M_1/l$ subsets $U_{i1}, \ldots, U^i{}_s$ of power l (one may assume that M_1 is dividable by l without limiting the generality). The subset U_{ij} $(i = 1, 2, \ldots, M_0;\ j = 1, 2, \ldots, s)$ will be called bad if for any vector $u \in U_{ij}$ there will be three vectors, u', v, v' such that $u' \in U_p$, $v \in V_i$, $v' \in V_p$, $u + v = u' + v'$. The sequence of pairs of codes $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ will be called bad if there is at least one bad subset U_{ij}, $1 \leq i \leq M_0$; $1 \leq j \leq s$. It may be easily seen that the number of bad sequences is upper-bounded by

$$\frac{M_0 M_1}{l} [2 \cdot 6^n M_1 l M_2^2]^l \cdot 2^{[(M_1+M_2)M_0 - 4l]n} <$$

$$< 2^l M_0^{l+1} M_1^{l+1} M_2^{2l} 6^l \cdot 2^{[(M_1+M_2)M_0-4l]n} = B.$$

Consequently, if $B \leq A$, i.e. $M_2^2(M_0M_1)^{1/l}M_0M_1 \leq 8^n/2 \cdot 3^n$, there will be at least one sequence of code pairs which is not bad. Assume that $l = n$, $R_0 = \frac{1}{n}\log_2 M_0$, $R_1 = \frac{1}{n}\log_2 M_1$, $R_2 = \frac{1}{n}\log_2 M_2$. The inequality (1) may be rewritten as $R_0 + R_1 + 2R_2 + \frac{1}{n}(R_0 + R_1) \leq 3 - \log_2 3 - \frac{1}{n}$. Thus, for rather large n and $R_0 + R_1 + 2R_2 < 3 - \log_2 3$ there is a sequence of code pairs which is not bad and has rates R_0, R_1, R_2.

Let $(U_1, V_1), \ldots, (U_{M_0}, V_{M_0})$ be a sequence of code pairs which is not bad. Denote $u_i = \{u_{ij} | u_{ij} \in U_{ij}$ such that $\neg \exists v', u', v', v \in V_i, u' \in U_p, v' \in V_p$ and that $u_{ij} + v = u' + v'$, $p = 1, 2, \ldots, M_0\}$. Note that such a vector u_{ij} will be necessarily in each set u_{ij} because the sequence of code pairs $(U_1, V_1), \ldots$ $\ldots, (U_{M_0}, V_{M_0})$ is not bad, Moreover, without restricting the generality one may assume that only one vector is taken from each subset U_{ij} even if more may be taken.

It is easy to see that codes the U'_i, $i = 1, \ldots, M_0$, obtained in this way have power $\frac{M_1}{n}$, and that the sequence $(U', V_1), \ldots, (U'_{M_0}, V_{M_0})$ is a separable sequence of code pairs, its rate triple being $R_0, R_1 - \frac{\log n}{n}, R_2$. Thus, for any triple of rates satisfying the inequalities $0 \leq R_1 < 1, 0 \leq R_2 < 1$, $0 \leq R_0$, $R_0 + R_1 + 2R_2 < 3 - \log 3$ there is under sufficiently great n a separable sequence of code pairs with triple of rates R_0, R_1, R_2.

Q.E.D.

The lower bound of the domain of allowable rate triples may be constructed through Theorems 1 and 2, Corollary of Lemma 2 and an evident reason that under the conditions of Theorem 2 rates R_1 and R_2 may be interchanged (Fig. 1). Coordinates of

point A: $R_0(A) = \log_2 3$, $R_1(A) = R_2(A) = 0$;

point B: $R_0(B) = \log_2 3 - \frac{2}{3}$, $R_1(B) = \frac{2}{3}$, $R_2(B) = 0$;

point C: $R_0(C) = 2 - \log_2 3$, $R_1(C) = 1$, $R_2(C) = 0$;

point D: $R_0(D) = 0$, $R_1(D) = 1$, $R_2(D) = 0$;

point E: $R_0(E) = 0$, $R_1(E) = 1$, $R_2(E) = 1 - \frac{1}{2}\log_2 3$.

Coordinates of points B', C', D', E' may be obtained from the coordinates of points B, C, D, E, respectively, by substitution of R_2 and R_1 for R_1 and R_2.

4. Conclusion

Let us dwell now on possible improvement of the derived lower bound of the domain of allowable rate triples. Evidently the estimate cannot be improved within the interval $R_0 \geq \log_2 3 - \frac{2}{3}$ because it coincides with the evident upper bound $R_0 + R_1 + R_2 \leq \log_2 3$. Under other values of R_0 the estimate may be improved, for example, through consideration within the plane $(R_1, R_2), (R_0 = 0)$ of allowable rate pairs for conventional adder-channel [2, 3]. For $R_1 < 1$, $R_2 \geq 1 - \frac{1}{2}\log_2 3$ code pairs were obtained [2] lying above the straight line EE'. Time-sharing of these code pairs with code pair sequences whose rates are depicted in Fig. 1 improves the obtained lower bound within the interval $BEE'B'$. Thus, the procedure proposed in the paper for code construction enables either complete or sufficiently good use of channel capacity. The technique proposed may be extended to the case when the number of sources is $2^k - 1$ and the number of coders is k.

References

1. *Slepian, D., Wolf, J. K.*, A coding theorem for multiple access channels with correlated sources, BSTJ, 1973, **52**, 7, pp. 1037–1076.
2. *Kasami, T., Lin, S.*, Bounds on the Achievable Rates of Block Coding for a Memoryles. Multiple-Access Channel, IEEE Trans. IT, 1978, **24**, 2, pp. 187–197.
3. *Weldon, E. J.*, Coding for a Multiple-Access Channel, I & C, **36**, pp. 256–274 (1978)

AN APPLICATION OF ADAPTIVE KALMAN FILTERING TO SHORT-TERM LOAD FORECASTING FOR POWER SYSTEMS

H. HINKEL

The problem of short-term load forecasting in power systems for forecasting periods from 10 minutes to 1 hour is considered. Special adaptive Kalman filters are used for prediction. Computer programs were tested by using both simulated and real data for the frequency of the electrical power system.

1. Introduction

An efficient production and distribution of electrical energy is a problem of great economic importance. For a fast, purposeful and possibly automatic influence on the production and distribution of electrical energy, it is necessary to know the future demand of energy in the whole system or in parts of the system.

This problem is called the problem of short-term load forecasting. The purpose of this paper is to show how Kalman filters can be successfully used in dealing with problems of prediction. Since data describing the load of the system have not been available so far, prediction was carried out for a real sequence of data describing the frequency of the electrical power system, where the deviation of this frequency from the nominal frequency ($f = 50\,\mathrm{Hz}$) is a measure for the difference between consumption and production of electrical energy.

2. State estimation in the case of unknown covariances of the noise processes

For the application of discrete Kalman filters to prediction it is necessary to know a discrete-time state model (2.1)

$$\begin{aligned} x_{k+1} &= Fx_k + u_k \\ y_k &= Hx_k + v_k \end{aligned} \tag{2.1}$$

which describes the process to be predicted, where $\{x_k\}$ is the state of the system and $\{y_k\}$ the observed process. It will be assumed that the disturbances $\{u_k\}$ and $\{v_k\}$ are independent zero-mean stationary discrete white noise sequences.

For the construction of a Kalman filter not only the matrices F and H but also the covariances $Q = E(u_k u_k^T)$ and $R = E(v_k v_k^T)$ of the disturbances have to be known.

In the case of short-term load forecasting it can be assumed that the matrices F and H are given in a very simple form (see Section 4). However, the main problem is that the matrices Q and R are unknown. For this reason it is impossible to construct a Kalman filter according to the well-known relation

$$\hat{x}_{k+1} = F\hat{x}_k + L_k[y_k - H\hat{x}_k] \tag{2.2}$$

with

$$L_k = FM_k H^T [HM_k H^T + R]^{-1}, \quad \hat{x}_0 = Ex_0 \tag{2.3}$$

where the covariance $M_k = E(e_k e_k^T)$ of the estimation error $e_k = x_k - \hat{x}_k$ is governed by the discrete matrix Riccati equation

$$M_{k+1} = FM_k F^T - FM_k H^T [HM_k H^T + R]^{-1} HM_k F^T + Q. \tag{2.4}$$

Such problems and ways of how to solve them are known from the literature. Some papers exist which deal with so-called adaptive filters. For application to short-term load forecasting, correlation methods seem to be the best of the known ways of adaptive filtering, because in this practical case the demands of these methods are approximately satisfied.

3. The innovation correlation method

The basic idea of correlation methods is to correlate the output of the system either directly or after a known linear operation on it [2]. A set of equations is derived relating the system parameters to the observed auto-correlation function.

A comparative analysis shows that the so-called innovation correlation method which correlates the innovations

$$v_i = y_i - \hat{y}_i = y_i - Hx_i$$

of the output process $\{y_i\}$ is especially suitable for short-term load forecasting.

This method requires system (2.1) to be completely controllable and completely observable.

Instead of filter (2.2) the steady-state filter

$$\hat{x}_{k+1} = F\hat{x}_k + L(y_k - H\hat{x}_k) \tag{3.1}$$

will be considered, where L is the constant steady-state filter gain ($\lim_{k\to\infty} L_k = L$). Moreover, filter (3.1) will be considered in the steady-state. Then we have $M_k = M = \text{const.}$ for the covariance M of the estimation error. If the covariances Q and R are approximately known, a suboptimal filter gain L can be determined and it is possible to construct a corresponding suboptimal filter

$$x^*_{k+1} = Fx^*_k + L^*(y_k - Hx^*_k).$$

The aim of the application of the innovation correlation method to adaptive filtering is to determine an estimate of the optimal filter gain based on the knowledge of estimates of the covariances of the innovation sequence of given suboptimal filter. For the solution of this problem two different methods are known from the literature [2, 3]. The first purpose of the analysis carried out here was to compare the efficiency of both methods for application to short-term load forecasting. But it could be shown that the two methods are not essentially different; therefore only one of them was programmed for computation and used for prediction.

In Sections 3.1 and 3.2 the basic ideas of both methods are described and in Section 3.3 the connection of these methods is shown.

3.1. *Innovation correlation method* I [2]

Consider a suboptimal filter for system (2.1)

$$x^*_{i+1} = Fx^*_i + L^*(y_i - Hx^*_i). \tag{3.2}$$

The innovation sequence $\{v^*_i\}$ of the suboptimal filter

$$v^*_i = y_i - y^*_i = y_i - Hx^*_i = He^*_i + v_i$$

is not a white noise process.

For the covariance of this sequence we have in the steady-state case

$$T^*_k = E\{v^*_i\, v^{*T}_{i-k}\} = \begin{cases} HM^*H^T + R & \text{for } k = 0 \\ HE^*_i\, e^{*}_{i-k} H^T + HE\, e^*_i\, v^T_{i-k} & \text{for } k > 0 \end{cases}$$

M^* is the covariance of the estimation error $e^*_i = x_i - x^*_i$ of the suboptimal filter.

Transformed in a suitable way, the system of equations governing T^*_k ($k = 1, 2, \ldots, n$) yields the following relation

$$AM^*H^T = \begin{bmatrix} T^*_1 + HL^*T^*_0 & & \\ T^*_2 + HL^*T^*_1 & + HFL^*T^*_0 & \\ \vdots & & \\ T^*_n + HL^*T^*_{n-1} + & \ldots + \mathrm{HF}^{n-1}L^*T^*_0 & \end{bmatrix} = E, \tag{3.3}$$

where

$$A^T = [(HF)^T, (HF^2)^T, \ldots, (HF^n)^T].$$

By denoting the pseudoinverse of A by A^+ the following relation will be obtained

$$M^* H^T = (A^T A)^{-1} A^T E = A^+ E .$$

Asymptotically unbiased and consistent estimates $\hat{T}_k^*$ are given by formula (3.4)

$$\hat{T}_k^* = \frac{1}{N} \sum_{i=k}^{N} \nu_i^* \nu_{i-k}^{*T}, \qquad k = 1, 2, \ldots \tag{3.4}$$

If the notation $\delta M = M - M^*$ is introduced for the difference between the covariance of the estimation error M of the wanted optimal filter and the covariance M^* we shall obtain the following relations for $\hat{L}$

$$\hat{L} = F(\widehat{M^* H^T} + \delta M H^T)\ (\hat{T}_0^* + H \delta M H^T)^{-1}. \tag{3.5}$$

It can be shown that δM is governed by the following matrix difference equation

$$\delta M = F[\delta M - (\widehat{M^* H^T} + \delta M H^T)(\hat{T}_0^* + H \delta M H^T)^{-1}(\widehat{H M^*} + H \delta M)]\, F^T + C, \tag{3.6}$$

where

$$C = L^* H M^* F^T + F M^* H^T L^{*T} - L^* \hat{T}_0^* L^{*T}.$$

If δM which solves this equation is inserted into the relation (3.5), asymptotically unbiased and consistent estimates of the optimal filter gain will be obtained.

3.2. *Innovation correlation method* II [3]

In a known way a state space model (2.1) can be transformed into another equivalent state space model (3.7), where the optimal estimates $\hat{x}$ play the role of the state variables.

$$\begin{aligned} \hat{x}_{i+1} &= F\hat{x}_i + L\nu_i \\ y_i &= H\hat{x}_i + \nu_i . \end{aligned} \tag{3.7}$$

The sequence $\{\nu_i\}$ of the disturbances is identical here with the innovation sequence of the optimal filter.

For system (3.7) a suboptimal Kalman filter (3.2) can be constructed which yields suboptimal estimates x_{i+1}^* for $\hat{x}_{i+1}$. The steady-state covariance P^* of the estimation error $\tilde{e}_i^* = \hat{x}_i - x_i^*$ is essential for this method. This matrix is governed by the following relation

$$\begin{aligned} P^* &= \{(\hat{x}_{i+1} - x_{i+1}^*)(\hat{x}_{i+1} - x_{i+1}^*)^T\} = \\ &= (F - L^*H)P^*(F - L^*H)^T + (L^* - L)T_0(L^* - L)^T \end{aligned} \tag{3.8}$$

where

$$T_0 = E\, v_i v_i^T = HMH^T + R.$$

Analogously to the innovation correlation method I the covariances T_k^* of the innovation sequence of the suboptimal filter (3.2) are determined for the steady-state. These covariances are functions of P^*. By using suitable estimates $\hat{T}_k^*$ $(k = 1, \ldots, n)$ analogous to formula (3.4), a relation can be obtained which connects the desired matrix $\hat{L}$ with T_0 and P^*.

If the relation

$$T_0 = T_0^* - HP^*H^T$$

is taken into consideration three simultaneous matrix equations are given governing P^*, T_0 and $\hat{L}$. These equations can be solved by means of a special converging algorithm.

3.3. *Connection between the innovation correlation methods* I *and* II

The analysis carried out here has shown that the results of both methods must be identical, because the matrix P^* does not differ essentially from the matrix δM.

Both methods start from the suboptimal filter (3.2). It can be shown that the covariance M^* of the estimation error $e_i^* = x_i - x_i^*$ and the covariance P^* of the estimation error $\tilde{e}_i^* = \hat{x}_i - x_i^*$ are connected in a very simple way. M^* can be represented by the following relation

$$\begin{aligned} M^* &= Ex_i^* x_i^{*T} = E(x_i - x_i^*)(x_i - x_i^*)^T = \\ &= E[(x_i - \hat{x}_i) + (\hat{x}_i - x_i^*)][(x_i - \hat{x}_i) + (\hat{x}_i - x_i^*)] = \\ &= E(x_i - \hat{x}_i)(x_i - \hat{x}_i)^T + E(\hat{x}_j - x_i^*)(\hat{x}_i - x_i^*)^T. \end{aligned} \tag{3.9}$$

Hence

$$M^* = M + P^* \tag{3.10}$$

and

$$P^* = M^* - M = -\delta M, \text{ respectively.}$$

The two innovation correlation methods differ only in the way they determine P^* and $\hat{L}$ and δM and $\hat{L}$, respectively. The system of simultaneous equations for P^*, T_0, $\hat{L}$ can be transformed immediately into equation (3.6) which governs δM. Therefore it is possible to restrict oneself to the computerized implementation of one of these methods. The computer program to be used for short-term load forecasting is realizing the innovation correlation method I.

4. Test of computer programs for simulated as well as real data

In order to test the program intended for short-term load forecasting it is necessary to know a corresponding model of the system. Only prediction intervals from 10 minutes to 1 hour are considered. For such intervals a simple system model (4.1) was given in [4].

It is reasonable for two state variables to exist

- load, ξ_k
- increment of load, Δ_k.

The structure of state and observation is given by the following model

$$\begin{bmatrix} \xi_{k+1} \\ \Delta_{k+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \xi_k \\ \Delta_k \end{bmatrix} + \begin{bmatrix} u_k^1 \\ u_k^2 \end{bmatrix}$$

$$y_k = [1 \quad 0] \begin{bmatrix} \xi_k \\ \Delta_k \end{bmatrix} + v_k. \tag{4.1}$$

The disturbances $\{u_k^1\}$ and $\{u_k^2\}$ are discrete stochastic processes. It will be assumed that they are independent and dependent of the noise process $\{v_k\}$. The covariances of the disturbances have the structure

$$E(u_i^1 u_j^1) = q_i^{11} \delta_{ij} \qquad E(v_i v_j) = R_i \delta_{ij}$$
$$E(u_i^2 u_j^2) = q_i^{22} \delta_{ij}$$

where δ_{ij} is the Kronecker delta. The covariance matrices Q_i and R_i are a priori unknown.

4.1. *Test for simulated data*

Innovation correlation methods start from the innovation sequence of a suboptimal steady-state filter. Two demands have to be satisfied for constructing such an innovation sequence.

a) A sequence of observed values $\{y_n\}$ must be known for the system (2.1).
b) A suboptimal filter gain L^* must be known.

A sequence $\{y_n\}$ can be generated by simulation of a corresponding state model. Model (4.1) is unsuited for simulation (the roots of the characteristic equation $|\lambda I - F| = 0$ are on the unit circle). Therefore, model (4.1) was modified in such a way that the resulting system will be stable, completely controllable and completely observable.

The system matrices were given by

$$F^M = \begin{bmatrix} 0.8 & 1 \\ 0 & 0.8 \end{bmatrix} \qquad H^M = [1 \quad 0] \tag{4.2}$$

and the covariances of the stationary white noise processes were chosen in the following way

$$Q^M = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \qquad R^M = 0.5. \tag{4.3}$$

The digital simulation provided the sequence $\{y_n^M\}$ of data satisfying demand a.

For the system given by formulas (4.2) and (4.3) the optimal filter gain can be determined by means of formulas (2.3) and (2.4). If covariances Q^{M*} and R^* are used, which differ only a little from Q^M and R^M, a suboptimal filter gain L^* satisfying demand b can be determined, again by means of formulas (2.3) and (2.4).

On the basis of simulated data $\{y_n^M\}$ for different optimal filter gains L^* the corresponding filter gain $\hat{L}$ was computed by means of programs realizing the innovation correlation method I. A comparison of $\hat{L}$ with the optimal filter gain showed that the quality of the results obtained does not depend so much on the suboptimal filter used. By way of experiment also $L^* = 0$ was used as a "suboptimal" filter gain. In this case the results were indeed a little worse than in case of theoretically determined suboptimal filter gains. But for the special simulation example considered here the results were so good that it seems admissible to use $L^* = 0$ (as an experiment) in the case of short-term load forecasting, where a priori information on the disturbances is not available.

As a whole it could be shown that by means of the program realizing an adaptive Kalman filter (innovation correlation method) a very good approximation to the theoretically optimal filter gain can be achieved. It could be demonstrated that the filter working with the estimated filter gain $\hat{L}$ produces an innovation sequence which is approximately a white noise process. Thus it has been verified that by means of the computed filter gain an almost optimal

filter or a corresponding Kalman predictor according to the relation

$$\hat{y}(k+\delta) = H\hat{x}(k+\delta) = HF^{\delta-1}\hat{x}(k), \quad \delta \geq 1 \tag{4.4}$$

can be computed.

In Fig. 2 both the simulated sequence $\{y^M\}$ (steady-state case) and the results of the one-step prediction obtained by means of adaptive filtering are depicted.

4.2. *Test for real data*

The test of the program for short-term prediction by means of adaptive Kalman filtering was carried out at first by using a sequence of data for the frequency of the electrical power system. Such a sequence $\{y_n\}$, $n=1, \ldots, 5000$, measured at a distance of ten minutes (period of the measurement: about 35 days) was available. For prediction the simple system model (4.1) was used. It was assumed that the disturbances are approximately stationary.

For judging in which way the number K of working steps of the suboptimal filter up to the steady state and the number K of values of the innovation sequence necessary for the estimation of T_k^* have to be chosen, different combinations of parameters were used ($K = 20$: $N = 1000$, $N = 1500$; $K = 50$: $N = 1000$, $N = 2000$). As a suboptimal filter gain $L^* = 0$ was used.

However, simple statistical tests have already shown that the hypothesis of the innovation sequence of the optimal filter being a white noise process must be rejected. Therefore it is impossible to construct an optimal Kalman filter by means of the computed filter gains $\hat{L}$.

On the one hand, this effect may be caused by a nonstationarity of the real sequence of data describing the frequency of the electrical power system. On the other hand, the chosen model (4.1) can be too simple.

Nevertheless, it could be shown that by means of the most favourable filter gain (obtained for $K = 50$, $N = 15000$) satisfactory results can be achieved by filtering and by one-step prediction, respectively.

A comparison between the measured values and the values obtained by one-step prediction can be carried out by using Fig. 2. The results of the prediction for more than one step ahead are not so favourable.

If it were possible to use the predicted values of the frequency in such a way that the predicted deviation from the nominal frequency is compensated by a corresponding influence on the production of electrical energy, a frequency difference within the range of the prediction error would be obtained as a result of this measure.

Therefore the following mean prediction error

$$\sigma_\delta^2 = \frac{1}{S-\delta} \sum_{i=\delta+1}^{S} (y_i - \tilde{y}_i^\delta)^2, \qquad \delta = 1, 2, 3, 4$$

has been used for judging the results. By $\{\tilde{y}^\delta\}$ the sequence of values obtained by δ-step prediction is denoted. The mean deviation of the measured frequency from the nominal frequency (50 Hz) is $\sigma_{\Delta f} = 162$ mHz for the whole sequence of data. This value is directly proportional to the mean balance of real power.

In the future it is intended to test the prediction program for real data describing the system load as soon as such sequences are available.

An essential assumption for the application of innovation correlation methods is the stationarity of the process to be predicted. This assumption is not satisfied by the process describing the development of the system load. However, it could be shown that it is possible to obtain suitable additional information, which can be used for a transformation of this process in such a way that the process to be predicted is approximately stationary.

References

1. *Mehra, R. K.*, "On the identification of variances and adaptive Kalman-filtering", IEEE Trans. on Automatic Control, vol. **AC-15**, 1970, pp. 175—184.
2. *Mehra, R. K.*, "Approaches to adaptive filtering", IEEE Trans. on Automatic Control, vol. **AC-17**, 1972, pp. 693—698.
3. *Carew, B., Bélanger, P. R.*, "Identification of optimum filtering steady-state gain for systems with unknown noise covariances", IEEE Trans on Automatic Control, vol. **AC-18**, 1973, pp. 582—587.
4. *Toyoda, J., Chen, M., Inoue, Y.*, "An application of state estimation to short-term load forecasting. Part I: Forecasting modeling", IEEE Trans. on Power Apparatus and Systems, vol. **PAS-89**, 1970, pp. 1678—1682.

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

L. Györfi

Technical University of Budapest

H–1111 Budapest XI., Stoczek u. 2, 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 (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 and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) 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 статей в четырех экземплярах на русском языке или в двух на английском и в двух на русском языке следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском и на английском языке (в двух экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Дьячков А. Г., Рыков В. В.: О нижней границе длины B_4-кода (*Dyachkov, A. G., Rykov, V. V.*: On lower bound of the length of B_4-code) 299

Дынькин В. Н., Кацман Г. Л., Курдюков А. П.: Нижняя граница области допустимых скоростей кодов для суммирующего канала с коррелированными источниками (*Dynkin, V. N., Katsman, G. L., Kurdyukov A. P.*: Lower bound of the region of acceptable code rates for a summary channel with correlated sources) 309

Kuznetsov, N. A., Yashin, A. I.: On consistent estimation in adaptive filtering (*Кузнецов Н. А., Яшин, А. И.*: О состоятельности оценивания параметров в схеме адаптивной фильтрации) 317

Хинкель Х.: Применение адаптивных фильтров Калмана для предсказания нагрузки в электроэнергетических системах на кратковременные интервалы (*Hinkel, H.*: An application of adaptive Kalman filtering to short-term load forecasting for power systems) 329

Kowalewski, A., Kotarski, W.: On the application of conical approximations to an optimal control problem for systems described by partial differential equations of parabolic type with time delay (*Ковалевски А., Катарски В.*: Применение конических аппроксимаций в задачах оптимального управления системами, описываемыми уравнениями с частными производными параболического типа с запаздыванием) 341

Bakule, L.: On a new hierarchical approach to the large-scale system solution (*Бакуле Л.*: О новом иерархическом подходе к решению больших систем) 353

HU ISSN 0370—2529

Index: 26 660

316.920

VOL. TOM 10 * NUMBER HOMER 6

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1981

AKADÉMIAI KIADÓ, BUDAPEST
DISTRIBUTED OUTSIDE THE COMECON-COUNTRIES
BY PERGAMON PRESS, OXFORD

PROBLEMS OF CONTROL AND INFORMATION THEORY

An international bi-monthly sponsored jointly by the Presidium of the Academy of Sciences of the USSR, of the Hungarian Academy of Sciences and of the Czechoslovak Academy of Sciences. The six issues published per year make up a volume of some 480 pp. It offers publicity for original papers and short communication of the following topics:

- theory of control processes
- general system theory
- information theory
- theory of decisions and estimates; queueing theory
- theory of adaptive system, identification, learning and pattern recognition.

While this bi-monthly is mainly a publication forum of the research results achieved in the socialist countries, also papers of international interest from other countries are welcome.

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Международный журнал Академии наук СССР, Венгерской Академии наук и Чехословацкой Академии наук выходит 6 раз в год общим объемом 480 печатных страниц.

В журнале публикуются оригинальные научные статьи и статьи обзорного характера по следующим проблемам управления и теории информации:

- теория процессов управления,
- общая теория систем,
- теория информации,
- теория принятия решений и оценивания; теория массового обслуживания,
- теория адаптивных систем, идентификации, обучения и распознавания образов.

Целью журнала является ознакомление научной общественности различных стран с важнейшими проблемами, имеющими актуальный и перспективный характер, научными достижениями ученых социалистических и других стран.

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czechoslovakia, German Democratic Republic, Korean People's Republic, Mongolia, Poland, Roumania, Viet-Nam and Yugoslavia

KULTURA Hungarian Foreign Trading Co.
P. O. Box 149 H–1389 Budapest, Hungary

For all other countries

PERGAMON PRESS LTD, Headington Hill Hall, Oxford OX3 OBW, England

or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA

1981 Subscription Rate US Dollars 110.00 per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS	РЕДАКТОРЫ ЖУРНАЛА
B. N. PETROV (Moscow)	Б. Н. ПЕТРОВ (Москва)
G. BOGNÁR (Budapest)	Г. БОГНАР (Будапешт)
COORDINATING EDITORS	ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ
S. V. EMELYANOV (Moscow)	С. В. ЕМЕЛЬЯНОВ (Москва)
N. N. KRASOVSKII (Sverdlovsk)	Н. Н. КРАСОВСКИЙ (Свердловск)
E. P. POPOV (Moscow)	Е. П. ПОПОВ (Москва)
V. S. PUGACHEV (Moscow	В. С. ПУГАЧЕВ (Москва)
V. I. SIFOROV (Moscow)	В. И. СИФОРОВ (Москва)
E. D. TERYAEV (Moscow)	Е. Д. ТЕРЯЕВ (Москва)
T. VÁMOS (Budapest)	Т. ВАМОШ (Будапешт)
L. VARGA (Budapest)	Л. ВАРГА (Будапешт)
A. PRÉKOPA (Budapest)	А. ПРЕКОПА (Будапешт)
S. CSIBI (Budapest)	Ш. ЧИБИ (Будапешт)
I. CSISZÁR (Budapest)	И. ЧИСАР (Будапешт)
J. KOCSIS (Budapest)	Я. КОЧИШ (Будапешт)
J. BENEŠ (Prague)	И. БЕНЕШ (Прага)
V. STREJC (Prague)	В. СТРЕЙЦ (Прага)



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

$$J' = \frac{\lambda H_1}{1-H_2} + \frac{\nu\alpha_1}{f(1-H_2)}\left(\sum_{k=1}^{m}(F_{k-1} - F_k z_k - \frac{H_1(2-a-H_2)}{(1-H_2+H_1)(1-o)}\right) + o(1)$$

$$\pi = 1 - \frac{(1-H_2)\nu\alpha_1}{(1-H_2+H_1)f(1-a)\lambda} + o\left(\frac{1}{\lambda}\right).$$

When $H_1 = H_2 = H$, for the probabilistic characteristics J_1 and $\nu\pi + J_1'\mu_1 + J_2'\mu_2$ one can find three terms of expansion into a series of the power of λ.

$$J_1 = \frac{\lambda H}{\mu_1(1-H)} + \frac{\nu\alpha_1}{f\mu_1(1-H)}\left(\sum_{k=1}^{m-1}(F_{k-1} - F_k)Z_k + (F_{m-1} - F_m)Z_m H - H\right) -$$

$$- \frac{\nu\alpha_1}{f^2\mu_1(1-a)}\left(\sum_{k=1}^{m=1}(F_{k-1} = F_k)Z_k + (F_{m-1} - F_m)Z_m H - H\right)\frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right);$$

$$\lambda\pi + J_1'\mu_1 + J_2'\mu_2 = \frac{\lambda}{1-H} + \frac{\nu\alpha_1}{f(1-H)}\left(\sum_{k=1}^{m}(F_{k-1} - F_k)Z_k - \frac{1-aH}{1-a}\right) -$$

$$- \frac{\nu\alpha_1^2}{f^2(1-a)}\left(\sum_{k=1}^{m}(F_{k-1} - F_k)Z_k - \frac{1-aH}{1-a}\right)\frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right).$$

Note that for approximate computation of M, Q, π_3, π_2 and π_1 with large values of λ better results are obtained by using relations (6.7) and the definition of the probabilistic characteristics without expansion into series of the power of $1/\lambda$.

Let us consider an expansion of appropriate probabilistic characteristics of the i-th IAS input. Let $\lambda_i \to \infty$, $H_2 < 1$, and the remaining model parameters be fixed. Then acting as in the case of a fully accessible bunch one can see that the form of the coefficients $I_{k,0}, J_{1,-1}, J_{1,0}, J_{2,0}, J'_{-1}, J'_0, \bar{\pi}_0$ and $\bar{\pi}_1$ in expansion (6.7) coincides with that of the associated coefficients in the expansion of similar characteristics into a series of powers of $1/\lambda_i$ for the i-th IAS input. The parameters of the i-th input have the i-th indexing and accessibility d_i is used instead of the number of lines ν.

For illustration of the accuracy of the approximate method for computation of probabilistic model characteristics with large values of intensity of the primary load, Table 2 gives the results of accurate and approximate (by formulae (6.7)) computation of certain characteristics for a fully accessible bunch with a two phase service.

The input parameters take on the following numerical values $\nu = 1$; $\mu_1 = 20$; $\mu_2 = 20$; $H_1 = 0.6$; $H_2 = 0.7$; $\alpha_1 = 2$; $\alpha_2 = 1$; $P = 0.8$; $Z = 0.6$; $a = 0.3$.

Problems of Control and Information Theory, Vol. 10 (6), pp. 363–373 (1981)

AN APPLICATION OF THE AUGMENTED LAGRANGIAN APPROACH TO THE OPTIMAL CONTROL OF A BIHARMONIC SYSTEM WITH STATE-SPACE CONSTRAINTS

J. V. OUTRATA, J. JARUŠEK, Z. BERAN

(Prague)

(Received February 10, 1981)

For the solution of certain optimal control problems with elliptic systems and inequality state-space constraints the augmented Lagrangian approach is proposed. In order to ensure the existence of a solution of the appropriate dual problem, special class of augmented Lagrangians is constructed on such Sobolev spaces for which the continuous imbedding into C_0 is guaranteed (see e.g. [7]). This approach is demonstrated on the case of a system governed by a biharmonic equation.

0. Introduction

The application of the augmented Lagrangian approach, which is very well known and exploited for the finite-dimensional mathematical programming, meets various difficulties in the infinite-dimensional case. One of the main troubles is how to ensure the existence of a Kuhn–Tucker vector. Rockafellar in [15] discussed the infinite-dimensional programming problem, but the constraint space was supposed to be Euclidian. The first attempt to treat the case of the infinite-dimensional constraint space is [18], where the range space of the inequality constraints is Hilbert. By means of a generalized convexity concept, various augmented Lagrangians were obtained for constraints even in certain Banach spaces (reflexive or non-reflexive — cf. [2, 9]). Another class of augmented Lagrangians was applied to the optimal control of parabolic systems in [14].

In our paper we shall slightly modify the developed theory for the case of a general convex set of admissible observations. In Section 2 we apply it to a biharmonic system and we shall notice the numerical aspects of biharmonic problems, especially the numerical solution of the projection onto the considered convex set in the space $\overset{\circ}{H}{}^2(\Omega)$. The advantage of our method is that the existence of a Kuhn–Tucker vector is ensured due to the Slater condition, and because of using the space $\overset{\circ}{H}{}^2(\Omega)$, the numerical method is not so difficult as it would be the case for $H^2(\Omega)$.

1. Duality theory and numerical approximation

Let U and X be two reflexive Banach spaces, let H be a Hilbert space. Let us have a convex continuous functional J defined and lower bounded on $U \times X$. Let A be a continuous operator from $U \times X$ into X^* (the dual space of X). Let the derivatives A'_x, J'_x, A'_u, J'_u exist in the Fréchet sense, suppose J'_x, A'_x to be continuous as operators on $U \times X$. Let us denote by $\chi_{\mathfrak{D}}$ the indicatory function for an arbitrary convex set $\mathfrak{D}$, let $C \subset U$, $K \subset H$ be closed convex sets. Let us denote $\hat{J} = J + \chi_C$, assume $\hat{J}$ to be coercive with respect to u independently of $x \in X$ on $U \times X$. Let q be a continuous operator from X into H such that the functional $[p, u] \to \operatorname{dist}(p - q(\Pi(u + y_0)), K)$ is convex and continuous, where $\Pi(u + y_0)$ is the solution of the problem $A(u + y_0, x) = 0$ with prescribed u, y_0 is a fixed element of U and $p \in H$ (we suppose the existence and the uniqueness of that solution). The condition of the convexity and continuity of the function $[p, u] \to \operatorname{dist}(p - q(\Pi(u + y_0)), K)$ is fulfilled, e.g. if both q and Π are continuous and linear or if Π is continuous and linear, K is a closed convex cone with the vertex in the origin and q is K-convex.

Let us look for the solution of the following problem:

$$\operatorname{Inf}\{J(u, x);\ u \in C, q(x) \in K,\ A(u + y_0, x) = 0\}. \tag{P}$$

Firstly, we define the perturbed essential objective

$$F_r(u, p) = J(u, \Pi(u + y_0)) + \chi_C(u) + \chi_K(q(\Pi(u + y_0)) - p) + r\,||p||^2_H, \quad r > 0 \tag{1.1}$$

and we obtain by the usual way the augmented Lagrangian

$$L_r(u, p^*) = -(F_r)^*_u(p^*) = \hat{\hat{J}}(u) + \inf_{v \in H}[r||q||^3_H - \langle p^*, q\rangle + \\ + \langle p^* - 2rq, v\rangle + r||v||^2_H + \chi_K(v)],$$

where the convex function

$$\hat{\hat{J}}(u) \equiv \hat{J}(u, \Pi(u + y_0)),$$

$(F_r)^*_u(p^*)$ is the conjugate function to F_r with respect to p (for fixed u), the argument of q is dropped and $q - v$ is substituted for p. Denoting $\alpha = q - \dfrac{p^*}{2r}$, we obtain

$$L_r(u, p^*) = \hat{\hat{J}}(u) + r||q||^2_H - \langle p^*, q\rangle + 2r \inf_{v \in K}\left\langle -\alpha + \frac{v}{2}, v\right\rangle.$$

The infimum is attained for $w \in K$ iff $\langle -\alpha + w, v - w\rangle \geq 0$ for all $v \in K$, thus iff $w = \alpha^K$ (the projection of α onto K). So we have

$$L_r(u, p^*) = \hat{\hat{J}}(u) + r\left\|\left(q - \frac{p^*}{2r}\right) - \left(q - \frac{p^*}{2r}\right)^K\right\|_H^2 - r\left\|\frac{p^*}{2r}\right\|_H^2. \tag{1.2}$$

The dual objective function

$$g_r(p^*) = -F_r^*(0, p^*) = \inf_{u \in H} L_r(u, p^*) = -p_r^*(p^*),$$

where

$$p_r(p) = \inf\{\hat{\hat{J}}(u) + r||p||_H^2;\ q(\Pi(u + y_0)) \in p + K\} = p_0(p) + r||p||_H^2.$$

So we have

$$g_r(p^*) = \sup_{a^* \in H}\left(g_0(a^*) - \frac{1}{4r}\|p^* - a^*\|_H^2\right) = \tag{1.3}$$

$$= -\frac{1}{2r}\inf_{a^*}\left(\frac{1}{2}\|p^* - a^*\|_H^2 - 2rg_0(a^*)\right) \equiv -\frac{1}{2r}\left(\frac{1}{2}\|p^* - \tilde{p}^*\|_H^2 - 2rg_0(\tilde{p}^*)\right),$$

where $\tilde{p}^* \equiv \operatorname{prox}_{-2rg_0}(p^*)$. The dual problem has the form

$$\sup\{g_r(p^*);\ p^* \in H\}. \tag{D}$$

Lemma 1.1. Let the mentioned suppositions be satisfied, furthermore let $q \circ \Pi$ be a continuous mapping. Then the problem (P) is normal in the Ekeland and Temam sense (cf. [4]). Moreover, if K has a non-empty interior and there exists such $u \in C$ that $\Pi(u + y_0) \in \operatorname{Int} K$, then the problem (P) is stable (a Kuhn–Tucker vector exists) for every $r \geq 0$.

The proof is well known, the normality is a straightforward consequence of the lower weak semicontinuity and the coercivity of J, the convexity and closeness of K and the properties of Π and q. The stability is based on the obvious Slater condition.

Lemma 1.2. Let φ be a proper convex lower semicontinuous functional defined on some Hilbert space $\varkappa$. Let us denote $P_\varphi(x) \equiv \operatorname{prox}_\varphi(x)$. Then for the mapping $\beta(x) = \frac{1}{2}||x - P_\varphi(x)||_\varkappa^2 + \varphi(P_\varphi(x))$ there exists the Fréchet derivative $\beta'(x) = x - P_\varphi(x)$ and it holds

$$0 \leq \beta(x + h) - \beta(x) - \beta'(x)h \leq \frac{1}{2}||h||_\varkappa^2. \tag{1.4}$$

The proof is well known. It is a straightforward consequence of the variational formulation of the definition of $P_\varphi(x)$ and the convexity of β (cf. also e.g. [11]).

Corollary. The mapping $\vartheta : w \to \frac{1}{2}\,||w - w^{\mathcal{D}}||^2_{\varkappa}$, where $w^{\mathcal{D}}$ is the projection onto a closed convex set $\mathcal{D} \subset \varkappa$, has the Fréchet derivative in the form

$$\vartheta'(w) = w - w^{\mathcal{D}}. \tag{1.5}$$

Proof: Put the indicatory function $\chi_{\mathcal{D}}$ as φ into lemma 1.2.

By the corollary of lemma 1.2 it holds

$$(L_r)'_{p^*}(u, p^*) = \left(q(\Pi(u + y_0)) - \frac{p^*}{2r}\right)^K - q(\Pi(u + y_0)). \tag{1.6}$$

Numerical approximation of the solution of the problem (P) is based on the following schema (cf. e.g. [15])

k-th step:

k.1. given $p_k^* \in H$, determine $u_k \in C$ such that $L_r(u_k, p_k^*) \leq \inf_{u \in C} L_r(u, p_k^*) + \alpha_k$;

k.2. set $p_{k+1}^* = p_k^* + 2r\left[\left(q(\Pi(u_k + y_0)) - \frac{p_k^*}{2r}\right)^K - q(\Pi(u_k + y_0))\right]$.

Under suppositions mentioned above, the schema has the following properties:

Lemma 1.3. Suppose $\sum_{k=1}^{+\infty} \sqrt{\alpha_K} < +\infty$ and let the Slater condition hold. Then there exists a Kuhn–Tucker vector p^* such that $p_k^* \to p^*$.

The proof of this lemma is well known and is based on the following propositions:

Proposition 1.3.1. It holds

$$r\,||(L_r)'_{p^*}(u_k, p_k^*) - g_r'(p_k^*)||^2 \leq \alpha_k. \tag{1.7}$$

The proof consists in using of the mentioned properties of the function g_r and lemma 1.2.

Proposition 1.3.2. $\{p_k^*\}$ is a bounded sequence.

This proposition is a straightforward consequence of the preceding proposition, if we consider that

$$p_{k+1}^* - p_k^* = 2r(L_r)'_{p^*}(u_k, p_k^*) \text{ and } p_k^* - \operatorname{prox}_{-2rg_*}(p_k^*) = -2rg_r'(p_k^*)$$

(cf. lemma 1.2).

All optimal solutions $\hat{p}^*$ of (D) (cf. lemma 1.1) are fixed points of the mapping $\text{prox}_{-2rg_*}(p_k^*)$ which is non-expansive.

Proposition 1.3.3. $\{p_k^*\}$ is a maximizing sequence of the problem (D).

For the proof the inequality

$$g_r(p_{k+1}^*) \geq g_r(p_k^*) + r\,||g_r'(p_k^*)|| - 2c\sqrt{r\alpha_k} \tag{1.8}$$

is very essential, being a consequence of the properties of g_r (c is the local Lipschitz constant of g_r). Using the concavity of g_r and the fact $g_r'(p_k^*) \to 0$, we obtain the proved assertion.

Proposition 1.3.4. $\{p_k^*\}$ is convergent.

From propositions 1.3.3 and 1.3.4 we immediately obtain lemma 1.3. To prove proposition 1.3.4 we shall verify the Bolzano–Cauchy condition. From proposition 1.3.1 and the proof of proposition 1.3.2 and using (1.8) we obtain the following inequalities for $l < k$:

$$||p_k^* - p_l^*|| \leq \sum_{i=l}^{k-1} 2r\,||g_r'(p_i^*)|| + 2\sqrt{r\alpha_i} \leq 2(g_r(p_k^*) - g_r(p_l^*)) + \\ + 2(1 + 2c)\sqrt{r}\sum_{i=l}^{k-1}\sqrt{\alpha_i}. \tag{1.9}$$

Using proposition 1.3.3 and the summability of $\sum_{i=1}^{+\infty}\sqrt{\alpha_i}$, we obtain the required result.

Lemma 1.4. $\{u_k\}$ is a bounded asymptotically minimizing sequence for (P). Every weak cluster point of $\{u_k\}$ is an optimal solution of (P).

Proof: The boundedness of $\{u_k\}$ is implied by the form of $L_r(u_k, p_k^*)$ and the coercivity of $\widehat{J}$. Because of (1.8), (1.7) and (1.6) dist $(q(\Pi(u + y_0)), K) \to 0$. From this fact and lemma 1.3 we can see that $\{u_k\}$ is asymptotically minimizing for (P). Let u_0 be a weak cluster point of $\{u_k\}$. Due to weak lower semicontinuity of the functionals $u \to \hat{\hat{J}}(u)$ and $u \to \text{dist}(p - q(\Pi(u + y_0)), K)$ and the preceding assertion of our lemma, we can see that u_0 is an optimal solution of (P).

The solution of the intermediate problem

$$\min\{L_r(u, p_k^*);\, u \in C\} \tag{1.10}$$

(resp. of finding u_k such that the condition of step k.1 of our schema will be

satisfied) is substantially simplified under the assumption that $J(\cdot, \cdot)$ is continuously Fréchet differentiable on $U \times X$. In such a case, lemma 1.5 below can be utilized for the computation of the Fréchet derivatives of the function $L_r(u, p_k^*) - \chi_c(u)$ with respect to control.

Lemma 1.5. Let X, U, A, Π be as above, let Π be locally Lipschitz in the neighborhood of a point $u \in U$. Let $\mathfrak{F}$ be a functional on $U \times X$, let $\mathfrak{F}'_x$ exist in the Fréchet sense and be continuous as the operator from $U \times X$ into X^*. Define $\mathfrak{L}(u, x, \varkappa) = \mathfrak{F}(u, x) + \langle \varkappa, A(u + y_0, x) \rangle$, where $\varkappa \in X$ is a solution of the problem

$$\mathfrak{F}'_x(u, x) + \varkappa \circ A'_x(u + y_0, x) = 0 \in X^*. \tag{1.11}$$

Let $x = \Pi(u + y_0)$, $x' = \Pi(v + y_0)$. Then it holds

$$\mathfrak{F}(v, x) - \mathfrak{F}(u, x) = \mathfrak{L}(v, x, \varkappa) - \mathfrak{L}(u, x, \varkappa) + o(||u - v||). \tag{1.12}$$

The proof based on the direct calculation and obvious manipulations using all suppositions is made in [10].

In such a way, the intermediate problem is in fact a constrained minimization on the space U which can be solved by various existing methods. For simple control constraints such as upper and lower bounds in some spaces we recommend to apply some variant of the conjugate gradient method (cf. [7]) and to treat these constraints by the "clipping off" technique described in [13].

2. Optimal control of a system described by a biharmonic equation with the state-space and control constraints

Now, we shall apply the general theory described in the previous section to the following problem: Let Ω be a bounded domain in R^2 with a sufficiently smooth boundary Γ (our method is suitable also for the case $\Omega \subset R^3$). Let us take $H = X = \mathring{H}^2(\Omega)$, $U = L_2(\Omega)$, $q = Id_X$.

Let c_1, c_2, c_3, D be positive numbers. Let $\psi \in L_\infty(\Omega)$ fulfil the condition $\psi(x) \leq -c_3$ on some neighborhood of Γ. Let Ω' be a measurable subset of Ω. We suppose that $\varphi_0 \in L_2(\Omega)$ such that $\varphi_0(x) \in [c_1, c_2]$ a.e. on Ω. Let us assume

$$J(u) = \int_\Omega \varphi_0 u \, dy; \tag{2.1}$$

$$A(u, x) = D\Delta^2 x - u, \quad y_0 \equiv f \in L_2(\Omega), \tag{2.2}$$

$$K = \{x \in \mathring{H}^2(\Omega), x \geq \psi \text{ a.e. on } \Omega'\}, \tag{2.3}$$

$$C = \{u \in L_2(\Omega);\ 0 \leq u \leq \gamma \text{ a.e. on } \Omega\};\ \gamma \in L_\infty(\Omega). \tag{2.4}$$

On the space $\dot{H}^2(\Omega)$ we consider the norm $\|\cdot\|_2$ in the form

$$\|x\|_2 = \left(\int_\Omega \sum_{i,j=1}^{2} \left(\frac{\partial^2 x}{\partial y_i \, \partial y_j} \right)^2 dy \right)^{\frac{1}{2}} = \left(\int_\Omega (\Delta x)^2 \, dy \right)^{\frac{1}{2}}. \tag{2.5}$$

This system equation describes e.g. the deflection of the floor of a warehouse. This floor must not deflect from various reasons (safety, operation etc.), deeper than some previously stated limit. Function f represents the own weight of the floor, control u is the weight distribution of the stored material and the upper limit for u is given by the height of the ceiling and the density of the stored material. Function φ_0 appearing in the objective is the price for a unit of the storage area depending on the allocation in the warehouse. The number D (the modulus of the flexural rigidity) can be considered without any loss of generality to be equal to 1. Our task is to maximize the profit of the warehouse or, if $\varphi_0 \equiv$ const., the maximal capacity of the warehouse under preserving the given state-space constraint (feasible deflection of the floor).

We can see that the spaces, the operators and the set C satisfy the requirements of Section 1. Due to the declared properties of ψ and Sobolev's imbedding theorem for $H^2(\Omega)$ dim $\Omega \leq 3$, we have Int $K \neq \emptyset$. If we suppose the existence of such $u \in C$ that the corresponding solution of the state equation belongs to Int K, we shall ensure the stability of (P). To solve the problem (P) described by means of (2.1)–(2.5), the method derived in Section 1 can be applied. For the determination of u_k in the step k.1 of the numerical schema we exploit the assertion of lemma 1.5 for $\mathcal{F}(u, x) = L_r(u, x, p^*) - \chi_C(u)$ for the fixed parameter p_k^*. From (1.11) we state that the corresponding form of the co-state equation is

$$\varkappa = p_k^* - 2rx + 2r \left(x - \frac{p_k^*}{2r} \right)^K. \tag{2.6}$$

From (2.6) and (1.12) we can see that

$$\mathcal{F}'_u(u, \Pi u) = \frac{\partial \mathfrak{L}}{\partial u}(u, x, \varkappa) = \mathcal{F}'_u(u, x) + \varkappa \circ A'_u(u + f, x) =$$

$$= \varphi_0 + 2rx - p_k^* - 2r \left(x - \frac{p_k^*}{2r} \right)^K. \tag{2.7}$$

The minimization of step k.1 can be numerically carried out e.g. by the variant of the conjugate gradient method described for the case $U = L_\infty(\Omega)$ in [13]. The control constraints are respected in such a way that we involve

two disjoint subsets of Ω at the beginning of the iteration procedure as estimations of "active constraint sets"

$$\alpha_0 = \{y \in \Omega;\, u(y) = 0\}, \quad \beta_0 = \{y \in \Omega;\, u(y) = \gamma(y)\}.$$

They are currently updated in such a manner that they approach the real "active constraint sets" for u. The construction of a conjugate direction is correspondingly modified. Thus, we must be able to compute for every control $u \in C$ the corresponding value of the objective and the Fréchet derivative given by (2.7). For the evaluation of the cost value the system equation must be solved. If, moreover, the Fréchet derivative is to be evaluated, the co-state equation (2.6) must be solved. This is, in fact, still more complicated than the solution of the system equation, because the projection onto the set K is a complicated biharmonic problem with an obstacle on the domain Ω.

So, from the practical numerical point of view, we have two main auxiliary problems. The more difficult problem, namely to find the projection onto the set K (i.e. to find such $v' \in K$ that $\|v' - s\|_2^2 \leq \|w' - s\|_2^2$ for every $w' \in K$ for some previously given s), can be described (on setting v for $v' - s$ and $w = w' - s$) in the following form: We define $K_s = \{w \in \mathring{H}^2(\Omega);\, w \geq \psi - s$ a.e. on $\Omega\}$ and look for such $v \in K_s$ that for every $w \in K_s$ the following inequality holds

$$\int_\Omega \Delta v \Delta(w - v)\, dy \geq 0. \tag{2.8}$$

Among several possibilities of the numerical solution of the problem we look upon the following two methods as suitable:

1° The equivalent formulation of our problem is

$$\min \left\{ \frac{1}{2} \int_\Omega |\Delta w|^2\, dy,\, w \in K_s \right\}. \tag{2.9}$$

This problem can be solved directly using the finite element method. For the sake of simplicity we consider rectangular Ω. Let $\{\mathcal{R}_h\}$, $h \to 0_+$, $h = (h_1, h_2)$ be a regular system of rectangulations of $\bar{\Omega}$. Let $\mathcal{N}_h$ be the set of all vertices of rectangles of $\mathcal{R}_h$. We suppose $\mathcal{N}_h \subset \mathcal{N}_{h'}$ for $h > h'$. On $\mathcal{R}_h$ we define the Ahlin elements

$$Q_3(R_k) = \left\{ \mathfrak{a}(y_1, y_2);\, \mathfrak{a}(y_1, y_2) = \sum_{i,j=0}^{3} \alpha_{ij} y_1^i y_2^j,\, [y_1, y_2] \in R_k \right\},\ R_k \subset \mathcal{R}_h.$$

By the usual way we define the appropriate approximative spaces $V_h \subset \mathring{H}^2(\Omega)$

and the corresponding approximative convex sets $K_s^h \subset V^h$. For problem (2.9) the finite-dimensional Ritz approximation has the form

$$\min\left\{\frac{1}{2}\int_\Omega |\Delta w|^2\, dy;\ w \in K_s^h\right\} \tag{2.9}_h$$

and it is a quadratic programming problem. Using the Sobolev imbedding theorem ($\mathring{H}^2(\Omega) \hookrightarrow C_0(\bar{\Omega})$) and the coercivity of the functional in (2.9), we can prove the convergence of the method (see [6]). $(2.9)_h$ can be solved exploiting the well-known SOR method (for details see [5]). Similar results can be obtained by means of Ari–Aridini's elements.

2° We have $w - \psi + s \in L_\infty(\Omega)$. Let Λ be the dual positive cone in $(L_\infty(\Omega))^*$ and $\langle\langle \cdot, \cdot \rangle\rangle$ the duality pairing between $L_\infty(\Omega)$ and $(L_\infty(\Omega))^*$. Let us define the following Lagrangian on $\mathring{H}^2(\Omega) \times (L_\infty(\Omega))^*$

$$l(w, \check{c}) = \frac{1}{2}\int_\Omega |\Delta w|^2 dy + \langle\langle \check{c}, \psi - s - w \rangle\rangle.$$

Since Int $K_s \neq \emptyset$, l has a saddle point (v, τ), where v is a solution of (2.9) (cf. [4]). To find some saddle point, we use the following schema:

$$\lambda^0 \geq 0 \text{ arbitrary} \qquad \Delta^2 w^n = \lambda^n \text{ in } \Omega \tag{2.10}$$

$$w^n = \frac{\partial w^n}{\partial n} = 0 \text{ on } \Gamma, \qquad \lambda^{n+1} = [\lambda^n + g_n(\psi - s - w^n)]_+, \quad g_n > 0,$$

where $[\check{z}]_+ : y \to \max(\check{z}(y), 0)$ for a function $\check{z}$ on Ω. Then we solve the sequence of biharmonic equations. This direct way is advantageous especially for rectangular Ω; we can apply the same elements as in 1° and we use the standard procedure.

With help of a factorization we can convert the biharmonic equation into two 2nd order elliptic problems in $\mathring{H}^1(\Omega)$ as follows:

$$\lambda^0 \geq 0 \text{ arbitrary} \qquad -\Delta m^n = \lambda^n \text{ in } \Omega, \qquad -\Delta w^n = m^n \text{ in } \Omega \tag{2.11}$$

$$w^n = m^n = 0 \text{ on } \Gamma, \qquad \lambda^{n+1} = [\lambda^n + \varrho_n(\psi - s - w^n)]_+, \ \varrho_n > 0.$$

This procedure can be applied for polygons. We use the triangulation of Ω and the elements composed from linear or quadratic polynomials on triangles. If Ω is a general domain with the Lipschitz boundary, we can approach it by a certain polygon Ω^h from its interior and use the regular triangulation of Ω^h. The convergence is proved in [5]. The order of the convergence is independent of $S^h = \Omega \setminus \Omega^h$. For the problems with the operator Δ cf. also [1], where

some combination of procedures of 1° and 2° is described, or [17], where the external approximations are used.

Of course, the second main numerical difficulty, namely the state equation, can be solved by means of the same finite elements as in 2°. For details confront [5].

3. Conclusion

Our paper has two purposes: firstly to modify the primal–dual method for solving non-linear programs of [15], [16] and [18], in order to extend its area of applicability. As the space of perturbations a certain type of Sobolev spaces was chosen in order to guarantee the existence of a solution of the appropriate dual problem. The second aim is to introduce a non-trivial application of this method; namely the optimal control problem with a biharmonic equation and state–space and control constraints having a straightforward practical interpretation. The resulting numerical problems, especially the corresponding projection, are also considered. However, these numerical operations seem to be a little bit time-consuming at present. For a further study, the analysis of different generalized Lagrangians in this context supplies an attractive research field. The numerical solution of concrete problems of this type will be the subject of an intended paper.

References

1. *Brezzi, F., Hager, W. W., Raviart, P. A.*, Error estimates for the finite element solution of variational inequalities. Num. Math., Vol. **28,** *4*, 1977, 431–444 (Part I); Vol. **31,** *1*, 1978, 1–16 (Part II).
2. *Dolecki, S., Kurcyusz, S.*, On Φ-convexity in extremal problems. SIAM J. Contr. and Optimiz., Vol. **16,** 1978, 277–300.
3. *Duvaut, G., Lions, J. L.*, Les Inéquations en Mécanique et en Physique. Dunod, Paris, 1972.
4. *Ekeland, I., Temam, R.*, Analyse Convexe et Problèmes Variationnels. Dunod, Paris, 1974.
5. *Glowinski, R., Lions, J. L., Trémolières, R.*, Analyse Numérique des Inéquations Variationelles. Dunod, Paris, 1976.
6. *Haslinger, J.*, On numerical solution of a variational inequality of the 4th order by finite element method, Apl. mat., Vol. **23,** *5*, 1978, 334–345.
7. *Koivo, H. N., Mayorga, R. V.*, A robust conjugate-gradient algorithm for optimal control problems. Proc. 7-th. Congress IFAC, Helsinki, 1978.
8. *Kufner, A., John, O., Fučík, S.*, Function Spaces. Academia, Praha, 1977.
9. *Kurcyusz, S.*, Some remarks on generalized Lagrangians. Proc. 7-th IFIP Conference on Optimization Techniques, Nice, 1975.
10. *Luenberger, D. E.*, Optimization by Vector Space Methods. J. Wiley and Sons, N. Y., 1968.
11. *Moreau, J. J.*, Proximité et dualité dans un espace Hilbertien. Bull. Soc. Math. France, **93,** 1965, 273–299.
12. *Outrata, J. V., Schindler, Z.*, An augmented Lagrangian method for a class of convex continuous optimal control problems. This journal (to appear).

13. *Pagurek, B., Woodside, C. M.*, The conjugate gradient method for optimal control problems with bounded control variables. Automatica, Vol. **4,** 1968, 337–349.
14. *Pillo, G. di, Grippo, L.*, The multiplier method for optimal control problems of parabolic systems. Rap. Istit. Automatica Univ. Roma e Centro studio sist. contr. e calcolo autom. C.N.R., nov. 1978.
15. *Rockafellar, R. T.*, A dual approach to solving nonlinear programming problems by unconstrained optimization. Math. Programming, Vol. **3,** 1973, 354–373.
16. *Rockafellar, R. T.*, The multiplier method of Hestenes and Powel applied to convex programming. J. Optimiz. Theory and Appl., Vol. **9,** 1973, 555–562.
17. *Trémolières, R.*, Inéquation variationneles; Existence, approximations, résolution. Thèse, Univ. de Paris VI, 1972.
18. *Wierzbicki, A. P., Kurcyusz, S.*, Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space. SIAM J. Contr. and Optimiz., Vol. **15,** 1977, 25–56.

Применение обобщенных функций Лагранжа для оптимального управления бигармонической системой с ограничениями на состояние

Й. В. ОУТРАТА, Й. ЯРУШЕК, З. БЕРАН

(Прага)

Для решения одного класса выпуклых задач оптимального управления с эллиптическими системами и ограничениями типа неравенств на фазовые координаты предлагается подход, основанный на применении обобщенных функций Лагранжа. Чтобы обеспечить существование решения двойственной задачи, используется специальный класс этих функций на пространствах Соболева, для которых гарантировано вложение в пространство непрерывных функций. Этот подход применен к решению задачи с системой, описанной бигармоническим уравнением.

Jiří V. Outrata
Jiří Jarušek
Zdeněk Beran
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
182 08 Praha 8, Pod vodárenskou věží 4
Czechoslovakia

Problems of Control and Information Theory, Vol. 10 (6), pp. 375—386 (1981)

MODIFIED FUZZY SETS

J. SPAL

(Košice)

(Received January 15, 1981)

The modified fuzzy sets, defined as a mapping of the values of a qualitative indicator on a system of subsets on the basic set, serve as a means of semantic and pragmatic fuzzy classification, as compared with the syntactic significance of the random variable. Their properties are fully compatible with those of the classical well-defined sets.

KEY WORDS: Fuzzy set (Zadeh);
Modified fuzzy set;
Set operation;
Random variable;
Fuzzy classification

1. Classical fuzzy sets

1.1. Introductory remark

L. A. Zadeh [4] introduced the fuzzy sets (FS) as a means of modelling the imprecision inherent in human mental activity. Zadeh considers this imprecision as a totally distinct concept from uncertainty, the latter allowing modelling by probabilistic methods, the former requiring quite different ways of description and analysis.

1.2. Formal system of Zadeh's fuzzy sets

A universal set U, consisting of elements of a quite arbitrary nature is given. A membership function (MF) $m_S(b) \in \langle 0; 1 \rangle$ over an arbitrary basic set (carrier set) (BS) $B \subseteq U$, $b \in B$, defines a subset on the BS, the values $m_S(b)$ expressing the degree of "membership" of the element b in relation to S. The value $m_S(b) = 1$ designates the total inclusion of the element b into the set S; the value $m_S(b) = 0$ excludes the element from S, intermediate values indicating "partial membership" of b in S.

As compared with the formal definition of a WDS in the sense of the classical set theory:

$$(\forall\, b \in B): \qquad m_D(b) \in \{0; 1\}, \tag{1}$$

Table 1

Abbreviations and symbols

Abbreviation	General symbol	Significance	Section
		Sets:	
US	U	Universal set	1.2
BS	B	Basic set (carrier set)	1.2
WDS	D	Well-defined (classical) set	1.2
FS	F	Fuzzy set (Zadeh)	1.2
MFS	M	Modified fuzzy set	2.1
CMFS	$\overline{M}$	Complementary fuzzy set	2.2
LS	$D(k_i)$	Level set	2.2
CLS	$\overline{D}(k_i)$	Complementary level set	2.2
		Functions:	
MF	$m(b)$	Membership function (Zadeh)	1.2
CF	$h(b)$	Criterial function	2.2
		Other concepts:	
FCL		Fuzzy classification	2.3
QI		Qualitative indicator	2.3

the generalization, allowing moreover intermediate values of the MF:

$$(\forall\, b \in B)\colon \quad m_F(b) \in \langle 0; 1\rangle \tag{2}$$

gives the formal definition of the fuzzy set (FS).

The set calculus on a system of classical WDS's D_j with their respective MF $m_j(b)$ over the BS B may be reduced to the use of three basic operations:

I. Negation, assigning to any given set D of the system its complementary set $D_c = \text{Neg}\, D$ with the MF:

$$m(D_c) = 1 - m\,[D]. \tag{3}$$

II. Union of the sets $D_k (k = 1, \ldots, n)$, defining the set $D_u = \bigcup_{k=1}^{n} D_k$ by means of the formal definition:

$$(\forall\, b_u \in B;\ m_u(b_u) = 1)\colon\ \exists\, D_k\colon\ m_k(b_u) = 1. \tag{4}$$

III. Intersection of the sets $D_k\,(k = 1, \ldots, n)$, defining the set $D_i = \bigcap_{k=1}^{n} D_k$ by means of the formal definition:

$$(\forall\, b_i \in B;\ m_i(b_i) = 1)\colon\ \forall\, D_k\colon\ m_k(b_i) = 1. \tag{5}$$

Zadeh introduces for its FS the following definitions of set operations:

I. Negation, assigning to F its complementary FS $F_c = \text{Neg}\, F$ with the MF:

$$m[F_c] = 1 - m[F]. \tag{6}$$

II. Union of the FS's $F_k\,(k = 1, \ldots, n)$, $F_u = \bigcup_{k=1}^{n} F_k$ having the MF:

$$(\forall\, b \in B)\colon\ m_u(b) = \text{Max}\,(m_1(b), \ldots, m_n(b)). \tag{7}$$

III. Intersection of the FS's $F_k\,(k = 1, \ldots, n)$, $F_i = \bigcap_{k=1}^{n} F_k$ having the MF:

$$(\forall\, b \in B)\colon\ m_i(b) = \text{Min}\,(m_1(b), \ldots, m_n(b)). \tag{8}$$

Moreover, Zadeh and other authors introduce the operation of scaling (standardization), which consists in the modification of the original MF $m(b)$ to the scaled MF $m_{sc}(b)$ such that:

$$(\exists\, b \in B)\colon\quad m_{sc}(b) = 1 \tag{9}$$

which is obtained by multiplying the original MF by a suitable factor a:

$$(\forall\, b \in B\colon\ m_{sc}(b) = a \cdot m(b). \tag{10}$$

Such a formal system, in an appropriately developed form, has been tried in various branches of science and technology [2], in decision-making [1], etc. Zadeh [5] has developed the methodology by establishing the theory of the linguistic variable in order to adapt the FS to the analysis of the semantic and pragmatic information.

1.3. Discussion

1.3.1. Mathematical aspects

In order to be a self-consistent generalization of the classical set theory the formal system of the FS's has to meet two conditions:

1. Results identical with those of the classical set theory must be obtained by the application of the definitions of the FS to the special case of the classical WDS.

2. The basic properties of the classical set operations should be fulfilled even for intermediate values of the MF, i.e. others than 0 and 1; such properties are:

A) Complementarity

$$F \cup (\text{Neg } F) = B \tag{11}$$

where B, considered as a FS, is defined by the MF:

$$(\forall\, b \in B): \qquad m_B(b) = 1. \tag{12}$$

B) Mutual exclusion of the FS and its negation:

$$F \cap (\text{Neg } F) = \emptyset$$

the empty fuzzy set $\emptyset$ having the MF:

$$(\forall b \in B): \qquad m_{\emptyset}(b) = 0. \tag{14}$$

C) De Morgan's laws:

$$\text{Neg } (F_1 \cup F_2) = (\text{Neg } F_1) \cap (\text{Neg } F_2) \tag{15}$$

$$\text{Neg } (F_1 \cap F_2) = (\text{Neg } F_1) \cup (\text{Neg } F_2). \tag{16}$$

D) Idempotency

$$F \cup F = F \tag{17}$$

$$F \cap F = F. \tag{18}$$

Definitions (6), (7), (8) meet the first condition, but, for general values of the MF, do not fulfill the requirements of complementarity (11) and of mutual exclusion (13).

The situation is still worse with other definitions that have been tried, such as:

$$m[F_1 \cup F_2] = m[F_1] + m[F_2] - m[F_1] \cdot m[F_2] \tag{7a}$$

$$m[F_1 \cap F_2] = m[F_1] \cdot m[F_2]. \tag{8a}$$

These definitions fail moreover in the case of idempotency.

Thus neither Zadeh's system of the fuzzy-set operations (6), (7), (8) nor any other proposed alternatives known to the author form a self-consistent system of a fuzzy-set operations, completely compatible with the classical theory.

1.3.2. Conceptual aspects

The fundamental problem is that of distinction between "imprecision" and "uncertainty", which is claimed by Zadeh, but never has been formulated in an unambiguous way. Another side of the same problem is the mutual relation between the probabilistic concept of the random variable and the new concept of the fuzzy set (FS). Many authors got under strong influence of for-

mer probabilistic ideas and are treating the FS's as a kind of substitute for probabilistic formulations, especially in the sense of the subjective probability.

Zadeh gives a different interpretation by introducing the linguistic variable with a strong emphasis on the semantic and pragmatic aspect of the matter.

Probabilistic methods are preferably concerned with the syntactic aspects, expressing the inner structure of the systems, namely in the sense of the relative frequency of events. Fuzzy sets are aimed at the study of semantic and pragmatic relations, arising within the basic set, namely in connection with the problems of classification.

2. Modified fuzzy sets

2.1. Fundamental considerations

The modified fuzzy sets use a different system of definitions. They are introduced to obtain a self-consistent system, fully compatible with the classical set theory, moreover being better adapted to the semantic and pragmatic problems of classification.

2.2. Formal system of the modified fuzzy sets

Let us define a progression (scale of quality) K of real numbers k_i on an arbitrary interval $Q = \langle \inf k; \sup k)$, which is ordered according to the following rule:

$$(m < n): \qquad k_m < k_n. \tag{19}$$

Let us further define on the basic set B a system of WDS's $D(k_i)$ associated to all values k_i, this system fulfilling the following condition of strong inclusion:

$$(m < n): \qquad D(k_m) \supset D(k_n). \tag{20}$$

The set $D(k_i)$ will be called the level set (LS) at the level k_i.

To any $D(k_i)$ there exists a set $\bar{D}(k_i)$ complementary in respect of B, for which evidently the following strong inclusion holds:

$$(m < n): \qquad \bar{D}(k_m) \subset \bar{D}(k_n). \tag{21}$$

The set $\bar{D}(k_i)$ will be called the complementary level set (CLS) at the level k_i.

These set systems are still complemented by definitory relations:

$$D(\inf k) = B \qquad \bar{D}(\inf k) = \emptyset \tag{22}$$

$$D(\sup k) = \emptyset \qquad \bar{D}(\sup k) = B. \tag{23}$$

The progression K and the BS B may be finite, countable or uncountable, with the limitation that the power of the progression must not exceed that of the set B.

For any $b \in B$ there exists exactly one subscript $i(b)$ such that:

$$b \in D(k_{i(b)}) \qquad b \in \bar{D}(k_{i(b)+1}). \tag{24}$$

Thus exactly one value $k_i(b)$ is assigned to any b. This defines over B a criterial function (CF) $h(b)$ as a substitute for Zadeh's membership function (MF). The formal difference is that the CF may get values from the arbitrarily chosen interval Q, while the values of Zadeh's MF are limited to the interval $\langle 0; 1 \rangle$.

The progression K induces two systems of subsets on B: one system of LS $\mathbb{S}(D)$, and another of CLS, $\mathbb{S}(\bar{D})$. Each of these mappings is a bijection, attributing one subset in each of the systems to any value k_i. Moreover, to any value of the parameter $q \in Q$ the next higher value may be attributed:

$$k_q(q) = \min (k_i \geq q), \tag{25}$$

thus constituting a mapping between any value of the interval Q and a set of the system $\mathbb{S}(D)$ or $\mathbb{S}(\bar{D})$. Of course this mapping is generally no more a bijection.

Let us call the mapping from the progression K onto the set system $\mathbb{S}(D)$ the modified fuzzy set (MFS) M:

$$M: \qquad K \to \mathbb{S}(D) \tag{26}$$

and the mapping from the progression K onto the set system $\mathbb{S}(\bar{D})$ the complementary modified fuzzy set (CMFS) $\bar{M}$:

$$M: \qquad K \to \mathbb{S}(\bar{D}). \tag{27}$$

All properties, existing in the set systems $\mathbb{S}(D)$ and $\mathbb{S}(\bar{D})$, irrespectively of the level value k_i, are considered as the properties of the MFS, ev. CFMS themselves. Using this definition we can define on the BS:

1. The basic modified fuzzy set M_B with level sets equal to B at any level k_i.
2. The empty modified fuzzy set $M_\emptyset$ with empty level sets at any level k_i.

All LS's and CLS's are classical WDS's, obeying all general laws of the classical set theory (de Morgan's laws, idempotency, complementarity, mutual

exclusion of LS and its respective CLS etc.). Thus all laws of the classical set theory remain valid for the MFS's and CMFS's, too.

The set operations are defined for the MFS's as follows:

I. Negation; the MFS M and its negation $\overline{M} = \text{Neg } M$ have a common CF $h(b)$, differing only by the way they are derived from this CF.

II. Union of MFS's $M_u = \bigcup_{k=1}^{n} M_k$ and intersection of CMFS's $\overline{M}_i = \bigcap_{k=1}^{n} \overline{M}_k$ have the same CF:

$$(\forall\, b \in B): \qquad h_u(b) = h_i(b) = \text{Max } (\overline{h_1(b)}, \ldots, h_n(b)). \tag{28}$$

III. Intersection of MFS's $M_i = \bigcap_{k=1}^{n} M_k$ and union of CNFS's $\overline{M}_u = \bigcup_{k=1}^{n} \overline{M}_k$ have the same CF:

$$(\forall\, b \in B): \qquad h_i(b) = \overline{h_u(b)} = \text{Min } (h_1(b), \ldots, h_n(b)). \tag{29}$$

2.2. Conceptual and methodological aspects

There are two ways of defining a set:

1. Enumeration of its elements; this leaves little room for fuzziness and further this way is not considered.

2. Definition of characteristic properties of its elements; these properties are generally expressed by logical (dichotomic) or numerical values of a variable, called the qualitative indicator (QI).

Sets defined by the latter method can be "open", i.e. accessible to alterations as to the extent, as well as to the inner structure by adding or deleting elements or by varying the properties of the subsistent elements.

The inclusion or exclusion of an element of the classical WDS is the result of a logical (dichotomic) decision. The element is either accepted or rejected according to the respective logical value of the corresponding QI. The manifestation is the process of logical (dichotomic) classification.

If there exists a set of possible (admissible) values of the CF $h(b)$, then any of these values may be interpreted as a level of the QI. To any of these levels corresponds a quite definite partitioning of the BS into a complementary pair of LS and CLS. The whole system of such partitionings induced by the given CF $h(b)$ together with the set of possible levels describes the respective process of FCL.

The FCL can be based both on the MFS and on Zadeh's FS, but there is a difference between these alternatives.

In the MFS the QI, expressed by the CF $h(b)$, retains its significance in respect of the values, as well as of their physical interpretation.

In Zadeh's FS these values are reduced to the interval $\langle 0; 1 \rangle$. This is connected with difficulties of two kinds:

1. Any alteration accompanied by a change of the interval of the QI requires the rescaling over the whole BS.

2. The real physical sense of the value of the MF is wiped out by this rescaling; hence arise the well-known difficulties as to the determination and interpretation of the MF in Zadeh's FS (compare [1]).

The process of classification (logical or fuzzy) is accomplished in two steps:

1. The preparatory phase of obtaining the CF in the way of analysing the properties of the BS from the point of view of the QI.

2. The implementary phase, i.e. the proper procedure of classification. At any level corresponding to the respective value of the CF $h(b)$ (ev. of the MF $m(b)$) an exactly defined partitioning of the BS into the pair of mutually complementary subsets LS and CLS takes place.

2.4. Example

The following example illustrates the preceding concepts and definitions.

Two BS's are defined on the universal set of natural numbers U: the set B_1 of arbitrarily chosen numbers:

$$B_1 = \{88, 23, 13, 49, 64, 4, 96, 60, 19, 72, 43, 12, 48, 75\}$$

and the set B_2, complemented in respect of B_1 by the element "1".

The value of the CF indicates for any element its number of divisors (including the divisor "1" and the respective element itself).

Table 2 contains for both sets the values of the CF $h(b)$, as well as of Zadeh's MF $m(b)$. Table 3 specifies the LS's at the respective levels. The dependence of the values $m(b)$ on the definition of the BS is evident. On the contrary, the values $h(b)$ are independent of the respective BS and preserve for the values of the CF the meaning of the number of divisors.

Examples of the classical WDS's can be demonstrated on the respective BS's too. For example sets of prime numbers on the BS can be expressed in the form:

$$h(b) \leqq 2. \tag{30}$$

A similar simple expression cannot be given for the subset of prime numbers by means of Zadeh's MF $m(b)$.

Table 2

Criterial function (number of divisors) and Zadeh's membership function, defined on two basic sets B_1, B_2

b	Divisors	Basic set B_1		Basic set B_2	
		$h(b)$	$m(b)$	$h(b)$	$m(b)$
88	1, 2, 4, 8, 11, 22, 44, 88	8	0.6	8	0.636
23	1, 23	2	0	2	0.091
13	1, 13	2	0	2	0.091
1	1	missing		1	0
49	1, 7, 49	3	0.1	3	0.182
64	1, 2, 4, 8, 16, 32, 64	7	0.5	7	0.545
4	1, 2, 4	3	0.1	3	0.182
96	1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 96	12	1	12	1
50	1, 2, 5, 10, 25, 50	6	0.4	6	0.455
19	1, 19	2	0	2	0.091
72	1, 2, 3, 4, 6, 8, 9, 12, 18, 24, 36, 72	12	1	12	1
43	1, 43	2	0	2	0.091
12	1, 2, 3, 4, 6, 12	6	0.4	6	0.455
48	1, 2, 3, 4, 6, 8, 12, 16, 24, 48	10	0.8	10	0.818
75	1, 3, 5, 15, 25, 75	6	0.4	6	0.455

2.5. Modified fuzzy sets and random variables

Let us examine in more detail the mutual relation of FS's and random variables, which has been outlined in Section 1.3.2, starting with the conceptual point of view:

The random variable represents a description of a stochastic process by the indication of the probability of the occurrence of the respective element or group of elements of the BS. These elements or groups are determined in forward.

Table 3

Level sets defined on the basic sets B_1, B_2

Value k_i	Level set	Symbol of the level set	
		Basic set B_1	Basic set B_2
13	Empty set	sup k	sup k
12	96, 72	k_6	k_7
10	96, 72, 48	k_5	k_6
8	88, 96, 72, 48	k_4	k_5
7	88, 64, 96, 72, 48	k_3	k_4
6	88, 64, 96, 50, 72, 12, 48, 75	k_2	k_3
3	88, 49, 64, 4, 96, 50, 72, 12, 48, 75	k_1	k_2
2	88, 23, 13, 49, 64, 4, 96, 50, 19, 72, 43, 12, 48, 75	inf k	k_1
1	88, 23, 13, 1, 49, 64, 4, 96, 50, 19, 72, 43, 12, 48, 75	Out of the interval	inf k

The underlined numbers indicate new accretions at the respectivel level

In contradiction to this descriptive character of the random variable, the FS and MFS have a specifically normative purpose as a methodic means of classification.

From the formal side there exists a close analogy between the random variable and the preparatory phase of classification, both starting from a predetermined object or set of objects and attributing to this object a certain information. This information is descriptive for the random variable and sometimes may be normative for the purposes of classification.

The implementary phase of classification always proceeds in the opposite sense, being directed from the selected value of the QI towards the LS as a subset of the BS.

Both the random variable and the preparatory (descriptive) phase of classification represent mappings, directed from a set system on the BS towards an interval of real numbers, eventually inducing (as in the case of the random variable) a morphism from the set operations on the set system on the BS to the arithmetic (algebraic) operations on the interval of real numbers.

The MFS, on the contrary, is a mapping, directed from a set of values of the QI towards a set system on the BS.

No inverse mapping exists in either case. No single subset is determined by a fixed value of probability, neither any single value of the QI is defined by the choice of a subset on the BS, except the special case when incidentally the chosen subset happens to be element of the system of the LS's.

3. Conclusion

The new theory of modified fuzzy sets (MFS) defines the fuzzy set, as well as its complement, as a mapping from the space of values of the qualitative indicator (QI) onto a system of subsets of the basic set (relations (26), (27)). Complete compatibility with the classical set theory is obtained in this way. Zadeh's membership function with values from the interval $\langle 0; 1 \rangle$ is substituted by the criterial function, admitting arbitrary intervals of values and enabling to preserve the original significance of the values of the qualitative indicator.

While the random variable serves preferably the purposes of description, expressing the syntactic aspects of the structure of the basic set, the modified fuzzy set is applicable for normative purposes.

The main field of application of Zadeh's fuzzy sets, as well as of the modified fuzzy sets, are problems of classification. The classical well-defined sets are means of dichotomic (logical) classification. The fuzzy sets open the way towards the fuzzy classification, which is substantially more effective and is well suited for the problems of multicriterial decisions. The problems of fuzzy classification have been dealt with in a separate paper [3].

The basic ideas and the relation of the modified fuzzy sets to Zadeh's fuzzy sets are illustrated by an example.

References

1. *Freeling, A. N. S.*, Fuzzy sets and decision analysis. IEEE-Trans., Vol. **SMC-10,** No. *7* (1980), pp. 341–354.
2. *Peschel, M.*, Modellbildung für Signale und Systeme (in German). VEB-Verlag Technik, Berlin 1978.
3. *Spal, J.*, Neostrá klasifikáčia (Fuzzy classification) (in Slovak). Informačné systémy, Bratislava (Czechoslovakia), Vol. **9,** No. *5* (1980), pp. 413–429.
4. *Zadeh, L. A.*, Fuzzy sets — Information and Control, Vol. **8** (1965), pp. 338–353.
5. *Zadeh, L. A.*, Outline of a new approach to the analysis of complex systems and decision processes. IEEE-Trans., Vol. **SMC-3,** No. *1,* pp. 28–44.

Модифицированные расплывчатые множества

Й. СПАЛ

(Кошице)

Новая теория модифицированных расплывчатых множеств определяет расплывчатое множество и его дополнительное множество в виде отображения из области значений качественного индикатора на систему подмножеств основного множества (формулы (26), (27)). Так достигается полная совместимость с классической теорией строго определенных множеств. Функция принадлежности по Заде с интервалом значений $\langle 0; 1 \rangle$ замещается критериальной функцией, допускающей произвольный интервал значений, предоставляющей возможность сохранения первоначального физического смысла значений качественного индикатора.

Обсуждается соотношение между расплывчатыми множествами и случайными переменными. Случайная переменная имеет преимущественно описательный характер, выражая синтаксические качества структуры основного множества. Расплывчатое множество служит для нормативных целей.

Основным назначением расплывчатых множеств по Заде и модифицированных расплывчатых множеств, являются проблемы классификации. Классические строго определенные множества являются исходным пунктом строгой дихотомической (логической) классификации. Расплывчатые множества дают возможность расплывчатой классификации, которая существенно эффективное, потому что открывает путь для многокритериальных решений. Вопросам расплывчатой классификации посвящена [3].

Основные понятия модифицированных расплывчатых множеств и их отношения к расплывчатым множествам по Заде иллюстрированы примерами.

Prof. Ing. Jindřich Spal C.Sc.
Technical University
Švermova 9
040 01 Košice
Czechoslovakia

Problems of Control and Information Theory, Vol. 10 (6), pp. 387—401 (1981)

СВОЙСТВА ВЕРОЯТНОСТНЫХ ХАРАКТЕРИСТИК НЕПОЛНОДОСТУПНОЙ СИСТЕМЫ С МНОГОФАЗНЫМ ОБСЛУЖИВАНИЕМ И НЕСКОЛЬКИМИ ТИПАМИ ПОВТОРНЫХ ВЫЗОВОВ

С. Н. СТЕПАНОВ

(Москва)

(Поступила в редакцию 12 августа 1980 г.)

Для неполнодоступной системы с многофазным обслуживанием и несколькими типами повторных вызовов выведены соотношения, связывающие основные вероятностные характеристики модели и получены формулы для приближенного вычисления вероятностных характеристик при большой и малой нагрузке.

1. Введение

Одним из наиболее известных проявлений действия пользователя на систему обслуживания являются повторные попытки установить соединение после получения отказа в обслуживании. Теоретическое исследование моделей информационных систем с учетом эффекта повторных вызовов является актуальной научно-технической задачей, решение которой необходимо для более адекватного описания функционирования сложных информационно-вычислительных систем. Актуальность этой проблемы находит свое отражение в постоянно возрастающем числе публикаций, посвященных этой теме.

Расчет вероятностных характеристик для систем с повторными вызовами очень трудоемкая задача вследствие большой размерности случайных процессов, описывающих функционирование модели. Точные результаты получены для небольшого класса систем при простейших предположениях о ее структуре и поведении абонента. В связи с этим большое значение при исследовании моделей с повторными вызовами приобретают методы получения качественных результатов о работе системы, основанные на асимптотическом разложении вероятностных характеристик модели при малой и большой нагрузке и методы установления соотношений между вероятностными характеристиками, которые бы не зависели от структуры исследуемой модели и схемы поведения абонента. Наличие таких методов дает возможность изучать модели информационных систем при более реальных предположениях о поведении абонента и структуре системы, что имеет большое значение для проектирования и эксплуатации сложных коммутационных узлов. В статье на примере достаточно сложной модели неполнодоступной системы (НС) с мно-

гофазным обслуживанием и несколькими типами повторных вызовов рассмотрен один из таких методов, основанный на использовании законов сохранения.

2. Схема функционирования модели

На НС поступает n взаимно независимых пуассоновских потоков первичных вызовов с интенсивностями λ_i, $i = 1, 2 \ldots, n$ и n соответствующих потоков повторных вызовов. Вызовам i-го потока доступно d_i линий НС. Поступивший вызов занимает первую в порядке доступности свободную линию на случайное время, распределенное экспоненциально с параметром $\alpha_{i,1}$. Назовем это время первой фазой обслуживания. Будем предполагать, что для вызова из i-го потока поступающей нагрузки полное время обслуживания состоит из $(m_i + 1)$-й экспоненциально распределенной фазы. Первые m_i фаз обслуживания соответствуют времени установления соединения в НС, последняя $(m_i + 1)$-я фаза обслуживания — время разговора. После окончания k-й ($k = 1, 2, \ldots, m_i - 1$) фазы обслуживания вызова из i-го потока с вероятностью $P_{i,k}$ наступает $(k + 1)$-я фаза обслуживания (длительность ее распределена экспоненциально с параметром $\alpha_{i,k+1}$); с вероятностью $(1 - P_{i,k})$ $(1 - Z_{i,k})$ абонент покидает систему, и с вероятностью $(1 - P_{i,k})\, Z_{i,k}$ абонент повторит попытку к соединению ($j_{i,1}$ — случайное число абонентов i-го потока нагрузки, повторяющих попытку соединения из-за потерь в процессе установления соединения, увеличивается на единицу). С вероятностью P_{i,m_i} после окончания m_i-й фазы обслуживания (последней фазы установления соединения) линия занимается на время разговора, распределенное экспоненциально с параметром α_{i,m_i+1}, с вероятностью $(1 - P_{i,m_i}) \times (1 - Z_{i,m_i})$ абонент покинет систему, с вероятностью $(1 - P_{i,m_i})\, Z_{i.m_i}$ абонент повторит попытку к соединению ($j_{i,2}$ — случайное число абонентов i-го потока нагрузки, повторяющих попытку соединения из-за неответа вызываемого абонента, увеличивается на единицу).

Интенсивность потока повторных вызовов, поступающих на i-й вход равна $j_{i,1}\, \mu_{i,1} + j_{i,2}\, \mu_{i,2}$, где $\mu_{i,1}$, $\mu_{i,2}$ соответственно параметры экспоненциально распределенного времени между последовательными повторными вызовами абонента из очереди $j_{i,1}$ и $j_{i,2}$. Если в момент поступления вызова все доступные линии заняты, то с вероятностью $H_{i,1}$ (если это был первый отказ в обслуживании) абонент повторит попытку к соединению ($j_{i,1}$ в этом случае увеличивается на единицу) и с вероятностью $(1 - H_{i,1})$ абонент покинет систему. Когда отказ получен в повторной попытке установить соединение, с вероятностью $H_{i,2}$ абонент предпримет следующую попытку соединения (в этом случае состояние системы не меняется, если повторный вызов поступил

из очереди $j_{i,1}$; если вызов поступил из очереди $j_{i,2}$, то $j_{i,1}$ увеличивается на единицу, а $j_{i,2}$ уменьшается на единицу) и с вероятностью $(1 - H_{i,2})$ абонент покидает систему ($j_{i,1}$ или $j_{i,2}$ соответственно уменьшаются на единицу). Обычно $H_{i,1}$ и $H_{i,2}$ называют функцией настойчивости абонента из i-го потока первичной нагрузки. В реальных системах, прежде чем попасть на обслуживание, вызов проходит несколько этапов соединения, на каждом из которых он может получить отказ. Чтобы учесть эти потери будем предполагать, что первичный или повторный вызов, поступающий на i-й вход НС, может с вероятностью a_i получить отказ в обслуживании до попадания на i-ый вход НС. В этом случае считается, что абонент ведет себя в соответствии с функцией настойчивости.

Чтобы ограничить число неизвестных в системе уравнений статистического равновесия, будем предполагать, что $j_{i,1} \leq N_{i,1}$, $j_{i,2} \leq N_{i,2}$ таким образом, если первичный или повторный вызов должен по схеме функционирования модели стать в очередь, а соответствующая очередь заполнена, он с вероятностью, равной единице, теряется. Здесь $N_{i,1}$ и $N_{i,2}$ — некоторые целые числа, $i = 1, \ldots, n$.

Обозначим через M общее число линий в НС. Занумеруем их произвольным образом. Состояние НС определим матрицей занятости x размерности $(2 \times M)$. Значение $x(1, k)$ равно l, если k-ая линия занята на l-ую фазу обслуживания и равно 0, если k-я линия свободна, $k = 1, \ldots, M$. Когда $x(1, k)$ отлично от 0, то $x(2, k)$ будет равняться номеру потока, чей вызов обслуживается k-й линией НС. Обозначим через S множество всех возможных состояний НС; X_i — множество состояний НС, когда все d_i линий, доступных i-у потоку, заняты.

Для i-го потока первичной нагрузки обозначим через $j_{i,1}(t)$ и $j_{i,2}(t)$ число абонентов, производящих повторные вызовы в момент времени t соответственно из-за отказов в процессе установления соединения и из-за неответа вызываемого абонента, $x(t)$ — состояние НС во время t. Функционирование введенной модели описывается марковским процессом

$$\xi(t) = \{j_{1,1}(t), j_{1,2}(t), \ldots, j_{n,1}(t), j_{n,2}(t), x(t)\}.$$

При условии выполнения свойства эргодичности для $\xi(t)$ можно выписать систему уравнений статистического равновесия. Обозначим через $p_{j,x}$ вероятность стационарных состояний модели. Здесь j — вектор состояний очередей $j = \{j_{1,1}, j_{1,2}, \ldots, j_{n,1}, j_{n,2}\}$. Уравнения статистического равновесия получаются достаточно просто, но имеют громоздкий вид, поэтому приводить их не будем.

3. Определение вероятностных характеристик. Их эквивалентность

Введем обозначения

$$\pi^i = \sum_j \sum_{x \in X_i} p_{j,x}, \quad I_{i,k} = \sum_j \sum_{x \in S} p_{j,x}\, l_{x,k,i},$$

$$J_{i,1}, = \sum_j \sum_{x \in S} p_{j,x}\, j_{i,1}, \quad J_{i,2} = \sum_j \sum_{x \in S} p_{j,x}\, j_{i,2}, \tag{3.1}$$

$$J'_{i,1} = \sum_j \sum_{x \in X_i} p_{j,x}\, j_{i,1}, \quad J'_{i,2} = \sum_j \sum_{x \in X_i} p_{j,x}\, j_{i,2},$$

где $\sum_j$ обозначает суммирование по всем $j_{i,k}$, меняющимся от 0 до $N_{i,k}$, $i = 1, \ldots, n$, $k = 1, 2$, а $l_{x,k,i}$ — число линий НС, занятых в состоянии x на k-ую фазу обслуживания вызовов из i-го потока.

Поскольку все дальнейшие рассуждения будут касаться свойств и поведения вероятностных характеристик i-го потока вызовов, $i = 1, \ldots, n$, то для сокращения записи формул в обозначениях для вероятностных характеристик i-го потока вызовов опустим индекс i. Таким образом вместо $H_{i,1}$, $P_{i,1}$, $\mu_{i,1}$, $J_{i,1}$, ... будем писать H_1, P_1, μ_1, J_1, ...

При исследовании систем с повторными попытками представляют интерес следующие вероятностные характеристики, связанные с обслуживанием i-го потока вызовов:

— вероятность того, что первичный вызов получит отказ в обслуживании

$$\pi_0 = a + (1 - a)\pi$$

— вероятность того, что повторный вызов получит отказ в обслуживании

$$\pi_1 = (J'_1\mu_1 + J'_2\mu_2)(1 - a)/(J_1\mu_1 + J_2\mu_2) + a$$

— вероятность того, что первичный или повторный вызов получит отказ в обслуживании

$$\pi_2 = (\lambda\pi + J'_1\mu_1 + J'_2\mu_2)(1 - a)/(\lambda + J_1\mu_1 + J_2\mu_2) + a$$

— общая вероятность того, что попытка вызова окончится неудачно (разговор не состоится)

$$\pi_3 = \Big((\lambda\pi + J'_1\mu_1 + J'_2\mu_2)(1 - a) + \sum_{k=1}^{m} I_k(1 - P_k)\,\alpha_k\Big)/(\lambda + J_1\mu_1 + J_2\mu_2) + a$$

— среднее число линий, занятых на k-ю фазу обслуживания вызовов i-го потока I_k

— среднее число абонентов, повторяющих из-за потерь в процессе установления соединения J_1 и из-за неответа вызываемого абонента J_2

— среднее число повторных вызовов на один первичный

$$M = (J_1\mu_1 + J_2\mu_2)/\lambda \tag{3.2}$$

— среднее число отказов на одно установленное соединение

$$Q = ((\lambda\pi + J_1'\mu_1 + J_2'\mu_2)(1-a) + (\lambda + J_1\mu_1 + J_2\mu_2)a + \sum_{k=1}^{m} I_k(1-P_k)\alpha_k)/I_{m+1}\alpha_{m+1}$$

— вероятность того, что требование покинет систему, не дождавшись обслуживания

$$\begin{aligned} D = {} & ((\lambda(1-H_1) + J_1\mu_1(1-H_2) + J_2\mu_2(1-H_2))\,a + \\ & + (\lambda\pi(1-H_1) + J_1'\mu_1(1-H_2) + J_2'\mu_2(1-H_2))\,(1-a) + \\ & + \sum_{k=1}^{m} I_k(1-P_k)\,\alpha_k(1-z_k))/\lambda\,. \end{aligned}$$

Рассмотрим несколько других эквивалентных определений вероятностных характеристик модели НС, использование которых в некоторых случаях облегчает их вычисление. Возьмем, например, такую характеристику i-го входа как среднее число повторных вызовов на один первичный (M в определении (3.2)). Для i-го потока нагрузки обозначим через $a(t)$ общее число повторных вызовов, а через $b(t)$ — общее число первичных вызовов, поступивших в полуинтервале $(0, t)$. Верна следующая

Лемма 1. Почти наверное

$$\lim_{t\to\infty} a(t)/b(t) \tag{3.3}$$

существует и равен $(J_1\mu_1 + J_2\mu_2)/\lambda$.

Утверждение леммы доказывается аналогично доказательству эквивалентности двух определений вероятности блокировки для НС с потерями (теорема 8.2, гл. 8 [1]).

Дадим другое эквивалентное определение M. Обозначим через $b_{j,x}(t)$ число первичных вызовов, поступивших в состояние (j, x) на i-ый вход НС в полуинтервале $(0, t)$, через $v_{j,x}(t)$ — число повторных вызовов, сделанных в $(0, t]$ k-ым в порядке поступления в состояние (j, x) первичным вызовом $k = 1, 2, \ldots, b_{j,x}(t)$. Поскольку интервалы времени между последовательными попаданиями процесса $\xi(t)$ в состояние (j, x) независимы, то при $t \to \infty$, $v_{j,x}^k(t)$ при различных k будут асимптотически одинаково распределенными независимыми случайными величинами со средним $v_{j,x}$.

Лемма 2. Почти наверное

$$\lim_{t\to\infty} \frac{\nu_{j,x}^1(t) + \ldots + \nu_{j,x}^{b_{j,x}(t)}(t)}{b(t)} \tag{3.4}$$

существует и равен $\nu_{j,x}\, p_{j,x}$.

Доказательство. Результат леммы доказывается аналогично доказательству леммы 1.

Таким образом,

$$M = (J_1\mu_1 + J_2\mu_2)/\lambda = \lim_{t\to\infty} a(t)/b(t) =$$

$$= \lim_{t\to\infty} \sum_j \sum_{x\in S} (\nu_{j,x}^1(t) + \ldots + \nu_{j,x}^{b_{j,x}(t)}(t))/b(t) = \sum_j \sum_{x\in S} \nu_{j,x}\, p_{j,x}.$$

Каждое из определений (3.2), (3.3), (3.4) дает возможность решать задачу о вычислении вероятностных характеристик систем с повторными вызовами при различных входных условиях. Если можно рассчитать вероятности стационарных состояний случайного процесса, описывающего функционирование модели, то для вычисления вероятностных характеристик используются определения (3.2), при статистическом моделировании или измерениях используются определения (3.3). Иногда (в частности, при малых или больших нагрузках) для систем с повторными вызовами можно получить приближенные формулы для вычисления некоторых вероятностных характеристик, рассматривая судьбу одного поступившего вызова, используя определение (3.4).

4. Законы сохранения. Формулы, связывающие основные вероятностные характеристики модели

Получим соотношения, связывающие основные вероятностные характеристики модели. С учетом (3.1) справедлива

Лемма 3

$$I_k\,\alpha_k P_k = I_{k+1}\,\alpha_{k+1}, \quad k = 1, \ldots, m; \tag{4.1}$$

$$J_1\mu_1 = (\lambda\pi\, H_1 + J_1'\mu_1 H_2 + J_2'\mu_2 H_2)(1-a) +$$

$$+ (\lambda H_1 + J_1\mu_1 H_2 + J_2\mu_2 H_2)\,a + \sum_{k=1}^{m-1} I_k(1-P_k)\,\alpha_k Z_k - \gamma_1, \tag{4.2}$$

$$J_2\mu_2 = I_m(1-P_m)\,\alpha_m Z_m - \gamma_2, \tag{4.3}$$

$$(\lambda(1-\pi)+(J_1-J_1')\mu_1+(J_2-J_2')\mu_2)\ (1-a)=I_1\alpha_1, \tag{4.4}$$

где

$$\gamma_1=\sum_{j,j_1=N_1}\Big(\sum_{x\in X_i}(\lambda H_1+j_2\mu_2H_2)(1-a)p_{j,x}+$$

$$+\sum_{x\in S}\Big((\lambda H_1+j_2\mu_2H_2)a+\sum_{k=1}^{m-1}l_{x,k}\alpha_k(1-P_k)Z_k\Big)p_{j,x}\Big),$$

$$\gamma_2=\sum_{j,j_2=N_2}\sum_{x\in S}l_{x,m}\alpha_m(1-P_m)Z_m\,p_{j,x}.$$

Доказательство. Результат леммы доказывается способом, использованным в [2]. Можно показать, что когда $\lambda\to 0$, а N_1, N_2 фиксированы, то γ_1, γ_2 ведут себя соответственно как $0(\lambda^{N_1})$ и $0(\lambda^{N_2})$, и $\gamma_1=\gamma_2=0$, когда $N_1=N_2=\infty$. Интуитивно понятно также, что при фиксированных входных параметрах модели γ_1, γ_2 стремятся к нулю при $N_1\to\infty$, $N_2\to\infty$.

Полученные соотношения носят характер законов сохранения и имеют важное значение при исследовании сложных моделей систем с повторными вызовами. Они могут, в частности, применяться для косвенной оценки тех вероятностных характеристик модели, прямое измерение которых затруднительно из-за сложностей, возникающих при отделении первичных вызовов от повторных (к таким характеристикам относятся: интенсивность поступления первичных вызовов, среднее число повторяющих абонентов и т. д.). Для i-го входа НС можно мерить: π (в силу пуассоновского характера поступления первичных вызовов среднее значение этой характеристики совпадает с долей времени занятости всех доступных i-му потоку линий НС);

$$L_1=(\lambda+J_1\mu_1+J_2\mu_2)(1-a)$$

интенсивность общего потока вызовов, поступивших на i-ый вход НС;

$$L_2=(\lambda\pi+J_1'\mu_1+J_2'\mu_2)(1-a)$$

интенсивность общего потока вызовов, поступивших на i-ый вход и получивших отказ из-за отсутствия свободных доступных линий НС. Пусть $N_1=\infty$, $N_2=\infty$.

Тогда все введенные характеристики находятся последовательно из следующих формул, которые выведены из соотношений (4.1)–(4.4). Обозначим через

$$x_k=(L_1-L_2)\,P_1,\ldots,P_k,\quad k=0,\,1,\ldots,m,\ (x_0=L_1-L_2);$$

$$\lambda = \frac{L_1(1 - H_2 a)/(1 - a) - L_2 H_2 - \sum_{k=1}^{m} x_{k-1}(1 - P_k) Z_k}{1 - (H_2 - H_1)(\pi + a(1 - \pi))};$$

$$\pi_1 = \frac{(L_2 - \lambda\pi(1 - a))(1 - a)}{L_1 - \lambda(1 - a)} + a; \quad \pi_2 = \frac{L_2(1 - a)}{L_1} + a;$$

$$\pi_3 = 1 - P_1, \ldots, P_m (1 - \pi_2); \; I_k = x_{k-1}/\alpha_k, \quad k = 1, \ldots, m + 1;$$

$$J_2 = x_{m-1}(1 - P_m) Z_m/\mu_2; \; M = L_1/\lambda(1 - a) - 1;$$

$$J_1 = (L_1/(1 - a) - \lambda - J_2 \mu_2)/\mu_1; \; Q = \frac{L_1}{x_m(1 - a)} - 1;$$

$$D = (\lambda - I_{m+1} \alpha_{m+1})/\lambda.$$

5. Поведение вероятностных характеристик при малой нагрузке

Обозначим через $\lambda^* = \max_k \lambda_k$. Пусть $\lambda^* \to 0$. Найдем первый ненулевой член разложения по степеням λ вероятностной характеристики M. Воспользуемся определением (3.4). Возьмем один поступивший первичный вызов и рассчитаем среднее число повторных вызовов, которые он сделает, находясь в системе. При $\lambda^* \to 0$ поступающий первичный или повторный вызов получает отказ в обслуживании из-за занятости всех доступных линий с вероятностью $o(1)$, поскольку эта вероятность пропорциональна появлению по крайней мере одного первичного вызова в течение конечного промежутка времени. Тогда среднее число повторных вызовов для одного первичного можно записать в виде суммы

$$M = D_1 Q_2 \sum_{k=0}^{\infty} D_2^k (k + 1) + 0(1).$$

Здесь

$$D_1 = aH_1 + (1 - a)((1 - P_1)Z_1 + P_1(1 - P_2)Z_2 + \ldots + P_1 \ldots P_{m-1}(1 - P_m)Z_m);$$

$$D_2 = aH_2 + (1 - a)((1 - P_1)Z_1 + P_1(1 - P_2)Z_2 + \ldots + P_1 \ldots P_{m-1}(1 - P_m)Z_m);$$

$$Q_2 = a(1 - H_2) + (1 - a)((1 - P_1)(1 - Z_1) + P_1(1 - P_2)(1 - Z_2) + \ldots + P_1 \ldots P_{m-1}(1 - P_m)(1 - Z_m) + P_1 \ldots P_m).$$

Проведя суммирование и воспользовавшись определением (3.2), получим

$$M = \frac{J_1 \mu_1 + J_2 \mu_2}{\lambda} = \frac{D_1}{Q_2} + o(1).$$

Используя определение вероятностных характеристик и результат леммы 3, найдем для остальных вероятностных характеристик i-го входа НС следующие приближенные формулы, справедливые при $\lambda \to 0$.

$$J_1 = \frac{\lambda(D_1 - (1 + a(H_1 - H_2))(1 - a) P_1 \dots P_{m-1}(1 - P_m) Z_m)}{Q_2 \mu_1} + o(\lambda);$$

$$J_2 = \frac{\lambda(1 + a(H_1 - H_2))(1 - a) P_1 \dots P_{m-1}(1 - P_m) Z_m}{Q_2 \mu_2} + o(\lambda); \qquad (5.1)$$

$$I_k = \frac{\lambda(1 + a(H_1 - H_2)) P_1 \dots P_{k-1}(1 - a)}{Q_2 \alpha_k} + o(\lambda), \quad k = 1, \dots, m + 1;$$

$$D = 1 - (1 + a(H_1 - H_2))\ P_1 \dots P_m(1 - a)/Q_2 + o(1);$$

$$\pi_3 = a + (1 - a)(1 - P_1 \dots P_m) + o(1);\ \ Q = -1 + 1/(1 - a) P_1 \dots P_m + o(1);$$

$$\pi_0 = a + o(1);\ \pi_1 = a + o(1);\ \pi_2 = a + o(1).$$

В таблице 1 приведены результаты точного и приближенного (по формулам (5.1)) расчета вероятностных характеристик частного случая НС — полнодоступного пучка из v линий. Имеется один поступающий поток первичной нагрузки, обслуживание состоит из двух фаз. Входные параметры модели принимают следующие численные значения:

$v = 3$; $\mu_1 = 10$; $\mu_2 = 5$; $H_1 = 0{,}7$; $H_2 = 0{,}9$; $\alpha_1 = 10$; $\alpha_2 = 2$; $P = 0{,}5$; $Z = 0{,}5$; $a = 0{,}1$; значение λ приведено в таблице 1, $N_1 = N_2 = 7$.

Из анализа численных данных следует, что при малых значениях λ точность асимптотического метода вычисления вероятностных характеристик довольно велика. При возрастании λ точность падает, но остается приемлемой для практического использования даже при тех λ, которые не являются малыми (например, $\lambda = 1$, в таблице 1).

6. Поведение вероятностных характеристик модели при большой нагрузке

Рассмотрим вначале частный случай НС-модель v — линейного полнодоступного пучка с многофазным обслуживанием и двумя типами повторяющих абонентов. Входными параметрами модели будут: $\alpha_1, \dots, \alpha_{m+1}, P_1, \dots$ $\dots, P_m, Z_1, \dots, Z_m, H_1, H_2, a, \mu_1, \mu_2, \lambda$. Определение и обозначение входных параметров модели и вероятностных характеристик совпадает с соответствующими характеристиками исходной модели НС только без индекса i, номера потока нагрузки (см. разделы 2–3). Предположим, что $H_2 < 1$, $N_1 = \infty$, $N_2 = \infty$, $\lambda \to \infty$, а остальные параметры модели фиксированы. В силу ана-

Таблица 1

Сравнение точного и приближенного метода расчета вероятностных характеристик пучка при $\lambda \to 0$

λ	M		I_1		I_2		Q		π	
	точно	прибл.	точно	прибл.	точно	прибл.	точно	прибл.	точно	
1	0,4504	0,4307	0,1282	0,1288	0,3204	0,3219	1,263	1,222	0,1039	10^{-1}
0,5	0,4336	0,4307	0,06434	0,06438	0,1608	0,1600	1,228	1,222	0,1571	10^{-2}
0,1	0,4307	0,4307	0,01288	0,01288	0,03219	0,03219	1,222	1,222	0,1467	10^{-4}

λ	J_1		J_2		D		π_3		π_2	
	точно	прибл.	точно	прибл.	точно	прибл.	точно	прибл.	точно	прибл.
1	0,01299	0,01088	0,06409	0,06438	0,3592	0,3562	0,5582	0,55	0,1163	0,1
0,5	0,005593	0,005440	0,03217	0,03219	0,3566	0,3562	0,5512	0,55	0,1024	0,1
0,1	0,001088	0,001088	0,006438	0,006438	0,3562	0,3562	0,5500	0,55	0,1000	0,1

литичности I_k, J_k, π, $k = 1, 2$, и справедливости соотношений (4.1)–(4.4), разложение основных характеристик пучка при $\lambda \to \infty$ можно искать в виде:

$$
\begin{aligned}
&I_k = I_{k,0} + I_{k,1} \cdot \frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right), \quad k = 1, \ldots . m+1;\\
&J_1 = J_{1,-1}\,\lambda + J_{1,0} + o(1);\\
&J_2 = J_{2,0} + J_{2,1}\frac{1}{\lambda} + o(1);\\
&\pi = \bar{\pi}_0 + \bar{\pi}_1 \frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right);\\
&J'_1\mu_1 + J'_2\mu_2 = J' = J'_{-1}\,\lambda + J'_0 + 0(1).
\end{aligned}
\tag{6.1}
$$

Выведем некоторые соотношения для коэффициентов в разложении (6.1). Обозначим через $f = 1 + \alpha_1 \sum_{k=1}^{m} \frac{P_1 \ldots P_k}{\alpha_{k+1}}$. Справедлива

Лемма 4

$$
\bar{\pi}_0 = 1, \sum_{k=1}^{m+1} I_{k,0} = \nu, I_{k,0} = \frac{\nu\alpha_1 P_1 \ldots P_{k-1}}{f\alpha_k}; \tag{6.2}
$$

$$
\bar{\pi}_1 = -(1 - H_2)\,\nu\alpha_1/(1 - H_2 + H_1)\,f(1 - a); \tag{6.3}
$$

$$
I_{k,1} = -(1 - H_2)\,\nu\alpha_1^2 P_1 \ldots P_{k-1}/(1 - H_2 + H_1)\,f^2\alpha_k(1 - a). \tag{6.4}
$$

Соотношения (6.2) очевидны. Докажем соотношение (6.3). Обозначим через $y(\lambda)$ ту часть потока первичной нагрузки, которая сразу попадает на первую фазу установления соединения. Ясно, что

$$
\lim_{\lambda\to\infty} y(\lambda) = \lim_{\lambda\to\infty} \lambda(1 - \pi)\,(1 - a) = -\bar{\pi}_1(1 - a). \tag{6.5}
$$

Рассмотрим поток повторных вызовов, поступающий на вход пучка. При $\lambda \to \infty$ его можно считать асимптотически пуассоновским с интенсивностью $\lambda H_1(1 - a)/(1 - H_2)$.

В этом случае справедливо соотношение

$$
\lim_{\lambda\to\infty} \frac{y(\lambda)}{I_1\alpha_1 - y(\lambda)} = \frac{1 - H_2}{H_1}. \tag{6.6}
$$

Из (6.2), (6.5), (6.6) следует (6.3). Соотношение (6.4) вытекает из равенства

$$\sum_{k=1}^{m+1} I_k = \pi\nu + (1-\pi)(\nu-1) + o\left(\frac{1}{\lambda}\right),$$

справедливого для данной модели при $\lambda \to \infty$.

Подставим (6.1) в (4.1)–(4.4) и приравняем коэффициенты при одинаковых степенях λ. Используя результат леммы 4, решим полученные системы уравнений относительно неизвестных коэффициентов разложения в ряд (6.1). Обозначим $F_k = P_1 \dots P_k$, ($F_0 = 1$). В результате формулы (6.1) будут иметь вид:

$$I_k = \frac{\nu\alpha_1 F_{k-1}}{f\alpha_k} - \frac{(1-H_2)\nu\alpha_1^2 F_{k-1}}{(1-H_2+H_1)f^2\alpha_k(1-a)\lambda} + o\left(\frac{1}{\lambda}\right), \quad k = 1, \dots, m+1;$$

$$J_1 = \frac{\lambda H_1}{\mu_1(1-H_2)} + \frac{\nu\alpha_1}{f\mu_1(1-H_2)}\left[\sum_{k=1}^{m-1}(F_{k-1}-F_k)Z_k + \right.$$

$$\left. + (F_{m-1}-F_m)Z_m H_2 - \frac{H_1}{1-H_2+H_1}\right] + o(1); \tag{6.7}$$

$$J_2 = \frac{\nu\alpha_1(F_{m-1}-F_m)Z_m}{f\mu_2} - \frac{(1-H_2)\nu\alpha_1^2(F_{m-1}-F_m)Z_m}{f^2\mu_2(1-H_2+H_1)(1-a)\lambda} + o\left(\frac{1}{\lambda}\right);$$

$$J' = \frac{\lambda H_1}{1-H_2} + \frac{\nu\alpha_1}{f(1-H_2)}\left[\sum_{k=1}^{m}(F_{k-1}-F_k)Z_k - \frac{H_1(2-a-H_2)}{(1-H_2+H_1)(1-a)}\right] + o(1);$$

$$\pi = 1 - \frac{(1-H_2)\nu\alpha_1}{(1-H_2+H_1)f(1-a)\lambda} + o\left(\frac{1}{\lambda}\right).$$

Когда $H_1 = H_2 = H$, для вероятностных характеристик J_1, $\lambda\pi + J_1'\mu_1 + J_2'\mu_2$ можно найти три члена разложения по степеням λ.

$$J_1 = \frac{\lambda H}{\mu_1(1-H)} + \frac{\nu\alpha_1}{f\mu_1(1-H)}\left[\sum_{k=1}^{m-1}(F_{k-1}-F_k)Z_k + (F_{m-1}-F_m)Z_m H - H\right] -$$

$$- \frac{\nu\alpha_1^2}{f^2\mu_1(1-a)}\left[\sum_{k=1}^{m-1}(F_{k-1}-F_k)Z_k + (F_{m-1}-F_m)Z_m H - H\right]\frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right);$$

$$\lambda\pi + J_1'\mu_1 + J_2'\mu_2 = \frac{\lambda}{1-H} + \frac{\nu\alpha_1}{f(1-H)}\left[\sum_{k=1}^{m}(F_{k-1}-F_k)Z_k - \frac{1-aH}{1-a}\right] -$$

$$- \frac{\nu\alpha_1^2}{f^2(1-a)}\left[\sum_{k=1}^{m}(F_{k-1}-F_k)Z_k - \frac{1-aH}{1-a}\right]\frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right).$$

Таблица 2

Сравнение точного и приближенного метода расчета вероятностных характеристик пучка при $\lambda \to \infty$

λ	M		I_1		I_2		Q		π	
	точно	прибл.	точно	прибл.	точно	прибл.	точно	прибл.	точно	прибл.
5	1,761	1,718	0,3425	0,3564	0,5480	0,5703	24,19	22,74	0,8905	0,9267
10	1,870	1,859	0,3654	0,3705	0,5846	0,5928	48,10	47,19	0,9500	0,9634
20	1,932	1,930	0,3761	0,3776	0,6018	0,6041	96,46	95,97	0,9779	0,9817
30	1,954	1,953	0,3793	0,3799	0,6068	0,6078	145,1	144,7	0,9861	0,9878

λ	J_1		J_2		D		π_1		π_2	
	точно	прибл.	точно	прибл.	точно	прибл.	точно	прибл.	точно	прибл.
5	0,4361	0,4253	0,004110	0,004277	0,8904	0,8859	0,9603	0,9542	0,9504	0,9438
10	0,9308	0,9253	0,004384	0,004446	0,9416	0,9407	0,9796	0,9784	0,9745	0,9732
20	1,928	1,925	0,004513	0,004531	0,9699	0,9698	0,9897	0,9895	0,9872	0,9869
30	2,927	2,925	0,004551	0,004559	0,9798	0,9797	0,9932	0,9930	0,9914	0,9913

Отметим, что для приближенного вычисления $M, Q, \pi_3, \pi_2, \pi_1$ при больших значениях λ более точные результаты дает использование соотношений (6.7) и определений вероятностных характеристик без разложения по степеням $1/\lambda$ (таким образом были рассчитаны приближенные значения для Q, π_3, π_2, в таблице 2).

Рассмотрим разложение соответствующих вероятностных характеристик i-го входа НС. Пусть $\lambda_i \to \infty$, $H_2 < 1$, а остальные параметры модели фиксированы. Тогда, действуя аналогично случаю полнодоступного пучка, можно убедиться, что вид коэффициентов $I_{k,0}$, $J_{1,-1}$, $J_{1,0}$, $J_{2,0}$, J'_{-1} J'_0, $\overline{\pi}_0$, $\overline{\pi}_1$ в разложении (6.7) будет совпадать с видом соответствующих коэффициентов в разложении подобных характеристик по степеням $1/\lambda_i$ для i-го входа НС. При этом учитывается то, что параметры i-го входа имеют i-ю индексацию и вместо числа линий v берется доступность d_i.

В качестве иллюстрации точности приближенного метода вычисления вероятностных характеристик модели при больших значениях интенсивности первичной нагрузки в таблице 2 приведены результаты точного и приближенного (по формулам (6.7)) расчета значений некоторых характеристик для полнодоступного пучка с двухфазным обслуживанием.

Входные параметры принимают следующие численные значения:

$v = 1$; $\mu_1 = 20$; $\mu_2 = 20$; $H_1 = 0{,}6$; $H_2 = 0{,}7$; $\alpha_1 = 2$; $\alpha_2 = 1$; $p = 0{,}8$; $Z = 0{,}6$; $a = 0{,}3$; значение λ дано в таблице 2.

Из анализа численных данных следует, что наиболее точно рассчитываются характеристики поведения абонентов (M, Q, D, π_3, π_2 в таблице 2) и менее точно характеристики пропускной способности пучка (π, I_1, I_2 в таблице 2). Если аналогичное сравнение произвести с данными, приведенными в таблице 1, то можно заметить, что при большой нагрузке наблюдается обратная картина в сравнении точности приближенных формул для различных вероятностных характеристик, чем при малой нагрузке.

Можно надеяться, что рассмотренный метод исследования систем с повторными вызовами расширит область применения сложных теоретических моделей для описания функционирования реальных информационно-вычислительных систем.

Литература

1. *Бенеш В. Э.* Математические основы теории телефонных сообщений. М., «Связь», 1968.
2. *Степанов С. Н.* Интегральные соотношения равновесия для неполнодоступной системы с повторными попытками и их применение. Проблемы передачи информации, 1980, т. **16,** вып. *2*.

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

Probabilistic characteristics of an incompletely accessible multi-phase service system with several types of repeated calls

S. N. STEPANOV

(Moscow)

A non-full available system with repeated calls is investigated. The primary traffic offered has a Poisson distribution, the holding time consists of n phases each having exponential distribution. After finishing phase number i with probability p_i begins next phase, with probability $(1 - p_i)(1 - z_i)$ call leaves system and with probability $(1 - p_i) z_i$ it makes repeated call. The intervals between repeated calls are mutually independent, exponentionally distributed and depend on failure. Primary calls and repeated calls may be blocked before entering system and because busy of available trunks, in this case call may leave system with some probability and with additional probability makes another call. For this model relations between different probabilistic characteristics and approximate formulas for calculation characteristics in the cases of small and big intensities of primary calls are given. The accuracy of approximate formulas is illustrated by examples of exact calculations for some models.

С. Н. Степанов
Институт проблем передачи информации АН СССР
101447 Москва ГСП—4, ул. Ермоловой, 19

Problems of Control and Information Theory, Vol. 10 (6), pp. 403—415 (1981)

MINIMAL REALISATION PROCEDURE FROM PARTIAL FRACTION REPRESENTATION OF A RATIONAL TRANSFER MATRIX

J. BOKOR, J. PETŐ

(Budapest)

(Received February 1, 1981)

Minimal realisation procedures of multivariable systems utilising the partial fraction representation of the rational transfer matrix have been introduced by Gilbert [1] and Kalman [2]. This paper derives a minimal realisation from the partial fraction representation of the transfer matrix having poles (real and complex) of multiplicity one. The procedure can be considered an extension of Gilbert's method but because of the restriction on the pole's multiplicity it is less general than Kalman's method. This procedure is based on a minimal dyadic decomposition of partial fraction matrices and therefore much simpler than that of Kalman's. This simplicity offers an easy application on digital computers and in many practical cases.

1. Introduction

In this paper we derive a minimal realisation procedure for multivariable systems. We deal with discrete time systems which can be described in some alternative ways (differential operator representation, matrix fraction description). We consider here the

(i) n-order State Space Representation $(\mathbf{F}, \mathbf{G}, \mathbf{H}, \mathbf{D})$ or

$$\mathbf{x}(t+1) = \mathbf{F}\mathbf{x}(t) + \mathbf{G}\mathbf{u}(t) \tag{1.1}$$

$$\mathbf{y}(t) = \mathbf{H}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)$$

where $\mathbf{x}(t)$ is an $(n \times 1)$ vector called the state $\mathbf{y}(t)$ is a $(p \times 1)$ output vector, $\mathbf{u}(t)$ is an $(m \times 1)$ input vector and $\mathbf{F}$, $\mathbf{G}$, $\mathbf{H}$, $\mathbf{D}$ are real matrices of appropriate dimensions.

(ii) the Transfer Matrix Representation $\mathbf{Z}(z)$ or

$$\mathbf{y}(z) = \mathbf{Z}(z)\,\mathbf{u}(z) \tag{1.2}$$

with $\mathbf{y}(z)$, $\mathbf{u}(z)$ the Z-transform of the system output and input respectively, assuming zero initial condition on the state. The transfer matrix is said to be rational if every element of $\mathbf{Z}(z)$ is a ratio of polynomials in z with coeffitients over the field of real numbers $\mathcal{R}$.

$\mathbf{Z}(z)$ is regular if $\mathbf{Z}(\infty) \neq \infty$ and proper if $\mathbf{Z}(\infty) = 0$. The transfer from state-space to transfer matrix representation is unique:

$$\mathbf{Z}(z) = \mathbf{H}(z\mathbf{I} - \mathbf{F})^{-1}\mathbf{G} + \mathbf{D} \tag{1.3}$$

but the converse is not true.

The realisation problem is: given $\mathbf{Z}(z)$ find $(\mathbf{F}, \mathbf{G}, \mathbf{H}, \mathbf{D})$ such that these four matrices realise $\mathbf{Z}(z)$ so that (1.3) be held.

We deal with proper matrices ($\mathbf{D} = 0$) because the presence of $\mathbf{D}$ has no effect on the realisation procedure.

Every proper matrix has a finite dimensional realisation. A realisation of $\mathbf{Z}(z)$ is irreducible or minimal if the dimension of the state is minimal over the dimensions of all possible realisations.

There are various approaches to obtain a minimal realisation of a transfer matrix (see e.g. Gilbert[1], Kalman [2], Mayne [4], Ho [3], Rosenbrock [5]).

The method proposed in Section 3 of this paper is close to Gilbert's and Kalman's approach. It starts with the partial fraction representation (PFR) of the transfer matrix $\mathbf{Z}(z)$ (see the definition later), derives a minimal realisation for each matrices from this expansion, and builds up the minimal realisation of $\mathbf{Z}(z)$ as the direct sum of these realisations.

Our method can be considered either as the extension of Gilbert's method, since it can treat both real and conjugate complex poles, or as a special case of Kalman's method. The resulting procedure, however, is different and much simpler than that of Kalman's [2]. Moreover it can easily be applied either analytically or on digital computers. In the authors opinion it can be useful in various practical problems.

2. Preliminaries

First we give some definitions which will be used in the sequel and shortly review Kalman's minimal realisation procedure.

Let $(\mathbf{F}_\mu, \mathbf{G}_\mu, \mathbf{H}_\mu)$, $\mu = 1, \ldots, q$ be realisations of proper rational transfer matrices $\mathbf{Z}_\mu(z)$ having dimension $n(\mathbf{Z}_\mu)$. By the direct sum of these realisations we mean the realisation of

$$\mathbf{Z}(z) = \sum_{\mu=1}^{q} \mathbf{Z}_\mu(z) \tag{2.1}$$

given by

$$\mathbf{F} = \begin{bmatrix} \mathbf{F}_1 & & \\ & \ddots & \\ & & \mathbf{F}_q \end{bmatrix} \qquad \mathbf{G} = \begin{bmatrix} \mathbf{G}_1 \\ \vdots \\ \mathbf{G}_q \end{bmatrix} \tag{2.2}$$

$$\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_q].$$

The dimension of (2.2) is $n(\mathbf{Z}) = \sum_{\mu=1}^{q} n(\mathbf{Z}_\mu)$.

Proposition 2.1. Consider q irreducible realisations $(\mathbf{F}_\mu, \mathbf{G}_\mu, \mathbf{H}_\mu)$ of $\mathbf{Z}(z)$. If $\mathbf{F}_\mu$ and $\mathbf{F}_\nu$ have no eigenvalues in common for $\mu \neq \nu$, then the direct sum of these realisations is also irreducible [2].

Let the polynomial $\psi(z)$ denote the monic least common denominator of the elements of $\mathbf{Z}(z)$. It can be factored as a multiple of irreducible polynomials over the field of complex numbers $\mathbf{C}$.

$$\psi(z) = \prod_{\mu=1}^{q} \psi_\mu(z); \quad \psi_\mu(z) = (z - \lambda_\mu)^{r_\mu}; \quad \lambda_\mu \in \mathbf{C} \tag{2.3}$$

where λ_μ are the poles of $\mathbf{Z}(z)$ and r_μ is a positive integer called the multiplicity of pole λ_μ.

By the partial fraction representation (PFR) of a rational matrix $\mathbf{Z}(z)$ we mean the following. Let $\psi(z)$ be factored as in (2.3). Take the partial fraction decomposition of every element of $\mathbf{Z}(z)$ and collect all terms associated with λ_μ into the matrix $\mathbf{Z}_\mu(z)$ called as a partial fraction matrix (PFM) of $\mathbf{Z}(z)$. $\mathbf{Z}(z)$ is then the sum of $\mathbf{Z}_\mu(z)$.

$$\mathbf{Z}(z) = \sum_{\mu=1}^{q} \mathbf{Z}_\mu(z). \tag{2.4}$$

Notice that $\mathbf{Z}_\mu(z)$ have only one pole at $z = \lambda_\mu$ and these poles are different.

In [2] Kalman derived a minimal realisation procedure based on the PFR of $\mathbf{Z}(z)$. The essence of his procedure is the following.

Construct the partial fraction representation (2.4) of $\mathbf{Z}(z)$. As the matrices $\mathbf{Z}_\mu(z)$ have different poles, by Proposition 2.1 it is possible to build up a minimal representation of $\mathbf{Z}(z)$ as the direct sum of minimal realisation for each $\mathbf{Z}_\mu(z)$.

To get a minimal realisation for $\mathbf{Z}_\mu(z)$, transform $\mathbf{Z}_\mu(z)$ into its Smith–McMillan form:

$$\mathbf{Z}_\mu(z) = \mathbf{C}_\mu(z)\,\psi_\mu(z)\,\mathbf{B}_\mu(z); \quad \mathbf{C}_\mu(z) = \mathbf{A}_\mu(z)\,\mathbf{E}_\mu(z) \tag{2.5}$$

where $\mathbf{A}_\mu(z)$, $\mathbf{G}_\mu(z)$ are unimodular matrices of dimensions $(p \times p)$ and $(m \times m)$ respectively:

$$\mathbf{E}_\mu(z) = \begin{bmatrix} \varepsilon_{\mu 1}(z) & & & & \\ & \ddots & & & \\ & & \varepsilon_{\mu R_\mu}(z) & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \quad \Psi_\mu(z) = \begin{bmatrix} \psi_{\mu 1}(z) & & & & \\ & \ddots & & & \\ & & \psi_{\mu r_\mu}(z) & & \\ & & & 1 & \\ & & & & \ddots \\ & & & & & 1 \end{bmatrix}. \tag{2.6}$$

The dimensions of $\mathbf{E}_\mu(z)$, $\Psi_\mu(z)$ are $(p \times m)$ and $(m \times m)$, furthermore $\varepsilon_{\mu i}(z)$, $\psi_{\mu i}(z)$ are relative prime polynomials with real or complex coefficients. The integer R_μ is the rank of $\psi_\mu(z)\mathbf{Z}_\mu(z)$, and $r_\mu \leq R_\mu$ is called the subrank of $\mathbf{Z}_\mu(z)$ Kalman has shown that the polynomials $\psi_{\mu i}(z)$, $\mu = 1, \ldots, q$, $i = 1, \ldots, r$ are the elementary divisors of $\mathbf{Z}(z)$.

Using (2.5) $\mathbf{Z}_\mu(z)$ can be written as

$$\mathbf{Z}_\mu(z) = \sum_{i=1}^{r_\mu} \mathbf{Z}_{\mu i}(z); \quad \mathbf{Z}_{\mu i}(z) = \psi_{\mu i}^{-1}(z)\, \mathbf{c}_{\mu i}(z)\, \mathbf{b}_{\mu i}^T(z) \tag{2.7}$$

where $\mathbf{c}_{\mu i}(z)$, $\mathbf{b}_{\mu i}^T(z)$ denote the i-th column and row vectors of $\mathbf{C}_\mu(z)$ and $\mathbf{B}_\mu(z)$ respectively.

According to (2.5) $\mathbf{Z}(z)$ can be realised as the direct sum of minimal realisation of $\mathbf{Z}_{\mu i}(z)$. In general, however, even if $\mathbf{Z}_\mu(z)$ are proper, $\mathbf{Z}_{\mu i}(z)$ need not be proper. For this reason $\mathbf{Z}_{\mu i}(z)$ are decomposed into the sum of their proper and polynomial part and $\mathbf{Z}_\mu(z)$ is realised as the direct sum of minimal realisation of the proper parts. Kalman shows that in the minimal realisation the $\mathbf{F}_{\mu i}$ matrices are Jordan blocks having dimension $r_{\mu i}$ and the $\mathbf{H}_{\mu i}$, $\mathbf{G}_{\mu i}$ matrices can be determined from the solution of simultaneous linear equations.

This realisation may have complex coeffitients, so Kalman applies a linear transformation resulting in a representation with real coefficients.

The difficulties in the application of the procedure arise from the transformation of $\mathbf{Z}(z)$ into the Smith–McMillan form, furthermore from the determination of the proper part for each $\mathbf{Z}_{\mu i}(z)$ and from the computation with complex numbers.

In the next section we show that if the poles (real and conjugate complex pairs) have multiplicity one, a minimal realisation procedure can be derived without the above difficulties.

This procedure will utilise also a partial fraction representation of transfer matrix $\mathbf{Z}(z)$.

3. Minimal realisation procedure using minimal dyadic decomposition of partial fraction matrices

In the foregoing discussion we suppose that the poles of $\mathbf{Z}(z)$ (the zeros of $\psi(z)$) have multiplicity one. If $\psi(z)$ is factored as a multiple of irreducible polynomials over the field of real numbers $\mathfrak{R}$, we get:

$$\psi(z) = \prod_{\mu=1}^{n_r} \psi_{r\mu}(z) \prod_{\nu=1}^{n_c} \psi_{c\nu}(z) \tag{3.1}$$

$$\psi_{r\mu}(z) = z - \lambda_\mu; \quad \psi_{c\nu}(z) = z^2 + \alpha_{\nu 1} z + \alpha_{\nu 2}; \quad \lambda_\mu, \alpha_{\nu 1}, \alpha_{\nu 2} \in \mathfrak{R}$$

where n_r, n_c denote the number of different real and complex conjugate pairs of poles respectively.

Now take a partial fraction decomposition of every element of $\mathbf{Z}(z)$ over $\mathfrak{R}$, collect all terms associated with $\psi_{r\mu}(z)$ into the matrix $\mathbf{Z}_{r\mu}(z)$ and all terms associated with $\psi_{c\nu}(z)$ into the matrix $\mathbf{Z}_{c\nu}(z)$.

$\mathbf{Z}_{r\mu}(z)$ and $\mathbf{Z}_{c\nu}(z)$ are called the partial fraction matrices, and the PFR of $\mathbf{Z}(z)$ is given by

$$\mathbf{Z}(z) = \sum_{\mu=1}^{n_r} \mathbf{Z}_{r\mu}(z) + \sum_{\nu=1}^{n_c} \mathbf{Z}_{c\nu}(z). \tag{3.2}$$

Notice that all $\mathbf{Z}_{r\mu}(z)$ and $\mathbf{Z}_{c\nu}(z)$ have distinct poles, and by Proposition 2.1, $\mathbf{Z}(z)$ can be realised as the direct sum of minimal realisations of $\mathbf{Z}_{r\mu}(z)$, $\mathbf{Z}_{c\nu}(z)$.

This minimal realisation will be constructed from a minimal dyadic decomposition of the matrices $\psi_{r\mu}(z)\mathbf{Z}_{r\nu}(z)$, $\psi_{c\nu}(z)\mathbf{Z}_{c\nu}(z)$ as the direct sum of realisations corresponding to their dyads.

Definition 3.1. Let $\mathbf{c}$ be an $(p\times 1)$ dimensional column and $\mathbf{b}^T$ be a $(1\times m)$ dimensional row vector over a field $\mathfrak{K}$. The product $\mathbf{cb}^T$ (it is a $(p\times m)$ matrix) is called the dyadic product (or simple dyad) of vectors $\mathbf{c}$, $\mathbf{b}$.

Theorem 3.1. Let $\mathbf{W} = \{w_{ij}\}$ be a $(p\times m)$ matrix with $w_{ij} \in \mathfrak{K}$ and rank r_w. Then $\mathbf{W}$ can be decomposed as the sum of dyads. The minimum number of dyads in this decomposition is r_w.

The proof of this Theorem is well known from standard linear algebra textbooks.

A decomposition of $\mathbf{W}$ as the sum of r_w dyads is called minimal dyadic decomposition of $\mathbf{W}$.

To show how this minimal dyadic decomposition can be performed, let us introduce the following notations. Denote $\overset{(1)}{\mathbf{W}} = \mathbf{W}$ and $\overset{(1)}{\mathbf{v}}_{j_1}$, $\overset{(1)}{\mathbf{w}}_{i_1}^T$ the j_1-th column and i_1-th row vector of $\overset{(1)}{\mathbf{W}}$. Find the first nonzero element in the first nonzero row of $\overset{(1)}{\mathbf{W}}$, say $\overset{(1)}{\mathbf{w}}_{i_1,j_1}$.

The $\overset{(1)}{w}_{i_1,j_1}$ is called the generating element of the first dyad, and its inverse in $\mathfrak{K}$ will be denoted by $\overset{(1)}{w}{}^{-1}_{i_1,j_1}$.

Construct the vectors $\mathbf{c}_1, \mathbf{b}_1^T$ as:

$$\begin{aligned} \mathbf{c}_1 = \overset{(1)}{w}{}^{-1}_{i_1,j_1} \overset{(1)}{\mathbf{v}}_{j_1} &= [0, \ldots, 0, 1, \overset{(1)}{w}{}^{-1}_{i_1,j_1} \overset{(1)}{w}_{i_1+1,j_1}, \ldots, \overset{(1)}{w}{}^{-1}_{i_1,j_1} \overset{(1)}{w}_{p,j_1}]^T \\ \mathbf{b}_1^T = \overset{(1)}{\mathbf{w}}_{i_1}^T &= [0, \ldots, 0, \overset{(1)}{w}_{i_1,j_1}, \ldots, \overset{(1)}{w}_{i_1,m}]. \end{aligned} \tag{3.3}$$

The first dyad is the product $\mathbf{c}_1\mathbf{b}_1^T$. Compute $\overset{(2)}{\mathbf{W}}$ as $\overset{(2)}{\mathbf{W}} = \overset{(1)}{\mathbf{W}} - \mathbf{c}_1\mathbf{b}_1^T$.

The matrix $\overset{(2)}{\mathbf{W}}$ will have only zero element in its first i_1 rows and the j_1-th column. Now find the first nonzero element in the first nonzero row of $\overset{(2)}{\mathbf{W}}$, say $\overset{(2)}{w}_{i_1,j_2}$ clearly $i_2 > i_1$ and $j_2 \neq j_1$ and compute the second dyad as $\mathbf{c}_2 \mathbf{b}_2^T = \overset{(2)}{w}{}_{i_2,j_2}^{-1} \overset{(2)}{\mathbf{v}}_{j_2} \overset{(2)}{\mathbf{w}}{}_{i_2}^T$. By this construction the first i_2 elements of $\mathbf{c}_2$ and the first j_2 elements of $\mathbf{b}_2^T$ will be zeros.

In general the above procedure is given by the recursion

$$\overset{(k+1)}{\mathbf{W}} = \overset{(k)}{\mathbf{W}} - \mathbf{c}_k \mathbf{b}_k^T; \quad \overset{(1)}{\mathbf{W}} = \mathbf{W} \tag{3.4}$$

and it ends after r_w steps, when $\overset{(r_w+1)}{\mathbf{W}}$ becomes a zero matrix.

Consider now the matrices $\mathbf{Z}_{r\mu}(z)$ corresponding to a real pole λ_μ. Since $\psi_{r\mu}(z)\mathbf{Z}_{r\mu}(z)$ is a matrix over the field of real numbers $\mathfrak{R}$, its minimal dyadic decomposition can be constructed according to (3.4) as

$$\psi_r(z)\mathbf{Z}_{r\mu}(z) = \sum_{i=1}^{r_{r\mu}} \mathbf{c}_{\mu i} \mathbf{b}_{\mu i}^T; \quad \psi_{r\mu}(z) = z - \lambda_\mu \tag{3.5}$$

where $r_{r\mu}$ is the rank of $\psi_r(z)\mathbf{Z}_{r\mu}(z)$ (in vectors $\mathbf{c}_{r\mu i}, \mathbf{b}_{r\nu i}$ we omitted the index r for simplicity). From (3.5) we can write

$$\mathbf{Z}_{r\mu}(z) = \sum_{i=1}^{r_{r\mu}} \frac{1}{z - \lambda_\mu} \mathbf{c}_{\mu i} \mathbf{b}_{\mu i}^T = \mathbf{H}_\mu (z\mathbf{I} - \mathbf{F}_\mu)^{-1} \mathbf{G}_\mu \tag{3.6}$$

where

$$\mathbf{F}_\mu = \begin{bmatrix} \lambda_\mu & & \\ & \ddots & \\ & & \lambda_\mu \end{bmatrix} \qquad \mathbf{G}_\mu = \begin{bmatrix} \mathbf{b}_{\mu 1}^T \\ \vdots \\ \mathbf{b}_{\mu r_{r\mu}}^T \end{bmatrix} \tag{3.7}$$

$$\mathbf{H}_\mu = [\mathbf{c}_{\mu 1}, \ldots, \mathbf{c}_{\mu r_{r\mu}}].$$

The dimensions of $\mathbf{F}_\mu$, $\mathbf{G}_\mu$, $\mathbf{H}_\mu$ are $(r_{r\mu} \times r_{\mu r})$, $(r_{r\mu} \times m)$ and $(p \times r_{r\mu})$.

To illustrate the structure of $\mathbf{H}_\mu$, $\mathbf{G}_\mu$, let us assume that the generating element of dyads $\mathbf{c}_{\mu i} \mathbf{b}_{\mu i}^T$ were on the main diagonal. Then from the decomposition procedure follows that $\mathbf{H}_\mu$, $\mathbf{G}_\mu$ have the form:

$$\mathbf{H}_\mu = \begin{bmatrix} 1 & & & \\ x & 1 & & \\ x & x & \ldots & 1 \\ \cdot & \cdot & & x \\ \cdot & \cdot & & \vdots \\ \cdot & \cdot & & \\ x & x & & x \end{bmatrix} \qquad \mathbf{G}_\mu = \begin{bmatrix} x & x & x & \ldots & x \\ & x & x & \ldots & x \\ & & \cdot & & \\ & & \cdot & & \\ & & \cdot & & \\ & & x & \ldots & x \end{bmatrix} \tag{3.8}$$

where x denotes a possible nonzero element.

It can be seen that rank $\mathbf{H}_\mu$ = rank $\mathbf{G}_\mu = r_{r\mu}$, and $\mathbf{H}_\mu$ has a certain "echelon" structure. Moreover it is not difficult to conclude that the realisation $(\mathbf{F}_\mu, \mathbf{H}_\mu, \mathbf{G}_\mu)$ of $\mathbf{Z}_\mu(z)$ with dimension $n(\mathbf{Z}_\mu) = r_{r\mu}$ is minimal.

We want to derive a similar minimal realisation from the minimal dyadic decomposition of $\psi_{cv}(z)\mathbf{Z}_{cv}(z)$. But the dyadic decomposition procedure (3.4) was defined for matrices over a field, and the elements of $\psi_{cv}(z)\mathbf{Z}_{cv}(z)$ belong to the ring of polynomials where there is no multiplicative inverse.

Since the elements of $\psi_{cv}(z)\mathbf{Z}_{cv}(z)$ are polynomials (of first order) they can represent an element in the residue class ring of polynomials modulo $\psi_{cv}(z)$. Denote $R[z]/\psi_{cv}(z)$ this residue class ring which is a field because $\psi_{cv}(z)$ is an irreducible polynomial over $\mathfrak{K}$. Then the matrix $\psi_{cv}(z)\mathbf{Z}_{cv}(z)$ can be considered as a matrix over the field $R[z]/\psi_{cv}(z)$ and procedure (3.4) can be applied.

We show that it is very easy to perform the necessary operations (multiplication, computation of an inverse element) in $R[z]/\psi_{cv}(z)$.

The first step in procedure (3.4) is to determine the inverse of $w^{(1)}_{i_1,j_1}$ which is a polynomial of degree less than or equal to one. The inverse of a first order polynomial $\varphi(z) = az + b$ can be easily obtained in $R[z]/\psi(z)$, $\psi(z) = z^2 + \alpha_1 z + \alpha_2$ as

$$\varphi^{-1}(z) = \frac{a}{ab\alpha_1 - a^2\alpha_2 - b^2} z + \frac{a\alpha_1 - b}{ab\alpha_1 - a^2\alpha_2 - b^2}.$$

It can be seen that if $a = 0$, $\varphi(z) = b$ then $\varphi^{-1}(z) = \dfrac{1}{b}$. To compute the first dyad, the vector $\mathbf{c}_1 = w^{-1}_{i_1,j_1}\mathbf{v}^{(1)}_{j_1}$ has to be computed. Its elements can be second order polynomials of form $Az^2 + Bz + C$. To determine a first order representation of $\mathbf{c}_1$ in $R[z]/\psi_{cv}(z)$, its elements have to be replaced by the residue obtained from the Euclidean division of each element by $\psi_{cv}(z)$. This residue will have the form $(B - A\alpha_1)z + (C - A\alpha_2)$. If $A = 0$, the residue is $Bz + C$.

Computing the product $\mathbf{c}_1\mathbf{b}_1^T$, the elements will be second order polynomials. Divide these products by $\psi_{cv}(z)$ and replace them by their residue. Now the first dyad is to be subtracted from $\mathbf{W}$ which means the subtraction of first order polynomials. These steps have to be repeated to compute the further dyads. The procedure ends after r_{cv} steps where r_{cv} = rank $\psi_{cv}(z)\mathbf{Z}_{cv}$ over $R[z]/\psi_{cv}(z)$.

Suppose that the minimal dyadical decomposition of $\psi_{cv}(z)\mathbf{Z}_{cv}(z)$ over $R[z]/\psi_{cv}(z)$ has already been constructed.

This is denoted by:

$$\psi_{cv}(z)\mathbf{Z}_{cv}(z) \equiv \sum_{i=1}^{r_{cv}} \mathbf{c}_i(z)\,\mathbf{b}_i^T(z) \mod \psi_{cv}(z). \tag{3.9}$$

Here we omitted the indices cv in $\mathbf{c}_{cvi}$, $\mathbf{b}_{cvi}$ for simplicity. Since the elements of $\mathbf{c}_i(z)$, $\mathbf{b}_i(z)$ are first order polynomials in z, we can write

$$\mathbf{c}_i(z) = \mathbf{c}_i^1 z + \mathbf{c}_i^0 \quad \mathbf{b}_i(z) = \mathbf{b}_i^1 z + \mathbf{b}_i^0. \tag{3.10}$$

Let us introduce the following notation. If $\mathbf{W}(z) = \{w_{ij}(z)\}$ is a polynomial matrix, denote $P(\psi_{cv})\{\mathbf{W}(z)\}$ a matrix of the same dimension whose (i, j)-th element is the residue obtained from the Euclidean division of $w_{ij}(z)$ by $\psi_{cv}(z)$.

Now (3.9) can be written as

$$\psi_{cv}(z)\,\mathbf{Z}_{cv}(z) = P(\psi_{cv})\,\{\psi_{cv}(z)\,\mathbf{Z}_{cv}(z)\} = P(\psi_{cv})\left\{\sum_{i=1}^{r_{cv}} \mathbf{c}_i(z)\,\mathbf{b}_i^T(z)\right\}$$

$$\psi_{cv}(z)\,\mathbf{Z}_{cv}(z) = \sum_{i=1}^{r_{cv}} P(\psi_{cv})\,\{\mathbf{c}_i(z)\,\mathbf{b}_i^T(z)\}$$

and

$$\mathbf{Z}_{cv}(z) = \sum_{i=1}^{r_{cv}} \psi_{cv}^{-1}(z)\,P(\psi_{cv})\,\{\mathbf{c}_i(z)\,\mathbf{b}_i^T(z)\} = \sum_{i=1}^{r_{cv}} \mathbf{Z}_{cvi}(z). \tag{3.11}$$

To proceed we need the following Lemma.

Lemma 3.1. Denote $\mathbf{F}_{c_v}$ the companion matrix of $\psi_{cv}(z) = z^2 + \alpha_{v1}z + \alpha_{v2}$, and introduce the vectors $\mathbf{q}(z)$, $\mathbf{k}(z)$:

$$\mathbf{F}_{cv} = \begin{bmatrix} 0 & \\ -\alpha_{v2} & -\alpha_{v1} \end{bmatrix}; \quad \mathbf{q}(z) = \begin{bmatrix} 1 \\ z \end{bmatrix}; \quad \mathbf{k}(z) = \begin{bmatrix} z+\alpha_{v1} \\ 1 \end{bmatrix}.$$

Then

$$(z\mathbf{I} - \mathbf{F}_{cv})^{-1} = \psi_{cv}^{-1}(z)\,P(\psi_{cv})\,\{\mathbf{q}(z)\,\mathbf{k}^T(z)\}. \tag{3.12}$$

The validity of (3.12) can be easily checked by computing both side of (3.12).

From (3.10), (3.11) we can write:

$$\mathbf{Z}_{cvi}(z) = \psi_{cv}^{-1}(z)\,P(\psi_{cv})\,\{\mathbf{c}_i(z)\,\mathbf{b}_i^T(z)\} = \psi_{cv}^{-1}(z)\,P(\psi_{cv})\,\{[\mathbf{c}_i^1 z + \mathbf{c}_i^0]\,[\mathbf{b}_i^1 z + \mathbf{b}_i^0]^T\} =$$

$$= \psi_{cv}^{-1}(z)\,P(\psi_{cv})\left\{[\mathbf{c}_i^0,\ \mathbf{c}_i^1]\begin{bmatrix} 1 \\ z \end{bmatrix}[z + \alpha_{v1},\ 1]\begin{bmatrix} \mathbf{b}_i^{1T} & \\ \mathbf{b}_i^{0T} & -\alpha_{v1}\mathbf{b}_i^{1T} \end{bmatrix}\right\}.$$

Denote

$$\mathbf{H}_{cvi} = [\mathbf{c}_i^0,\ \mathbf{c}_i^1]; \quad \mathbf{G}_{cvi} = \begin{bmatrix} \mathbf{b}_i^{1T} & \\ \mathbf{b}_i^{0T} & -\alpha_{v1}\mathbf{b}_i^T \end{bmatrix} \tag{3.13}$$

then from the Lemma follows that

$$\mathbf{Z}_{cvi}(z) = \psi_{cv}^{-1}(z)\,P(\psi_{cv})\,\{\mathbf{H}_{cvi}\,\mathbf{q}(z)\,\mathbf{k}^T(z)\,\mathbf{G}_{cvi}\} = \psi_{cv}^{-1}(z)\,\mathbf{H}_{cvi}\,P(\psi_{cv})\,\{\mathbf{q}(z)\,\mathbf{k}^T(z)\}\,\mathbf{G}_{cvi} =$$

$$= \mathbf{H}_{cvi}(z\mathbf{I} - \mathbf{F}_{cvi})^{-1}\,\mathbf{G}_{cvi}. \tag{3.14}$$

From (3.9) and (3.14) it can be seen that a realisation of $\mathbf{Z}_{cv}(z)$ can be constructed as the direct sum of realisation $(\mathbf{F}_{cvi}, \mathbf{G}_{cvi}, \mathbf{H}_{cvi})$ $i = 1, \ldots, r_{cv}$:

$$\mathbf{F}_{cv} = \begin{bmatrix} \mathbf{F}_{cv1} & & \\ & \ddots & \\ & & \mathbf{F}_{cvr_{cv}} \end{bmatrix} \quad \mathbf{G}_{cv} = \begin{bmatrix} \mathbf{G}_{cv1} \\ \vdots \\ \mathbf{G}_{cvr_{cv}} \end{bmatrix} \tag{3.15}$$

$$\mathbf{H}_{cv} = [\mathbf{H}_{cv1} \ldots \mathbf{H}_{cvr_{cv}}]$$

where

$$\mathbf{F}_{cvi} = \begin{bmatrix} 0 & 1 \\ -\alpha_{2vi} & -\alpha_{1vi} \end{bmatrix}$$

and $\mathbf{H}_{cvi}$, $\mathbf{G}_{cvi}$ are given by (3.13). The dimensions of $\mathbf{F}_{cv}$, $\mathbf{G}_{cv}$, $\mathbf{H}_{cv}$ are $(2r_{cv} \times 2r_{cv})$, $(2r_{cv} \times m)$ and $(p \times 2r_{cv})$.

To show the structure of $\mathbf{H}_{cv}$, $\mathbf{G}_{cv}$, let us assume first that the generating elements of dyads were on the main diagonal of $\psi_{cv}(z)\,\mathbf{Z}_{cv}(z)$. From the decomposition procedure and (3.13) follows that $\mathbf{H}_{cv}$, $\mathbf{G}_{cv}$ have the form:

$$\mathbf{H}_{cv} = \begin{bmatrix} 1 & 0 & & & & & \\ x & x & 1 & 0 & & & \\ x & x & x & x & & & \\ x & x & x & x & \ldots & 1 & 0 \\ \vdots & & \vdots & & & \vdots & \vdots \\ x & x & x & x & & x & x \end{bmatrix}; \quad \mathbf{G}_{cv} = \begin{bmatrix} x & x & \ldots & x \\ x & x & \ldots & x \\ 0 & x & \ldots & x \\ 0 & x & & x \\ & \vdots & & \\ & x & \ldots & x \\ & x & \ldots & x \end{bmatrix}$$

where x denotes a possible nonzero element.

We can see, that $\mathbf{H}_{cv}$ consists of r_{cv} block columns of dimension $(p \times 2)$, and $\mathbf{G}_{cv}$ consists of r_{cv} block rows of dimension $(2 \times m)$. Furthermore $\mathbf{H}_{cv}$ is in a certain "echelon" form.

If the dyad-generating elements are not on the main diagonal, the following rule is valid. Let the k-th generating element be $w^{(k)}_{i_k, j_k}$. Then the 1 in the k-th block column of $\mathbf{H}_{cv}$ will be in the i_k-th row, and the first nonzero element in the k-th block row of $\mathbf{G}_{cv}$ will be in the j_k-th column. So $\mathbf{H}_{cv}$ will remain in "echelon" form, but the block rows of $\mathbf{G}_{cv}$ will permute.

What remained is to show that the realisations $(\mathbf{F}_{cv}, \mathbf{G}_{cv}, \mathbf{H}_{cv})$ of $\mathbf{Z}_{cv}(z)$ with dimension $n(\mathbf{Z}_{cv}) = 2r_{cv}$ are minimal. It can be performed e.g. by examining the rank of the observability and controllability matrix of these realisations, utilising the special structure of $\mathbf{F}_{cv}$ and that the rank $\mathbf{H}_{cv} \geq r_{cv}$ and

rank $\mathbf{G}_{cv} \geq r_{cv}$. The proof for the observability matrix is simple, but for the controllability matrix is somewhat more involved.

Summarising the previous results, we are now able to construct a minimal realisation of $\mathbf{Z}(z)$ as the direct sum of realisations $(\mathbf{F}_{r\mu}, \mathbf{G}_{r\mu}, \mathbf{H}_{r\mu})$, $(\mathbf{F}_{cv}, \mathbf{G}_{cv}, \mathbf{H}_{cv})$ $\mu = 1, \ldots, n_r$, $\nu = 1, \ldots, n_c$. The dimension of this realisation is given by

$$n(\mathbf{Z}) = \sum_{\mu=1}^{n_r} n(\mathbf{Z}_{r\mu}) + \sum_{\nu=1}^{n_c} n(\mathbf{Z}_{cv}) = \sum_{\mu=1}^{n_r} r_{r\mu} + 2 \sum_{\nu=1}^{n_c} r_{cv}.$$

Example. To illustrate how this procedure works, an example will be shown. Suppose we have already the partial fraction expansion of $\mathbf{Z}(z)$ as:

$$\mathbf{Z}(z) = \left(z - \frac{1}{2}\right)^{-1} \begin{bmatrix} 2 & 3 & 4 \\ 6 & 4 & 1 \\ 3 & 5 & 2 \end{bmatrix} + (z^2 - z + 1)^{-1} \begin{bmatrix} 2z - 1 & z + \frac{1}{2} & 2z + 1 \\ z - \frac{1}{2} & \frac{3z}{2} + \frac{3}{4} & 2z + 1 \\ \frac{5z}{2} & -\frac{17z}{2} + \frac{8}{3} & \frac{7z}{6} + \frac{35}{6} \end{bmatrix} =$$

$$= \mathbf{Z}_r(z) + \mathbf{Z}_c(z).$$

Clearly $\psi_r(z) = z - \frac{1}{2}$; $\psi_c(z) = z^2 - z + 1$. Now using the minimal dyadic decomposition as described in Section 3 we get:

$$\psi_r(z)\,\mathbf{Z}_r(z) = \begin{bmatrix} 1 \\ 3 \\ \frac{3}{2} \end{bmatrix} [2 \;\; 3 \;\; 2] + \begin{bmatrix} 0 \\ 1 \\ -\frac{1}{10} \end{bmatrix} [0 \;\; -5 \;\; -11] + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \left[0 \;\; 0 \;\; -\frac{51}{10}\right]$$

which shows that $r_r = \operatorname{rank} \psi_r(z)\,\mathbf{Z}_r(z) = 3$. From (3.6) follows, that

$$\mathbf{F}_r = \begin{bmatrix} \frac{1}{2} & & \\ & \frac{1}{2} & \\ & & \frac{1}{2} \end{bmatrix} \quad \mathbf{G}_r = \begin{bmatrix} 2 & 3 & 4 \\ 0 & -5 & -11 \\ 0 & 0 & -\frac{51}{10} \end{bmatrix} \quad \mathbf{H}_r = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ \frac{3}{2} & -\frac{1}{10} & 1 \end{bmatrix}.$$

The next step is to construct a minimal dyadic decomposition of $\mathbf{W}(z) = \psi_c(z)\mathbf{Z}_c(z)$. We find that the first generating element is $2z - 1$, its inverse

is $-\frac{2}{3}z+\frac{1}{3}$. Subtracting the first dyad from $\mathbf{W}(z)$, the second generating element will be $\overset{(2)}{w}_{22}(z)=z-1$; its inverse is $\overset{(2)}{w}_{22}^{-1}=-z$. Subtracting the second dyad generated by $\overset{(2)}{w}_{22}(z)$ we find that $\overset{(3)}{\mathbf{W}}(z)=0$, and $r_c=\operatorname{rank}\psi_c(z)\,\mathbf{Z}_c(z)=2$ over $R[z]/\psi_c(z)$.

The resulting dyadic decomposition is:

$$\psi_c(z)\,\mathbf{Z}_c(z)=\begin{bmatrix}1\\ \frac{1}{2}\\ -\frac{5}{6}z+\frac{5}{3}\end{bmatrix}\begin{bmatrix}2z-1 & z+\frac{1}{2} & 2z+1\end{bmatrix}+$$

$$+\begin{bmatrix}0\\ 1\\ -2z+1\end{bmatrix}\begin{bmatrix}0 & z-1 & z+\frac{1}{2}\end{bmatrix}.$$

Using equation (3.12) the minimal realisation of $\mathbf{Z}_c(z)$ is given by

$$\mathbf{F}_c=\begin{bmatrix}0&1&&\\-1&1&&\\&&0&1\\&&-1&1\end{bmatrix}\quad \mathbf{G}_c=\begin{bmatrix}2&1&2\\1&\frac{3}{2}&3\\0&1&1\\0&0&\frac{3}{2}\end{bmatrix}\quad \mathbf{H}_c=\begin{bmatrix}1&0&0&0\\\frac{1}{2}&0&1&0\\\frac{5}{3}&-\frac{5}{6}&1&-2\end{bmatrix}.$$

The minimal realisation of $\mathbf{Z}(z)$ is the direct sum of realisations $(\mathbf{F}_r, \mathbf{G}_r, \mathbf{H}_r)$ and $(\mathbf{F}_c, \mathbf{G}_c, \mathbf{H}_c)$:

$$\mathbf{F}=\begin{bmatrix}\frac{1}{2}&&&&&\\&\frac{1}{2}&&&&\\&&\frac{1}{2}&&&\\&&&0&1&\\&&&-1&1&\\&&&&&0&1\\&&&&&-1&1\end{bmatrix}\quad \mathbf{G}=\begin{bmatrix}2&3&4\\0&-5&-11\\0&0&-\frac{51}{10}\\2&1&2\\1&\frac{3}{2}&3\\0&1&1\\0&0&\frac{3}{2}\end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & \frac{1}{2} & 0 & 1 & 0 \\ \frac{3}{2} & \frac{1}{10} & 1 & \frac{5}{3} & -\frac{5}{6} & 1 & -2 \end{bmatrix}.$$

The dimension of this minimal realisation is $n = r_r + 2r_c = 7$.

Conclusions

We have derived a minimal realisation procedure for multivariable systems, using the partial fraction representation of rational transfer matrix having real or complex poles of multiplicity one. The method is based on minimal dyadic decomposition of the partial fraction matrices. In an earlier paper [6] we showed the so-called ESS representation based on partial fraction expansion of scalar transfer function can be successfully applied in the identification of the structure and parameters of single variable systems. This identification procedure can be generalized for multivariable case if the representation derived in this paper is used. The simplicity of this realisation procedure offers an easy application on digital computer. A list of a FORTRAN program is available by the authors.

Acknowledgements

The authors are grateful to Dr. L. Keviczky who called their attention to the usefulness of this representation in multivariable system identification.

References

1. *Gilbert, E. G.*, Controllability and observability in multivariable control systems J. SIAM Control, Ser. A. Vol. **2,** No. *1,* 1963
2. *Kalman, R. E.*, Irreducible realisations and the degree of a rational matrix. J. Soc. Indust. Appl. Math. Vol. **13,** No. *2,* 1965
3. *Ho, B. L.—Kalman, R. E.*, Allerton Conf. on Circuit and System theory. p. 449, 1965
4. *Mayne, D. Q.*, Computational procedure for a minimal realisation of transfer function matrices. Proc. IEEE. **115,** p. 1363, 1968
5. *Rosenbrock, H. H.*, Computation of minimal realisation of transfer function matrices. Proc. IEEE. **115,** p. 325, 1968
6. *Keviczky, L.—Bokor, J.—Bányász, Cs.*, A new identification method with special parametrization for model structure determination. Fifth IFAC Symp. on Ident. and System. Par. Est., Darmstadt, 1979

Метод получения минимальной реализации передаточной матрицы в многомерных системах автоматического управления

Й. БОКОР, Й. ПЕТО

(Будапешт)

В работах Гильберта [1] и Калмана [2] были разработанны методы получения минимальной реализации систем многих переменных, использующие разложение рациональной матрицы переноса в сумму элементарных дробей. В данной статье приводится метод для получения минимальной реализации матрицы переноса, имеющей однократные действительные и комплексные корни, исходя из формы разложения её в сумму элементарных дробей. С одной стороны метод является далнейшим развитием метода Гильберта, а с другой, из-за ограничения однократности корней, его можно считать менее общим, чем метод Кальмана.

Процедура основана на минимальном диадическом разложении матриц в сумму элементарных дробей и поэтому она намного проще метода Калмана.

Эта простота даёт возможность широкого применения метода на ЭВМ и во многих практических случаях.

J. Bokor, J. Pető
Institute for Management and Computer Technique
Csepel Iron and Metal Works
Budapest
Hungary

Problems of Control and Information Theory, Vol. 10 (6), pp. 417—425 (1981)

A SIMPLE ASYNCHRONOUS RESERVATION PROTOCOL FOR A MULTIPLE ACCESS CHANNEL

L. JEREB, CS. SZABÓ

(Budapest)

(Received February 1, 1981)

A simple reservation protocol is proposed according to which reservations are made asynchronously on the basis of simple ALOHA and positive acknowledgements are used for indicating the beginning and the end of a particular reserved interval for all users. A heuristic approach allows to obtain an expression for channel utilization. Simulation study is given investigating the channel utilization and the total packet delay. It is demonstrated that for a set of system parameters of practical interest a channel utilization of about 0.6 can be achieved with acceptable packet delay performance.

Introduction

The basis idea of reservation algorithms is the following. The random access channel is divided (either in frequency or in time) into two parts, the reservation and the "information" subchannels. A particular user ready to transmit an information packet at some instant first bids for the channel sending a reservation (request-to-send) packet via the reservation subchannel. The reservation subchannel is usually a random access channel but may be, of course, a fixed time-division channel, too. Information packets are transmitted via the information subchannel and cannot be disturbed by other packets.

Many reservation protocols are described in the literature and we refer here only to some of them. The first proposals were made by Crowther et al. [1] and by Roberts [2]. Both are synchronous in the sense that the time axis is divided into slots and the slots are organized into frames.

Kleinrock and Tobagi [3] proposed and investigated a "dynamic" reservation technique using a separate frequency-divided channel for reservations. Borgonovo and Fratta [4] considered also a time divided synchronous channel but they proposed to make reservations only when collisions occur.

Tsybakov and Berkovsky in [5] proposed and investigated a time divided, synchronous protocol in which one of Tsybakov's conflict-resolution-type protocols was used together with reservation.

Reservation protocols permit to improve significantly the performance

of the simple random access protocols and asymptotically, to provide a channel utilization of 1. They may be either stable or unstable algorithms according to the basic random access protocol, which is used for reservations.

The protocol

It is assumed in the sequel that a radio channel C_1 is available for communication from the population of users to the central station. This is a multiple access channel. A channel C_2 carries the traffic from the central station to the users. This is a broadcast channel (see Fig. 1).

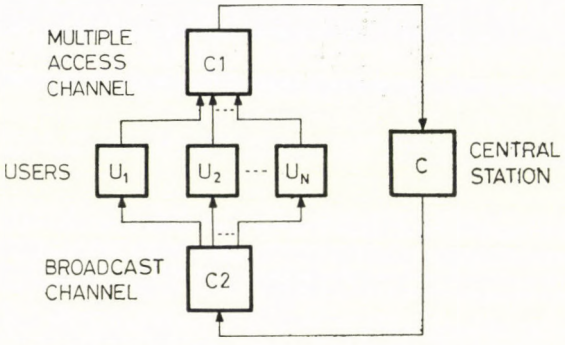


Fig. 1. The system model

The protocol is formulated as follows.

1) A user is "ready" when it has a message packet in its buffer of length L bits, $L-h$ of which are information bits and h bits contain address information and some bits serve for error detection. This latter is called head. A ready user first transmits a (possibly short) request-to-send (reservation-) packet of length T bits (these packets also contain header and error detection bits) on the C_1 channel. Request packets are allowed to be sent at any instant like the pure (asynchronous) Aloha scheme.

2) The attempt of a user may be successful if no other user transmit request packet at the same time (no arrivals occur within the interval of length T before and after the start time of a given packet). Otherwise the attempt is unsuccessful.

3) The central station transmits an acknowledgement packet on the broadcast channel C_2 after receiving a successful request packet. If a user receives no acknowledgement within an appropriate waiting interval, it retransmits the request packet according to random delay.

4) The acknowledgement of the central station

(i) informs the given user that the channel is reserved from this moment and

(ii) informs the others that they are not allowed to transmit during the reserved time interval.

They have to reschedule their attempts according to a random delay having appropriate distribution function.

The protocol is illustrated by time diagram in Fig. 2.

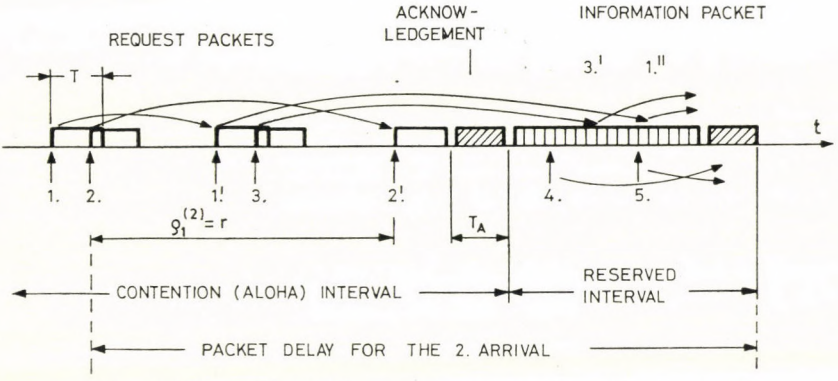


Fig. 2. Illustration to the asynchronous reservation protocol

Channel utilization

Wishing to analyze the reservation protocol described above, one meets the same or similar difficulties as in the case of a simple Aloha protocol [6]. Till now, no explicit expression has been found for the channel utilization factor and the packet delay distribution.

We show here one of the possible heuristic approaches which permits us to get a simple expression for the channel utilization.

Denote by λ the intensity of generation of new packets, in packets per seconds. Then the channel utilization for request packets is equal to λT which is denoted by S_R.

We can take into account the fact that during the reserved intervals no requests can be made (and consequently, served) in the following way: we can calculate with an intensity factor of

$$\lambda' = \frac{\lambda}{1 - \lambda L} \tag{1}$$

where the duration of reserved intervals, L is constant for the sake of simplicity.

In this case we determine the channel utilization for request packets as

$$S'_R = \lambda' T = \frac{\lambda T}{1 - \lambda L} \tag{2}$$

and the corresponding traffic in the Aloha-intervals can be written as

$$G_R = S'_R \cdot e^{2G_R}. \tag{3}$$

From this expression, the channel utilization factor for message packets which we denote by $S = \lambda \cdot L$ is given by the following formula:

$$S = \frac{G_R \cdot e^{-2G_R}}{a + G_R \cdot e^{-2G_R}} \quad \text{where} \quad a = \frac{T}{L}. \tag{4}$$

This expression is plotted in Fig. 5 by dashed line.

Packet delay performances

The packet delay time is defined as the time interval between the moment when a particular user becomes ready and the end of the transmission of this user's message packet. The packet delay which we denote by ϑ is a random variable and is determined by the following formula:

$$\vartheta = T + L + 2T_A + \sum_{i=1}^{\nu} \varrho_i \tag{5}$$

where ν is the random number of retransmissions, ϱ_i, $i = 1, 2, \ldots, \nu$ are the random delays having the distribution function $F_\varrho(t)$ and T_A is the waiting time for acknowledgement. (The first three terms obviously represent the minimum of the packet delay. Any propagation and processing time is neglected.)

In Fig. 2 we illustrated the packet delay time for the arrival denoted by 2, for which $d^{(2)} = T + L + 2T_A + \varrho_1^{(2)}$. Unfortunately, Eq. (5) does not permit to calculate the packet delay, for example, the channel utilization factor as a function of S, because the number of retransmissions is not independent of the retransmission time distribution.

The simulation method

A special-purpose simulation package for studying of asynchronous concurrent processes was developed. The most important properties which can be studied by this method are as follows:

(i) the behaviour of the system is determined by a finite number of subprocesses,

(ii) each subprocess resides in a state of an own finite space at a given moment,

(iii) the subprocesses have no direct effect on each other, but they collectively have a deterministic or stochastic effect on the direction of their change of state and on the distribution of occupancy time in the new state.

On the basis of these conditions, an event oriented simulation program was built up by means fo the E-net representation. This method is very useful for the description and analysis of asynchronous concurrent systems and is explained in detail in [7].

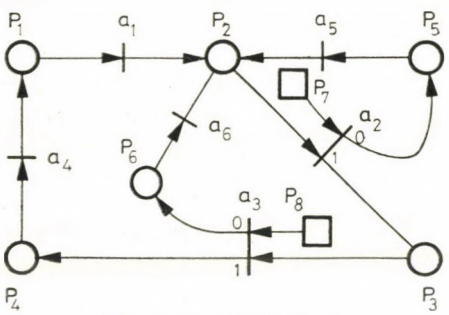


Fig. 3. Terminal E-net

Employing the graph model of E-nets, a special state diagram was derived for each terminal (see Fig. 3). It contains three kinds of nodes and directed edges. The nodes represented by circles ($P_1 - P_6$) define conditions for the operation of system and are called locations. This condition may exist for a period of time, and a marking of the place called token represents an existing condition. A location has at most one edge directed in and one edge directed out, and the tokens may move along directed paths shown by arrows.

Transitions are the nodes in the net at which the determination of token flow and modification takes place, they are represented by bars ($a_1 - a_6$). A deterministic or stochastic time delay belongs to each transition. The square places ($P_7 - P_8$), called resolution location, work like switches and depending on the result of the resolution procedure shunt the token from the input location of the transition to one of the output locations.

In Fig. 3, the notations of locations and transitions are the following:

P_1: the terminal may generate a message
P_2: the message is ready to be sent

P_3: the request packet may be sent
P_4: the message packet may be sent
P_5: the channel is occupied
P_6: the request packet was unsuccessful
P_7: resolution location; the setting of the switch is 1, if no message packet is on the channel
P_8: resolution location; the setting of the switch is 1, if the request packet was successful
a_1: the terminal generates a message
a_2: the terminal tests channel state
a_3: the terminal transmits a request packet and waits for an acknowledgement
a_4: the terminal sends a message packet
a_5: the terminal delays the beginning of retransmission of the request packet waiting for the end of the reserved period
a_6: the terminal delays the beginning of retransmission of request packet getting no acknowledgement.

According to the simulation conditions only the changes of the terminal states may modify the resolution results. It means that we have the possibility to represent the resolution conditions by means of a special state diagram (see Fig. 4).

In Fig. 4 the places may correspond to the states of the channel and the edges may correspond to the locations of Fig. 3. The meanings of the places in Fig. 4 are:

R_1: the channel is empty
R_2: a successful request packet is transmitted on the channel
R_3: a message packet is transmitted
R_4: an unsuccessful request packet is on the channel
$R_5, R_6, \ldots$: two, three, etc. request packets are transmitted.

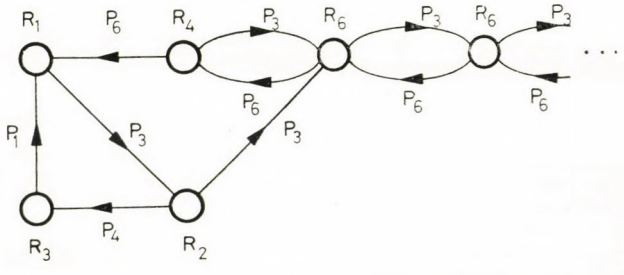


Fig. 4. Representation of the change of the resolution condition

Simulation results

In the course of experiences the following characteristics were chosen:

- the distribution of the generation time of new messages (a_1 transition) was exponential,
- the transmission time for both the request and the message packets, respectively were fixed (a_3, a_4 transition) with $a = 1/15$,
- the distributions of retransmission delays (a_5 and a_6 transition) were identical,
- the terminals were uniform.

Modifying the parameters of distributions and the number of terminals the following results were derived:

a) The channel utilization factor was determined at various intensity of generation of new packets and at various mean retransmission delays (see Fig. 5).

In Fig. 5 the dotted line shows the results of the heuristic approach. It is well discernible that the mean retransmission delay is long enough ($M\varrho > 10\ T$) as compared to the request packet time, the simulation results well correspond to the heuristic approach.

b) In Fig. 6 packet delay performances are shown at various conditions. It is visible that the protocol can provide good results in a wide range of channel utilization although at short retransmission delay and increasing traffic the system may become unstable.

c) The effect of the number of terminals was examined as well, at constant mean message generation time to terminal number ratio. It could be ascertained that over 16 terminals this effect on the channel traffic and time delay was negligible (1–13%) even at high channel utilization.

(Remark: each point of the simulation results was obtained from the data of at least two thousand successful messages.)

Conclusions

The proposed asynchronous reservation-Aloha algorithm is a natural combination of the reservation principle with the simple Aloha protocol. It has advantages against the reservation protocols referred above because of its simplicity in implementation. (There is no need in either slotting the time axis and organizing frames, or dividing the frequency band of the radio channel into parts.)

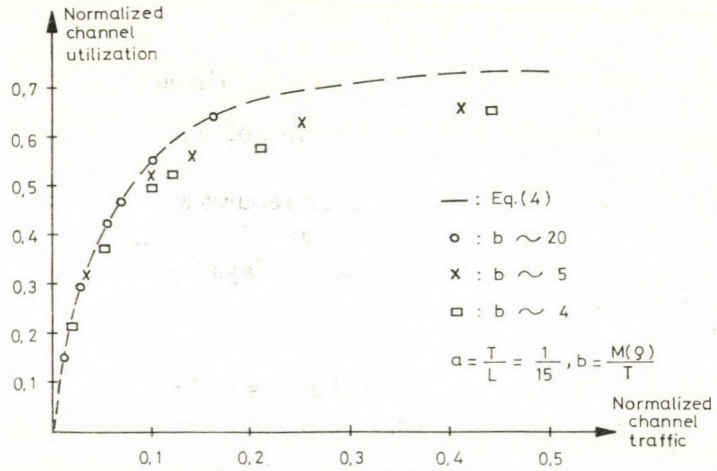


Fig. 5. The channel utilization factor

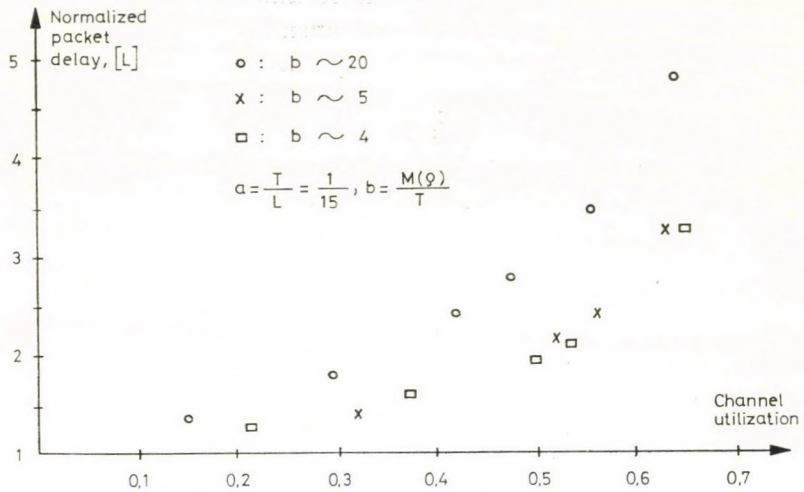


Fig. 6. Simulation results for packet delay

Simulation studies show that for a set of system parameters of practical interest a channel utilization factor of about 0.6 can be achieved with acceptable packet delays.

One has to emphasize that this system can be unstable like the simple Aloha system, so some kind of control is needed to maintain stability at larger channel utilization.

References

1. *Crowther, W. et al.*, "A system for Broadcast Communication: Reservation Aloha" Proc. 6th Hawaii International Conference on System Sciences, Jan. 1973.
2. *Roberts, L. G.*, "Dynamic Allocation of Satellite Capacity Through Packet Reservation", AFIPS Conf. Proc., **42,** June 1973.
3. *Kleinrock, L., Tobagi, F.*, Packet Switching in Radio Channels: Part III — Polling and (Dynamic) Split-Channel Reservation Multiple Access, IEEE Tr., Vol. **COM-24,** No. *8*, August 1976.
4. *Borgonovo, F., Fratta, L.*, A New Technique for Satellite Broadcast Channel Communication, Proc. Fifth Data Communication Symposion, Snowbird, Sept. 1977.
5. *Tsybakov, B. S., Berkovsky, M. A.*, Mnozestvenniy dostup s reservirovaniem. Problemy peredachi informacii, 1980. No. *1*.
6. *Abramson, N.*, Throughput of Packet Broadcasting Channels. IEEE Tr., Vol. **COM-25,** No. *1*, January 1977.
7. *Noe, J. D., Nutt, G. J.*, Macro *E*-nets for Representation of Parallel Systems, IEEE Tr. on Computers, Vol. **C-22,** No. *8*, August 1973.

Простой асинхронный алгоритм множественного доступа с резервированием

Л. ЙЕРЕБ, Ч. САБО

(Будапешт)

Предлагается простой алгоритм с резервированием, по которому пакеты-заявки передаются по каналу в соответствии с простым алгоритмом ALOHA, и положительные квитанции используются для индикации начала и конца данного резервированного периода. Эвристическим путем найдено выражение для коэффициента использования канала в равновесном состоянии. Проведены исследования коэффициента использования и времени задержки пакетов с применением специального метода моделирования. Показано, что для целесообразных с точки зрения практики значений параметров при коэффициенте использования канала около 0,6 средняя задержка пакетов получается весьма приемлемой (в 3–4 раза больше времени передачи одного пакета).

L. Jereb
Cs. Szabó
Technical University of Budapest
H-1111 Budapest
Stoczek u. 2.

Problems of Control and Information Theory, Vol. 10 (6), pp. 427—437 (1981)

AN ALGORITHM FOR MAXIMIZING A FINITE SUM OF POSITIVE FUNCTIONS AND ITS APPLICATION TO CLUSTER ANALYSIS

J. GRIM

(Prague)

(Received January 10, 1981)

An iterative procedure for maximizing a finite sum of positive functions is described which may be used in a wide variety of problems [1]. The applicability of this algorithm depends on the form of the positive functions. Roughly speaking, the computation is efficient in case of exponential — or product functions. The algorithm converges monotoneously to some possibly local maximum and may be used also in a discrete case.

1. Introduction

The original idea of the described algorithm is due to Shlesinger [6]. Estimating the parameters of general mixtures of probability density functions he suggested an iterative scheme for maximizing the corresponding likelihoop function. An analogous scheme has later been described also by other authors cf. [2], [5], [7], but only for a special case of mixture of normal densities. Unlike these papers, Shlesinger proved that the maximized likelihood functiou increases in each iteration.

The present algorithm may be viewed, in certain sense, as a generalizəp form of Shlesinger's algorithm. Its properties are formulated and proved ɪu a different way and the principle of the algorithm is compared with the gradieuɥ method which is shown to be qualitatively different.

2. Formulation of the algorithm

Let

$$f_m(\mathbf{x}) > 0;\ \mathbf{x} = (x_1, \ldots, x_d)^T \in R_d;\quad m = 1, \ldots, M \tag{2.1}$$

be positive functions defined in a d-dimensional real space R_d. We denote by

$$f(\mathbf{x}) = \sum_{m=1}^{M} f_m(\mathbf{x}) \tag{2.2}$$

the sum of these functions which is to be maximized. Obviously the local

maxima of $f(\mathbf{x})$ are identical with that of the function

$$Q(\mathbf{x}) = \ln f(\mathbf{x}) = \ln\left[\sum_{m=1}^{M} f_m(\mathbf{x})\right]. \tag{2.3}$$

A point $\mathbf{x}^* \in R_d$ of absolute maximum of $f(\mathbf{x})$ may be therefore characterized equivalently by the inequality

$$Q(\mathbf{x}^*) \geq Q(\mathbf{x}); \quad \forall\, \mathbf{x} \in R_d \tag{2.4}$$

which, in case of local maximum, is satisfied only in some neighbourhood of $\mathbf{x}^*$.

To make the principle of the algorithm more clear we express first the function $Q(\mathbf{x})$ in a special form [6]. Denoting W_M the set of all M-dimensional weight vectors

$$W_M = \left\{\mathbf{w} = (w_1, \ldots, w_M)^T \in R_M : \sum_{m=1}^{M} w_m = 1;\ w_m \geq 0\right\} \tag{2.5}$$

we can write

$$Q(\mathbf{x}) = \sum_{m=1}^{M} w_m \ln f_m(\mathbf{x}) - \sum_{m=1}^{M} w_m \ln \frac{f_m(\mathbf{x})}{f(\mathbf{x})} \tag{2.6}$$

for any $\mathbf{x} \in R_d$ and $w \in W_M$. Denoting further by $\mathbf{p}(\mathbf{x})$ the weight vector

$$\mathbf{p}(\mathbf{x}) = (p(1\,|\,\mathbf{x}), \ldots, p(M\,|\,\mathbf{x}))^T \in W_M;\ p(m\,|\,\mathbf{x}) = \frac{f_m(\mathbf{x})}{f(\mathbf{x})} \tag{2.7}$$

and making the substitution $\mathbf{w} = \mathbf{p}(\mathbf{x})$ in (2.6) we obtain

$$Q(\mathbf{x}) = \sum_{m=1}^{M} p(m\,|\,\mathbf{x}) \ln f_m(\mathbf{x}) - \sum_{m=1}^{M} p(m\,|\,\mathbf{x}) \ln p(m\,|\,\mathbf{x}). \tag{2.8}$$

Let us recall that the terms in (2.6) and (2.8) are always defined because of assumption (2.1).

To avoid any ambiquity the algorithm described in the following will be called $\mathcal{A}_0$.

Algorithm $\mathcal{A}_0$

Step 0. Choose an initial value $\mathbf{x}_0 = (x_{01}, \ldots, x_{0d})^T \in R_d$.

Step 1. For a given vector $\mathbf{x}_t = (x_{t1}, \ldots, x_{td})^T \in R_d$; $(t = 0, 1, \ldots)$ compute the values $f_m(\mathbf{x}_t)$, $m = 1, \ldots, M$; $f(\mathbf{x}_t)$; $Q(\mathbf{x}_t)$ and the corresponding parameters

$$p(m\,|\,\mathbf{x}_t) = f_m(\mathbf{x}_t)/f(\mathbf{x}_t);\ m = 1, 2, \ldots, M. \tag{2.9}$$

Step 2. Using parameters (2.9) compute a new vector $\mathbf{x}_{t+1} \in R_d$ by the formula

$$\mathbf{x}_{t+1} = \arg \max_{\mathbf{x} \in R_d} \left\{ \sum_{m=1}^{M} p(m|\mathbf{x}_t) \ln f_m(\mathbf{x}) \right\} \tag{2.10}$$

and return to step 1.

Remark 2.1. We shall assume that the vector $\mathbf{x}_{t+1}$ maximizing the function of the form

$$q(\mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} w_m \ln f_m(\mathbf{x}); \ \mathbf{x} \in R_d; \ \mathbf{w} \in W_M \tag{2.11}$$

in (2.10) is uniquely defined or uniquely choiced when the maximum of $q(\mathbf{x}, \mathbf{w})$ is not unique. From equation (2.10) it follows

$$\sum_{m=1}^{M} p(m|\mathbf{x}_t) \ln f_m(\mathbf{x}_{t+1}) \geq \sum_{m=1}^{M} p(m|\mathbf{x}_t) \ln f_m(\mathbf{x}); \ \mathbf{x} \in R_d. \tag{2.12}$$

Obviously, an efficient use of the algorithm is possible only when the implicit form of the relation (2.10) may be replaced by a simple explicit formula. From the point of view of applicability it is an important limiting feature.

To characterize the principle of the algorithm let us note that the vector $\mathbf{p}(\mathbf{x}_t)$, computed in step 1, plays a role of an internal dependent parameter. Step 2 corresponds to the maximization of the expression (2.8) under fixed values of the parameter vector $\mathbf{p}(\mathbf{x})$.

The solution obtained by an iterative scheme represents usually a value which does not change in further iterations. In this sense we introduce

Definition 2.1. A point $\mathbf{x}^* \in R_d$ satisfying the equation

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in R_d} \left\{ \sum_{m=1}^{M} \frac{f_m(\mathbf{x}^*)}{f(\mathbf{x}^*)} \ln f_m(\mathbf{x}) \right\} \tag{2.13}$$

will be called the stationary point.

By iterating steps 1 and 2, the algorithm $\mathfrak{A}_0$ generates three infinite sequences

$$\dot{\mathbf{x}} = \{\mathbf{x}_t\}_{t=0}^{\infty}; \ \dot{Q} = \{Q_t\}_{t=0}^{\infty}; \ \dot{\mathbf{p}} = \{\mathbf{p}_t\}_{t=0}^{\infty} \tag{2.14}$$

whereby, to simplify the notation, we denote

$$Q_t = Q(\mathbf{x}_t); \ \mathbf{p}_t = \mathbf{p}(\mathbf{x}_t); \qquad t = 0, 1, 2, \ldots$$

It is natural to expect that the limit of the sequence $\dot{\mathbf{x}}$, if it exists, simultaneously satisfies the equation of the stationary point. A simple sufficient condition of this property is given in the following theorem.

Theorem 2.1. Let $\mathbf{x}^* \in R_d$ be the limit of the seque nce $\dot{}$, i.e.

$$\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^*, \tag{2.15}$$

the functions $f_1(\mathbf{x}), \ldots, f_M(\mathbf{x})$ be continuous at $\mathbf{x}^*$ and the function $q(\mathbf{x}, \mathbf{p}(\mathbf{x}^*))$ has a unique maximum. Then the sequence $\dot{\mathbf{p}}$ has the limit

$$\lim_{t\to\infty} \mathbf{p}_t = \mathbf{p}(\mathbf{x}^*) = (p(1\,|\,\mathbf{x}^*), \ldots, p(M\,|\,\mathbf{x}^*))^T \in W_M \tag{2.16}$$

and $\mathbf{x}^*$ satisfies equation (2.13).

Proof. Considering the continuity of the functions $f_m(\mathbf{x})$ at $\mathbf{x}^*$, we obtain equation (2.16) from (2.9) and (2.15) since ($m = 1, 2, \ldots, M$)

$$\lim_{t\to\infty} p(m\,|\,\mathbf{x}_t) = \lim_{t\to\infty} f_m(\mathbf{x}_t)/f(\mathbf{x}_t) = p(m\,|\,\mathbf{x}^*). \tag{2.17}$$

Recall further that the inequality (2.12) holds for all t ($t = 0, 1, \ldots$). Combining (2.12) and (2.17) we obtain

$$\sum_{m=1}^{M} p(m\,|\,\mathbf{x}^*) \ln f_m(\mathbf{x}^*) \geq \sum_{m=1}^{M} p(m\,|\,\mathbf{x}^*) \ln f_m(\mathbf{x}); \quad \mathbf{x} \in R_d. \tag{2.18}$$

Since the function $q(\mathbf{x}, \mathbf{p}(\mathbf{x}^*))$ has a unique maximum, inequality (2.18) implies the equation (2.13) and the proof is complete.

In the following section we make use of a discrete version of an inequality known in information theory. We bring the proof here since it is brief. The corresponding general assertion may be found e.g. in [3].

Lemma 2.1. Any two discrete probability distributions $\mathbf{a} = (a_1, \ldots, a_M)^T \in$ $\in W_M$; $\mathbf{b} = (b_1, \ldots, b_M)^T \in W_M$ satisfy the inequality

$$\sum_{m=1}^{M} a_m \ln \frac{a_m}{b_m} \geq 0, \tag{2.19}$$

whereby the equality sign in (2.19) holds if and only if $\mathbf{a} = \mathbf{b}$.

Proof. Making the convention

$$0 \ln 0 = \lim_{\xi\to 0} \xi \ln \xi = 0 \tag{2.20}$$

we can consider only positive terms in the sum in (2.19). In other words, without any loss of generality we may assume $a_m > 0$ for all $m = 1, 2, \ldots, M$. If we denote

$$\alpha_m = \frac{b_m}{a_m}; \qquad m = 1, 2, \ldots, M \tag{2.21}$$

then, using Jensen's inequality, we obtain the relation

$$\sum_{m=1}^{M} a_m \ln \alpha_m \leq \ln \left[\sum_{m=1}^{M} a_m \alpha_m \right] = 0 \tag{2.22}$$

which may be rewritten in the form (2.19). Obviously the left-hand side of (2.19) is zero if $\mathbf{a} = \mathbf{b}$. Finally, since logarithm is a strictly concave function the left-hand side of (2.22) equals zero only if $\alpha_1 = \ldots = \alpha_M$, i.e. only if $\mathbf{a} = \mathbf{b}$.

3. Properties of the algorithm

We first formulate a relation between the stationary point and the classical solution obtained from the necessary conditions for a maximum.

Theorem 3.1. A stationary point $\mathbf{x}^* \in R_d$ satisfies the necessary condition for a maximum of $Q(\mathbf{x})$

$$\operatorname{grad} Q(\mathbf{x}) = 0 \tag{3.1}$$

provided that the corresponding derivatives exist.

Proof. Since the stationary point $\mathbf{x}^*$ satisfies eq. (2.13), it satisfies also the appropriate necessary condition

$$\sum_{m=1}^{M} \frac{f_m(\mathbf{x}^*)}{f(\mathbf{x}^*)} \operatorname{grad} [\ln f_m(\mathbf{x})]|_{\mathbf{x}=\mathbf{x}^*} = 0 \tag{3.2}$$

which may be rewritten in the form (3.1):

$$\frac{1}{f(\mathbf{x}^*)} \sum_{m=1}^{M} \operatorname{grad} f_m(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^*} = Q(\mathbf{x})|_{\mathbf{x}=\mathbf{x}^*} = 0. \tag{3.3}$$

The proof is complete.

The most important property of the algorithm $\mathfrak{A}_0$ is its monotoneous convergence, more precisely expressed by the following theorem.

Theorem 3.2. If $\dot{Q}, \dot{\mathbf{x}}$ and $\dot{\mathbf{p}}$ are sequences produced by the algorithm $\mathfrak{A}_0$ then it holds

$$Q_{t+1} - Q_t > 0; \qquad t = 0, 1, 2, \ldots \tag{3.4}$$

as long as $\mathbf{p}_{t+1} \neq \mathbf{p}_t$. Conversely, if $Q_{t+1} = Q_t$ then $\mathbf{p}_{t+1} = \mathbf{p}_t$.

Proof. By substituting $\mathbf{x} = \mathbf{x}_{t+1}$; $\mathbf{w} = \mathbf{p}_t$ in (2.6) we obtain

$$Q_{t+1} = \sum_{m=1}^{M} p(m|\mathbf{x}_t) \ln f_m(\mathbf{x}_{t+1}) - \sum_{m=1}^{M} p(m|\mathbf{x}_t) \ln p(m|\mathbf{x}_{t+1}). \tag{3.5}$$

Further, if we express Q_t in the form (2.8), we can write

$$Q_{t+1} - Q_t = \sum_{m=1}^{M} p(m \mid \mathbf{x}_t) \ln \frac{f_m(\mathbf{x}_{t+1})}{f_m(\mathbf{x}_t)} + \sum_{m=1}^{M} p(m \mid \mathbf{x}_t) \ln \frac{p(m \mid \mathbf{x}_t)}{p(m \mid \mathbf{x}_{t+1})} . \quad (3.6)$$

By the inequality (2.12) the first sum in (3.6) is non-negative. Hence we have

$$Q_{t+1} - Q_t \geq \sum_{m=1}^{M} p(m \mid \mathbf{x}_t) \ln \frac{p(m \mid \mathbf{x}_t)}{p(m \mid \mathbf{x}_{t+1})} . \quad (3.7)$$

By Lemma 2.1 the right-hand side of (3.7) is nonnegative for any two vectors $\mathbf{p}_t$, $\mathbf{p}_{t+1}$ and equals zero only if $\mathbf{p}_t = \mathbf{p}_{t+1}$. Concerning the second part of the assertion, note that both sums in (3.6) must equal zero if $Q_{t+1} = Q_t$. Consequently, the proof is complete since, by Lemma 2.1 it follows $\mathbf{p}_{t+1} = \mathbf{p}_t$.

As it may be seen from eq. (3.6) and the proof above, the increment of the function $Q(\mathbf{x})$ consists of two nonnegative parts. The first one may be written in the form

$$\sum_{m=1}^{M} p(m \mid \mathbf{x}_t) \ln \frac{f_m(\mathbf{x}_{t+1})}{f_m(\mathbf{x}_t)} = q(\mathbf{x}_{t+1}, \mathbf{p}_t) - q(\mathbf{x}_t, \mathbf{p}_t)$$

and is achieved in step 2 while computing the new vector $\mathbf{x}_{t+1}$. The second part is expressed as a measure of dissimilarity between $\mathbf{p}_t$ and $\mathbf{p}_{t+1}$ known as relative entropy or minimum discrimination information [2]. It is achieved in step 1 by introducing the new parameter vector $\mathbf{p}_{t+1}$.

Now, from the practical point of view it would be desirable to have sufficient conditions for the convergence of the sequence $\dot{\mathbf{x}}$ which is assumed in theorem 2.1. Note first that if $Q(\mathbf{x})$ is a bounded function then the sequence $\dot{Q}$ converges since it is nondecreasing by theorem 3.2. However, the convergence of $\dot{Q}$ does not generally imply that of the sequence $\dot{\mathbf{x}}$.

Remark 3.1. Without any change the algorithm $\mathcal{A}_0$ may be also used in a discrete case when the set of possible values of the vector $\mathbf{x}$ is finite. Only equation (2.13), defining the stationary point, and theorem 2.3 are then sufficient to express the simplified situation. The infinite sequence $\dot{\mathbf{x}}$ must then contain at least two vectors satisfying

$$\mathbf{x}_t = \mathbf{x}_{t+k}; \quad t \geq 0; \quad k \geq 1 \quad (3.8)$$

whereby it holds $Q_t = Q_{t+k}$. By theorem 3.2 the intermediate elements of $\dot{Q}$ are constant and the same is true for the sequence $\dot{\mathbf{p}}$. Thus if t_0 is the smallest index with property (3.8) then the sequence $\dot{\mathbf{x}}$ is constant for $t > t_0$. Simulta-

neously we can write

$$\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}_{t_0+1};\ \lim_{t\to\infty} Q_t = Q_{t_0};\ \lim_{t\to\infty} \mathbf{p}_t = \mathbf{p}_{t_0} \tag{3.9}$$

and equation (2.13) holds for $\mathbf{x}_t = \mathbf{x}_{t_0+1}$.

Without making any further assumptions the following assertion may be verified:

Theorem 3.3. If the sequence $\dot{Q}$ has a finite limit Q^*, i.e.

$$\lim_{t\to\infty} Q_t = Q^*;\ Q^* < \infty \tag{3.10}$$

then the sequence $\dot{\mathbf{p}}$ satisfies the necessary condition of convergence:

$$\lim_{t\to\infty} ||\mathbf{p}_{t+1} - \mathbf{p}_t|| = 0. \tag{3.11}$$

Proof. To prove the theorem we make use of the inequality

$$\sum_{m=1}^{M} p(m|\mathbf{x}_t) \ln \frac{p(m|\mathbf{x}_t)}{p(m|\mathbf{x}_{t+1})} \geq \frac{1}{4}\left[\sum_{m=1}^{M} |p(m|\mathbf{x}_{t+1}) - p(m|\mathbf{x}_t)|\right]^2 \tag{3.12}$$

derived in [3]. Combining (3.7) and (3.12) we can write

$$Q_{t+1} - Q_t \geq \frac{1}{4}\left[\sum_{m=1}^{M} |p(m|\mathbf{x}_{t+1}) - p(m|\mathbf{x}_t)|\right]^2 \geq \frac{1}{4}\,||\mathbf{p}_{t+1} - \mathbf{p}_t||^2. \tag{3.13}$$

The proof is complete since the left-hand side of (3.13) tends to zero by (3.10).

Property (3.11) may, under additional assumptions, imply the convergence of $\dot{\mathbf{x}}$, e.g. when the number of stationary points is finite. However, the proof of such theorems is tedious and their meaning is rather limited.

4. Comparison of algorithm $\mathcal{A}_0$ with the gradient method

The fundamental recursive formula of the gradient method is in our case given by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \gamma \operatorname{grad} Q(\mathbf{x})\big|_{\mathbf{x}=\mathbf{x}_t};\ \gamma > 0;\quad t = 0, 1, \ldots \tag{4.1}$$

Using relation (3.3), (4.1) may be rewritten in the following equivalent form

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \gamma\left[\sum_{m=1}^{M} p(m|\mathbf{x}_t) \operatorname{grad} (\ln f_m(\mathbf{x}))\big|_{\mathbf{x}=\mathbf{x}_t}\right]. \tag{4.2}$$

The parenthesized expression in (4.2) is the gradient of the function $q(\mathbf{x}, \mathbf{p}_t)$ with respect to $\mathbf{x}$ at the point $\mathbf{x} = \mathbf{x}_t$. Now, comparing equations (2.10) and

(4.2), we can see that the new vector $\mathbf{x}_{t+1}$ computed in (2.10) maximizes the function $q(\mathbf{x}, \mathbf{p}_t)$ whereas in (4.2) the same function is used only to compute the steepest direction. The maximization of $q(\mathbf{x}, \mathbf{p}_t)$ is, however, only a partial problem having a complex relation to the original function $Q(\mathbf{x})$. Probably, in a particular case, any of the methods above may appear to be better. Concerning the computational complexity we have a similar situation.

To emphasize the principal difference between the gradient method and algorithm $\mathcal{A}_0$ note that if both procedures start from the same point $\mathbf{x}_t$, the new points $\mathbf{x}_{t+1}$ in (2.10) and (4.2) will coincide only in a special case when the direction computed in (4.2) crosses the maximum of $q(\mathbf{x}, \mathbf{p}_t)$ and the step size has a proper value.

5. Application to nonparametric cluster analysis

One of the most exact approaches in cluster analysis is based on a statistical interpretation of the analysed set of points. If we assume that

$$\mathbb{S} = \{\mathbf{y}_1, \ldots, \mathbf{y}_M\} \subset R_d \tag{5.1}$$

is a set of realizations of a real random vector with an unknown density $p(\mathbf{x})$, then the clusters in $\mathbb{S}$ asymptotically ($M \to \infty$) correspond to the local maxima of $p(\mathbf{x})$. In this sense, if $f(\mathbf{x})$ is an estimate of $p(\mathbf{x})$, the clusters may be estimated e.g. in a way suggested by Schnell [11].

It is assumed that $f(\mathbf{x})$ is a differentiable function having a finite set of local maxima

$$\mathcal{C} = \{C_1, C_2, \ldots, C_K\} \subset R_d. \tag{5.2}$$

For each point $\mathbf{y}_m \in \mathbb{S}$ there is a curve (ascending trajectory) connecting $\mathbf{y}_m$ with a local maximum

$$\xi_m(\tau) \in R_d;\ 0 \leq \tau \leq \tau_m;\ \xi_m(0) = \mathbf{y}_m \in \mathbb{S};\ \ \xi_m(\tau_m) \in \mathcal{C} \tag{5.3}$$

and having at each point the direction of grad $f(\mathbf{x})$:

$$\frac{d\xi_m(\tau)}{d\tau} = \operatorname{grad} f(\mathbf{x})\big|_{\mathbf{x}=\xi_m(\tau)};\ \ \tau \in \langle 0, \tau_m). \tag{5.4}$$

Consequently, we have the following natural partition of $\mathbb{S}$ into clusters

$$\mathbb{S} = \bigcup_{k=1}^{K} \mathbb{S}_k;\ \mathbb{S}_k = \{\mathbf{y}_m \in \mathbb{S} : \xi_m(\tau_m) = C_k \in \mathcal{C}\}. \tag{5.5}$$

Possible ties occurring by zero gradient may be solved arbitrarily.

Note that the clusters obtained in (5.5) have asymptotic properties implied by the used density estimate. It is therefore advantageous to use asymptotically unbiased and consistent kernel estimate [10].

$$f(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \frac{1}{h} \varphi \left(\frac{\mathbf{x} - \mathbf{y}_m}{h} \right) \tag{5.6}$$

which can be made differentiable by a proper choice of the kernel φ. Obviously, in this case the number of local maxima is influenced by smoothing of the kernel estimate ($1 \leq k \leq M$). For this reason the smoothing parameter h should be optimized [9] or varied [11].

The trajectories ξ_m, $m = 1, \ldots, M$ may be constructed using gradient method [11] with an accuracy depending on the chosen step size. The application of algorithm $\mathcal{A}_0$ in this context is motivated by an obvious similarity between the sequence of points obtained by the gradient method and the sequence $\dot{\mathbf{x}}$ which also converges to a local maximum. Moreover, using the discrete version of algorithm $\mathcal{A}_0$ and a generalized notion of the kernel estimate [8], the described approach is applicable to binary data. The advantages of algorithm $\mathcal{A}_0$ are partly counter-balanced by the fact that there is no clear relation between the sequence $\dot{\mathbf{x}}$ and condition (5.4).

To illustrate the method of cluster analysis based on algorithm $\mathcal{A}_0$ we introduce the particular form of the fundamental relation (2.10) both in a continuous and a discrete case.

a) *Sum of normal densities*

First we choose the positive functions $f_m(\mathbf{x})$ in (2.2) in the normal form (cf. (5.6))

$$f_m(\mathbf{x}) = \frac{a_0}{M} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{y}_m)^T A (\mathbf{x} - \mathbf{y}_m) \right\}; \quad \mathbf{x} \in R_d; \ \mathbf{y}_m \in \mathcal{S} \tag{5.7}$$

where A is a positive definite matrix including the smoothing parameter h and a_0 the corresponding positive constant. The function $q(\mathbf{x}, \mathbf{w})$ (cf. (2.11)) which is to be maximized in step 2 of the algorithm $\mathcal{A}_0$ is then given by

$$q(\mathbf{x}, \mathbf{w}) = \sum_{m=1}^{M} w_m \ln f_m(\mathbf{x}) = \alpha - \frac{1}{2} \sum_{m=1}^{M} w_m (\mathbf{x} - \mathbf{y}_m)^T A (\mathbf{x} - \mathbf{y}_m). \tag{5.8}$$

Denoting

$$\alpha = \sum_{m=1}^{M} w_m \ln \frac{a_0}{M}; \quad \tilde{\mathbf{y}} = \sum_{m=1}^{M} w_m \mathbf{y}_m \tag{5.9}$$

we can write

$$\sum_{m=1}^{M} w_m(\mathbf{x}-\mathbf{y}_m)^T A(\mathbf{x}-\mathbf{y}_m) = (\mathbf{x}-\tilde{\mathbf{y}})^T A(\mathbf{x}-\tilde{\mathbf{y}}) + \sum_{m=1}^{M} w_m(\tilde{\mathbf{y}}-\mathbf{y}_m)^T A(\tilde{\mathbf{y}}-\mathbf{y}_m) \tag{5.10}$$

and it is easily verified that the function $q(\mathbf{x}, w)$ is uniquely maximized by $\mathbf{x} = \tilde{\mathbf{y}}$. Now we can write the corresponding particular form of equation (2.10)

$$\mathbf{x}_{t+1} = \sum_{m=1}^{M} p(m|\mathbf{x}_t)\,\mathbf{y}_m. \tag{5.11}$$

b) *Sum of Bernoulli distributions*

Let $\mathbf{x} = (x_1, \ldots, x_d)$ be a vector of binary variables $x_i \in \{0, 1\}$, $i = 1, 2, \ldots, d$ and $f_m(\mathbf{x})$ positive functions defined by

$$f_m(\mathbf{x}) = \frac{1}{M} \prod_{i=1}^{d} \lambda_i^{1-|x_i-y_{mi}|}(1-\lambda_i)^{|x_i-y_{mi}|}; \quad m = 1, \ldots, M \tag{5.12}$$

where $\boldsymbol{\lambda}$ is a vector of parameters satisfying

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d); \quad \frac{1}{2} \leq \lambda_i < 1. \tag{5.13}$$

The function $f(\mathbf{x})$ represents then an analogy of the nonparametric kernel estimate in the sense of the paper [8]. Using definition (2.11) we can write

$$q(\mathbf{x}, \mathbf{w}) = -\ln M + \sum_{i=1}^{d} \ln \lambda_i + \sum_{i=1}^{d}\left[\sum_{m=1}^{M} w_m |x_i - y_{mi}|\right] \ln \frac{1-\lambda_i}{\lambda_i}. \tag{5.14}$$

If we define

$$\psi(\zeta) = \begin{cases} 1; & \zeta \geq 0 \\ 0; & \zeta < 0 \end{cases} \tag{5.15}$$

then it follows from (5.14) that the function $q(\mathbf{x}, \mathbf{w})$ is maximized when

$$x_i = \psi\left\{\sum_{m=1}^{M} w_m y_{mi} - \frac{1}{2}\right\}; \quad i = 1, 2, \ldots, d. \tag{5.16}$$

The particular form of equation (2.10) corresponding to (5.12) is given by

$$x_{t+1,i} = \psi\left\{\sum_{m=1}^{M} p(m|\mathbf{x}_t)\, y_{mi} - \frac{1}{2}\right\}; \quad i = 1, \ldots, d. \tag{5.17}$$

Both special versions of the algorithm $\mathfrak{A}_0$ obtained by introducing formulas (5.11) and (5.17), respectively have been succesfully tested using real data.

6. Conclusion

Concerning the applicability of algorithm $\mathcal{A}_0$ let us note that each function $f_m(\mathbf{x})$ may be defined as a product of functions of the type (5.12) and (5.7).

The maximized function $f(\mathbf{x})$ may also be given as a finite product of functions of the form (2.2). As an example we can consider the likelihood function of a mixture of densities [6]. The corresponding eq. (2.10) has then a more complex form since a sum of functions (2.11) is to be maximized. However, in this particular case, the solution may be decomposed into independent problems of the form (2.10).

References

1. *Beneš, J.*, Teorie systémů. (System Theory). Academia, Praha (1974).
2. *Day, N. E.*, Estimating the components of a mixture of normal distributions. Biometrika, **56,** 463—474 (1969).
3. *Kullback, S.*, Information Theory and Statistics. New York, Dover 1968.
4 *Kullback, S.*, SIAM Control **4,** 454—459 (1966).
5. *Peters, B. C., Walker, H. F.*, An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. SIAM Journal on Applied Mathematics, **35,** *2*, 362—378 (1978).
6. *Shlesinger, M. I.*, Vzaimosvjaz obutcheniya i samoobutcheniya v rozpoznavanii obrazov. (On the relation between learning and nonsupervised learning in pattern recognition.) Kibernetika, *2*, 81—88 (1968).
7. *Wolfe, J. H.*, Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research. **5,** 329—350 (1970).
8. *Aitchinson, J., Aitken, C. G. C.*, Multivariate binary discrimination by the kernel method. Biometrika, **63,** 413—420 (1976).
9. *Duin P. W.*, On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans. on computers **C-25,** (1976), *11*, 1175—1179.
10. *Parzen, E.*, On the estimation of a probability density function and the mode. Annals of Mathematical Statistics, **33,** (1962), 1065—1076.
11. *Schnell, P.*, Eine Methode zur Auffindung von Gruppen. Biometrische Zeitschrift, **6,** *1*, 47—48 (1964).

Алгоритм для максимизацим конечной суммы позитивных функций и его применение для кластерного анализа

И. ГРИМ

(Прага)

В работе описывается итеративный алгоритм для нахождения максимума суммы позитивных функций. Его применимость или вычислительная эффективность строго зависят от вида максимизируемых функций. Алгоритм сходится монотонно в точке абсолютного или локального максимума и применим также в дискретном случае. Рассматривается применение алгоритма для кластерного анализа и его взаимосвязь с градиентным методом.

J. Grim
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
Praha 8, Pod vodárenskou věží 4
Czechoslovakia

BOOK REVIEW

Distributed Parameter Systems: Modelling and Identification
Edited by **A. V. Balakrishnan** and **M. Thoma**, Proceedings of the IFIP Working Conference, Rome, Italy, June 21–24, 1976. vol. **1** in "Lecture Notes in Control and Information Sciences", Springer Verlag, Berlin–Heidelberg–New York

This book contains the papers presented at the Working Conference on Distributed Parameter Systems: Modelling and Identification held in Rome (Italy), June 21–24, 1976. This Conference was the first dealing with D. P. S.

The Conference was mainly devoted to recent advances in the following research areas: mathematical description of D. P. S. and their structural properties; identification, with practical applications; filtering and state estimation; control problems concerning the practical implementation of control schemes. A feature of the Conference was the participation of both theoretical and application oriented researchers.

This volume can be divided into two main parts. In the first part there are three survey lectures on different topics:

- Identification of Distributed Parameter Systems: Non-Computational Aspects (by A. V. Balakrishnan),
- Some Aspects of Modelling Problems in Distributed Parameter Systems (by J. L. Lions),
- Numerical Implementation of Distributed Parameter Filters with Application to Problems in Air Pollution (by J. H. Seinfeld and M. Koda).

The second part of the book contains 23 research reports presented during the Conference. These papers give a glimpse into the recent state of research in the field of distributed parameter systems. The papers in this part of the book cover a wide range in modeling, identification and control of distributed parameter systems. Consequently the volume is recommended for experts and researchers dealing with such problems.

Miklós Vajta, Jr.

Lecture Notes in Computer Science **76**
Codes for Boundary-Value Problems in Ordinary Differential Equations
Proceedings of a Working Conference, May 14–17, 1978
Edited by **B. Childs, M. Scott, J. W. Daniel, E. Denman** and **P. Nelson**

The first survey paper by J. W. Daniel explains briefly what each of a variety of methods for two-point boundary-value problems is, how methods relate to one another, and where are the difficulties today that stimulate interesting research problems. The author viewed each method as a solution technique for some discrete model of a transformed problem. He indicates in what presentations included in the proceedings one can find more details on these questions. The second paper by H. A. Watts looks at initial value integrators and their role in BVP codes. In particular, it examines initial value methods, describes algorithmic matters which are important for the design of quality

software, mentions some available codes. The state-of-the-art of solving nonlinear equations arising in BVP codes is surveyed in the third paper. It is pointed out that the structure of the underlying BVP should be transparent in the selected realization of the NLEQ algorithm. The Newton's method is examined thoroughly. The superposition-orthonormalization procedure for solving linear two-point BVP which forms the basis of the code SUPORT is prosented by M. R. Scott and H. A. Watts. The paper by G. W. Reddien is intended to be a survey of some of the theoretical developments on projection methods and to provide a general introduction to them. The mesh selection considerations given by R. D. Russel are important to understand the different efficiency of different computer implementations. The invariant imbedding algorithms, namely the state-transition and transfer matrix algorithms are described by E. D. Denman. Several variants of those are derived and the stability of the Ricatti differential equation is investigated.

Papers on working codes that are available for problems-solving are given in these proceedings. They are the following: 1. a shooting code for Sturm–Liouville problems (P. B. Bailey and L. F. Shampine), 2. a finite-difference code for first-order systems (V. Pereyra), 3. BVP codes in NAG Library (I. Gladwell), 4. a collocation code for BVP-s (U. Ascher, J. Christiansen and R. D. Russell), 5. a code for multi-point BVP-s in ordinary differential or difference equations (B. Childs and H. R. Porter), 6. a software package designed for problems in hydrodynamic stability (J. M. Gersting Jr.), 7. a subroutine for solving a system of DE in Chebyshev series (E. L. Albasiny).

Numerous papers describe particular aspects of solving BVP or discuss results of a numerical experiment in which BVP-s are encountered.

The workshop summaries are a valuable part of the proceedings. They concern at the following topics: 1. selection of shooting points, 2. demonstration of working codes, 3. basis selection, 4. mesh selection, 5. benchmark problems and code availability.

A combined bibliography of 290 entries is added at the end.

K. Balla

A. V. Balakrishnan and **M. Thoma** (Eds.): Global and Large Scale System Models. Proceedings of the International Summer Seminar, Dubrovnik, Yugoslavia, Aug. 21–26, 1978. Lecture Notes in Control and Information Sciences Series, Vol. **19**, Springer Verlag, Berlin–Heidelberg–New York, 1979, 232 pp.

The volume contains a set of lectures given at the Summer Seminar on Global and Large Scale System Models held in Dubrovnik in 1978. The aim of the seminar was to discuss fundamental problems and different approaches to large scale system modeling with applications to corporate models, national economy models and world models.

In the first paper R. E. Kalman gives a system-theoretic critique of dynamic economic models. He emphasizes the growing importance of the "system aspect" of things, mainly those characterized by complexity, interactions and man-made structures, where the classical conception of "experiment-theory-confirmation" does not work. He gives some points of view for model-building from the aspect of system theory.

R. Tomovič defines the term large system as a real object having multilevel organization and multigoal criterion function. The investigation of such systems is complicated also by the fact that the behaviour of the large majority of real-life large systems may be described only by a combination of measurable and nonmeasurable attributes.

A. V. Balakrishnan deals with problems of system identification in global modeling. An illustrative example is taken from the field of flight.

Several papers represent the problems and results of economic system modeling. (There was a decided emphasis on the Third World in respect to world models). The problematics of the translation of real world problems into a mathematical programing formulation was dealt with.

M. Mesarovic analyses the future evolution of the world system as a practical application of global modeling through a world system computer model. A program package was developed to investigate the probable effects of alternative decisions for the purpose of predicting the future. As significant uncertainties are involved in the models, numerous variants are available for investigation. The world model represents several regions of the world by their basic components such as economics, population, food

production and consumption, energy and material resources, etc. Changes in prices and other characteristic quantities and several constraints were taken also into account. The paper gives as an example some trends of the USA economics.

References at the end of every paper inspire the interested reader in deeper study of the topic.

R. Bars

Methods and Applications in Adaptive Control. Proc. of an International Symposium, Bochum, 1980. Edited by **H. Unbehauen**. Lecture Notes in Control and Information Sciences Series, Vol. **24,** Ed.: **A. V. Balakrishnan** and **M. Thoma**. Springer-Verlag, Berlin–Heidelberg–New York

An International Symposium on Adaptive Systems was held at the Department of Electrical Engineering of the Ruhr-University Bochum during March 20th and 21st 1980. The aim of this symposium was the discussion of the actual situation and the future development of adaptive systems. About 180 specialists from 13 countries came together. All papers presented at the above conference are published in this volume.

The papers of the symposium are partly concerned with methods in adaptive control and partly with applications. Three of the 24 presented papers represent survey papers. The survey paper of K. J. Åström gives a unfied description of many types of self-tuning regulators and their design principles. K. S. Narendra and B. B. Peterson point out in their survey the recent developments in adaptive control, especially the solution of stability problems in adaptive systems. The survey paper of P. C. Parks et al. reviews the application of adaptive control in three areas: aircraft control systems, process control and electrical drives.

The primary intention of the conference was to bring together researchers and practising engineers. Theorists on the one side could learn from the relevant features of practical applications, whereas on the other side practising engineers could learn what possibilities are offered by adaptive control theories, especially how to apply specific theoretical methods. The main result of the round table panel discussion held at the end of the symposium was that adaptive control is now ready for practical application which is well demonstrated by the papers of this book.

Csilla Bányász

D. L. Iglehart, G. S. Shedler: Regenerative Simulation of Response Times in Networks of Queues. Lecture Notes in Control and Information Sciences Series, Vol. **26.** Ed: **A. V. Balakrishnan** and **M. Thoma**. Springer-Verlag, 1980

The monograph deals with mathematical and statistical methodologies for discrete event simulation of networks of queues. These models can be applied in several fields, in particular, they are widely used in studies of computer and communication system performance analysis.

As the complexity of models for most real systems has reached a level where we are unable to study them analytically, computer simulation becomes the only alternative. So far a great deal of efforts has been devoted to the software engineering side (simulation languages, compact simulation packages), but relatively little has been done to develop the theoretical foundations of simulation methodologies disregarding the widely used general statistical disciplines. This monograph is to fill this gap by providing design methodologies based on underlying stochastic structures.

The emphasis of the study is on recently developed methods for the estimation of some characteristics of the so-called "passage times" in closed networks of queues which are of fundamental importance in most communication system models representing response times.

The estimation methods developed provide strongly consistent point estimates and asymptotically valid confidence intervals for general characteristics of the limiting passage times.

The monograph consists of eleven chapters. The first five ones explain basic concepts and methodologies defining passage times and the marked job method amongst

them. Chapter 6 deals with finite capacity open networks of queues based on two formulations of the finite capacity constraint. Chapter 7 considers marked job simulation via hitting times while chapter 8 discusses decomposition methods. Chapter 9 deals with the statistical efficiency of the marked job and decomposition methods. Estimation of passage times in closed networks of queues with multiple job types is the subject of chapter 10. Finally, aspects of uniform and nonuniform random number generation pertinent to implementation of a passage time simulation are considered throughout the last section.

Előd Knuth

International Symposium on Programming. Proceedings of the Fourth "Colloque International sur la Programmation", Paris, April 1980. Ed. by **B. Robinet**

The book refers to some major developments of theoretical programming, namely semantics, theory of consurrent systems, program design and meaning.

Algebraic and logical approaches to the theory of programming are in the focus of attention. Some papers of the book concern these areas. In the paper of D. Bert is shown that the language semantics is composed by a fixed part: the control structure semantics and by an extensible part: the data type semantics. A. Bertoni, G. Mauri and P. Miglioli devote an extensive analysis to the problem of the recursiveness of abstract data types from model theoretic point of view. H. Ehrig, H. J. Kreowski, P. Padawitz present a new implementation concept for algebraic specification languages. F. Nourani gives a new constructive method of extension of abstract specification. The paper of M. Broy, P. Pepper and M. Wirsing is about relations between nondeterministics programs. F. Ermine and I. Guessarian study two usual properties of program schemes: termination and simplification property. P. Guerrio present a relational model for nondeterministic programs. P. R. F. Cunha, T. S. E. Maibaum resent a communication data type for message oriented programming. R. Barbuti and A. Martelli present an algorithm for static type checking for languages with parametric types and polymorphic procedures.

The paper of A. Koester and that of W. M. Lippe and F. Simon concern the programming language LISP.

The papers concerned with the theory of concurrent systems present powerful tools for the study of it of both theoretical and practical significance. R. Janicki gives a formal approach to the problem of treating concurrent systems as the superposition of sequential subsystems or primitive concurrent subsystems. In the paper of J. A. Barnden a new style of language is proposed for the programming of systems exhibiting complex concurrent or nondeterministic behaviour. Y. S. Kwong and D. Wood present a new solution to the problem of supporting concurrent operations in B-trees. P. E. Lauer, P. R. Torrigiani and R. Devillers describe a highly parallel strategy for well-known bankers problem using COSY notation.

The last part of the book concerns program design and meaning. M. Sintzoff proposes in his paper to express program designs by hierarchical specifications of design decisions. J. L. Weiner, R. M. Burstall show how the readability of the program can be improved by having its text linearly ordered so that each sentence motivates upcoming ones.

E. Csuhaj-Varjú

W. E. Riddle, R. E. Fairley: Software Development Tools. Springer-Verlag, Berlin, Heidelberg, New York 1980

The book contains the proceedings of the workshop on software development tools, held at Pingree Park, Colorado, May 1979.

After an introduction (Chapter I) the book presents the prepared papers, as well as summarizes the comments prepared by the authors and summaries of the informal discussions.

Chapter II assesses of tool needs from "traditional" software systems, distributed software and industrial software development. For the particular steps of software development and for the whole software development process special language constructions are proposed. Chapter III deals with the importance of experiences learned from

previous efforts, pointing out lack of collecting of the evaluative data. In Chapter IV, two papers in Section "Development Support System" address several questions concerning tool-collections, particularly what capabilities belong into a collection and how should these capabilities be provided as a collection of complementary tools?

Chapters V and VI are concerned with procedural and non-procedural notational tools for use during the pre-implementation phase of software development.

In Chapter VII the appliccation of automated program verification and validation tools and the interplay between formal specification and these tools are investigated.

Finally, Chapter VIII presents summaries of discussions on "where to next?" with respect to support systems, specification languages and analysis techniques. Each summary contains recommendation for short-term and long-term research directions.

Tamás Sztanó

S. Eisenbach, C. Sadler: PASCAL for Programmers. Springer Verlag, Berlin, Heidelberg, New York, 1981, 201 pp.

This text-book on Niklaus Wirth's famous structured programming language, PASCAL, is meant to be read by a new class of programmers — the microcomputer enthusiasts, both amateur and professional, about whom two assumptions are made. The first is that they have already learned to program (probably in BASIC). The second is that they will not be adverse to an occasional glimpse of how the underlying machine played its part in executing these programs. As a result of these, no attempt is made to teach the "problem-solving" aspects of programming (although the top-down philosophy for program design forms a key feature) and no apology is made for the repeated references to the way in which a PASCAL compiler views some particular code fragment.

The first five chapters

Introduction
Fundamentals: Action and Data
Control Structures 1: Loops
Data Structures 1: Simple Data Types, Arrays and Sets
Control Structures 2: Branches

contain all the material one would need to cover in an initial course in programming while chapters 6 to 10

Data Structures 2: Records and Files
Procedures and Functions
Top-down Design
Advanced Programming Techniques
The Finishing Touches

tackle some of the more sophisticated techniques available to the Pascal programmer. For illustrative purposes some fairly extensive language for which "readibility" was a design goal, and this is exploited here in the sense that very little explanatory material appears outside of the program listing. Instead, the reader is expected to read the program to ascertain its function and the tactics behind its design, and only refer to the accompanying text for particular features and additional details.

László Király

Iréne Guessarian: Algebraic Semantics. Lecture Notes in Computer Science **99** (1981)

The book deals with one of the most important areas of theoretical programming, namely algebraic semantics.

The book is divided into six chapters. After an exhaustive introduction, in chapter 2 the author defines her framework, namely ordered magmas, their representation as trees and tree grammars.

Chapter 3 is devoted to the definition of program schemes. It is shown that program schemes can be viewed as a particular kind of tree grammars, generating tree-languages which are fully described by some infinite tree belonging to a complete free algebra. After the definition of computation rules and their correctness it is proved that the parallel-outermost and full-substitution rules are correct.

In chapter 4 the notion of interpretation is introduced. After the definition of operational, denotational and algebraic semantics it is proved that they are equivalent.

Chapter 5 is devoted to the study of classes of interpretations. First, it is proved that any class C admits a C-free interpretation, from which all others can be obtained; the free interpretation H associated with the class of all interpretations has as domain the free algebra of all formal trees. The C-free interpretation and the C-equivalence $\equiv_C$ for various usual classes interpretations are characterized, too. Moreover, it is shown that most of the usual classes of interpretations are algebraic.

Chapter 6 is a first exploration of the applications of algebraic semantics. It is shown how can one simplify program schemes by detecting and deleting all useless branches from a program scheme. Then the results of the previous chapters are applied to proofs of program correctness, program equivalence and program transformations.

The approach of this book is modern and it is written in an elegant style. In addition to the precise definitions and proofs a lot of remarks and examples lead to better understanding.

The book is recommended to everybody who are interested in algebraic semantics.

E. Csuhaj-Varju

Distributed Systems — Architecture and Implementation. An Advanced Course, editedby **B. W. Lampson, M. Paul,** and **H. J. Siegert.** Springer Verlag, Berlin–Heidelberg–New York, 1981, 510 pp.

The 20 papers comprising this book were prepared and presented at the Advanced Course on Distributed Systems — Architecture and Implementation, held from March 4 to March 13, 1980 at the Technical University, Munich (Technische Universität München). The goal of the course was to concentrate on the total system view of distributed systems and not merely to present isolated single aspects.

The authors of the papers are Gerald LeLann, Richard W. Watson, Kenneth J. Thurber, Donald W. Davies, E. Douglas Jensen, Butler W. Lampson, Elmar Holler, Stephen R. Kimbleton and Pearl Wang.

In the first chapter LeLann gives a definition of the distributed system used throughout the book. User needs as major motivations for distributed computing and the objectives of distributed systems are discussed in detail. In chapter 2 Watson provides a three dimensional model of a distributed system and some common terminology to be used in the following chapters which elaborate this model and present different approaches to its realization. Hardware interconnection technology, link level protocols, network hierarchies and end to end protocols which constitute the interprocess communication layer of the model are considered in chapter 3. The next chapter discusses the hardware structure upon which all layers will be built. Both papers are presented by Thurber. Chapter 5 is devoted to the link level which is the next one of the model. Specific link protocols as HDLC, the Arpanet IMP—IMP and Cyclades MV8 protocols are also described. The author of this chapter is LeLann. In the next chapter Davies and Watson deal with the problems of hierarchy which can be summarized as addressing, routing, flow control and congestion avoidance. Interprocess Communication Interface and end to end protocol design issues are the subject of chapter 7 presented by Watson. In the next chapter Jensen provides a conceptual model of the spectrum of control alternatives. The characteristics and design goals for identification in a distributed operating system are outlined in chapter 9 by Watson. In the following chapter Davies discusses several protection methods in a distributed system. The concept of atomic transaction is introduced in the next chapter by Lampson which can be used to perform complex updates to distributed data. Some synchronization mechanism in decentralized environment are characterized by LeLann in chapter 12. Solutions of problems concerning to the update of multiple copies of data bases and their verifications are outlined in the following chapter

by Holler. In chapter 14 Kimbleton, Wang and Lampson describe protocols which are required for effective support of network data access. LeLann defines the basic concepts of error recovery and reliability in the next chapter. Some specific hardware features that could be utilized to support distributed processing and the synergism between hardware and software in distributed systems are discussed in chapters 16 and 17 by Thurber and Jensen respectively. The aim of the National Software Works was to design and implement a network operating system. Holler provides the concept and facilities of the NSW in chapter 18. The Ethernet communication system, the PUP internetwork architecture integrating the Ethernet and the Violet decentralized information systems which is constructed for users connected to the PUP internet are described in chapter 19 by Lampson. In the last chapter Thurber provides final remarks to the course. Useful references and a subject index complete the book.

Tibor Gyires

The Programming Language ADA. Reference Manual. Lecture Notes in Computer Science. Eds: **G. Goos**; **J. Hartmanis,** Vol. **106.** Springer-Verlag, 1981.

It was in the late seventies that the US Department of Defense decided to support the development of a new programming language which would serve as a standard programming tool for the Department.

First the requirements were developed and after a long procedure the language elaborated by the CII Honeywell Bull Research Center under the leadership of Jean D. Ichbiah was accepted. It was named after Ada Byron, a coworker of Charles Babbage.

The first version of the language was published in the Sigplan Notices in 1979 June. This preliminary Reference Manual had been slightly altered and the final version published in July 1980 by the Department of Defense. This book is the copy of the mentioned DoD publication.

Although for the time being there is no compiler available — the DoD strictly prohibits any subset of the language to call Ada — it is pretty sure that in the near future the language will spread out in a very big extent. This is partly due to the fact that the probably biggest consumer of software the DoD is directly interested in the language partly to the relatively abundant finacial tools supporting its implementation.

According to the philosophy of the language security is preferred before effectivity. For example there is a strict typing in the language and mixing of objects of different types must be detected in compile time even when these objects have the same representation within the computer. The violation of index boundaries must be checked if not possible in compile time than in run time.

The language is capable of handling parallel and real time programming. These are summarized in the chapter tasking.

There is a sophisticated built in mechanism for exception handling.

An effort has been made to insure separate compilation and still keep security.

Many interesting features cannot be treated within a short summary. Anyway it is worthwhile to study Ada not only because it has become very fashionable but having some real new and interesting language-constructs, too.

However it has to be mentioned that the s[illegible]ftware word reacted upon the language in a pretty controversial way. It is enough to me[illegible]n Hoare's Turing lecture award lecture where he pointed out that Ada is too big and to[illegible]omplex to achieve its original goals.

Iván Bach

M. Vidyasagar: Input Analysis of Large-Scale Interconnected Systems. Springer Verlag, Berlin, Heidelberg, New York, 1981, 221 pp.

In the past decade there has been a great progress in the theory of large-scale systems. To meet the interest appearing in this field, the book is a high-level research monograph that sets forth most of the available results on the decomposition, well-posedness, stability and instability of large-scale systems that can be obtained by input–output methods.

Throughout this book, the emphasis is on treating the large-scale system at hand as an interconnected system, consisting of several subsystems interacting various interconnection operators.

The volume consists of nine chapters.

Chapter 1 is a brief introduction.

Chapter 2 presents the concepts of truncation on extended spaces, which provide the mathematical setting for input–output analysis and the definitions of well-posedness and stability.

Chapter 3 deals with the concepts of gain and dissipativity playing an important role in the various criteria for stability and instability, and with the explicit methods for computing gains and testing dissipativity.

Chapter 4 presents a few graph-theoretic techniques for efficient decompositions of large-scale interconnected systems. It is shown that the well-posedness and stability can be determined by studying only the so-called strongly connected components (SCC's).

Chapter 5 treats some sufficient conditions for the well-posedness of systems.

Chapter 6 introduces some generalization of the single-loop "small gain" theorem to arbitrary interconnected systems.

Chapter 7 is devoted to state and to prove several generalizations of the single-loop "passivity" theorem.

Chapter 8 deals withderivation several L_2-instability criteria for large-scale systems.

Chapter 9 shows how the technique of exponential weighting can be used to study L_∞-stability and L_∞-instability using the previous results.

The book is recommended for students and researchers interested in the theory of large-scale systems, and for those who have the necessary background in elementary functional analysis.

T. Kovács

W. K. Giloi: Computer Architecture. Springer Verlag, Berlin, Heidelberg, New York, 1981, 380 pp. (Published in German. Original title: Rechnerarchitectur)

The development in the field of computer architecture has been an ad hoc process in the last 30 years. It has been more an art than a science. It is the same for every new field of science for systematization and generalization a lot of experience and information is needed.

W. K. Giloi, the author of this book, made an attempt to systematize the field of computer architecture. He wrote this book on the basis of his lectures held at the University of Minnesota and at the Technical University, Berlin for several years.

The reviewer's opinion is that Prof. Giloi has made a fine job. He has succeeded in his aim, his book gives a systematic and very clear treatment of computer architecture. He found it possible to cover a broad area in a relatively short book without becoming superficial.

Computer architecture is defined by different ways by different authors. Chapter 1 as an introduction to this topic shows how the author defines and uses it in this book: Computer architecture is defined with the functional behaviour of the computer at the level of the control and information structure and with the structure of the functionally separated hardware units.

Chapter 2 deals with the classical von Neumann-machine. It defines the notions which are most closely related to the computer architecture such as computing, algorithm, program and data flow control, etc. Emphasis is put on the discussion of program structures, concurrent processes and storage organisation.

In chapter 3 the structures of one processor system are discussed. Among them the minimal von Neumann-machine, the multiregister machines and the principle of pipelining. It discusses the storage organisation (physical and logical addresses, memory manager and memory mapping, virtual memory, cache storage, capability addressing), bus organisation, input/output organisation and microprogramming.

Chapter 4 is devoted to the concepts of parallelism. The level of parallel operations, programming of parallel operations, data structures, arrays and the efficiency of the multiprocessor systems are discussed.

MTA Könyvtára
Periodika 1981/3842

Chapter 5 summarises the innovating mechines formed by von Neumann-machines. Detailed discussion is given to the Holland machine, to the ILLIAC IV and to the CDC 6600. The next chapter discusses the principle of vector processing, some computers with vector processing architecture (STAR 100, CYBER 200, CRAY-1), the associative storages and systems (STARAN).

In chapter 7 the self identificating information and data type architectures are treated. Their advantages, the hierarchy of type identification and the DRAMA principle (Descriptor Referenced Autonomous Memory Allocation) are discussed.

Next chapter is devoted to the high-level language architecture. It is dealt with here with the stack machines, the SYMBOL-machine, the Borrough B6700 and some APL-machines.

Chapter 9 describes the data flow multiprocessor architectures while the last chapter deals with the general purpose multiprocessor systems. Author's Index and Subject Index complete the book.

T. Gál

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

ENGLISH TRANSLATION OF THE PAPER IN RUSSIAN

Problems of Control and Information Theory, Vol. **10**, *No. 6 (1981)*

PROBABILISTIC CHARACTERISTICS OF AN INCOMPLETELY ACCESSIBLE MULTI-PHASE SERVICE SYSTEM WITH SEVERAL TYPES OF REPEATED CALL

S. N. STEPANOV

For an incompletely accessible multi-phase service systems where severa types of repeated calls are possible relations between probabilistic characteristics of the model and formulae for approximate computation of probabilistic characteristics with low and high load are derived.

1. Introduction

A most obvious effect the user has on a service system is produced by a repeated attempt to establish a connection following refusal. A theoretical study of this situation will hopefully lead to a better insight into the functioning of complex computer-based data processing systems. The number of papers on tbis topic is on the increase.

Computation of probabilistic characteristics for systems of repeated calls is very difficult because of high dimensionality of random processes which describe the model functioning. Accurate results have been obtained only for a narrow range of systems with simplest assumptions on the structure and the subscriber's behaviour. Therefore in studying models which allow for repeated calls those methods may lead to qualitative results on the system operation which involve asymptotic expansion of the model probabilistic characteristics with low and high load and those which lead to relations between probabilistic characteristics and are independent of the model structure or the subscriber's behaviour pattern. These methods permit making more realistic assumptions on the subscriber's behaviour and the system structure, which is important for design and operation of sophisticated switchboards. One of such methods which essentially relies on conservation laws is described in the article in application to a fairly advanced model of an incompletely accessible system (IAS) where the service takes several stages and several types of repeated calls possible.

2. Model functioning

An IAS receives n independent Poisson flows of primary at rates of λ_i, $i = 1, 2, \ldots, n$ and n flows of repeated calls. Calls of the i-th flow can access d_i IAS lines. The arriving call takes the first free line in the order of accessibility for a random duration which is exponentially distributed with a parameter $\alpha_{i,1}$. Let us refer to this time as the first phase. Assume that for a call from the i-th flow of the incoming load the full service time consists of $(m_i + 1)$ exponentially distributed phases. The first m_i service phases take up the time needed for establishing a connection in the IAS, the last, $(m_i + 1)$-st service phase is the phase of conversation. Once the k-th $(k = 1, 2, \ldots, m_i - 1)$ phase of serving a call is over, the $(k + 1)$-st phase of serving a call from the i-th flow with a probability of $P_{i,k}$ starts (its duration is distributed exponentially with a parameter $\alpha_{i,k+1}$); with a probability $(1 - P_{i,k})(1 - Z_{i,k})$ the subscriber leaves the system; with a probability of $(1 - P_{i,k})Z_{i,k}$ the subscriber tries again to connect ($j_{i,1}$, a random number of subscribers to the i-th flow of the load who retry a connection because of losses in establishing the connection, is incremented by a unity). With a probability of P_{i,m_i} following the end of the m_i-th service phase (the last one in establishing a connection) the line is busy for the duration of the conversation; this duration is distributed exponentially with a parameter α_{i,m_i+1}; with a probability of $(1 - P_{i,m_i}) \times (1 - Z_{i,m_i})$ the subscriber leaves the system; with a probability of $(1 - P_{i,m_i}) Z_{i,m_i}$ the subscriber retries a connection ($j_{i,2}$, a random number of subscribers of the i-th flow who retry a connection because the subscribers whom they call do not reply, is incremented by a unity).

The rate of retries at the i-th input is $j_{i,1}\, \mu_{i,1} + j_{i,2}\, \mu_{i,2}$ where $\mu_{i,1}$ and $\mu_{i,2}$ are parameters of the exponentially distributed duration between successive calls of a subscriber from queue $j_{i,1}$ and $j_{i,2}$, respectively. If at the time of the call all accessible lines are busy, then with a probability of $H_{i,1}$ (if it was the first refusal) the subscriber retries a connection ($j_{i,1}$ is in the case incremented by a unity) and, with a probability of $(1 - H_{i,1})$ the subscriber hangs up. If the refusal comes in a retry, then with a probability of $H_{i,2}$ the subscriber makes another attempt at connection (in this case the system state does not change if the repeated call comes from queue $j_{i,1}$; if the call comes from queue $j_{i,2}$, then $j_{i,1}$ is increased and j_{i2} decreased by a unity) and the subscriber leaves the system with a probability of $(1 - H_{i,2})$ ($j_{i,1}$ or $j_{i,2}$ decreases by a unity). Usually $H_{i,1}$ and $H_{i,2}$ are referred to as the insistence function of a subscriber from the i-th flow of the primary load. Before being served, a call can be refused at any stage of connection in real systems. To allow for these losses

let us assume that a primary or repeated call can be refused with a probability of a_i before it arrives at the i-th IAS input. In this case the subscriber is believed to behave as his insistence function dictates.

To reduce the number of unknowns in the statistical equilibrium equation set let us assume that $j_{i,1} \leq N_{i,1}$ and $j_{i,2} \leq N_{i,2}$ so that if a primary or repeated call should queue and the appropriate place in the queue is busy the call is lost with a probability of one. Here $N_{i,1}$ and $N_{i,2}$ are some integers, $i = 1, \ldots, n$.

Let M denote the total number of lines in the IAS and denumerate them arbitrarily. The IAS state is defined by a $2 \times M$ engagement matrix x. The value of $x(1, k)$ is l if the k-th line is engaged for the l-th service phase and 0 if the k-th line is free, $k = 1, \ldots, M$. When $x(1, k)$ is nonzero, then $x(2, k)$ is equal to the ordinal number of the flow whose call is served by the k-th line of the IAS. Let S denote the set of all possible IAS states; X_i is the set of IAS states when all d_i lines accessible to the i-th flow are engaged.

For the i-th flow of primary load the number of subscribers who make repeated calls at time t because of refusals in establishing a connection and because the subscriber who is called does not respond is denoted as $j_{i,1}(t)$ and $j_{i,2}(t)$, respectively; $x(t)$ is the IAS state at time t. The functioning of the model is described as a Markov process

$$\xi(t) = \{j_{1,1}(t), j_{1,2}(t), \ldots, j_{n,1}(t), j_{n,2}(t), x(t)\}.$$

If the ergodicity condition is met for $\xi(t)$ a static equilibrium set can be written. Let $P_{j,x}$ denote the probability of stationary model states; here j is the state vector of queues, $j = \{j_{1,1}, j_{1,2}, \ldots, j_{n,1}, j_{n,2}\}$. The equations are reasonably easy to obtain but cumbersome and so will not be written.

3. Determining the probabilictic characteristics. Their equivalence

Let us introduce the notation

$$\pi_i = \sum_j \sum_{x \in X_i} P_{j,x}, \; I_{i,k} = \sum_j \sum_{x \in S} P_{j,x} l_{x,k,i},$$

$$J_{i,1} = \sum_j \sum_{x \in S} P_{j,x} j_{i,1}, \; J_{i,2} = \sum_j \sum_{x \in S} P_{j,x} j_{i,2}, \tag{3.1}$$

$$J'_{i,1} = \sum_j \sum_{x \in X_i} P_{j,x} j_{i,1}, \; J'_{i,2} = \sum_j \sum_{x \in X_i} P_{j,x} j_{i,2}$$

where $\sum_j$ denotes summation over all $j_{i,k}$ which vary from 0 to $N_{i,k}$, $i = 1, \ldots, n$ $k = 1, 2$, and $l_{x,k,i}$ is the number of IAS lines busy in state x for the k-th phase of serving calls from the i-th flow.

Since all subsequent reasoning will apply to properties and behaviour of probabilistic characteristics of the i-th flow, $i = 1, \ldots, n$, the subscript i will be omitted. Consequently, rather than $H_{i,1}, P_{i,1}, \mu_{i,1}, J_{i,1} \ldots$ we will write $H_1, P_1, \mu_1, J_1 \ldots$

In investigating systems where repeated attempts are made to establish a connection, the following probabilistic characteristics of serving the i-th flow are important:

— the probability that a primary call is refused service

$$\pi_0 = a + (1 - a)\,\pi$$

— the probability that a repeated call is refused service

$$\pi_1 = (J_1' \mu_1 + J_2' \mu_2)\,(1 - a)/(J_1 \mu_1 + J_2 \mu_2) + a$$

— the probability that a primary or repeated call is refused

$$\pi_2 = (\lambda\pi + J_1' \mu_1 + J_2' \mu_2)\,(1 - a)/(\lambda + J_1 \mu_1 + J_2 \mu_2) + a$$

— the total probability that no connection is established:

$$\pi_3 = \Big((\lambda\pi + J_1' \mu_1 + J_2' \mu_2)\,(1 - a) + \sum_{k=1}^{m} I_k(1 - P_k)\,\alpha_k\Big)/(\lambda + J_1 \mu_1 + J_2\mu_2) + a$$

— the mean number of lines busy for the k-th service phase of the i-th flow $= I_k$

— the mean number of subscribers who repeat their calls because of losses in connection J_1 and because they receive no response $= J_2$.

— the mean number of repeated calls for one primary call

$$M = (J_1 \mu_1 + J_2\mu_2)/\lambda; \tag{3.2}$$

— the mean number of refusals per connection

$$Q = \Big((\lambda\pi + J_1' \mu_1 + J_2' \mu_2)\,(1 - a) + (\lambda + J_1 \mu_1 + J_2 \mu_2)\,a + \\ + \sum_{k=1}^{m} I_k(1 - P_k)\,\alpha_k\Big)/I_{m+1}\,\alpha_{m+1};$$

— the probability that the call leaves the system before it gives by

$$D = \big((\lambda(1 - H_1) + J_1 \mu_1(1 - H_2) + J_2 \mu_2(1 - H_2))\,a + \\ + (\lambda\pi(1 - H_1) + J_1' \mu_1(1 - H_2) + J_2' \mu_2(1 - H_2))\,(1 - a) + \\ + \sum_{k=1}^{m} I_k(1 - P_k)\,\alpha_k(1 - Z_k)\big)/\lambda.$$

Let us consider several other equivalent definitions of IAS model probabilistic characteristics which may be helpful in computing them. One of them is the mean number of repeated calls per primary one (M in definition (3.2)). For the i-th flow of load let $a(t)$ denote the total number of repeated calls and $b(t)$, the total number of primary calls received over a half-interval $(0, t]$. The following Lemma holds.

Lemma 1. Almost surely

$$\lim_{t\to\infty} a(t)/b(t) \tag{3.3}$$

exists and is equal to $(J_1\mu_1 + J_2\mu_2)/\lambda$.

The Lemma is proved in a way similar to the equivalence of two definitions of blocking probability for an IAS with losses (Theorem 8.2, Ch. 8, [1]).

Let us give another equivalent definition of M. Let $b_{j,x}(t)$ denote the number of primary calls which arrived when the system was in state (j, x) at the i-th IAS input over the half-interval $(0, t]$ and $\nu^k_{j,x}(t)$ the number of repeated calls made over $(0, t]$ by a k-th primary call in the order of arrival, in state (j, x) $k = 1, 2, \ldots, b_{j,x}(t)$. Because time intervals between successive arrivals of the process $\xi(t)$ in state (j, x) are independent, then as $t \to \infty$ and with different k the values of $\nu^k_{j,x}(t)$ are asymptotically identically distributed random values with mean $\nu_{j,x}$.

Then holds

Lemma 2. Almost surely

$$\lim_{t\to\infty} \frac{\nu^1_{j,x}(t) + \ldots + \nu^{b_{j,x}(t)}_{j,x}(t)}{b(t)}$$

exists and is equal to $\nu_{j,x}P_{j,x}$.

Proof. The Lemma is proved as Lemma 1 is. Consequently

$$M = (J_1\mu_1 + J_2\mu_2)/\lambda = \lim_{t\to\infty} a(t)/b(t) =$$

$$= \lim_{t\to\infty} \sum_j \sum_{x\in S} (\nu^1_{j,x}(t) + \ldots + \nu^{b_{j,x}(x)}_{j,x}(t))/b(t) = \sum_j \sum_{x\in S} \nu_{j,x}P_{j,x}.$$

Each of the definitions, (3.2), (3.3), and (3.4) permits computing the probabilistic characteristics of a system in which repeated calls are made with different input conditions. If one can compute the probabilities of stationary states for a random process which describes the model action, then definitions

of type (3.2) are used; in statistical modeling or measurements definitions (3.3) are used. In some cases (in particular, with low or high loads) approximate formulae can be obtained through following up the fate of one call with the use of definition (3.4).

4. Conservation laws. Formulae which relate basic probabilistic characteristics of the model

Let us obtain relations of basic probabilistic characteristics of the model. With (3.1) holding, the following Lemma is true

Lemma 3

$$I_k \alpha_k P_k = I_{k+1} \alpha_{k+1}, \quad k = 1, \ldots, m \tag{4.1}$$

$$J_1 \mu_1 = (\lambda\pi H_1 + J_1' \mu_1 H_2 + J_2' \mu_2 H_2)(1 - a) +$$

$$+ (\lambda H_1 + J_1 \mu_1 H_2 + J_2 \mu_2 H_2)\, a + \sum_{k=1}^{m-1} I_k (1 - P_k) \alpha_k Z_k - \gamma_1, \tag{4.2}$$

$$J_2 \mu_2 = I_m (1 - P_m) \alpha_m Z_m - \gamma_2, \tag{4.3}$$

$$(\lambda(1 - \pi) + (J_1 - J_1') \mu_1 + (J_2 - J_2') \mu_2)(1 - a) = I_1 \alpha_1 \tag{4.4}$$

where

$$\gamma_1 = \sum_{j, j_1 = N_1} \Big(\sum_{x \in X_i} (\lambda H_1 + j_2 \mu_2 H_2)(1 - a) P_{j,x} +$$

$$+ \sum_{x \in S} \Big((\lambda H_1 + j_2 \mu_2 H_2)\, a + \sum_{k=1}^{m-1} l_{x,k} \alpha_k (1 - P_k) Z_k \Big) P_{j,x} \Big),$$

$$\gamma_2 = \sum_{j, j_2 = N_2} \sum_{x \in S} l_{x,m} \alpha_m (1 - P_m) Z_m P_{j,x}.$$

Proof. The Lemma is proved as in [2]. One can prove that as $\lambda \to 0$ and N_1 and N_2 are fixed, then γ_1 and γ_2 behave as $0(\lambda^{N_1})$ and $0(\lambda^{N_2})$, respectively and $\gamma_1 = \gamma_2 = 0$ when $N_1 = N_2 = \infty$. It is also intuitively understandable that with fixed model parameters γ_1 and γ_2 tend to zero as $N_1 \to \infty$ and $N_2 \to \infty$.

The resultant relations act as conservation laws and are helpful in investigating complex models of systems which are repeatedly accessed. In particular, they are applicable to indirect estimation of those probabilistic model characteristics which are difficult to measure because primary calls should

be separated from repeated ones (such characteristics include the of primary calls, the mean number of re-calling subscribers, etc.). For the i-th IAS input one can measure π (by virtue of the Poisson nature of arrival of primary calls the mean value of this characteristic coincides with the time fraction during which all IAS lines accessible to the i-th flow are busy);

$$L_1 = (\lambda + J_1\mu_1 + J_2\mu_2)(1 - a),$$

the rate of the overall flow of calls at the i-th ISA input:

$$L_2 = (\lambda\pi + J_1'\mu_1 + J_2'\mu_2)(1 - a),$$

the rate of the overall flow of calls at the i-th input which were refused for want of free accessible lines. Let $N_1 = \infty$ and $N_2 = \infty$.

Then all the characteristics are derived in succession from the following formulae which have been derived from relations (4.1)–(4.4). Denote

$$x_k = (L_1 - L_2)P_1 \ldots P_k, \quad k = 0, 1, \ldots, m, \quad (x_0 = L_1 - L_2);$$

$$\lambda = \frac{L_1(1 - H_2 a)/(1 - a) - L_2 H_2 - \sum_{k=1}^{m} x_{k-1}(1 - P_k)Z_k}{1 - (H_2 - H_1)(\pi + a(1 - \pi))};$$

$$\pi_1 = \frac{(L_2 - \lambda\pi(1 - a))(1 - a)}{L_1 - \lambda(1 - a)} + a; \quad \pi_2 = \frac{L_2(1 - a)}{L_1} + a;$$

$$\pi_3 = 1 - P_1 \ldots P_m(1 - \pi_2); \; I_k = x_{k-1}/\alpha_k, \quad k = 1, \ldots, m + 1;$$

$$J_2 = x_{m-1}(1 - P_m)Z_m/\mu_2; \; M = L_1/\lambda(1 - a) - 1;$$

$$J_1 = ((L_1/(1 - a) - \lambda - J_2\mu_2)/\mu_1;$$

$$Q = \frac{L_1}{x_m(1 - a)} - 1;$$

$$D = (\lambda - I_{m+1}\alpha_{m+1})/\lambda.$$

5. The behaviour of probabilistic characteristics with low load

Let $\lambda^* = \max_k \lambda_k$ and $\lambda^* \to 0$. Let us find the first non-zero term of the power expansion series of λ for the probabilistic characteristic M. Let us make use of definition (3.4). Let us take up one call and compute the mean number of repetitions it will make while in the system. As $\lambda^* \to 0$ a primary or repeated call is refused service because all accessible lines are busy with a probability of $o(1)$ because this probability is proportional to the arrival of

at least one primary call during a finite time interval. Then the mean number of repeated calls per primary one can be represented as a sum

$$M = D_1 Q_2 \sum_{k=0}^{\infty} D_2^k (k+1) + o(1).$$

Here

$$D_1 = aH_1 + (1-a)((1-P_1)Z_1 + P_1(1-P_2)Z_2 + \ldots + P_1 \ldots P_{m-1}(1-P_m)Z_m);$$

$$D_2 = aH_2 + (1-a)((1-P_1)Z_1 + P_1(1-P_2)Z_2 + \ldots + P_1 \ldots P_{m-1}(1-P_m)Z)_m;$$

$$Q_2 = a(1-H_2) + (1-a)((1-P_1)(1-Z_1) + P_1(1-P_2)(1-Z_2) + \ldots + P_1 \ldots P_{m-1}(1-P_m)(1-Z_m) + P_1 \ldots P_m).$$

Summation and using definition (3.2) give

$$M = \frac{J_1\mu_1 + J_2\mu_2}{\lambda} = \frac{D_1}{Q_2} + o(1).$$

Using the definitions of probabilistic characteristics and Lemma 3, let us find for the remainder of the probabilistic characteristics of the i-th IAS input the following approximate formulae which hold as $\lambda \to 0$

$$J_1 = \frac{\lambda(D_1 - (1 + a(H_1 - H_2))(1-a)P_1 \ldots P_{m-1}(1-P_m)Z_m)}{Q_2\mu_1} + o(\lambda);$$

$$J_2 = \frac{\lambda(1 + a(H_1 - H_2))(1-a)P_1 \ldots P_{m-1}(1-P_m)Z_m}{Q_2\mu_2} + o(\lambda); \qquad (5.1)$$

$$I_k = \frac{\lambda(1 + a(H_1 - H_2))P_1 \ldots P_{k-1}(1-a)}{Q_2\alpha_k} + o(\lambda), \quad k = 1, \ldots, m+1$$

$$D = 1 - (1 + a(H_1 - H_2))P_1 \ldots P_m(1-a)/Q_2 + o(1);$$

$$\pi_3 = a + (1-a)(1 - P_1 \ldots P) + o(1);$$

$$Q = -1 + 1/(1-a)P_1 \ldots P_m + o(1);$$

$$\pi_0 = a + o(1); \ \pi_1 = a + o(1); \ \pi_2 = 1a + o(1).$$

Table 1 summarizes the results of accurate and approximate (by formulae (5.1)) computation of probabilistic characteristics for a particular case of an IAS, a fully accessible bunch of lines. There is a single incoming flow of primary load; the service takes two phases. The model input parameters take on the following numerical values

Table 1

$v = 3$; $M_1 = 10$; $M_2 = 5$; $H_1 = 0.7$; $H_2 = 0.9$; $\alpha_1 = 10$;
$\alpha_2 = 2$; $P = 0.5$; $Z = 0.5$; $a = 0.1$

the value of λ is given in Table 1, $N_1 = N_2 = 7$.

Analysis of numerical data suggests that with low values of λ the accuracy of the asymptotic method for computing the probabilistic characteristics is rather high. With λ increasing the accuracy falls but remains acceptable for practical use even for λ other than small (e.g. $\lambda = 1$ in Table 1)

6. Behaviour of probabilistic model characteristics with high load

Let us first take up a particular case, a model of v-linear fully accessible bunch with multi-phase service and two types of repeatedly calling subscribers. The model input parameters are $\alpha_1, \ldots, \alpha_{m+1}$; $P_1, \ldots, P_m$; $Z_1, \ldots, Z_m$; H_1; H_2; a; μ_1; μ_2 and λ. The definition and notation of model input parameters and probabilistic characteristics coincide with those of the original IAS model, except the subscript i which denotes the ordinal number of the load flow (see Sections 2 and 3). Assume that $H_2 < 1$, $N_1 = \infty$, $N_2 = \infty$, $\lambda \to \infty$, and other model parameters are fixed. Using the analytic property of I_k, J_k, π, $k = 1, 2$ and relations (4.1)–(4.4) the expansion of the basic probabilistic characteristics of the bunch as $\lambda \to \infty$ may be found in the form

$$
\begin{aligned}
&I_k = I_{k,0} + I_{k,1} \cdot \frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right), \quad k = 1, \ldots, m+1; \\
&J_1 = J_{1,-1}\lambda + J_{1,0} + o(1); \\
&J_2 = J_{2,0} + J_{2,1}\frac{1}{\lambda} + o(1); \qquad (6.1) \\
&\pi = \bar{\pi}_0 + \bar{\pi}_1 \frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right) \\
&J_1'\mu_1 + J_2'\mu_2 = J' = J_1'\lambda + J_0' + o(1);
\end{aligned}
$$

Let us derive certain relations for coefficients in the expansion (6.1)

Let $f = 1 + \alpha_1 \sum_{k=1}^{m} \frac{P_1 \cdots P_k}{\alpha_{k+1}}$. Then holds

Lemma 4

$$\overline{\pi}_0 = 1, \sum_{k=1}^{m+1} I_{k,0} = \nu, I_{k,0} = \frac{\nu\alpha_1 P_1 \dots P_{k-1}}{f\alpha_k} \tag{6.2}$$

$$\overline{\pi}_1 = -(1 - H_2)\nu\alpha_1/(1 - H_2 + H_1) f(1 - a); \tag{6.3}$$

$$I_{k,1} = -(1 - H_2)\nu\alpha_1^2 P_1 \dots P_{k-1}/(1 - H_2 + H_1) f^2 \alpha_k (1 - a). \tag{6.4}$$

Relation (6.2) is obvious. Let us prove relation (6.3). Let $y(\lambda)$ denote the part of the primary load flow which immediately reaches the first stage of connection. It is obvious that

$$\lim_{\lambda\to\infty} y(\lambda) = \lim_{\lambda\to\infty} \lambda(1 - \pi)(1 - a) = -\pi_1(1 - a). \tag{6.5}$$

Let us consider the flow of repeated calls which reaches the bunch input. As $\lambda \to \infty$, it can be regarded as asymptotic Poisson with a rate of $\lambda H_1(1 - a)/(1 - H_2)$. In that case the relation holds

$$\lim_{\lambda\to\infty} \frac{y(\lambda)}{I_1\alpha_1 - y(\lambda)} = \frac{1 - H_2}{H_1}. \tag{6.6}$$

From (6.2), (6.5), and (6.6) follows (6.3). Relation (6.4) follows from the equality

$$\sum_{k=1}^{m+1} I_k = \pi\nu + (1 - \pi)(\nu - 1) + o\left(\frac{1}{\lambda}\right)$$

which holds for this model as $\lambda \to 0$.

Substitute (6.1) into (4.1)–(4.4) and equate coefficients with identical power of λ. By using Lemma 4 let us solve the resultant equation systems for unknown coefficients of the series expansion (6.1). Denote $F_k = P_1 \dots P_k$, $(F_0 = 1)$. As a result formulae (6.1) take the form

$$I_k = \frac{\nu\alpha_1 F_{k-1}}{f\alpha_k} - \frac{(1 - H_2)\nu\alpha_1^2 F_{k-1}}{(1 - H_2 + H_1) f^2 \alpha_k (1 - a)\lambda} + o\left(\frac{1}{\lambda}\right), \qquad k = 1, \dots, m+1;$$

$$J_1 = \frac{H_1}{\mu_1(1 - H_2)} + \frac{\nu\alpha_1}{f\mu_1(1 - H_2)} \left[\sum_{k=1}^{m-1} (F_{k-1} - F_k) Z_k + \right.$$

$$\left. + (F_{m-1} - F_m) Z_m H_2 - \frac{H_1}{1 - H_2 + H_1}\right] + o(1); \tag{6.7}$$

$$J_2 = \frac{\nu\alpha_1(F_{m-1} - F_m) Z_m}{f\mu_2} - \frac{(1 - H_2)\nu\alpha_1^2(F_{m-1} - F_m) Z_m}{f^2 \mu_2 (1 - H_2 + H_1)(1 - a)\lambda} + o\left(\frac{1}{\lambda}\right);$$

Analysis of numerical data suggests that the characteristics of subscriber behaviour (M, Q, D, π_3, and π_2) are computed most accurately and those of the bunch throughput, least accurately. If a similar comparison is made with the data of Table 1, one can note that with large load the picture is reversed.

This method, applied to systems with repeated calls, will hopefully expand the scope of theoretical models for description of the way when real computer-based data retrieval systems operate.

References

1. *Beneš, V. J.*, Matematicheskie osnovi teorii telefonnikh soobshenii. Moscow, 1968.
2. *Stepanov, S. N.*, Integral Equilibrium Relations for an Incompletely Accessible System with Repeated Attempts at Connection and Their Application. Problemmy peredachi informatsii, Vol. **16**, No. *2*, 1980.

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

L. Györfi

Technical University of Budapest

H – 1111 Budapest XI., Stoczek u. 2, 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 (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 and address of that author who manages the proof-reading should also be given.

An abstract should head the paper.

The authors are encouraged to use the following headings: Introduction (outlining the problem), Methods and results, should not exceed 15 pages (25×50 characters) 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 статей в четырех экземплярах на русском языке или в двух на английском и в двух на русском языке следует направлять по адресу: 117901 ГСП-1 Москва-В-71, Ленинский проспект, 14, корп. 4, комн. 18. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова, тел. 237-99-53).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50 – 100 слов и приложено резюме – реферат объемом не менее 10 – 15% объема статьи на русском и на английском языке (в двух экземплярах каждый), на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями в традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Outrata, J. V., Jarušek, J., Beran, Z.: An application of the augmented Lagrangian approach to the optimal control of a biharmonic system with state-space constraints (*Outrata И. В., Ярушек И., Беран З.* Применение обобщенных функций Лагранжа для оптимального управления бигармонической системой с ограничениями на состояние) 363

Spal, J.: Modified fuzzy sets (*Спал Й.* Модифицированные расплывчатые множества) 375

Степанов С. Н. Свойства вероятностных характеристик неполнодоступной системы с многофазным обслуживанием и несколькими типами повторных вызовов (*Stepanov, S. N.*: Probabilistic characteristics of an incompletely accessible multi-phase service system with several types of repeated calls) 387

Bokor, J., Pető, J.: Minimal realisation procedure from partial function representation of a rational transfer matrix (*Бокор И., Пето Й.* Метод получения минимальной реализации передаточной матрицы в многомерных системах автоматического управления) 403

Jereb, L., Szabó, Cs.: A simple asynchronous reservation protocol for a multiple access channel (*Йереб Л., Сабо Ч.* Простой асинхронный алгоритм множественного доступа с резервированием) 417

Grim, J.: An algorithm for maximizing a finite sum of positive functions and its application to cluster analysis (*Грим И.* Алгоритм для максимизации конечной суммы позитивных функций и его применение для кластерного анализа) 427

Book review 439

HU ISSN 0370—2529

Index: 26 660