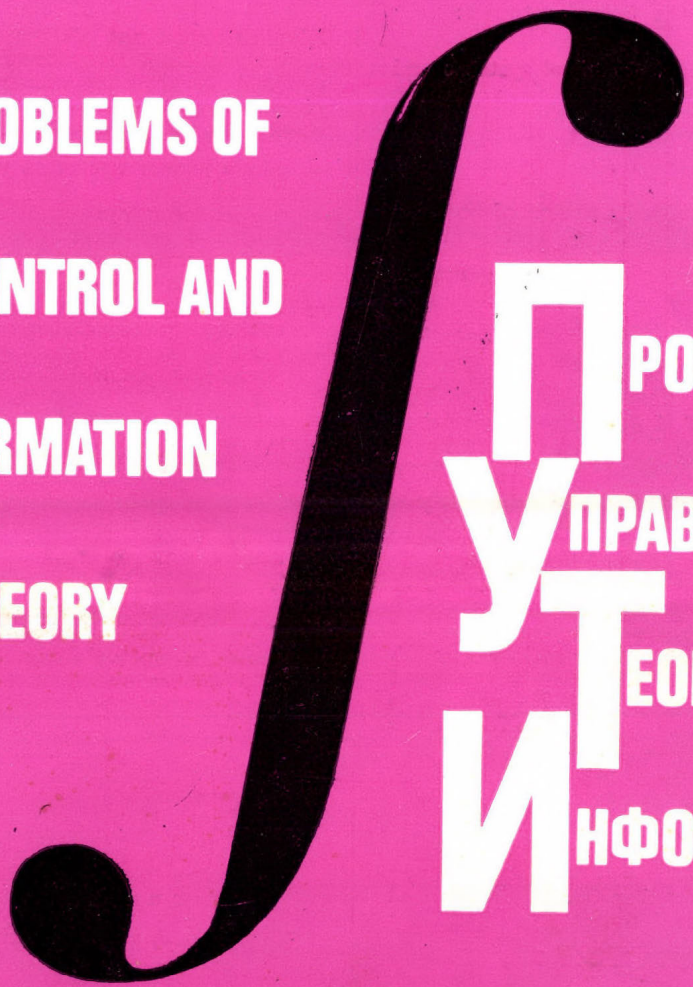
316.920
2
1973

VOL. 2 * NUMBER 1
TOM ТОМ * NUMBER НОМЕР 1

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY



ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1973

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

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

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

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

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

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

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

Budapest 62, P.O. Box 149, Hungary

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

Budapest 502, P.O BOX 24, Hungary

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал создан для обеспечения быстрейшей публикации материалов о новейших результатах научных исследований, разработок и кратких сообщений о достижениях в следующих областях науки:

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

Журнал издается четыре раза в год общим объёмом около 320 печатных страниц.

Журнал в первую очередь будет публиковать научные достижения обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

Будапешт 62, п. о. 149, Венгрия

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии Наук Венгрии

Будапешт 502, п. о. 24, Венгрия

PROBLEMS OF CONTROL AND INFORMATION THEORY

Volume 2, 1973

PAPERS

Apokorin, D. S.: Asymptotic properties of Markov chain parameter estimation. *2*, 3—4, pp. 219—234 219
Bassalygo, L. A.: The generalization of Lloyd's theorem on an arbitrary alphabet. *2*, 2, pp. 133—137 133
Bunich, A. L.—Raibman, N. S.: Dispersional identification equation of nonlinear plants. *2*, 3—4, pp. 257—272 257
Burnashev, M. V.—Has'minsky, R. Z.: Asymptotical expressions for the information in the channel with discontinuous noise density. *2*, 2, pp. 107—122 107
Csiszár, I.: On the capacity of noisy channels with arbitrary signal costs. *2*, 3—4, pp. 283—304 283
Fritz, J.: On a pattern classification algorithm of R. L. Kashyap. *2*, 1, pp. 81—93 ... 81
Gács, P.—Körner, J.: Common information is far less than mutual information. *2*, 2, pp. 149—162 149
Gavrilov, M. A.—Ostianu, V. M.—Bivol, L. G.: On properties of elements suitable for the synthesis of d-faultless structures. *2*, 2, pp. 139—147 139
Lagovier, B. A.: An analysis technique for asynchronous automata. *2*, 1, pp. 27—40.. 27
Markovskii, A. V.—Shipilina, L. B.: Realization of partially symmetric functions by networks of functional elements. *2*, 3—4, pp. 283—304 283
Meerov, M. V.—Litvak, B. L.: The optimization of multivariable systems and Markovian programming. *2*, 2, pp. 95—105 95
Mityugov, V. V.: Entropy, information and work in quantum statistics. *2*, 3—4, pp. 243—256 243
Ohera, L.: Basic considerations concerning frequency dictionaries. *2*, 3—4, pp. 305—314 305
Petrov, B. N.—Popov, E. P.: Fundamental problems of technical plants control science. *2*, 1, pp. 3—15 3
Przybylo, S.: Relations entre les morphismes et l'equivalence des automates indéterminés. *2*, 2, pp. 163—171 163
Rejtő, L.—Révész, P.: Density estimation and pattern classification. *2*, 1, pp. 67—80 67
Siforov, V. I.: Third International Symposium on Information Theory. *2*, 2, pp. 173—178 173
Somló, J.: On the computer aided design of control systems. *2*, 1, pp. 41—65 41
Vittikh, V. A.: Synthesis of algorithms for data compression. *2*, 3—4, pp. 235—241 .. 235
Willems, J. L.: Mean square stability criteria for linear white noise stochastic systems. *2*, 3—4, pp. 199—217 199
Yadrenko, M. I.: On optimal linear estimates of mean value and regression coefficients of homogeneous and isotropic random field. *2*, 1, pp. 17—26 17
Yu, P. L.—Leitmann, G.: Compromise solutions, domination structure and Salukvadze's solution. *2*, 3—4, pp. 183—197 183
Zinovjev, V. A.—Leontjev, V. K.: The nonexistence of perfect codes over Galois fields. *2*, 2, pp. 123—132 123

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

AUTHORS

Apokorin, D. S. *2*, 3—4, pp. 219—234
Bassalygo, L. A. *2*, pp. 133—137
Bivol, L. G. *2*, 2, pp. 139—147
Bunich, A. L. *2*, 3—4, pp. 257—272
Burnashev, M. V. *2*, 2, pp. 107—122
Csiszár, I. *2*, 3—4, pp. 283—304
Fritz, J. *2*, 1, pp. 81—93
Gács, P. *2*, 2, pp. 149—162
Gavrilov, M. A. *2*, 2, pp. 139—147
Has'minsky, R. Z. *2*, 2, pp. 107—122
Körner, J. *2*, 2, pp. 149—162
Lagovier, B. A. *2*, 1, pp. 27—40
Leitmann, G. *2*, 3—4, pp. 183—197
Leontjev, V. K. *2*, 2, pp. 123—132
Litvak, B. L. *2*, 2, pp. 95—105
Markovskii, A. V. *2*, 3—4, pp. 273—281
Meerov, M. V. *2*, 2, pp. 95—105
Mityugov, V. V. *2*, 3—4, pp. 243—256
Ohera, L. *2*, 3—4, pp. 305—314
Ostianu, V. M. *2*, 2, pp. 139—147
Petrov, B. N. *2*, 1, pp. 3—15
Popov, E. P. *2*, 1, pp. 3—15
Przybylo, S. *2*, 2, pp. 163—171
Raibman, N. S. *2*, 3—4, pp. 257—272
Rejtő, L. *2*, 1, pp. 67—80
Révész, P. *2*, 1, pp. 67—80
Siforov, V. I. *2*, 2, 173—178
Shipilina, L. B. *2*, 3—4, pp. 273—281
Somló, J. *2*, 1, pp. 41—65
Vittikh, V. A. *2*, 3—4, pp. 235—241
Willems, J. L. *2*, 3—4, pp. 199—217
Yadrenko, M. I. *2*, 1, 17—26
Yu, P. L. *2*, 3—4, pp. 183—197
Zinovjev, V. A. *2*, 2, pp. 123—132

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Том 2, 1973

СТАТЬИ

Апокорин Д. С.: Оценка параметра марковской цепи. *2*, 3—4, стр. 219—234. 219

Бассалыго Л. А.: Обобщение теоремы Ллойда на случай произвольного алфавита. *2*, 2, стр. 133—137. 133

Бунич А. Л., Райбман Н. С.: Дисперсионное уравнение идентификации нелинейных объектов. *2*, 3—4, стр. 257—272. 257

Бурнашев М. В., Хасьминский Р. З.: Асимптотические выражения для информации в канале с разрывной плотностью шума. *2*, 2, стр. 107—122. 107

Виллемс Й. Л.: Квадратичные критерии устойчивости для линейных стохастических систем с белыми шумами. *2*, 3—4, стр. 199—217. 199

Виттих В. А.: Синтез алгоритмов сжатия данных. *2*, 3—4, стр. 235—241. 235

Гаач П., Корнер Й.: Общая информация существенно больше, чем взаимная информация. *2*, 2, стр. 149—162. 149

Гаврилов М. А., Остиану В. М., Бивол Л. Г.: О свойствах элементов, удобных для синтеза d-безотказных структур. *2*, 2, стр. 139—147. 139

Зиновьев В. А., Леонтьев В. К.: Несуществование совершенных кодов над полями Галуа. *2*, 2, стр. 123—133. 123

Йу П. Л., Лейтманн Г.: Компромиссные решения, доминирующие структуры и решение Салуквадзе. *1*, 3—4, стр. 183—197. 183

Лаговиер Б. А.: Метод анализа асинхронных автоматов. *2*, 1, стр. 27—40. 27

Марковский А. В., Шипилина Л. Б.: Реализация частично-симметрических функций схемами из функциональных элементов. *2*, 3—4, стр. 273—281. 273

Мееров М. В., Литвак Б. Л.: Оптимизация многосвязных систем и марковское программирование. *2*, 2, стр. 95—105. 95

Митюгов В. В.: Энтропия, информация и работа в квантовой статистике. *2*, 3—4, стр. 243—256. 243

Охера Л.: Общие методы составления частотных словарей. *2*, 3—4, стр. 305—314. . 305

Петров Б. Н., Попов Е. П.: Основные проблемы науки об управлении техническими объектами. *2*, 1, стр. 3—15. 3

Пршибило С.: О связях между морфизмом и эквивалентностью недетерминированных автоматов. *2*, 2, стр. 163—171. 163

Рейто Л., Ревес П.: Оценка плотности вероятностей и классификация образов. *2*, 1, стр. 67—80. 67

Сифоров В. И.: Третий Международный симпозиум по теории информации. *2*, 2, стр. 173—178. 173

Фриц Й.: Об алгоритме распознавания образов по Р. Л. Кашиап. *2*, 1, стр. 81—93. .. 81

Чисар И.: О проблеме емкости каналов с помехами при любых затратах на сигналы. *2*, 3—4, стр. 283—304. 283

Шомло Я.: Проектирование систем автоматического регулирования при помощи ЦВМ. *2*, 1, стр. 41—65. 41

Ядренко М. И.: Об оптимальных линейных оценках математического ожидания и коэффициентов регрессии однородных и изотропных случайных полей. *2*, 1, стр. 17—26. 17

АВТОРЫ

Апокорин Д. С. *2*, 3–4, стр. 219–234
Бассалыго Л. А. *2*, 2, стр. 133–137
Бивол Л. Г. *2*, 2, стр. 139–147
Бунич А. Л. *2*, 3–4, стр. 257–272
Бурнашев М. В. *2*, 2, стр. 107–122
Виллемс Й. Л. *2*, 3–4, стр. 199–217
Виттих В. А. *8*, 3–4, стр. 235–241
Гаач П. *2*, 2, стр. 149–162
Гаврилов М. А. *2*, 2, стр. 139–147
Зиновьев В. А. *2*, 2, стр. 123–132
Йу П. Л. *2*, 3–4, стр. 183–197
Лаговиер Б. А. *2*, 1, стр. 27–40
Лейтманн Г. *2*, 3–4, стр. 183–197
Леонтьев В. К. *2*, 2, стр. 123–132
Литвак Б. Л. *2*, 2, стр. 95–105
Корнер Й. *2*, 2, стр. 149–163
Марковский А. В. *2*, 3–4, стр. 273–281
Мееров М. В., *2*, 2, стр. 95–105
Митюгов В. В. *2*, 3–4, стр. 243–256
Остиану В. М. *2*, 2, стр. 139–147
Охера Л. *2*, 3–4, стр. 305–314
Петров Б. Н. *2*, 1, стр. 3–15
Попов Е. П. *2*, 1, стр. 3–15
Пршибило С. *2*, 2, стр. 163–171
Райбман Н. С. *2*, 3–4, стр. 257–272
Ревес П. *2*, 1, стр. 67–80
Рейто Л. *2*, 1, стр. 67–80
Сифоров В. И. *2*, 2, стр. 173–178
Фриц Й. *2*, 1, стр. 81–93
Хасьминский Р. З. *2*, 2, стр. 107–122
Чисар И. *2*, 3–4, стр. 283–304
Шипилина Л. Б. *2*, 3–4, стр. 273–281
Шомло Я. *2*, 1, стр. 41–65
Ядренко М. И. *2*, 1, стр. 17–26

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

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

DEPUTY EDITORS

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

CO-ORDINATING EDITORS

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

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

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

B. S. SOTSKOV (Moscow)
L. VARGA (Budapest)

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

РЕДАКТОРЫ ЖУРНАЛА

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

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

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

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

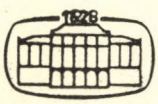
С. В. ЕМЕЛЬЯНОВ (Москва)
Л. КАЛМАР (Будапешт)

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

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

Б. С. СОТСКОВ (Москва)
Л. ВАРГА (Будапешт)

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



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

Problems of Control and Information Theory, Vol. 2 (1), pp. 3—15 (1973)

FUNDAMENTAL PROBLEMS OF TECHNICAL PLANTS CONTROL SCIENCE

B. N. PETROV, E. P. POPOV

(Moscow)

(Received November 22, 1972)

1. Introduction

Scientific problems of technical plants control are extremely various. They are connected with many complex modern problems of mathematics, mechanics, physics, chemistry, physiology and engineering psychology. The number of technical plants to be scientifically controlled is great. The set of tasks, more and more complex, are also numerous. In this connection there exist many scientific problems both in theory and technique of control each being a subject for many special scientific experiments. In order to bring this scientific research complex under purposeful control, to forecast and plan it, to single out the most urgent problems and to call attention to the "blank spaces" in this branch of science, it is expedient to somehow systematize all these problems. This task might be accomplished only by a great number of scientists and engineers of different specialities. We shall confine

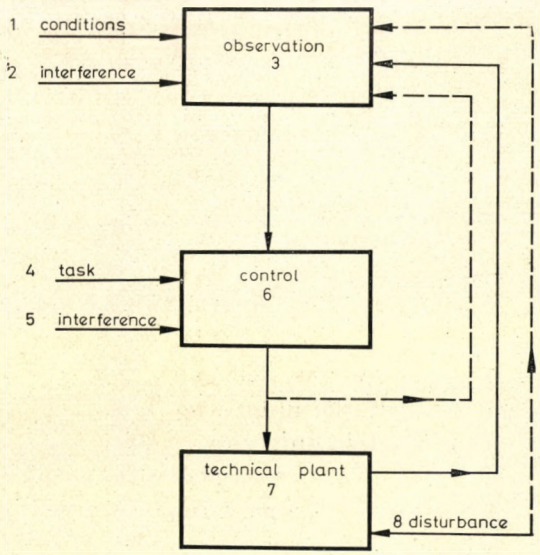


Fig. 1

MAGYAR
TUDOMÁNYOS AKADÉMIA
— KÖNYVTÁRA

ourselves to stating only general trends as feasible basis for the future systematized description and estimation of separate problems. The picture presents a generalized block diagram of any technical plant control system, divided into three large blocks: technical plant, observation and control.

Solid arrows indicate main internal and external influences and flows of information, and dotted arrows show additional possible influences and information. Technical plant is a technical system to be controlled. Observation includes the whole complex of processes and means of gaining and handling information on the technical plant condition and outside situation in order to provide the data necessary for future shaping of control signals.

Control comprises the whole complex of processes and means of control signals shaping and execution, correcting and intensifying mechanisms and units effecting the technical plant according to the determined optimal criteria and the selected regularities as its algorithms.

2. Controls

Different types of control are composed as combination of various means and methods of observation and control.

Observation complex (information-measuring part of the system) may vary in physical and technical design but their functioning principles usually are unique. One of the main observation tasks is measurement of actual controlled variables in control process (solid lines in the diagram) and sometimes control or disturbance variables (dotted arrows) and their feeding to control block.

Alongside with direct measurement of control or disturbance variable by some instrument often goes measurement with data handling. It might be applied for different aims:

a) strong interference bampers to make out the needed signal;

b) direct variable (or disturbance) control is impossible and it is to be controlled indirectly by measuring some other variable and data handling;

c) calculating a definite quality criterion for self-adjustment of control block by proper data handling.

In terminal control system as well as in other systems in question, measurement is permanent but the result is only accumulated and processed without feeding to control block in order to provide a command for the control block at some definite moment or at intermediate moments of movement or work. Proper pilotage system is the most important part of many movable technical plant observation blocks (inestial, radiotechnical, astronomic, magnetic and the like).

Coordinators of different orientators (optical, radiotechnical etc.) form another group of observation devices fixing necessary changes and external factors and feeding the data to control block. It is connected with the most general task of reflecting external factors and internal conditions of a technical plant, if it is necessary for control especially in case of human participation when visual observation is impossible.

Observation with adoptation presupposes possibility of automated changes in system structure or parameters or algorithmic methods for control purposes in case of environment or internal factors changed. All the before-mentioned types of observation may be continuous or granular, that is essential for the theory of the problem. Let us proceed to control block characteristics. Elementary types of control are stabilization and following with fixed parameter constant in the whole process or following reference input.

In terminal control it is important to realize the definite control task at the end of the process while all transient conditions of the process may be arbitrary in definite boundaries, providing no breaks-down will happen and the fixed final task will be realized. Three main cases may be distinguished: the endpoint is given in time and it is given in space (eg. the task of the cosmic apparata is the smooth landing to the surface of a planet) or a restricted task is given which is not in connection with time and with a given space (eg. coupling of space-ships). Controls are classified according to program types:

a) time control program, when controlled variable is programmed as definite time function;

b) parameter control program when controlled variable is a definite function of another variable as it is with the laws of approach control;

c) flexible control program (especially with terminal control) when the task is to realize programmed condition of the system at the end and there is no permanent control of transient conditions.

In this case no program is set beforehand but it is worked out in the process of control depending on transient conditions of the system, related to the final task. Through this great control flexibility is achieved self optimalization of control programs with recognition of actual transient conditions, occasionally changing.

Any type of control mentioned may be applied under permanent structure and parameters of control system but they may vary as well. Thus we may distinguish the following systems:

a) with programmed change of system parameters, and may be structure also; such change usually goes by jumps at definite moments (especially in case of variable parameters);

b) with variable structure when structure (and parameters) vary depending for example on divergence degree and controlled coordinate speed or on definite logical complex of indexes of the system conditions;

c) with adaptation transient quality on exponents of control process; these are self-adjusting systems analysing transient quality of control process with automatically changing structure or parameters, depending on final task;

d) with adaptation on external conditions and internal disturbances, these are self-adjusting systems analysing external factors according to some fixed criterion (for example general efficiency of the system, or power consumption) or environment changes, or disturbances (inside and outside) according to control task that brings about automatic change of system parameters and structure for the achievement of terminal task.

Guidance and docking are also specific types of control when there are two objects, one of them (or sometimes both) controlled by the system. In this case terminal control purpose might be clash or close enough passing as for soft docking or close approach (for refuelling, testing, etc.).

The task of following and escaping presupposes two or more objects.

Here may be such situations when pursuer must choose the most important object among several objects by some special feature. And the second object must escape with the help of interference or manoeuvre etc.

Sometimes group pursuit is possible with rational arrangement of escaping objects among following ones under the common control system. Here game and conflict situations are realized in their dynamic aspect. In any case control may be pure automatic or automatized with human participation in combination with control instruments which provide people with data sources for control and then transform and reinforce control functions of human-operator.

Any type of the above systems may be based on continuous control principles or granular control which is important for the theory as well as for the practice of control system design.

In control system design almost any type of observation in combination with any type of control may be applied to different types of technical plants which results in many different concrete systems.

3. Basic Trends of Control Theory Development

All existing and developing theories of control and observation may be divided into three large classes: determined, stochastic and game. According to an other principle they are classified in the following way: theory of stabi-

lity and quality of control; theory of dynamic precision of control and observation processes in the presence of interferences and external influence. The theory of invariance control and observation in relation to external influences and interferences belongs to here whose task is to guarantee the minimal dependence of the controllable quantities of these effects as well as the theory of perceptibility that deals with the influence of changes in system parameters upon quality and accuracy of control processes including systems with casual changes of system parameters and also for non-linear and discrete systems. The theory of optimal data control is very important in processing and transformation of information in case of several control devices as well, also for estimation of controlled variable or disturbance by indirect signs for environment reflection, for adaptation of purposes etc.

For all the afore-mentioned trends the most important element of modern control theory is the theory of optimal control. Approximative optimalization of control is necessary in many variants for strict optimalization of control very often brings about non-existing results thus it is often unsuitable for complicated systems or inexpedient for other reasons.

So the next problem is control system with approximative optimalization and also different algorithmic methods of optimalization with digital computers for complex controls. In both cases the most important task of applied sciences is selection of expedient criteria of optimalization for purpose of the best system realzation and satisfactory design procedure. A primitive criterion, chosen for the sake, of simplicity cannot bring proper solution, no matter how close the optimalization is. Another important task of this aspect is limit fornalization laid on control system in the process of optimalization. Proper limits application can meet some requirements to control system, thus preventing excessive complication of optimal criterion. Applied theory of the problem allows to find the best distribution of the whole complex of requirements to the system, between criteria and limits.

The problem of identification is of independent importance both in solving the tasks of proper mathematical models of real plants and systems of control and observation and also other tasks such as laboratory and full-scale testing. Theory of reliability helps to investigate another important characteristics of control system. It combines mathematical theories with some physical aspects which are indispensable for achievement of desirable results and also theory of general control efficiency based on realiability, exactness, efficiency, productivity etc.

The most important task of applied theory is here the problem of preferable reasonable criteria of control system and observation applied to technical plants of different types. In case of human participation in control or

observation system a well-grounded definition of human functions must be provided in combination with technical design of observation and control taking into consideration engineering psychology data. This is the task of all human-operator features detailed research considering psychological physiological and other factors under the circumstances. So there must be a kind of mathematical description of human operator functions as a part of the system. This allows optimalization of man-machine system based on definite criteria for each special control system. Theory of planning and prognosis, recently applied, became an indispensable element of design. Its application allows to take into account future development and perfection of controlled technical plants in connection with inevitable development and modification of tasks performed and probable change of control purposes.

Those problems are related to the tasks of control over economy and human collectives. Thus a new theory of organized control appears based on new approach to control tasks and mathematical interpretation alongside with all other methods of control theory spoken above. Great variety of principles and methods of control theory that are applied to several aspects of theory and human activities and the importance of tasks, brings the necessity of generali zation of principles that must provide a common basis for all those numerous principles and methods. This is the aim of modern general system theory.

4. Questions of Control System Realization

Control system realization includes tasks of design, manufacturing, testing and exploitation of systems. All technical devices used in observation and control systems based on various physical principles may be divided into three large classes according to definite control processes:

a) continuous systems;

b) granular and logical systems;

c) digital computers. with fully developed memory.

The design of new technical means of control and observation is often connected with application of new physical principles. This is the case with manufacturing of new perceptible elements for observation systems, such as navigation, new computing, transmitting, intensifying and executing mechanisms. Those physical principles are not nessessarily new in general science but new in application for the tasks mentioned. Chmical principles may be well used alongside with physical ones.

Nowadays some bionic principles are developd — study of processes and mechanisms of observation and control of living creatures for future application in technics.

Control and observation systems design presents a very important and labour-consuming task, which requires a considerable time and greatly determines the spreading of the new technique.

Design automatization with digital computer is one of the most important problems of engineering progress, for possible contacts of operators with digital computer developed nowadays allows to transfer to digital computers a number of stages in modeling of a new system that previously have been performed by models. This is of great importance for it allows to avoid expensive and long modeling. Employing displays in systems of computer-aided design makes it possible that we may demonstrate on the screen a great variety of the operation of designed systems (eg. technical chains, diagrams). With the reflection of the process before his eyes the operator can immediately improve diagrams and the change in the process is immediately reflected on the screen.

This helps to avoid unnecessary modeling.

In the process of design the system must also be provided with special device that will put it into starting action in a definite period of time.

Principle of reliability based on physical and mathematical aspects is also connected with a complex of processes requiring special scientific studying and organization. It is necessary to elaborate this principle in the earliest stage of the design.

The process of design also envisages maintenance, adjustment, control and repair conveniences based on scientific experiment. One of the important criteria of design is the economic index. Versatile process of design and control system realization and its complexity requires scientific elaboration, providing strict consideration of multiple factors, theoretical and practical demands and also optimal distribution of resources and stages of elaboration (considering time and man-power). The subject of the thorough basic and applied science experiment presents technology and organization laboratory and fullscale testing. System approach to control over large groups of technical plants (machine-tools, shops, experimental stands etc.) exercised by a digital computer.

Scientific investigation is also indispensable for generalization of observation and control system exploitation process to provide most objective information that is of great importance not only for the present-day but also for the future technical systems.

And finally true technical revolution in industry, as well as in exploitation and testing of technical systems must be accomplished by robots and manipulators, that will complete complex automation. Nowadays more or less satisfactory automation and control is achieved over main process of produc-

tion, exploitation and testing. But multiple junction and "subsidiary" process connected with rearrangement and erection of machine parts and blocks maintenance and emergency repairs are very often outside the stage of automation. Here manipulators and teleoperators present the main link of automation. It is of special importance in case of unhealthy trade and in special circumstances (vacuum, high temperature, high pressure). Besides manipulators will replace human operators in many production lines and other process. The so-called supervisory remote controls and robots are of a higher stage and multifunctional technical mechanism with elements artificial intellect and accordingly this fact yields wide range of possibilities to perform complex operation with the help of men, sometimes without it greater amount of operations performed then today's manipulators.

It results absolutely new qualities in high mobility and flexibility. Besides such self-contained systems will find great application for investigation of areas which are difficult to reach, namely the opposite part of the moon surface, other planets and bottoms of seas and then Ocean.

5. Objects of Complex Studies of Control Systems

Problems of complex studies of control systems are most difficult and new issues from both scientific and practical points of view. Those are not only problems of technical sciences but also great problems of scientific organization. This may be applied not only to design control systems through separate channels and isolated instruments and technical devices although many new scientific issues appears here too. We speak now about complex control systems for complicated tasks and groups of technical plants where it is especially important to realize the necessity and essence of such scientifically based complex study of control systems. With the development of digital computers its importance and practical application are constantly growing. It is impossible to concretize many perspective scientific problems of this sphere within the limits of an article. So let us take up only general principles. Complex control systems must as a rule be worked out not for separate technical plants but for large-scale tasks, for the solution of which technical objects themselves are projected.

Nowadays one cannot hope to quickly provide control system of ready made instruments and devices for already erected plant even in case we actually dispose of such. Control systems must be developped simultaneously with the plant and may have considerable effect upon their design and general characterictics. Absence of such preliminary stage of complex studies of control systems of a large technical plant as a unit might bring about joining of

system parts in the process of testing of controlled plants. It results in additional transient passes, transforming, encoding and discoding devices. Thus appears small realiability of joints of the system parts and too bulky apparatus. Alongside inevitably appears reduction of general efficiency of the system.

System approach — complex principle and technical studies of system as a whole is the main thing for progress in the field of current complex control system and also for efficient application of large technical plants of any kind. The organization of control systems, complex studies itself has become a subject of scientific investigation for not yet genius could manage it depending only on his experience and intuition. Hence general complex studies of control systems and optimal erection of a unity should always be based on-up to-day scientific methods of complex systems control.

Here we may applie (and this is practical) the afore-mentioned terminal control principles, with corrections in the process of development for speeding and optimalization of quality and efficiency of the whole process. It is an important problem to develop the technical means of joint information processing and exchange inside the system itself, and also between the system and outer world.

Human participation in control loop brings about the necessity of unique principles of elaborating logical demonstrative systems, which simplify the decision and and transmits human operators's commanding signals. Another important feature of the system under discussion is its orientation not only on one particular narrow purpose but with due repair for possible analogues, and consequently with the account of future development of controlled technical plants.

It will help to avoid so called "deadlock" programs that can be applied to the given concrete tasks only and provide possibility of future development of the system itself to meet the requirements of growing problems. Besides technical reasons it is also important in terms of economy, for modern complex control systems are very expensive. In this connection it is indispensable to provide a large quantity of independent control system programs for definite groups of tasks in parallel to programs of development of technical plants for the same groups of tasks. The development of programs for unified base elements origination both for electronic and electromechanic and other mechanisms and their miniature models to heighten the general technical level reliability and lower cost of control systems production.

The problem of the most effective system testing, control and exploitation of the whole technical plant control system complex including rational distribution of laboratory and full-scale testing is also important.

The development of science and electronic computing machines enlarges the application of modelling in all controlled processes in complex, even in desing before testing. It results in rapid extension of laboratory experiment and provides great reliability and accident — free first full-scape testing, acceleration of the process and high efficiency. Rational determination of parameters is total, that must be reduced to minimum quantity of generalized indices of the control process quality is very important for the least charge on trail tie lines and quick processing of the data (for some parameters — in the tempo of experiment process itself and for other — with possibly shortened time delay).

Future orientation of many technical plants general control task on to days' methods is impossible because of diversity, high speed and responsibility of the whole control system functions.

Making effective decisions for management requires usually handling of great amount of data. It requires not only computers but, which is much more important, great amount of mathematical supply with a lot of special technical instruments, often more expensive then computers themselves.

As a whole control problems are of particular significance in the general problem of large engineering plants creation and also testing and exploitation. Technical systems perfection results in the increase of the control systems' role. And so the importance of future control system complex study and its scientifically based organization is increased.

Reliability of the whole complex is extremely important and is achieved through great variety of arrangements at all stages of construction, production and exploitation.

One can say that perfection of technical plants control systems and assemblies determines in the long run practical efficiency of the plants, for in case of poor control system no machine, be even perfect and powerful, can be efficient enough.

And finally problems of concrete technical plants control in different aspects, as applied sciences, are closely linked with the general control theory as a fundamental science. This connection is essential for the best application of general theory achievements as well as for reciprocal statement of new problems before the general control theory.

Основные проблемы науки об управлении техническими объектами

Б. Н. ПЕТРОВ, Е. П. ПОПОВ

(Москва)

Статья содержит пять разделов: 1) введение, 2) виды управления, 3) основные направления развития теории управления, 4) вопросы реализации систем управления, 5) задачи комплексной проработки систем управления.

1. Научные проблемы управления техническими объектами чрезвычайно разнообразны. Развитие вычислительной техники приводит к новым возможностям совершенствования и комплексирования систем управления, причем возникают совершенно новые принципы управления, включающие распознавание образов и создание искусственного интеллекта. Становится возможным более полное использование различной информации с оптимальной ее обработкой в интересах управления.

Использование вычислительной техники коренным образом меняет также организацию и методы проектирования, испытания и эксплуатации систем управления. Большое значение приобретают человеко-машинные проблемы.

2. Различные виды управления складываются как комбинации разнообразных средств и методов наблюдения и управления.

В комплексе задач наблюдения (информационно-измерительная часть системы), кроме непосредственного измерения управляемой величины используется измерение с обработкой информации при самых различных алгоритмах обработки в зависимости от поставленных целей и предъявляемых требований к выдаче информации, а также от природы источников информации.

Классические виды управления — стабилизация и слежение — обогащаются идеями различных форм адаптации с гибкими программами управления, с перестройкой параметров и структуры систем в зависимости от поставленных задач, изменения внешней обстановки, внутренних качественных факторов и т. п.

Особое значение приобретает терминальное управление, когда цель состоит в осуществлении заданного состояния системы в конце (или в другой определенный момент) и отсутствует непрерывное регулирование промежуточных состояний, лишь бы только система не попадала в аварийные режимы. При этом достигается наибольшая гибкость управления с самооптимизацией процесса управления по каким-либо важным характеристикам, связанным с общей эффективностью действия всей системы в целом при изменяющихся внешних условиях.

Важными видами управления являются также наведение и сближение, в том числе в задачах преследования и уклонения, где реализуются игровые и конфликтные ситуации в их динамическом аспекте.

Существенным является крупное комплексирование различных средств управления и наблюдения в единые большие системы, как автоматические, так и с участием человека.

3. Основные направления развития теории управления определяются указанными прикладными задачами, а также возможностями математических (аналитических и алгоритмических) методов в сочетании с цифровым и аналоговым моделированием.

Все используемые теории можно разделить на три больших класса: детерминированные, стохастические и игровые. По другим признакам различаются: теория устойчивости и качества процессов управления, теория динамической точности (включая инвариантность, чувствительность и др.) для линейных, нелинейных, непрерывных и дискретных систем.

Важнейшим элементом является теория оптимального управления не только со строгой оптимизацией, но и с различными методами приближенной оптимизации для достижения лучшей реализуемости системы на данном объекте. Столь же важна и теория оптимальной обработки информации, в том числе и для целей адаптации.

Проблема назначения целесообразных критериев оптимальности, надежности и общей эффективности управления приобретает важное, а часто и решающее значение в связи с необходимостью удовлетворения целому комплексу различных требований, но без излишнего усложнения процесса проектирования и самой проектируемой системы управления.

Самостоятельное значение приобретает проблема идентификации как в решении задачи правильного построения математических моделей реальных объектов и систем, так и при постановке лабораторных и натурных испытаний.

Все сказанное относится не только к полностью автоматическим системам, но также и к системам, включающим человека в контур управления. При этом процессу оптимизации его деятельности в технической системе предшествует детальное изучение всех свойств человека-оператора как звена системы с привлечением данных инженерной психологии и других факторов.

Использование теории планирования и прогнозирования позволяет учитывать будущее развитие и совершенствование управляемых технических объектов, что составляет важный элемент проектирования.

Теория организационного управления и общая теория систем содержат обобщения принципов и методов теории управления на системы, отличающиеся от чисто технических систем, но и для последних эти теории позволяют найти новые подходы к решению задач.

4. К вопросам реализации систем управления относятся задачи проектирования, изготовления, испытания и эксплуатации систем.

Проектирование новых технических средств управления и наблюдения часто связано с использованием новых физических, а иногда и бионических принципов.

Сами методы и организация проектирования включают комплекс прикладных научных исследований. Автоматизация проектирования с использованием ЦВМ — одна из важнейших проблем технического прогресса. При этом на экране дисплея можно видеть различные стороны построения и функционирования создаваемой системы, что позволяет частично заменить дорогостоящее и затяжное макетирование.

Испытания и эксплуатация систем немыслима без научной проработки и научной организации.

Все эти процессы в целом требуют научно поставленного управления сложными разработками и их реализацией, позволяющего учесть многочисленные требования и оптимально распределить этапы работы и имеющиеся ресурсы.

Большую роль в производстве, эксплуатации, ремонте и т. п. призваны сыграть работы и манипуляторы, проблемы управления которыми стоят сейчас в числе важнейших. Они придают системам высокую мобильность и автономность, в том числе и при исследовании и освоении трудно доступных пространств.

5. Задачи комплексной проработки систем управления являются сейчас важнейшими и решающими для наилучшего управления сложными техническими объектами и группами объектов, особенно находящихся в процессе движения. Если управление по отдельным каналам и частям общего комплекса разработано, это еще не гарантирует качественного функционирования всего комплекса в целом. Именно с этим часто бывают связаны неудачи, заставляющие дорабатывать и состыковывать отдельные части системы уже в процессе испытаний, что ведет к появлению лишних переходных устройств, к снижению надежности и общей эффективности комплекса.

Системный подход — комплексная идейная и техническая проработка системы в целом — самое главное для технического прогресса в области сложных современных систем управления. Кроме научных проблем здесь важно решение и технических проблем создания единых средств согласования устройств, единых принципов и языка обработки информации.

Важнейшей стороной при этом является учет перспектив развития задач управления в данной области техники, чтобы создавать дорогостоящий комплекс не только на данную узкую задачу, но с учетом ее будущего развития, чтобы избежать так называемых «тупиковых программ» и закладывать в систему возможности ее дальнейшего совершенствования в желаемых направлениях.

Другой важной стороной комплексной проработки систем управления является создание системы не под готовый уже объект, а совместная разработка системы управления с самого начала проектирования или модернизации данного объекта или группы объектов под крупные новые задачи. Нельзя надеяться быстро сделать удачное управление даже из готовых приборов под готовый объект. Это надолго затягивает окончание разработки в целом и может заставить частично переделывать и сам объект.

В процессе испытаний и эксплуатации систем для принятия эффективных решений руководством требуется, как правило, обработка больших массивов различной информации. Это предполагает не только наличие вычислительных машин, но и самое главное — значительный объем математического обеспечения, а также и наличие ряда специальных дополнительных устройств на входе и выходе машины. Это сейчас стало более важной проблемой, чем сама вычислительная машина.

В целом проблемы управления выходят в настоящее время на передний край в общем комплексе проблем создания крупных технических объектов. Совершенством управления в значительной мере определяется общая эффективность действия любого технического объекта или комплекса.

Problems of Control and Information Theory, Vol. 2 (1), pp. 17—26 (1973)

ОБ ОПТИМАЛЬНЫХ ЛИНЕЙНЫХ ОЦЕНКАХ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ И КОЭФФИЦИЕНТОВ РЕГРЕССИИ ОДНОРОДНЫХ И ИЗОТРОПНЫХ СЛУЧАЙНЫХ ПОЛЕЙ

М. И. ЯДРЕНКО

(Киев)

(Поступила в редакцию 10 сентября 1971 г.)

Рассматриваются оптимальные в смысле минимума среднеквадратической ошибки линейные несмещенные оценки математического ожидания и коэффициентов регрессии однородного и изотропного случайного поля, наблюдаемого на некоторой окружности. Получены оптимальные оценки среднего значения однородного и изотропного поля в том случае, когда измерения проведены в вершинах некоторого правильного многоугольника или на системе концентрических окружностей.

1. Постановка задачи

При решении ряда задач теории автоматического управления, геологии, метеорологии, геофизики, теории распознавания образов, радиофизики, статистической оптики и голографии необходимо получать статистические оценки некоторых параметров случайных полей (например, математического ожидания, коэффициентов регрессии). Укажем еще упоминавшуюся в [1] задачу об оценке температуры во вращающейся цементной печи; результаты настоящей статьи могут быть использованы для решения этой задачи.

Пусть $\eta(p)$ — случайная функция, зависящая от точки плоскости (случайное поле на плоскости) с нулевым математическим ожиданием и известной корреляционной функцией. Предположим, что на некотором множестве D плоскости наблюдается случайное поле $\xi(p) = m + \eta(p)$ или $\xi(p) = \sum_{s=1}^{N} a_s g_s(p) + \eta(p)$, где параметры $m, a_1, \ldots, a_N$ неизвестны, а $g_s(p)$ $(s = 1, N)$ — некоторые известные функции. Требуется построить оценки $\hat{m}$, $\hat{a}_s$ параметров m, a_s, линейно выражающиеся через результаты наблюдений, несмещенные ($M\hat{m} = m$, $M\hat{a}_s = a_s$) и обладающие наименьшими среднеквадратическими погрешностями среди всех линейных и несмещенных оценок. В дальнейшем такие оценки будем называть кратко оптимальными линейными оценками.

Вид оптимальных оценок зависит, конечно, от множества D, на котором произведены наблюдения. Задачу построения оптимальной оценки неизвестного математического ожидания m будем рассматривать в том случае, когда множество D — система концентрических окружностей, или множество вер-

шин некоторого правильного многоугольника, а $\eta(p)$ — однородное и изотропное случайное поле. Тогда

$$M\eta(p_1)\eta(p_2) = B(\varrho_{p_1 p_2}) \tag{1.1}$$

зависит лишь от расстояния $\varrho_{p_1 p_2}$ между точками p_1 и p_2.

В § 4 оптимальные оценки коэффициентов регрессии приведены в том случае, когда множество D — окружность.

2. Спектральное разложение однородного и изотропного случайного поля

Функция $B(\varrho)$ для однородного и изотропного случайного поля имеет вид [2]

$$B(\varrho) = \int_0^\infty y_0(\lambda\varrho)\, d\Phi(\lambda) \tag{2.1}$$

где $\Phi(\lambda)$ — некоторая неубывающая функция на $(0, +\infty)$, $J_0(x)$ — бесселева функция первого рода нулевого порядка [3]. Само случайное поле представимо в виде [4], [5]

$$\eta(p) = \sum_{n=-\infty}^{\infty} e^{in\varphi} \int_0^\infty y_n(\lambda r)\, dZ_n(\lambda) \tag{2.2}$$

где (r, φ) — полярные координаты точки p, $J_n(x)$ — бесселева функция n-ого порядка, $Z_n(\lambda)$ — последовательность случайных процессов с некоррелированными приращениями таких, что

$$M\, dZ_n(\lambda)\, \overline{dZ_m}(\lambda) = \delta_n^m\, d\Phi(\lambda) \tag{2.3}$$

(черта означает символ комплексного сопряжения).

Представление (2.2) не стало еще, к сожалению, достаточно популярным среди специалистов, занимающихся приложениями теории случайных полей, хотя оно существенно описывает свойство однородности и изотропности и значительно облегчает решение основных статистических задач.

В дальнейшем важную роль будут играть функции

$$b_k(r) = \int_0^\infty y_k^2(\lambda r)\, d\Phi(\lambda) = \frac{1}{\pi}\int_0^\pi e^{-2ikx} B(2r\sin x)\, dx \tag{2.4}$$

которые мы будем называть спектральными коэффициентами случайного поля $\xi(r, \varphi)$.

3. Оценка математического ожидания

Предположим, что на множестве D точек плоскости наблюдается случайное поле $\xi(p) = m + \eta(p)$, где $M\eta(p) = 0$. Пусть $\mathcal{H}_D$ — замкнутое (в смысле среднеквадратической сходимости) линейное многообразие, порожденное случайными величинами $\xi(p)$, $p \in D$, то есть совокупность всевозможных линейных комбинаций величин $\xi(p)$, $p \in D$ и их пределов в среднем квадратичном. Пусть Γ — замкнутое линейное многообразие, порожденное случайными величинами вида $\sum\limits_{p_k \in D} \xi(p_k) c_k$, где $\Sigma c_k = 1$. Каждая случайная величина $\tilde{m}$ из Γ может рассматриваться в качестве линейной несмещенной оценки ($M\tilde{m} = m$). Как известно ([6], [7]), оценка с минимальной дисперсией $\hat{m}$ определяется из уравнения

$$M\eta(p)\,\hat{m} = c, \qquad p \in D, \tag{3.1}$$

где $c = D[\hat{m}]$.

Рассмотрим вопрос об оценке неизвестного среднего в том случае, когда множество D есть система концентрических окружностей.

Теорема 1. Пусть случайное поле $\xi(p) = m + \eta(p)$ наблюдается на концентрических окружностях с радиусами $R_1, \ldots, R_N$ и центром в начале координат. Тогда линейная несмещенная оценка m с минимальной дисперсией имеет вид

$$\hat{m} = \sum_{k=1}^{N} \frac{\Delta_k}{\sum\limits_{i=1}^{N} \Delta_i} \frac{1}{2\pi} \int_0^{2\pi} \xi(R_k, \varphi)\, d\varphi\,, \tag{3.2}$$

где Δ — определитель матрицы, (s, k)-ый элемент которой равен $\int\limits_0^{\infty} J_0(\lambda R_k) \cdot J_0(\lambda R_s)\, d\Phi(\lambda)$, а определитель Δ_i получается из определителя Δ заменой всех элементов i-ого столбца на 1. Дисперсия оценки $\hat{m}$ равна

$$\mathcal{D}\,\hat{m} = \frac{\Delta}{\sum\limits_{i=1}^{N} \Delta_i}. \tag{3.3}$$

Пусть

$$\xi_k(r) = \frac{1}{2\pi} \int_0^{2\pi} e^{-ik\varphi}\, \xi(r, \varphi)\, d\varphi\,.$$

В силу (2.2)

$$\xi_k(r) = \begin{cases} \int_0^\infty J_k(\lambda r)\, dZ_k(\lambda) \\ m + \int_0^\infty J_0(\lambda r)\, dZ_0(\lambda) \end{cases}. \tag{3.4}$$

Каждый элемент $\tilde{m}$ из $\mathcal{H}_D$ имеет вид:

$$\hat{m} = \sum_{i=1}^{N} \sum_{n=-\infty}^{\infty} c_{ni}\, \xi_n(R_i)\,, \tag{3.5}$$

где $\xi_n(R_i)$ определяется равенствами (3.4). Основное уравнение (3.1) принимает вид

$$\sum_{s=1}^{N} \sum_{n=-\infty}^{\infty} c_{ns}\, e^{in\varphi} \int_0^\infty J_n(\lambda R_s)\, J_n(\lambda R_k)\, d\Phi(\lambda) \tag{3.6}$$

при всех $0 \leq \varphi \leq 2\pi$ и $k = 1, \ldots, N$. Отсюда получаем систему уравнений для c_{0s}

$$\sum_{s=1}^{N} c_{0s} \int_0^\infty J_0(\lambda R_s)\, J_0(\lambda R_k)\, d\Phi(\lambda) = c\,. \tag{3.7}$$

Принимая во внимание условие несмещенности $\sum_{k=1}^{N} c_{0k} = 1$, получим

$$c_{0s} = \frac{c\,\Delta_s}{\Delta}\,, \quad c = \frac{\Delta}{\sum_{s=1}^{N} \Delta_s}\,. \tag{3.8}$$

В частном случае, когда наблюдения проводятся на одной окружности, из теоремы 1 получаем такое следствие:

среди всех несмещенных оценок m наименьшую дисперсию имеет оценка

$$\hat{m} = \frac{1}{2\pi} \int_0^{2\pi} \xi(r, \varphi)\, d\varphi\,, \tag{3.9}$$

представляющая собой среднее $\xi(r, \varphi)$ по окружности радиуса r; при этом

$$D\hat{m} = b_0(r) = \int_0^\infty J_0^2(\lambda r)\, d\Phi(\lambda)\,. \tag{3.10}$$

Интересно заметить, что в рассмотренном случае вид оптимальной оценки не зависит от корреляционной функции случайного поля $\eta(r, \varphi)$.

Естественно возникает вопрос о поведении оценки (3.9) при увеличении радиуса r. Ответ можно получить из результатов [8], где исследуется поведение дисперсий сферических средних. Оказывается, что, если $\Phi'(\lambda) = \lambda^{\alpha} g(\lambda)$, где $o \leq \alpha < 2$, $g(\lambda)$ ограничена и непрерывна в нуле, $g(o) \neq 0$, то

$$D\hat{m} = \frac{1}{R^{1+\alpha}} g(0)\, c_{\alpha} + 0\left(\frac{1}{R^{1+\alpha}}\right).$$

Предположим, что корреляционная функция случайного поля $\eta(p)$ известна. Поэтому при практическом применении полученных результатов часто нужно выбрать сначала некоторую подходящую априорную модель для $\eta(p)$. Укажем некоторые примеры однородных и изотропных случайных полей, приводя в каждом случае и выражения для спектральных коэффициентов $b_k(r)$.

Пример 1. Пусть

$$B(\varrho) = J_0(c\varrho). \tag{3.11}$$

Тогда

$$b_k(r) = J_k^2(cr). \tag{3.12}$$

Если λ_i — последовательность корней уравнения $J_0(x) = 0$, то из (3.10) и (3.12) следует, что по наблюдениям $\xi(p)$ с корреляционной функцией (3.11) на любой из окружностей с радиусами $r_i = \frac{\lambda_i}{c} m$ восстанавливается безошибочно. Заметим, что гауссовское случайное поле с корреляционной функцией (3.11) обладает следующимс войством марковости: для достаточно большого класса $\{K\}$ замкнутых жордановых контуров значения $\eta(p)$ в точках p_1 и p_2, разделенных контуром K, становятся независимыми, если известны значения $\eta(p)$ на контуре K [9].

Пример 2. Пусть

$$B(\varrho) = c\varrho\, K_1(c\varrho), \tag{3.13}$$

где $K_{\nu}(x)$ — модифицированная функция Бесселя третьего рода [3].

Тогда

$$b_0(r) = 2c^2 r[J_1(cr)\, K_0(cr) - K_1(cr)\, J_0(cr)],$$

где $I_{\nu}(x)$ — модифицированная функция Бесселя первого рода [3]. Интересно заметить, что гауссовское случайное поле с корреляционной функцией (3.3) обладает следующим марковским свойством: значения $\eta(p_1)$ и $\eta(p_2)$ становятся независимыми, если на контуре, разделяющем p_1 и p_2, известны значения поля и его нормальная производная.

Пример 3. Пусть

$$B(\varrho) = \exp\{-c\,\varrho^2\}\,. \tag{3.14}$$

Тогда

$$b_k(r) = \exp\{-2cr^2\}\,I_k(2cr^2)\,.$$

Рассмотрим теперь случай, когда множество D есть множество вершин правильного N-угольника.

Теорема 2. Предположим, что случайное поле $\xi(r, \varphi) = m + \eta(r, \varphi)$ наблюдается в точках $\left(r, k\,\frac{2\pi}{N}\right)$ $k = 0, \ldots, N-1$, являющихся вершинами правильного N-угольника, вписанного в окружность радиуса r. Тогда линейная оценка m с минимальной дисперсией имеет вид:

$$\hat{m} = \frac{1}{N}\sum_{k=0}^{N-1} \xi\left(r, k\,\frac{2\pi}{N}\right), \tag{3.15}$$

причем дисперсия этой оценки равна

$$\mathcal{D}\hat{m} = \sum_{k=-\infty}^{\infty} b_{kN}(r)\,. \tag{3.16}$$

Будем искать оценку $\hat{m}$ в виде $\hat{m} = \sum_{k=0}^{N=1} c_k \xi(r, \varphi_k)$, где $\varphi_k = \frac{2\pi}{N}k$, $\Sigma c_k = 1$. Из основного уравнения (3.1) получаем систему уравнений

$$\sum_{n=-\infty}^{\infty} e^{-in\varphi_s}\left(\sum_{k=0}^{N-1} c_k\, e^{in\varphi_k}\right) b_n(r) = c\,. \tag{3.17}$$

$$(s = 0, 1, \ldots, N-1)$$

Сложим все равенства (3.17), разделим на N и воспользуемся соотношением

$$\frac{1}{N}\sum_{s=0}^{N-1} e^{-in\varphi_s} = \begin{cases} 0 & n \neq Nl \\ 1 & n = Nl\,. \end{cases} \tag{3.18}$$

Получим

$$c = \sum_{l=-\infty}^{\infty} b_{Nl}(r)\,.$$

Теперь легко проверить, что $c_k = \frac{1}{N}$ $(k = 0, \ldots, N-1)$ есть решение системы (3.18), а это и доказывает, в силу единственности решения уравнения (3.1), теорему 2.

4. Об оценке коэффициентов регрессии

Пусть на окружности радиуса r наблюдается случайное поле $\xi(r, \varphi) = a g(r, \varphi) + \eta(r, \varphi)$, где a — неизвестный параметр, $g(r, \varphi)$ — известная функция. Поставим задачу об отыскании среди всех несмещенных оценок (т. е. оценок, удовлетворяющих условию $M\hat{a} = a$) оценки с наименьшей дисперсией. В частности, если $g(r, \varphi) = 1$ получаем задачу об оценке математического ожидания.

Пусть

$$g_n(r) = \frac{1}{2\pi} \int_0^{2\pi} e^{-in\varphi} g(r, \varphi)\, d\varphi\,. \tag{4.1}$$

Предположим, что выполнено условие

$$\sum_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)} < +\infty\,. \tag{4.2}$$

Теорема 3. Среди всех несмещенных оценок параметра a наименьшую дисперсию имеет оценка

$$\hat{a} = \frac{1}{\sum\limits_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)}} \int_0^{2\pi} l(r, \varphi)\, \xi(r, \varphi)\, d\varphi\,, \tag{4.3}$$

где

$$l(r, \varphi) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\varphi} \frac{g_n(r)}{b_n(r)}\,. \tag{4.4}$$

При этом

$$\mathcal{D}\hat{a} = M|\hat{a} - a|^2 = \frac{1}{\sum\limits_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)}}\,. \tag{4.5}$$

В частности, если $g(r, \varphi) = g(r)$ не зависит от φ, то

$$\hat{a} = \frac{1}{g(r)} \frac{1}{2\pi} \int_0^{2\pi} \xi(r, \varphi)\, d\varphi\,. \tag{4.6}$$

Будем искать оценку $\hat{a}$ в виде

$$\hat{a} = \sum_{n=-\infty}^{\infty} c_n\, \xi_n(r)\,,$$

где $\xi_n(r) = ag_n(r) + \eta_n(r)$ и $\Sigma c_n g_n(r) = 1$ (условие несмещенности). Заметим, что

$$\mathcal{D}\hat{a} = \sum_{n=-\infty}^{\infty} c_c^2 b_n(r). \tag{4.7}$$

Из неравенства

$$1 = \left[\sum_{n=-\infty}^{\infty} c_n g_n(r)\right]^2 \leq \sum_{n=-\infty}^{\infty} c_n^2 b_n(r) \sum_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)} \tag{4.8}$$

получаем

$$\mathcal{D}\hat{a} \geq \frac{1}{\sum\limits_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)}}, \tag{4.9}$$

причем минимум в (4.9) достигается, если

$$c_n = \frac{g_n(r)}{b_n(r)} \frac{1}{\sum\limits_{k=-\infty}^{\infty} \frac{g_k^2(r)}{b_k(r)}},$$

а это и доказывает теорему 3.

Заметим, что при нарушении условия (4.2) параметр a по наблюдениям на окружности восстанавливается безошибочно.

Из результатов [10] следует, что в гауссовском случае оценки (3.9) и (4.6) будут оценками максимального правдоподобия.

Предположим, что на окружности радиуса r наблюдается случайное поле $\xi(p) = \sum\limits_{s=1}^{N} a_s g_s(p) + \eta(p)$, где $g_s(p)$ — линейно независимые функции такие, что

$$\sum_{k=-\infty}^{\infty} \frac{g_{ks}^2(r)}{b_k(r)} < +\infty \qquad (s = 1, \ldots, N).$$

Обозначим

$$\gamma_{sj} = \sum_{k=-\infty}^{\infty} \frac{g_{ks}(r)\, g_{kj}(r)}{b_k(r)}.$$

Тогда справедлива следующая теорема.

Теорема 4. Линейная, несмещенная оценка с минимальной дисперсией параметра a_i имеет вид

$$\hat{a}_i = \sum_{s=1}^{N} \alpha_{si}\, l_s(r, \varphi),$$

где α_{si} решение системы линейных уравнений

$$\sum_{s=1}^{N} \alpha_{st} \gamma_{sj} = \delta_i^j \qquad (j = 1, \ldots, N),$$

а $l_s(r, \varphi)$ определяется равенством (4.4) с заменой $g_n(r)$ на $g_{sn}(r)$.

Доказательство может быть получено рассуждениями, близкими к тем, которые использованы при доказательстве теоремы 3.

Литература

1. *Виленкин С. Я., Дубенко Т. И.:* Об оптимальных линейных оценках математического ожидания однородного случайного поля. Известия Академии наук СССР, Техническая кибернетика, 1971, № 1, 134—141.
2. *Гихман И. И., Скороход А. В.:* Введение в теорию случайных процессов. «Наука», 1965.
3. *Бейтмен Г., Эрдейи А.:* Высшие трансцендентные функции. Функции Бесселя, функции параболического цилиндра, ортогональные многочлены, «Наука», 1966.
4. *Яглом А. М.:* Спектральные представления для различных классов случайных функций. Труды IV Всесоюзного математического съезда, 1963, т. I, 250—273.
5. *Ядренко М. И.:* Спектральні розклади ізотропних випадкових полів. Доповіді Академії наук УРСР, 1968, № 10, 835—837.
6. *Grenander, I.:* Stochastic processes and statistical inference. Arkiv für Math., 1951, B. 1, H. 3, 195—277. (русский перевод У. Гренандер: Случайные процессы и статистические выводы. ИЛ, 1961).
7. *Grenander, I.:* On empirical spectral analysis of stochastic processes. Arkiv für Math., B. 1, H. 6, 1952, 503—531.
8. *Ядренко М. Й.:* Сферичні середні ізотропних випадкових полів. Доповіді Академії наук УРСР, 1968, № 12.
9. *Ядренко М. И.:* Изотропные случайные поля марковского типа в эвклидовом и гильбертовом пространстве. Труды Всесоюзного совещания по теории вероятностей и математической статистике. Ереван, 19—25 сентября 1958, 263—279.
10. *Ядренко М. И.:* Об абсолютной непрерывности мер соответствующих гауссовским однородным случайным полям. Теория вероятностей и математическая статистика, 1972, вып. 7, 212—225.

On optimal linear estimates of mean value and regression coefficients of homogeneous and isotropic random field

M. I. YADRENKO

(Kiev)

Suppose that on the circle C_r with the radius r a homogeneous and isotropic random field $\xi(p)$ with unknown mean value m and known correlation function is observed. It has been proved that the average $\xi(p)$ by C_r has least dispersion among all the unbiased estimates of m. Estimates are considered if $\xi(p)$ is observed on the system of concentric circles or on the set apeaxes of rectlinear polygon.

Let a random field $\xi(p) = \sum_{s=1}^{N} a_s g_s(p) + \eta(p)$ is observed on C_r, a_s are unknown constants, $g_s(p)$ are known functions, $\eta(p)$ is homogeneous and isotropic random field with zero mean and known correlation function. Linear unbiased estimates of a_s are considered with minimum dispersion.

Ядренко М. И.
Киевский Государственный
Университет
СССР, Киев, Владимирская 60

Problems of Control and Information Theory, Vol. 2 (1), pp. 27—40 (1973)

МЕТОД АНАЛИЗА АСИНХРОННЫХ АВТОМАТОВ

Б. А. ЛАГОВИЕР

(Москва)

(Поступила в редакцию 30 декабря 1971 г.)

Рассматривается система функций возбуждения, описывающая схему управляющего устройства. Для этой системы предложен метод получения функций переходов в виде таблицы переходов с обобщенными состояниями входов. При этом функционирование каждого внутреннего элемента описывается таблицей переходов с ортогональной системой входных функций, а общая таблица получается перемножением таблиц внутренних элементов.

1. Введение. При разработке дискретных управляющих устройств возникает задача анализа автоматов, заданных, как правило, в виде схемы или списка функций возбуждения внутренних элементов. Для решения этой задачи необходимо получить алгоритм функционирования заданного устройства, т. е. построить его функции переходов и выходов. Распространенным способом описания алгоритмов функционирования являются автоматные таблицы переходов и выходов, задающие функцию переходов δ и выходов λ. Канонические автоматные таблицы для автомата, имеющего m входов и k внутренних состояний, содержат $2^m \times k$ клеток, в которых для каждой пары: входное состояние — внутреннее состояние, указывают следующее внутреннее состояние (переход) и состояние выхода. В [1] введено понятие обобщенного состояния входов и разработан метод построения таблиц переходов, в которых каждая клетка определяется парой: входная функция — внутреннее состояние. В [2, 3] определены необходимые и достаточные условия получения такой системы входных функций, доказано, что такая система должна быть ортогональной и разработаны методы ортогонализации. Там же описан метод получения таблицы переходов с обобщенными входами по заданным функциям возбуждения внутренних элементов. Однако этот метод требует довольно громоздких вычислений и пригоден для анализа схем с небольшим числом внутренних элементов. В настоящей работе описывается другой метод получения по заданной схеме таблицы переходов с обобщенными состояниями входов, более наглядный и, в некоторых случаях, более удобный для анализа схем с большим числом внутренних элементов.

2. Описание метода. Пусть задана система функций возбуждения внутренних элементов устройства

$$\begin{aligned} Y_1 &= \varphi_1(x_1, \ldots, x_m, y_1, \ldots, y_n), \\ Y_2 &= \varphi_2(x_1, \ldots, x_m, y_1, \ldots, y_n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Y_n &= \varphi_n(x_1, \ldots, x_m, y_1, \ldots, y_n), \end{aligned} \tag{2.1}$$

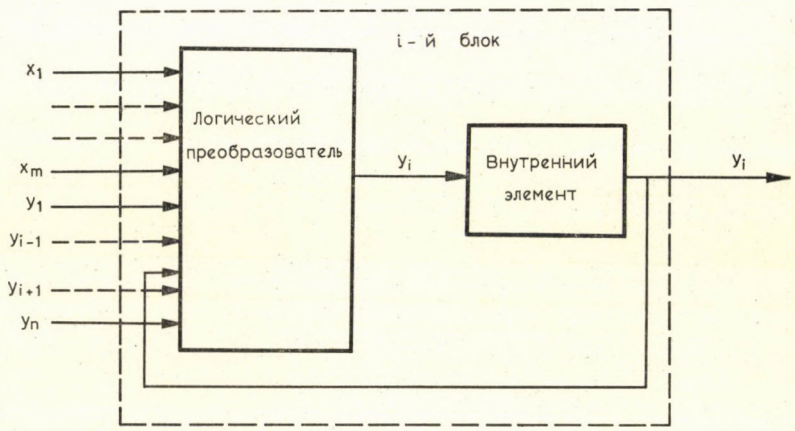


Рис. 1

где $x_1, x_2, \ldots, x_m$ — входные переменные, Y_i — вход внутреннего элемента, y_i — выход внутреннего элемента. Заданная система функций (2.1) описывает логическую сеть, каждый блок которой имеет два внутренних состояния и реализуется одним внутренним элементом. Рассмотрим i-й блок (рис. 1). Функция возбуждения Y_i полностью определена и будем считать, что она равна 1 на множестве M_{1i} рабочих состояний и равна 0 на множестве M_{0i} остальных состояний входных и внутренних переменных. При этом каждая функция возбуждения разбивает все множество 2^{m+n} состояний входных и внутренних переменных на две непересекающиеся части M_{1i} и M_{0i}, причем $M_{1i} \cap M_{0i} = \varnothing$. Построив функции переходов всех блоков в отдельности и перемножив их по описываемому ниже правилу, можно найти функцию переходов, реализуемую заданной логической сетью.

Разложим каждую функцию возбуждения из системы (2.1) по собственной переменной, так чтобы

$$Y_i = g_{0i}\bar{y}_i + g_{1i}y_i + g_{2i}, \qquad i = 1, 2, \ldots, n, \tag{2.2}$$

где g_{0i} — коэффициент при y_i; g_{1i} — коэффициент при y_i; g_{2i} — свободный член. Используя коэффициенты разложения Y_i, построим функцию переходов i-ого блока с единственным внутренним элементом. Представим функцию переходов этого блока таблицей с обобщенными состояниями входов.

Таблица 1

y_i	f_{0i}	f_{1i}	f_{2i}	f_{3i}
0	⟨0⟩	1	⟨0⟩	1
1	⟨1⟩	⟨1⟩	0	0

Здесь f_{0i} — функция удержания y_i, f_{1i} — функция включения, f_{2i} — функция сброса, f_{3i} — функция генерации.

Обобщенные входные функции (f_{0i}, f_{1i}, f_{2i}, f_{3i}) получим следующим образом. Из (2.2) ясно, что при $y_i = 0$ $Y_i^{(0)} = g_{0i} + g_{2i}$, а при $Y_i = 1$ $y_i^{(1)} = g_{1i} + g_{2i}$. Из первого столбца (табл. 1) видно, что при $y_i = 0$ также должно быть $Y_i = 0$, т. е. $f_{0i}^{(0)} = Y_i^{(0)} = \overline{g_{0i} + g_{2i}}$, и при $y_i = 1$ должно быть $Y_i = 1$, т. е. $f_{0i}^{(1)} = Y_i^{(1)} = g_{1i} + g_{2i}$. Аналогично из остальных столбцов табл. 1 имеем:

$$f_{1i}^{(0)} = Y_1^{(0)} = g_{0i} + g_{2i}\,;\quad f_{1i}^{(1)} = Y_i^{(1)} = g_{1i} + g_{2i}\,;$$

$$f_{2i}^{(0)} = \overline{Y_i^{(0)}} = \overline{g_{0i} + g_{2i}}\,;\quad f_{2i}^{(1)} = \overline{Y_i^{(1)}} = \overline{g_{1i} + g_{2i}}\,;$$

$$f_{3i}^{(0)} = Y_i^{(0)} = g_{0i} + g_{2i}\,;\quad f_{3i}^{(1)} = \overline{Y_i^{(1)}} = \overline{g_{1i} + g_{2i}}\,.$$

Используя эти выражения, получаем

$$\begin{aligned}
f_{0i} &= f_{0i}^{(0)} f_{0i}^{(1)} = (\overline{g_{0i} + g_{2i}})\,(g_{1i} + g_{2i}) = \bar{g}_{0i}\, g_{1i}\, \bar{g}_{2i}\,;\\
f_{1i} &= f_{1i}^{(0)} f_{1i}^{(1)} = (g_{0i} + g_{2i})\,(g_{1i} + g_{2i}) = g_{0i}\, g_{1i} + g_{2i}\,;\\
f_{2i} &= f_{2i}^{(0)} f_{2i}^{(1)} = (\overline{g_{0i} + g_{2i}})\,(\overline{g_{1i} + g_{2i}}) = \bar{g}_{0i}\, \bar{g}_{1i}\, \bar{g}_{2i}\,;\\
f_{3i} &= f_{3i}^{(0)} f_{3i}^{(1)} = (g_{0i} + g_{2i})\,(\overline{g_{2i} + g_{2i}}) = g_{0i}\, \bar{g}_{1i}\, \bar{g}_{2i}\,.
\end{aligned} \tag{2.3}$$

В непротиворечиво заданном асинхронном автомате функция генерации должна быть тождественно равна нулю $f_{3i} = 0$, а система входных функций f_{0i}, f_{1i}, f_{2i} обладает свойством полноты и ортогональности

Таблица 2

g_{0i}	g_{1i}	g_{2i}	f_{0i}	f_{1i}	f_{2i}	f_{3i}
0	0	0	0	0	1	0
0	0	1	0	1	0	0
0	1	0	1	0	0	0
0	1	1	0	1	0	0
1	0	0	0	0	0	1
1	0	1	0	1	0	0
1	1	0	0	1	0	0
1	1	1	0	1	0	0

Выполнение условий полноты и ортогональности позволяет проверить на самом начальном этапе анализа непротиворечивость задания каждой функции в системе (2.1). После такой проверки каждый блок может быть представлен функцией переходов (табл. 3), где входные функции f_{0i}, f_{1i}, f_{2i} получены из системы (2.3).

Таблица 3

y_i	f_{0i}	f_{1i}	f_{2i}
0	⟨0⟩	1	⟨0⟩
1	⟨1⟩	⟨1⟩	0

Пользуясь операциями объединения и умножения блоков, имеющих различные внешние входы, предложенными в [4] для канонических таблиц переходов, необходимо предварительно определить множество состояний входов. Для канонических таблиц переходов мощность этого множества равна произведению чисел состояний входов отдельных блоков. Входные наборы образуются конъюнкциями состояний входов объединяемых блоков.

Определим возможность применения тех же операций к таблицам переходов с обобщенными состояниями входов. Пусть система входных функций каждого отдельного блока полна и ортогональна. Для простоты рассмотрим объединение двух блоков. Все полученные доказательства легко распространяются на n блоков.

Пусть система входных функций первого блока состоит из функций $f_0, f_1, \ldots, f_k$, а второго — из $h_0, h_1, \ldots, h_l$. Система входных функций объединенного блока, согласно [4], будет содержать произведения функций

$$
\begin{aligned}
&f_0 h_0,\ f_0 h_1,\ \ldots,\ f_0 h_l,\\
&f_1 h_0,\ f_1 h_1,\ \ldots,\ f_1 h_l,\\
&\ldots\ldots\ldots\ldots\ldots\\
&\ldots\ldots\ldots\ldots\ldots\\
&f_k h_0,\ f_k h_1,\ \ldots,\ f_k h_l.
\end{aligned}
$$

Полученная система функций также полна и ортогональна. Действительно объединение всех функций можно представить в следующем виде: $f_0h_0 + f_0 h_1 + \ldots + f_k h_l = (f_0 + f_1 + \ldots + f_k)\ (h_0 + h_1 + \ldots + h_l)$, а каждое выражение в скобках в правой части этого равенства тождественно равно единице по условию полноты заданных систем функций. Отсюда и всё выражение равно единице, а значит система функций полна.

Для доказательства ортогональности возьмём конъюнкцию двух произвольных произведений $f_i h_j \wedge f_p h_q$ при $i \neq p$ или $j \neq q$. Известно, что две функции ортогональны, если их конъюнкция тождественно равна нулю. А $f_i h_j f_p h_q = f_i f_p h_j h_q \equiv 0$, так как: 1) при $i = p$ и $j \neq q$ $h_j h_q \equiv 0$; 2) при $i \neq p$ и $j = q$ $f_i f_p \equiv 0$; 3) при $i \neq p$, $j \neq q$, $f_i = h_j$ и $f_p = h_q$ $f_i f_i f_p f_p \equiv 0$.

Операция умножения блоков, описанных таблицами переходов с обобщенными входами, при отсутствии дополнительных внешних входов ничем не отличается от умножения канонических таблиц переходов. Таким образом, имея таблицы переходов отдельных внутренних элементов схемы и получив для каждого элемента ортогональную систему входных функций, можно их перемножением получить ортогональную систему входных функций всей сети.

Перейдем теперь к построению таблицы переходов. Пусть задана система (2.1) и для каждой функции Y_i получены обобщенные входные функции f_{0i}, f_{1i}, f_{2i}. Далее получаем обобщенную систему ортогональных функций:

$$
\begin{aligned}
F_1 &= f_{01} f_{02} \ldots f_{0n}\\
F_2 &= f_{01} f_{02} \ldots f_{0(n-1)} f_{1n}\\
&\ldots\ldots\ldots\ldots\ldots\\
&\ldots\ldots\ldots\ldots\ldots\\
F_{3^n} &= f_{21} f_{22} \ldots f_{2n}
\end{aligned}
\qquad (2.4)
$$

Эта система будет содержать не более 3^n функций. Фактически в системе (2.4) некоторые функции могут быть тождественно равны 0. В этом случае будем рассматривать систему из меньшего числа функций.

Опишем теперь, как по системе (2.4) построить таблицу переходов. Рассмотрим два случая. Первый, — когда в системе (2.1) все Y_i зависят лишь от входных и собственных внутренних переменных и второй, — когда в

системе (2.1) найдется Y_i, зависящий от несобственных внутренних переменных.

Первый случай соответствует параллельному соединению блоков. При этом в полученной системе (2.4) никакая из F_i не зависит от внутренних переменных по построению. Таблица переходов имеет 2^n строк и N столбцов, где N — число функций $F_i \not\equiv 0$ в системе (2.4). Строки закодированы конъюнкциями внутренних переменных $y_1^{\sigma_1} y_2^{\sigma_1} \ldots y_n^{\sigma_n}$, где $\sigma_i \in \{0, 1\}$. Пусть $F_i = f_{\alpha_1 1} f_{\alpha_2 2} \ldots f_{\alpha_n n}$, где $\alpha_i \in \{0, 1, 2\}$. На пересечении строки $y_1^{\sigma_1} y_2^{\sigma_2} \ldots y_n^{\sigma_n}$ и столбца F_i стоит код состояния в следущий момент $y_1^{\tau_1} y_2^{\tau_2} \ldots y_n^{\tau_n}$. Значение τ_j находим по следующему правилу. Если $\sigma_j = 0$ и $\alpha_j = 0$ или $\alpha_j = 2$, то $\tau_j = 0$. Если $\sigma_j = 0$ и $\alpha_j = 1$, то $\tau_j = 1$. Далее, если $\sigma_j = 1$ и $\alpha_j = 0$ или $\alpha_j = 1$, то $\tau_j = 1$. Если же $\sigma_j = 1$ и $\alpha_j = 2$, $\tau_j = 0$. Нетрудно видеть, что это правило действительно позволяет определить следующее состояние. Если для всех j, $1 \leq j \leq n$, $\sigma_j = \tau_j$, то состояние является устойчивым. Заметим, что если в функции $F_i = f_{\alpha_1 1} f_{\alpha_2 2} \ldots f_{\alpha_n n}$ ни одно α_j-ое не равно 0, то в столбце F_i содержится только одно устойчивое состояние, а именно состояние $y_1^{\tau_1} y_2^{\tau_2} \ldots y_n^{\tau_n}$, где

$$\tau_j = \begin{cases} 0, & \text{если } \alpha_j = 2 \\ 1, & \text{если } \alpha_j = 1 \end{cases}$$

Аналогично, если в функции F_i содержится одно α_j, равное 0, то в столбце F_i содержится два устойчивых состояния (в одном из них $\tau_j = 0$, в другом $\tau_j = 1$. Значения остальных определяются как выше), и т. д.

Для построения таблицы переходов, в которой учитываются все возможные внутренние состояния и переходы между ними, нет необходимости рассматривать каждое из 2^n внутренних состояний и определять является ли оно устойчивым. Рассмотрим систему (2.4) обобщенных входных функций. Распределим все функции в этой системе по группам, каждая из которых содержит $C_n^r 2^{n-r}$ функций, где r — число функций $f_{\alpha_i i}$, у которых $\alpha_i = 0$ в конъюнкции F_i.

Если в отдельно взятой такой группе функций ни одна из $F_i \not\equiv 0$, то эта группа определяет все 2^n внутренних состояний. Например, $r = 0$, т. е. возьмем группу F_i, в которой все $\alpha \in \{1, 2\}$. Пусть в этой группе нет ни одного $F_i \equiv 0$. Число функций в группе равно $C_n^0 2^{n-0} = 2^n$. Каждая функция F_i в этой группе, согласно сказанному выше, определяет по одному устойчивому состоянию внутренних элементов, и в каждом столбце таблицы переходов допустимы все переходы к одному (собственному) устойчивому состоянию. Так как $F_i F_j \equiv 0$ $(i \neq j)$, то устойчивые состояния в различных столбцах различны и их общее число равно 2^n. Если в каждой группе найдется $F_i \equiv 0$,

то каждая такая группа рассматривается отдельно и множество внутренних состояний автомата находится как сумма различных внутренних состояний, определяемых каждой группой.

Пример 1. Пусть $F_1 = f_{01}f_{02}$, $F_2 = f_{01}f_{12}$, $F_3 = f_{01}f_{22}$, $F_4 = f_{11}f_{02}$, $F_5 = f_{11}f_{12}$, $F_6 = f_{11}f_{22}$, $F_7 = f_{21}f_{02}$, $F_8 = f_{21}f_{12}$, $F_9 = f_{21}f_{22}$.

Тогда таблицей переходов объединенных блоков будет таблица 4.

Таблица 4

y_1y_2	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
00	⟨00⟩	01	⟨00⟩	10	11	10	⟨00⟩	01	⟨00⟩
01	⟨01⟩	⟨01⟩	00	11	11	10	⟨01⟩	⟨01⟩	00
10	⟨10⟩	11	⟨10⟩	⟨10⟩	11	⟨10⟩	00	01	00
11	⟨11⟩	⟨11⟩	10	⟨11⟩	⟨11⟩	10	01	01	00

Рассмотрим теперь второй случай системы (2.1), в которой некоторые или все Y_i зависят от несобственных внутренних переменных. При этом в полученной системе (2.3) некоторые функции F_i зависят от y_i ($i \neq j$). Представим $F_i = \Psi_i(x_1, \ldots, x_m, y_1, \ldots, y_n)$ в виде дизъюнкции $F_i = F_i' + F_i''$, где $F_i' = \Psi_i(x_1, \ldots, x_m)$, а $F_i = \Psi_i(x_1, \ldots, x_m, y_1, \ldots, y_n)$. Из $F_i F_j \equiv 0$ ($i \neq j$) следует, что $F_i' F_j = 0$ и $F_i'' F_j = 0$. По функциям F_i', как и по функциям F_i в рассмотренном выше случае, можно определить следующие состояния для всех 2^n внутренних состояний таблицы переходов в соответствующих столбцах. Для нахождения состояний, определяемых функциями F_i'', представим последние в ДНФ $F_i'' = k_{1i}^x k_{1i}^y + k_{2i}^x k_{2i}^y + \ldots + k_{s(i)i}^x k_{s(i)i}^y$, где k_{li}^y — l-ая конъюнкция входных переменных; k_{li}^y — l-ая конъюнкция внутренних переменных ($1 \leq l \leq s(i)$). Поместим в заголовке столбца k_{li}^x. Из $F_i F_j \equiv 0$ ($i \neq j$) следует, что $k_{li}^x k_{ti}^y \cdot F_j \equiv 0$. В частности, для функций типа F_i'' имеем $k_{li}^x k_{li}^y \cdot k_{pj}^x k_{pj}^y \equiv 0$. Пара конъюнкций k_{li}^x и k_{li}^y определяет следующие состояния функции переходов только для тех строк, где $k_{li}^y = 1$. Может оказаться, что для некоторых l и p ($1 \leq l \leq s(i)$, $1 \leq p \leq s(j)$, $i, j = 1, 2, \ldots, n$) $k_{li}^x \cdot k_{pj}^x \neq 0$. Очевидно, что при этом $k_{li}^y \ldots k_{pj}^y = 0$. В этом случае для каждой конъюнкции $k_{li}^x \ldots k_{pj}^x \neq 0$ вводится столбец в таблице переходов, в котором каждое следующее состояние в клетке $k_{li}^x k_{pj}^x k_{li}^y$ определяется по функции F_i, а состояние в клетке $k_{li}^x k_{pi}^x \cdot k_{pj}^y$ — по функции F_j описанным выше способом. Таким образом в рассматриваемом случае построение таблицы переходов выполняется по следующим этапам.

На первом этапе строится таблица переходов, состоящая как бы из двух групп столбцов. Первая группа — это столбцы, озаглавленные функциями типа F_i и F'_i, а вторая группа — столбцы, озаглавленные конъюнкциями k^x. На втором этапе рассматриваются только столбцы, озаглавленные конъюнкциями k^x и определяются все возможные пересечения этих конъюнкций. При этом выбираем первую по порядку конъюнкцию и определяем, нет ли у неё попарных пересечений со всеми последующими. Если таких пересечений нет, то заносим столбец, озаглавленный этой конъюнкцией в первую часть таблицы. Если же такие пересечения обнаружены, то все они записываются в виде новых столбцов этой же части таблицы. Столбец, озаглавленный проверенной конъюнкцией, из дальнейшего рассмотрения исключается. Далее берется вторая конъюнкция и проверяется её пересечение со всеми остальными, а также с вновь введенными конъюнкциями, полученными от пересечения первой конъюнкции с другими. При этом, как и раньше, если вторая конъюнкция не имеет пересечений с остальными, то соответствующий ей столбец сразу переносится в первую часть таблицы, если же такие пересечения есть, то они заносятся во вторую (проверяемую) часть таблицы, а проверенная конъюнкция из дальнейших проверок исключается.

Составление таблицы переходов заканчивается с перенесением последнего проверяемого столбца в первую часть таблицы.

Практический интерес представляют автоматы, у которых граф переходов сильно связан. Такие автоматы, как правило, применяют для управления циклическими производственными процессами.

В функциях переходов этих автоматов некоторые внутренние состояния могут оказаться недостижимыми и в работе автомата обычно не используются. Это происходит, когда часть входных функций равна нулю. Например, если в рассмотренном выше примере $F_3 = F_4 = F_6 = 0$, то состояние 10 ($y_1\bar{y}_2$) в табл. 4, оказывается недостижимым и может не указываться в полученной таблице переходов.

Ниже предлагается упрощенный алгоритм заполнения таблицы переходов таких автоматов. При этом из рассмотрения исключаются недостижимые внутренние состояния.

1. Выписываются заголовки столбцов.

2. Проставляются устойчивые состояния в столбцах, содержащих по одному такому состоянию.

3. Проставляются переходы (неустойчивые состояния) и соответствующие им устойчивые состояния в столбцах, содержащих два устойчивых состояния, затем три и так далее до столбцов, содержащих максимальное число устойчивых состояний.

4. От столбца, содержащего максимальное число устойчивых состояний, возвращаются к предыдущему и проставляют в нем недостающие переходы и так далее до столбцов, содержащих по одному устойчивому состоянию.

Описанный здесь метод поясняется ниже на двух примерах. Первый, более простой пример, наглядно иллюстрирует все этапы получения таблицы переходов. Второй пример заимствован из [2] и решен предложенным методом. Это позволяет сравнить оба алгоритма анализа.

Пример 2. Задана система функций

$$Y_1 = x_1\bar{y}_3(x_2 + y_1) = x_1\bar{y}_3y_1 + x_1x_2\bar{y}_3;$$
$$Y_2 = x_1(\bar{x}_2y_1 + y_2) = x_1y_2 + x_1\bar{x}_2y_1;$$
$$Y_3 = x_1(x_2y_2 + y_3) = x_1y_3 + x_1x_2y_2.$$

Вычислим системы функций отдельных блоков и построим блочные таблицы переходов.

I блок

$$f_{11} = x_1\bar{y}_3x_2$$
$$f_{21} = \bar{x}_1 + y_3$$
$$f_{01} = x_1\bar{y}_3\bar{x}_2$$
$$f_{31} = 0$$

Таблица 5

y_1	f_{01}	f_{11}	f_{21}
0	⟨0⟩	1	⟨0⟩
1	⟨1⟩	⟨1⟩	0

II блок

$$f_{12} = x_1\bar{x}_2y_1$$
$$f_{22} = \bar{x}_1$$
$$f_{02} = x_1(x_2 + \bar{y}_1)$$
$$f_{32} = 0$$

Таблица 6

y_2	f_{02}	f_{12}	f_{22}
0	⟨0⟩	1	⟨0⟩
1	⟨1⟩	⟨1⟩	0

III блок

$$f_{13} = x_1x_2y_2$$
$$f_{23} = \bar{x}_1$$
$$f_{03} = x_1(\bar{x}_2 + \bar{y}_2)$$
$$f_{33} = 0$$

Таблица 7

y_3	f_{03}	f_{13}	f_{23}
0	⟨0⟩	1	⟨0⟩
1	⟨1⟩	⟨1⟩	0

Определим общую систему входных функций всей логической сети. Для этого по алгоритму необходимо вычислить максимум 3^3 произведений входных функций блоков. Однако нетрудно проверить, что следующие произведения отличны от нуля.

$$F_1 = f_{01}f_{02}f_{03} = x_1\bar{x}_2\bar{y}_1y_3, \quad F_2 = f_{11}f_{02}f_{03} = x_1x_2\bar{y}_2\bar{y}_3,$$

$$F_3 = f_{21}f_{02}f_{03} = x_1x_2\bar{y}_2y_3 + x_1\bar{x}_2\bar{y}_1y_3, \quad F_4 = f_{01}f_{12}f_{03} = x_1\bar{x}_2y_1\bar{y}_3,$$

$$F_5 = f_{21}f_{12}f_{03} = x_1\bar{x}_2y_1y_3, \quad F_6 = f_{11}f_{02}f_{13} = x_1x_2y_2\bar{y}_3, \quad F_7 = f_{21}f_{02}f_{13} = x_1x_2y_2y_3,$$

$$F_8 = \bar{x}_1.$$

Строим таблицу переходов, в которой F_8 относится к первому типу функций, т. е. не зависит от внутренних переменных, а остальные семь функций — ко второму типу.

Для F_8 сразу же отводим столбец в таблице и озаглавливаем его соответственно $\bar{x}_1$.

Остальные семь функций разложим по конъюнкциям и упорядочим их следующим образом:

$k_{11} = x_1\bar{x}_2$, $k_{21} = x_1x_2$, $k_{31} = x_1x_2$, $k_{32} = x_1\bar{x}_2$, $k_{41} = x_1\bar{x}_2$, $k_{51} = x_1\bar{x}_2$, $k_{61} = x_1x_2$ $k_{71} = x_1x_2$.

Выбираем первую по порядку k_{11} и проверяем ее попарное пересечение со всеми остальными. k_{11} имеет пересечения, которые записываем в эту же таблицу, а k_{11} из дальнейшего рассмотрения исключаем.

$k_{21} = x_1x_2$, $k_{31} = x_1x_2$, $k_{32} = x_2\bar{x}_2$, $k_{41} = x_1\bar{x}_2$, $k_{51} = x_1\bar{x}_2$, $k_{61} = x_1x_2$, $k_{71} = x_1x_2$, $k_{11}k_{32} = x_1\bar{x}_2$, $k_{11}k_{41} = x_1\bar{x}_2$, $k_{11}k_{51} = x_1\bar{x}_2$.

На следующем шаге исключаем k_{21}, затем k_{31} и так далее, и в результате получим

$$k_{11}k_{32}k_{41}k_{51} = x_1\bar{x}_2 \quad \text{и} \quad k_{21}k_{31}k_{61}k_{71} = x_1x_2.$$

Этими конъюнкциями озаглавим столбцы в таблице переходов (табл. 8).

Для проверки ортогональности проверим не пересекаются ли соответствующие k^y.

$$k^y_{11} \cdot k^y_{32} = 0, \quad k^y_{11}k^y_{41} = 0, \quad k^y_{11}k^y_{51} = 0, \quad k^y_{32}k^y_{41} = 0, \quad k^y_{32}k^y_{51} = 0,$$

$$k^y_{41} \cdot k^y_{51} = 0, \quad k^y_{21}k^y_{31} = 0, \quad k^y_{21}k^y_{61} = 0, \quad k^y_{21}k^y_{71} = 0, \quad k^y_{31}k^y_{61} = 0,$$

$$k^y_{31} \cdot k^y_{71} = 0, \quad k^y_{61}k^y_{71} = 0.$$

Таким образом получим всего три обобщенных входных функции $x_1\bar{x}_2$, x_1x_2, $\bar{x}_1$ и следующие внутренние состояния $\bar{y}_1\bar{y}_3$, $\bar{y}_2\bar{y}_3$, $\bar{y}_2y_3$, $\bar{y}_1y_3$, $y_1\bar{y}_3$ y_1y_3, $y_2\bar{y}_3$, y_2y_3.

Построим таблицу переходов, содержащую эти состояния (табл. 8).

Заполнение первой строки таблицы переходов начинаем со столбца $\bar{x}_1$, содержащего одно устойчивое состояние ⟨000⟩. Это состояние можно поставить в одной из двух клеток $\bar{y}_1\bar{y}_3$ или $\bar{y}_2\bar{y}_3$. Выбираем первую и тем самым сразу доопределяем в этом состоянии значение y_2. Учитывая, что в состоянии $\bar{y}_1\bar{y}_2\bar{y}_3$ может быть $F_1 = 1$ (при $x_1\bar{x}_2 = 1$) и $F_2 = 1$ (при $x_1x_2 = 1$), а также, что $F_1 = f_{01}f_{02}f_{03}$ и $F_2 = f_{11}f_{02}f_{03}$, проставляем в столбце $x_1\bar{x}_2$ (в первой строке) устойчивое состояние ⟨000⟩, а в столбце x_1x_2 — неустойчивое состояние 100 и устойчивое состояние ⟨100⟩ в 5-й строке.

Таблица 8

y_1	y_2	y_3	$x_1\bar{x}_2$	x_1x_2	$\bar{x}_1$	
→0		0	⟨000⟩	100	⟨000⟩	1
—	0	0	—	—	—	2
	0	1	—	—	—	3
0		1	⟨011⟩	⟨011⟩	000	4
1		0	110	⟨100⟩	000	5
→1		1	011	011	000	6
	1	0	⟨110⟩	111	000	7
—	1	1	—	—	—	8

Вторую строку можно доопределить так, чтобы она стала эквивалентна первой или пятой. Так как это не влияет на дальнейшее, то выбор доопределения может быть произвольным. Доопределим ее так, чтобы вторая строка стала эквивалентна первой. Рассмотрим 5-ую строку. В столбце $x_1\bar{x}_2$ проставляем следующее состояние 110, так как при $y_1\bar{y}_2\bar{y}_3 = 1$ и $x_1\bar{x}_2 = 1$, $F_4 = f_{01}f_{12}f_{03} = 1$, а это означает, что y_1 и y_3 должны остаться в прежнем состоянии, а y_2 перейти в состояние $y_2 = 1$. Устойчивое состояние ⟨110⟩ проставляем в 7-й строке и тем самым доопределяем значение y_1 в ней. В этой же строке переходим в столбец x_1x_2 и с учетом того, что при $y_1y_2\bar{y}_3 = 1$ и $x_1x_2 = 1$, $F_6 = f_{11}f_{02}f_{13} = 1$, проставляем следующее состояние 111. Этот переход может привести как в 6-ую, так и в 8-ую строки. Объединяем их и проставляем в 6-й строке столбца x_1x_2 следующее состояние 011, учитывая, что при этом $F_7 = f_{21}f_{02}f_{13} = 1$. В соответствии с требованием этого неустойчивого состояния переходим в этом же столбце к 4-й строке, проставляя в ней устойчивое состояние ⟨011⟩. Перейдем в этой же строке к столбцу $x_2\bar{x}_2$ и проставим в нём то же самое устой-

чивое состояние ⟨011⟩. В этом же столбце в 6-й строке проставим неустойчивое состояние 011.

Отметим, что внутреннее состояние $\bar{y}_2 y_3$, указанное в третьей строке, оказалось недостижимым и в дальнейшем может не указываться. Для его достижимости необходимо, чтобы в одну из функций F_i входила хотя бы одна из перечисленных ниже конъюнкций: $f_{22}f_{13}$, $f_{22}f_{03}y_3$ или $f_{02}f_{13}\bar{y}_2$, а этого нет.

Теперь можно перейти к столбцу $\bar{x}_1$ и проставить в нем все недостающие неустойчивые состояния 000, приводящие в первую строку.

Окончательно полученная таблица переходов приведена в табл. 9.

Таблица 9

$y_1\ y_2\ y_3$	$x_1\bar{x}_2$	x_2x_2	$\bar{x}_1$	
000	⟨000⟩	100	⟨000⟩	1
011	⟨011⟩	⟨011⟩	000	2
100	110	⟨100⟩	000	3
111	011	011	000	4
110	⟨110⟩	111	000	5

Пример 3. Используя [2], возьмем две первоначальные функции в виде

$$Y_1 = (x_1 + y_1\bar{x}_5)x_2, \quad Y_2 = [(x_1 + y_1\bar{x}_5)x_2x_6 + x_4 + x_{12} + x_7y_2]x_3\bar{x}_5.$$

Определим функции переходов блоков:

I блок

$$f_{01} = \bar{x}_1x_2\bar{x}_5,$$
$$f_{11} = x_1x_2,$$
$$f_{21} = \bar{x}_2 + \bar{x}_1x_5,$$
$$f_{31} = 0.$$

Таблица 10

f_{01}	f_{11}	f_{21}
⟨0⟩	1	⟨0⟩
⟨1⟩	⟨1⟩	0

II блок

$$f_{02} = x_3\bar{x}_4\bar{x}_5x_7\bar{x}_{12}(\bar{x}_2 + \bar{x}_6 + \bar{x}_1\bar{y}_1),$$
$$f_{12} = x_3\bar{x}_5(x_4 + x_{12} + x_1x_2x_6 + x_2x_6y_1),$$
$$f_{22} = \bar{x}_3 + x_5 + \bar{x}_4\bar{x}_7\bar{x}_{12}(\bar{x}_2 + \bar{x}_6 + \bar{x}_1\bar{y}_1 + \bar{x}_1x_5)$$
$$f_{32} = 0$$

Таблица 11

f_{02}	f_{12}	f_{22}
⟨0⟩	1	⟨0⟩
⟨1⟩	⟨1⟩	0

Вычислим общую систему входных функций, состоящую из девяти произведений

$$F_1 = f_{01}f_{02} = \bar{x}_1x_2x_3\bar{x}_4\bar{x}_5x_7\bar{x}_{12}(\bar{x}_6 + \bar{y}_1) = F_1' + F_1'', \quad F_1' = f_{01}f_{02} = \bar{x}_1x_2x_3\bar{x}_4\bar{x}_5\bar{x}_6x_7\bar{x}_{12},$$
$$k_{21}^x = \bar{x}_1x_2x_3\bar{x}_4\bar{x}_5x_7\bar{x}_{12};$$

$$F_2 = f_{01}f_{12} = \bar{x}_1x_2x_3\bar{x}_5(x_4 + x_{12} + x_6y_1) = F_2' + F_2'', \quad F_2' = f_{01}f_{12} = \bar{x}_1x_2x_3\bar{x}_5(x_4 + x_{12}),$$
$$k_{32}^x = \bar{x}_1x_2x_3\bar{x}_5x_6;$$

$$F_3 = f_{01}f_{22} = \bar{x}_1x_2\bar{x}_5(\bar{x}_3 + \bar{x}_4\bar{x}_6\bar{x}_7\bar{x}_{12} + \bar{x}_4\bar{x}_7\bar{x}_{12}y_1) = F_3' + F_3'', \quad F_3' = f_{01}f_{22} =$$
$$= \bar{x}_1x_2\bar{x}_5(\bar{x}_3 + \bar{x}_4\bar{x}_6\bar{x}_7\bar{x}_{12}), \quad k_{33}^x = \bar{x}_1x_2\bar{x}_5\bar{x}_4\bar{x}_7\bar{x}_{12};$$

$$F_4 = f_{11}f_{02} = x_1x_2x_3\bar{x}_4\bar{x}_5\bar{x}_6x_7\bar{x}_{12};$$

$$F_5 = f_{11}f_{12} = x_1x_2x_3\bar{x}_5(x_4 + x_6 + x_{12});$$

$$F_6 = f_{11}f_{22} = x_1x_2(\bar{x}_3 + x_5 + \bar{x}_4\bar{x}_6\bar{x}_7x_{12});$$

$$F_7 = f_{21}f_{02} = x_2x_3\bar{x}_5\bar{x}_4x_7\bar{x}_{12};$$

$$F_8 = f_{21}f_{12} = \bar{x}_2x_3\bar{x}_5(x_4 + x_{12});$$

$$F_9 = f_{21}f_{22} = \bar{x}_2(\bar{x}_3 + x_5 + \bar{x}_4\bar{x}_7\bar{x}_{12}) + \bar{x}_1x_5;$$

$$k_{21}^x\,k_{32}^x \neq 0, \quad k_{21}^x\,k_{33}^x = 0, \quad k_{32}^x\,k_{33}^x \neq 0, \quad k_{21}^y\,k_{32}^y = 0, \quad k_{21}^y\,k_{33}^y \neq 0, \quad k_{32}^y\,k_{33}^y = 0,$$

С учетом этого строим следующую таблицу переходов.

Таблица 12

k_{21}^x	k_{32}^x	k_{33}^x	F_1'	F_2'	F_3'	F_4	F_5	F_6	F_7	F_8	F_9
⟨00⟩	–	⟨00⟩	⟨00⟩	01	⟨00⟩	10	11	10	⟨00⟩	01	⟨00⟩
⟨01⟩	–	00	⟨01⟩	⟨01⟩	00	11	11	10	⟨01⟩	⟨01⟩	00
–	11	–	⟨10⟩	11	⟨10⟩	⟨10⟩	11	⟨10⟩	00	01	00
–	⟨11⟩	–	⟨11⟩	⟨11⟩	10	⟨11⟩	⟨11⟩	10	01	01	00

Определим конъюнкции $k_{21}^x k_{32}^x$ и $k_{32}^x k_{33}^x$

$$k_{21}^x\,k_{32}^x = \bar{x}_1\,x_2\,x_3\,\bar{x}_4\,\bar{x}_5\,x_6\,x_7\,\bar{x}_{12}, \quad k_{32}^x\,k_{33}^x = \bar{x}_1\,x_2\,x_3\,\bar{x}_4\,\bar{x}_5\,x_6\,\bar{x}_7\,\bar{x}_{12}$$

и получим таблицу 13.

Таблица 13

k^x_{21} k^x_{32}	k^x_{32} k^x_{33}	F'_1	F_2'	F'_3	F_4	F_5	F_6	F_7	F_8	F_9
⟨00⟩	⟨00⟩	⟨00⟩	01	⟨00⟩	10	11	10	⟨00⟩	01	⟨00⟩
⟨01⟩	00	⟨01⟩	⟨01⟩	00	11	11	10	⟨01⟩	⟨01⟩	00
11	11	⟨10⟩	11	⟨10⟩	⟨10⟩	11	⟨10⟩	00	01	00
⟨11⟩	⟨11⟩	⟨11⟩	⟨11⟩	10	⟨11⟩	⟨11⟩	10	01	01	00

Нетрудно проверить, что найденное здесь множество функций F'_1, F'_2, F'_3, F_4, F_5, F_6, F_7, F_8, F_9, $k^x_{21}k^x_{32}$ и $k'_{32}k'_{33}$ совпадает с множеством функций, полученых в [2].

Литература

1. *Гаврилов М. А.:* Синтез таблиц переходов методом обобщенных состояний входов. Автоматика и телемеханика № 1, 1967.
2. *Гаврилов М. А., Кузнецов О. П., Хазацкий В. Е.:* Описание и анализ релейных устройств с большим числом входных переменных. Автоматика и телемеханика, № 10, 1969.
3. *Кузнецов О. П.:* Ортогональные системы булевых функций и их применение к анализу и синтезу логических сетей. Автоматика и телемеханика, № 10, 1970.
4. *Гаврилов М. А.:* Синтез таблиц состояний. Учебное пособие по курсу «Структурная теория релейных устройств», ч. IV, ВЗЭИ, 1964.

An analysis technique for asynchronous automata

B. LAGOVIER

(Moscow)

Design of digital control devices requires an analysis of their functional algorithms given in the schematic form or as a system of excitation functions of internal elements. The paper is concerned with a technique starting with a given system of excitation functions and arriving at transition functions written in the form of a table with generalized input states. The following steps are necessary: (1) each excitation function of a given system is transformed into a canonic transition function for one-memory-element block; (2) the function is written into the flow-table having generalized inputs; (3) multiplication of generalized input functions of individual blocks gives a system of generalized input functions of the automation as a whole. Some of these depend on input variables only, the rest depends both on input and internal ones; (4) a pseudo-flowtable is found having some columns with nonorthogonal input functions; (5) all possible intersections of nonorthogonal input functions are found, and (6) the pseudo-flowtable is transformed into the asynchronous automation flow-table.

As illustrated by examples, the proposed technique is gimpler and clearer than the known ones, and can tackle rather complex automata.

Б. А. Лаговиер
Институт проблем управления
(автоматики и телемеханики)
СССР Москва В—279
Профсоюзная ул., 81

Problems of Control and Information Theory, Vol. 2 (1), pp. 41—65 1973)

ON THE COMPUTER AIDED DESIGN OF CONTROL SYSTEMS

J. SOMLÓ

(Budapest)

(Received February 23, 1971)

A computer method is proposed for linear system design with an interactive aspect. The method combines a new root locus method with direct time response considerations. Computer methods for nonlinear system investigation are described. In particular, a computer implementation of harmonic linearization method when a finite number of higher harmonics is taken into consideration, is discussed.

Intruduction

To develop computer aided methods for control system design is an actual task of modern control theory.

Recently the theory of optimal and adaptive control systems has given an opportunity to construct highly perfect systems but, of course, with rather complicated structure.

In most of the practical tasks there is no need for such a sophisticated solution of the problems. Usually the structure of the system is given in advance and the design procedure is restricted to the proper (good, best) choice of several parameters; or to analyse the performance at several variants of a part of the system, and to choose the best with the proper parameters; or to determine the character of an element (possibly simple) which effects significantly the system performance, etc.

For solving such problems, generally speaking, theoretical methods are elaborated.

Their application, however, is not always very effective, because of high error probability and the need of many numerical computations.

The effectiveness of system investigation can significantly be improved using computer aided methods for system design.

In many cases the analog or digital simulation is the simplest way for the solution of a problem. But in other cases the simulation rather complements than replaces the theoretical investigation.

Simulation can be most effectively used by someone having a general picture of the processes in the system and being able to evaluate the system structure and the optimal values of its parameters.

For this the simple approximate methods, on one hand, and the computer implementation of existing or especially developed theoretical methods, on the other, can be used.

The second method seems very promising to us. The application of computer implementation of theoretical methods complemented by simulation in an interactive manner can be very effective in control system design. Methods developed by such aspect, even in different level of completeness, can find useful application.

In this paper we discuss some aspects of computer aided design of linear systems and in more detail the computer program developed by us for nonlinear system investigation.

On Computer Aided Design of Linear Systems

For the time being there are a number of works made in some or other respect connected with this problem. It is rather difficult to give their review.

One of the most advanced works is described in [1] by Coffey (see also [2]), where a system of programs (Linear Controls Analysis Program) is described for computer implementation of classical methods for continuous and sampled data system design. (Determination of transfer functions, frequency response, root locus, time response is carried out.) It is completed by an automatic frequency domain synthesis program. A similar problem for continuous systems is solved by programs described in [3] with rather sophisticated input forms. The solution of partial problems of the same character is described in [4, 5, 6] and many other papers.

The interactive aspects of the problem are discussed in [7—10].

As regards the optimal linear regulator with quadratic performance index, Letov, Kalman and others gave strong theoretical basis for the solution of this problem. Bass and Gura [11] developed a very effective method and a computer algorithm for optimal filter design.

When the problem of determining the optimal parameter values at a given integral criterion and at a given structure is to be solved the known methods (e.g. Parseval's indentity) can be used to express the performance index by system parameters. The optimal values can be determined by the search of the minimum of the performance integral. Computer methods for solving this task are proposed in [12, 13].

Another approach to this problem which can be used also in nonlinear system design is the simulation of the system and the search with a computer method for optimal parameter values, using the measured performance index.

Such a method can be realized in analog or hybrid simulation. The digital analog simulation languages, like MIMIC, CSMP, PACTOLUS etc. are especially suitable for the solution of problems in this way. This approach, using especially developed computer programs, is discussed in [14].

A New Design Philosophy

We propose a computer aided linear system design method which has a little bit different philosophy.

When someone uses a simulation method he can try a number of variants and choose the best.

For example if he wishes to choose two system parameters he simulates the system using different values from the admissable part of parameter plane (Fig. 1). Comparing the results, it is possible to solve the problem. When using this method the theoretical knowledge of the designer is the background which helps to carry out the design.

But very frequently it is rather difficult to evaluate all the possibilities.

Without the analysis of the problem as a theoretical one the results may be wrong, or anyway, not the best.

The combination of the simulation by the theoretical investigation of the problem (also performed by computer, in an interactive manner) may help to overcome this difficulty.

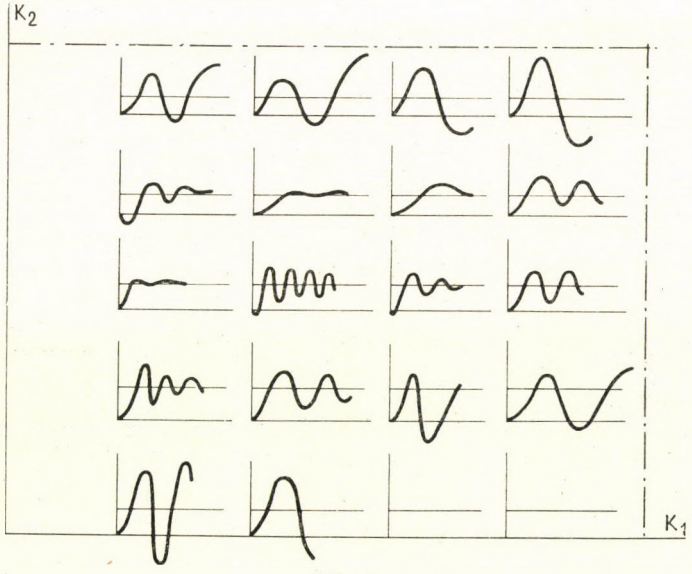


Fig. 1

For linear system design we propose, as a possible way, a method the idea of which will be outlined in the followings.

The root locus method is a powerful means for linear system investigation. The poles and zeros of the transfer function fully characterize the dynamical behaviour of the systems.

As regard the poles, the way of the theoretical investigation of their effect has been elaborated. The situation of the zeros is not the same but there are several rules for their desired location.

The root locus method gives much more compressed information than the system time responses.

So it seems to be a promising way to combine the root locus method and the simulation. The designer working in complex plane of roots (poles and zeros) controls his action by analysing time responses (Fig. 2). After a time his work

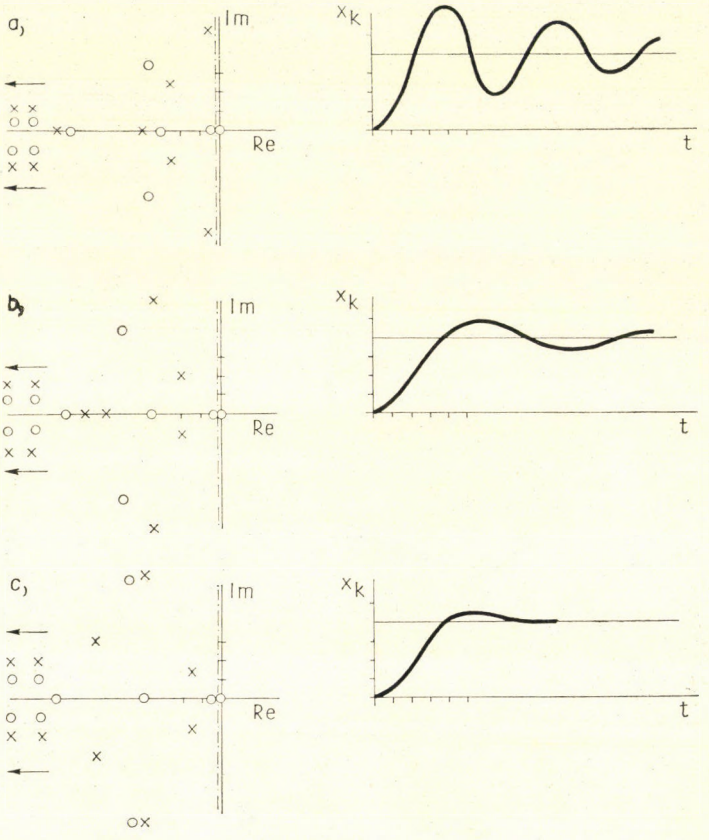


Fig. 2

will be more and more effective as he learns from the problems solved. There is an opportunity to control not only the time responses but the errors (position, velocity) in the system, the frequency characteristics (crossover frequency, maximum of absolute value of closed loop frequency response, etc.).

LSDP (Linear System Design Program)

To apply the outlined idea we have developed a system of programs for the case when the system differential equations is linear in, at least, two parameters.

Let us suppose that a system differential equation is the following

$$[D(p) + K_1 D_1(p) + K_2 D_2(p)]x(t) = [E(p) + K_1 E_1(p) + K_2 E_2(p)]g(t)$$

where

$D(p), D_1(p), D_2(p), E(p), E_1(p), E_2(p)$ — operational polynomials $\left(p = \frac{d}{dt}\right)$

$x(t)$ — error signal

$g(t)$ — input signal

K_1, K_2 — system parameters.

The goal is to design a system the performance of which is very close to that of a second-order system. This is the so-called dominant pole design problem [15, 16].

The K_1 and K_2 parameters can be determined in such a way that a pair of complex roots of the characteristic equation must be on the straight lines l_+, l_- of the complex plane (Fig. 3).

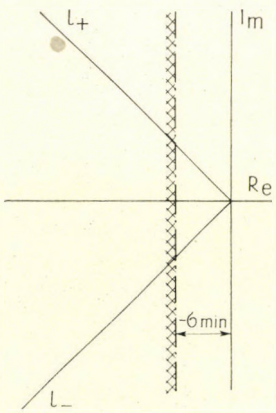


Fig. 3

Substituting

$$p = -\sigma + j\omega = -\eta\omega + j\omega;$$

into the characteristic equation

$$D(p) + K_1 D_1(p) + K_2 D_2(p) = 0$$

and separating the real and imaginary parts, two equations are obtained.

Solving these equations for $K_1 = K_1(\omega)$ and $K_2 = K_2(\omega)$ the lines l_+, l_- of the complex plane can be mapped into the K_1K_2 parameter plane, to get $K_2 = f(K_1)$.

Using the D-decomposition method for vertical line at value, $[-\sigma_{\min}]$ a part of the curve $K_2 = f(K_1)$ can be determined (if one exists) where, besides the pair of roots on l_+ and l_-, the others are left of this vertical.

In this way the number of parameter values to be considered are significantly reduced.

Several points of the given part of the $K_2 = f(K_1)$ curve can be investigated by determining the poles and zeros of the transfer function.

For the points seeming suitable the time response can be determined and compared. If there are K_1K_2 values where the error and frequency requirements etc. are also satisfied the design is finished.

It is characteristic for this root locus method that the poles and zeros are not determined for a given set of parameter values but for a set of values which already met a requirement, as it was emphasized.

Figure 4 demonstrates the parameter mapping. In this figure only that domain is considered where all the roots are left of $-\sigma_{\min}$ vertical. (In a concrete eight-order example analysing the poles in several points it was concluded that only the AB part of the curve $K_2 = f(K_1)$ resulted a dominant pole design. Several points of this part of the curve were investigated and the results were similar to that of given in Fig. 2. (The real scale could not be used.) The step response was very good at the point "c" of Fig. 4 where the velocity error constant was also suitable.

If there are other free parameters the result of the design may also be improved. Using two parameters a pair of complex poles can be fixed.

Let at this point $p_{1,2} = \eta\omega_0 \pm j\omega_0$.

Dividing the characteristic equation by $(p - p_1)(p - p_2) = p^2 + 2\,\eta\omega_0 p + \omega_0(\eta^2 + 1)$ an equation of power $n - 2$ is obtained.

In this way two parameters of the system can be expressed by the other parameters [16].

If the other roots of the characteristic equation are not far enough from the pair fixed, using the D-decomposition method for the reduced equation, a parameter domain can be determined where this requirement is met.

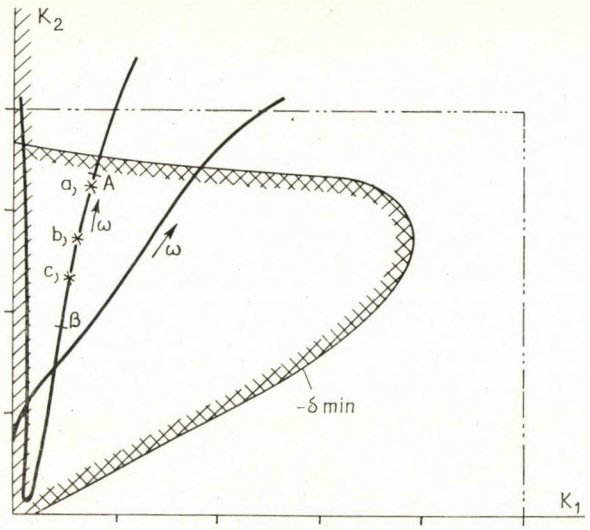


Fig. 4

The system of programs to realize this method consists of the following parts:

1) Subroutine for determining determinants with polynomial elements to get transfer functions.

2) Subroutine for parameter mapping to get $K_1 = K_1(\omega)$; $K_2 = K_2(\omega)$ (including a D-decomposition method for the investigation of stability).

3) Subroutine for determining the poles and zeros at given K_1, K_2 values.

4) Subroutine for determining time responses at given K_1, K_2 values using an inverse Laplace tranformation method.

5) Subroutine for determining a reduced equation from the characteristic equation and for implementing D-decomposition method for the reduced equation.

After sketching roughly the idea of a possible linear system design method we are discussing some questions of the nonlinear system design more fully.

On Computer Aided Design of Nonlinear Systems

It is a rather difficult task to develop computer aided methods for nonlinear control system design. This is due to the fact that, in contrary with the linear systems, general results are not available in this field.

For second-order nonlinear systems the phase plane method, or its computer implementation can be used with great effect. An advanced program for analysing nonlinear networks in the phase plane is published in [17].

For high-order systems a way of developing computer aided design method is the computer implementation of existing theoretical methods, in particular, of the method of harmonic linearization, of Popov's method of absolute stability, of the method of Lyapunov functions, etc.

In this paper the systems called in [18] as nonlinear systems of the first class will be analysed. The nonlinearities of these systems are the functions of one of the state variables. Some of the proposed computer methods (and developed programs) can be applied at the investigation of nonlinear systems of second and third class as well (where the nonlinearities are the functions of two or more system state variables connected by linear or nonlinear differential equations respectively.)

System Differential Equations

For developing a computer aided design method it is essential to set out from the form of the system equations in which they are usually given, and to perform all the necessary transformations by computer methods.

The differential equations of nonlinear automatic control systems of the first class, when the method of absolute stability or the method of harmonic linearization are used, are usually transformed to the following form:

$$Q(p)x_k = \sum_{i=1}^{m} R_i(p)\,[g_i(t) - F_i(x_k)] \tag{1}$$

where

x_k — a system state variable

$Q(p)$, $R_i(p)$ — operational polynomials

$F_i(p)$ — nonlinear function

$g_i(p)$ — external forcing signals

($i = 1, 2, \ldots m$; $m \leq n$; n is the order of the differential equation).

The system differential equations are usually not given in form (1). Frequently, a system block-diagram or the differential equations of the elements of the system are given.

The most general form of the differential equations of a nonlinear system of the first class is the following [18, 19]

$$D_{i1}(p)\,x_1 + D_{i2}(p)\,x_2 + \ldots + D_{im}(p)\,x_m + F_i(x_k) = g_i(t)$$
$$(i = 1, 2, \ldots m) \tag{2}$$

where $D_{ij}(p)$ are operational polynomials and have the following general form:

$$D_{ij}(p) = d_{ijn}p^n + d_{ij(n-1)}p^{n-1} + \ldots + d_{ij1}p + d_{ijo}$$
$$(i, j = 1, 2 \ldots m) \tag{3}$$

d_{ijh} are determined by system parameters (like gains and time constants) $(i, j = 1, 2, \ldots m;\ h = 0, 1 \ldots n)$.

If the system equations are not given in form (2) this form may easily be reconstructed from block diagrams.

Form (1) of differential equations may be obtained from (2) by formal operations. Formally, the system of equations (2) can be solved like an algebraic system of equations, by means of Cramer's rule. Thus $Q(p)$ can be obtained by the evaluation of a polynomial determinant

$$Q(p) = \begin{vmatrix} D_{11} & \ldots & D_{1k} & \ldots & D_{1m} \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ D_{i1} & \ldots & D_{ik} & \ldots & D_{im} \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ \cdot & & \cdot & & \cdot \\ D_{m1} & \ldots & D_{mk} & \ldots & D_{mm} \end{vmatrix} \tag{4}$$

$R_i(p)$ corresponds to the cofactor of the (i, k)-th element of the determinant (4).

Form (1) of the differential equations of the systems in rather simple cases can be obtained without any difficulties in traditional way, but to determine them for complicated, multiloop systems may be however a very complicated task.

This problem can easily be solved having a computer subroutine for the evaluation of the determinant of a matrix with polynomial elements.

On the basis of the idea published in [1] a subroutine (procedure) was developed by us in ALGOL for the evaluation of such determinants. The accuracy of the method is also due to a scaling procedure [1].

Let us suppose that in the system equations there is a parameter "K". Its effect on the system performance is going to be investigated. Let this parameter be only in one operational polynomial. Thus

$$D_{ef}(p) = D'_{ef}(p) + KD''_{ef}(p) \qquad (ef \in ij) \tag{5}$$

The sytem differential equation can be expressed in the form

$$[KQ_1(p)+Q_2(p)]\,x_k+\sum_{i=1}^{m}[KR_{1i}(p)+R_{2i}(p)]\,F_i(x_k)=$$

$$=\sum_{i=1}^{m}[KR_{1i}(p)+R_{2i}(p)]\,g_i(t) \qquad (i=1,2\ldots,m) \qquad (6)$$

The operational polynomials $Q_1(p)$, $Q_2(p)$, $R_{1i}(p)$, $R_{2i}(p)$ can be obtained by taking the f-th column as a sum of two columns, one with the only nonzero element D''_{ef}.

If there are more than one free parameters, then the operational polynomials can be divided into parts in a similar way. At the developed variant of the program at most four parameters can figure in the operational polynomials.

The polynomial matrix subroutine can also be used for linear system analysis to determine the transfer function between external forcing signals and any system state variable or between any of two system state variables. This very subroutine is used in linear system design method discussed in this paper earlier. Besides, it may be applied in the analysis of nonlinear systems of second and third class for obtaining the operational relationships connecting several state variables figuring in the nonlinear functions.

Implementation of the Method of Absolute Stability

After the system being designed as linear, the effect of the possible nonlinearities up on the absolute stability of the system [19—22] can be investigated.

In [23] a computer method to perform this task is described.

Having the transfer functions,

$$W_i(p)=\frac{R_i(p)}{Q(p)}, \qquad (i=1,2\ldots m) \qquad (7)$$

the modified frequency characteristics

$$W_i^*(\omega)=R_e\,W_i(j\omega)+j\omega\ \mathrm{Im}\ W_i(j\omega) \qquad (8)$$

$$(j=\sqrt{-1})$$

are determined by the developed computer program.

The frequency boundaries $\omega_{\min}$, $\Delta\omega_1$, ω_1, $\Delta\omega_2$, $\omega_2 \ldots \omega_{\max}$ and the step sizes $\Delta\omega$ have to be input. An automatic step size method had been considered but its proof is unnecessary. However, the program automatically computes the maximum absolute stability sector for one valued nonlinear characteristics.

The developed program can be cause also for the investigation of exponential stability [19].

A General Method for Harmonic Analysis of Piecewise Linear Systems

The harmonic linearization method has a very important role in solving practical problems. This is an approximate method the beauty of which is in its simplicity. However, as a matter of fact, this simple method sometimes needs quite a bit of computations and without having a computer aided method for system design, it becomes mainly of theoretical interest with no chance to be useful for solving rather complicated problems.

In this part of the paper piecewise linear systems will be discussed. Besides the usual describing function method a more advanced variant, proposed by E. P. Popov, is used for considering the effect of higher harmonics.

Using the method of harmonic linearization one is faced with the very real problem of determining the coefficients figuring at this method.

Detailed information about the describing functions of piecewise linear characteristics is available in works [18, 19, 24—27]. The first attempt to show some general features of their describing functions was made by K. Magnus [24] (see also [26, 27]), and an other approach by R. Sridar [25].

A general solution of this problem was given by the author in [28, 29], where a general describing function method was proposed for any single- or double-valued, symmetric or asymmetric piecewise linear characteristic. Moreover, the coefficients of higher harmonics were determined in the same manner. The method was extended to the determination of the coefficients of improvement of the approximate periodic solution.

Afterwards computer subroutines in FORTRAN and also in ALGOL were developed for the computation of the coefficients.

Let us consider a double-valued nonlinear characteristic — $F(x)$ — given in Fig. 5. It has a very general character and with the proper choice of its parameters $F_i^{(j)}$, $K_i^{(j)}$, $D_i^{(j)}$, $f^{(j)}(0_{-0})$, $j = 1, 2$) any piecewise linear characteristic can be realized. — Figure 5 shows that $F_i^{(j)}$ — are the bend point

coordinates, $K_k^{(j)}$ — are the slopes and $D_i^{(j)}$ — are the discontinuities of the characteristics, for the first branch $j = (1)$ (with x increasing) and second branch $j = (2)$ (with x decreasing). For single-valued characteristics the two branches are the same.

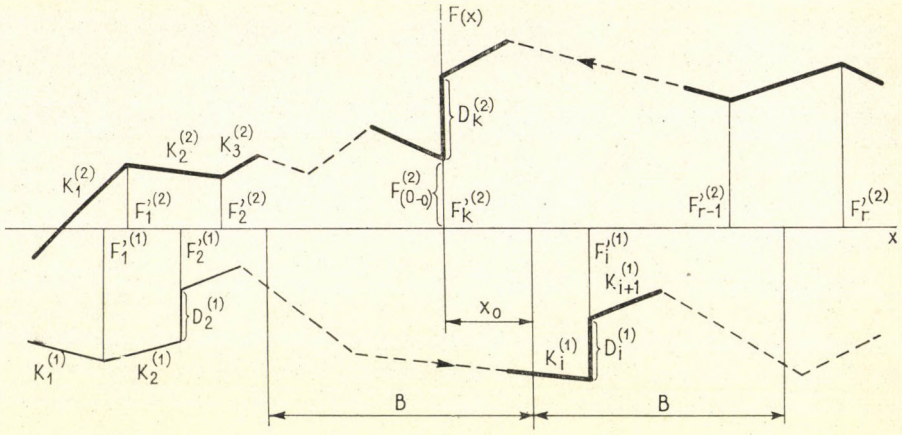


Fig. 5

Let us suppose that the input signal of this nonlinearity has the form:

$$x = x_0 + B \sin \omega t,$$

where

x_0 — constant component

B — amplitude

ω — frequency.

The relative coefficients of the Fourier expansion of the output signal are determined by the formulae:

constant component

$$A_0(x_0, B) = \frac{1}{2\pi} \int_0^{2\pi} F(x)\, d\varphi$$

first harmonic

$$b(x_0, B) = \frac{1}{\pi B} \int_0^{2\pi} F(x) \sin \varphi\, d\varphi$$

$$a(x_0, B) = \frac{1}{\pi B} \int_0^{2\pi} F(x) \cos \varphi\, d\varphi$$

higher harmonics

$$r_\nu(x_0,B) = \frac{1}{\pi B}\int_0^{2\pi} F(x)\sin\nu\varphi d\varphi$$

$$s_\nu(x_0,B) = \frac{1}{\pi B}\int_0^{2\pi} F(x)\cos\nu\varphi d\varphi.$$

For improving the approximate periodic solution the following coefficient are necessary

$$L_{s\nu}(x_0,B) = \frac{1}{2\pi}\int_0^{2\pi}\frac{dF(x)}{d(x)}\sin\nu\varphi d\varphi$$

$$L_{c\nu}(x_0,B) = \frac{1}{2\pi}\int_0^{2\pi}\frac{dF(x)}{dx}\cos\nu\varphi d\varphi$$

$$I_{s\nu}(x_0,B) = \frac{1}{\pi}\int_0^{2\pi}\frac{dF(x)}{dx}\sin\nu\varphi\sin\varphi d\varphi$$

$$I_{c\nu}(x_0,B) = \frac{1}{\pi}\int_0^{2\pi}\frac{dF(x)}{dx}\cos\nu\gamma\sin\varphi d\varphi$$

$$J_{s\nu}(x_0,B) = \frac{1}{\pi}\int_0^{2\pi}\frac{dF(x)}{dx}\sin\nu\varphi\cos\varphi d\varphi$$

$$J_{c\nu}(x_0,B) = \frac{1}{\pi}\int_0^{2\pi}\frac{dF(x)}{dx}\cos\nu\varphi\cos\varphi d\varphi\,.$$

In [28] and [29] general formulae are given for determining the coefficients $A_0, b, a, r_\nu, s_\nu, L_{s\nu}, L_{c\nu}, I_{s\nu}, I_{c\nu}, J_{s\nu}, J_{c\nu}$. (For ν even or odd the formulae are different.)

The developed computer program computes the coefficients for the given set of amplitudes by a given x_0 value according to data

$$F_i'^{(j)},\ K_i^{(j)},\ D_i^{(j)},\ f^{(j)}\ (\mathrm{O}_{-0})$$

($j = 1, 2$; $i = 1, 2 \ldots n$; n is the number of the linear sections).

Investigation of Approximate Periodic Solutions

In this paragraph the application of the idea of harmonic linearization for system investigation is discussed. In the lack of space no details, only the general idea can be given. More extensive information is available in [18, 19, 25, 29].

Let us have the differential equation of a system in form (1). Following the idea of [18] it is assumed that a periodic solution exists in the system. The investigation is accomplished for a properly chosen system state variable. The periodic solution can be sought in the form:

$$x_k = x_{k0} + \tilde{x}_k + \sum_{\nu=2}^{\infty} \tilde{x}_{k\nu} \tag{9}$$

where x_{k0} — constant component

x_k — first harmonic

$\tilde{x}_k = B \sin \omega t$

B — amplitude

ω — frequency

$\tilde{x}_{k\nu}$ — higher harmonics

$$\tilde{x}_{k\nu} = \delta_\nu B \sin (\nu\omega t + \alpha_\nu) \tag{10}$$

δ_ν — relative amplitude.

Let us assume that the harmonic linearization can be applied. Accordingly, the higher harmonics and their sum are of small value.

In the following the method proposed by E. P. Popov [18] for the improvement of the approximate periodic solution taking a finite number of higher harmonics into a consideration will be briefly outlined.

Let us expand the output signal of the nonlinearities in Taylor series at an input signal, corresponding to Eq. (8).

$$y_i = F_i(x_k) = F_i(x_{k_0} + \tilde{x}_k) + \sum_{\nu=2}^{\eta} \tilde{x}_{k\nu} \frac{dF_i(x_{k_0} + \tilde{x}_k)}{dx} + \ldots \tag{11}$$

Only the first two terms of the series are considered.

Then this signal is expanded in Fourier series. Truncating the series, only a finite number (η) of higher harmonics is taken into consideration.

$$y_i = \hat{A}_{0i} + \left(\hat{b}_i + \frac{\hat{a}_i}{\omega} P\right) \tilde{x}_k + \sum_{\nu=2}^{\eta} \left(\hat{r}_{\nu i} + \frac{dF_i(x_{k_0} + \tilde{x}_k)}{\nu\omega} P\right) B \sin \nu\omega t \quad (12)$$

$$P = \frac{d}{dt},$$

where A_{0i}, b_i, a_i, r_i, are relative Fournier coefficients of the expansion of the expression (11).

For example

$$b_i = b_i + \Delta b_i \quad (13)$$

$$b_i = \frac{1}{\pi B} \int_0^{2\pi} F_i(x_{k_0} + \tilde{x}_k) \sin \varphi \, d\varphi \qquad (\varphi = \omega t)$$

$$\Delta b_i = \frac{1}{\pi B} \int_0^{2\pi} \sum_{\nu=0}^{\infty} \tilde{x}_{k\nu} \frac{dF_i(x_{k_0} + \tilde{x}_k)}{dx} \sin \varphi \, d\varphi \,. \quad (14)$$

The other coefficients are determined by similar expressions. According to the system differential equation (1), substituting (12) we have:

$$Q(P)\left[x_{k_0} + \tilde{x}_k + \sum_{\nu=2}^{n} B \delta_\nu (\sin \nu\omega + \alpha_\nu)\right] +$$

$$+ \sum_{i=1}^{m} R_i(P) \sum_{j=i}^{li} \left[\hat{A}_{0i} + \left(\hat{b}_i + \frac{\hat{a}_i}{\omega} P\right) \tilde{x}_k + \sum_{\nu=2}^{n} \left(\hat{r}_{\nu i} + \frac{\hat{s}_{\nu i}}{\nu\omega} P\right) B \sin \nu\omega t\right] =$$

$$= \sum_{i=1}^{m} R_i(P) fi(t) \,. \quad (15)$$

The external forces are assumed to vary so slowly that during a period of the periodic solution they should be considered constants

$$\sum_{i=1}^{m} R_i(P) fi(t) = \sum_{i=1}^{m} M_i = M \,. \quad (16)$$

The M_i values can easily be determined according to [18]. Neglecting the small quantities of second-order from equation (15) the following computational relationships can be obtained [18, 28]

$$Q(j\omega) + \sum_{i=1}^{m} R_i(j\omega)(\hat{b}_i + j\hat{a}_i) = 0 \qquad (j = \sqrt{-1}), \quad (17)$$

where

$$b_i = b_i(x_{k0},\ B) + \Delta b_i(x_{k0}, B, \omega) \tag{18a}$$

$$a_i = a_i(x_{k0},\ B) + \Delta a_i(x_{k0}, B, \omega); \tag{18b}$$

$$Q(0)\, x_{k0} + \sum_{i=1}^{m} R_i(0)\, \hat{A}_{0i} = \sum_{i=1}^{m} M_i\,, \tag{19}$$

where

$$\hat{A}_{0i} = A_{0i}(x_{k0},\ B) + \Delta\, A_{0i}(x_{k0}.\ B, \omega)\,. \tag{20}$$

The negligence of small quantities is reflected in the fact that it is assumed that

$$\hat{r}_\nu(x_{k0},\ B) = r_\nu(x_{k0},\ B)$$

$$\hat{s}_\nu(x_{k0},\ B) = s_\nu(x_{k0},\ B)\,.$$

The coefficients ΔA_{0i}, Δb_i, Δa_i reflecting the effect of higher harmonics are determined by the following relationship

$$\Delta b_i(x_{k0},\ B,\ \omega) = \sum_{\nu=2}^{n} M_\nu\, I_{s\nu i} + N_\nu\, I_{c\nu i} \tag{21}$$

$$\Delta a_i(x_{k0},\ B,\ \omega) = \sum_{\nu=2}^{n} M_\nu\, J_{s\nu i} + N_\nu\, I_{l\nu i} \tag{22}$$

$$\Delta\, A_{0i}(x_{k0},\ B,\ \omega) = \sum_{\nu=2}^{n} M_\nu\, L_{s\nu i} + N_\nu\, L_{c\nu i} \tag{23}$$

where

$$M_\nu = \sum_{i=1}^{m} \delta_{\nu i} \cos \alpha_{\nu i} \tag{24a}$$

$$N_\nu = \sum_{i=1}^{m} \delta_{\nu i} \sin \alpha_{\nu i} \tag{24b}$$

$$\delta_{\nu i} = \left| \frac{R_i(j)}{Q(j)} \right| \sqrt{r_{\nu i}^2 + s_{\nu i}^2} \tag{24c}$$

$$\alpha_{\nu i} = \pi + \arg \frac{R_i(j)}{Q(j)} + \operatorname{arc\,tg} \frac{s_{\nu i}}{r_{\nu i}}\,. \tag{24d}$$

Formulae (21)—(24) are obtained on the basis of (15).

Separating the real and complex parts of Eq. (17), two equations can be obtained

$$\begin{aligned} X(x_{k0},\ B,\omega) &= 0 \\ Y(x_{k0},\ B,\omega) &= 0\,. \end{aligned} \tag{25}$$

From (25) the three unknowns (x_{k0}, B, ω) can be determined, taking Eq. (19) also into consideration.

A more realistic case is when x_{k0} is known (or assumed to be known) — then the two unknown (the B-amplitude and ω-frequency) can be determined from (25).

In a control system design the effect of the parameters on the system performance is of first interest. For system investigation the parameter plane method can be used [18, 19].

Let us suppose that there are two parameters — K_1 and K_2 — in the system, the effect of which we are to investigate and with respect to which Eqs (25) are linear. (It is a very realistic case, because linear part gains or time constant may be such parameters.)

Equation (25) will have the following form:

$$K_1 X_1(x_{k0}, B, \omega) + K_2 X_2(x_{k0}, B, \omega) + X_3(x_{k0}, B, \omega) = 0 \tag{26}$$

$$K_1 Y_1(x_{k0}, B, \omega) + K_2 Y_2(x_{k0}, B, \omega) + Y_3(x_{k0}, B, \omega) = 0.$$

On the basis of Eq. (26) taking a set of values of ω and B the constant ω and B lines can be determined in the K_1K_2 plane.

For the realization of the outlined methods the proper form of the differential equation of the system, the relative Fourier coefficients and the coefficients of improvement are necessary. The previous paragraphs were about a computer method for determining the proper form of differential equations on one hand, and for the case of piecewise linear characteristics computer method for determining the above-mentioned coefficients on the other. Using these methods a computer program for the investigation of the periodic solution can be developed.

The computer program developed by us is based on a principle somehow different from the one mentioned earlier in this paragraph. The parameter plane method is used. At first the approximate periodic solutions, obtained by harmonic linearization, are determined and then these are improved taking a finite number of higher harmonics into consideration.

The effect of one system parameter — K — is investigated. Let the system equations have the form

$$KX_{(1)}(x_{k0}, B, \omega) + X_{(2)}(x_{k0}, B, \omega) = 0$$

$$KY_{(1)}(x_{k0}, B, \omega) + Y_{(2)}(x_{k0}, B, \omega) = 0. \tag{27}$$

Such equation is obtained for example from Eq. (6), using the method of harmonic linearization.

At first we assume $\Delta b \equiv 0$, $\Delta a \equiv 0$.

It means that the usual describing function method is used.

Eliminating K from the two equations (27), an equation connecting the amplitudes B and frequencies ω is obtained. (x_{k0} is given.) E.g.

$$X_{(1)} Y_{(2)} - X_{(2)} Y_{(1)} = 0 . \tag{28}$$

For a set of amplitudes, for each amplitude, the positive real solutions of Eq. (28) are the possible frequencies of the oscillations. (The degenerated cases (at least one of $X_{(1)}$, $X_{(2)}$, $Y_{(1)}$, $Y_{(2)}$, is equal to zero) need special

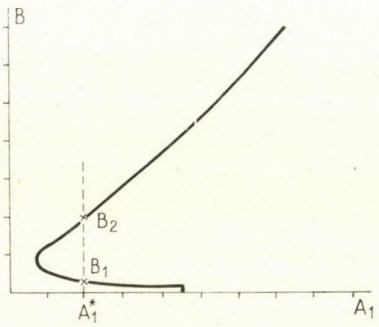


Fig. 6

attention.) At a given value B, and the values ω obtained in the above way, from one of the equations (27) the corresponding parameter values K can be determined. The functions $B = f_1(K)$, $\omega = f_2(K)$ can be built. In the case of asymmetrical oscillations from a type (19) equation ($\Delta A \equiv 0$) the external forces $\sum_{i=1}^{m} M_i$ causing the given x_{k0} value, can be determined.

(As an example a graphic $B = f(K)$ is given in Fig. 6.)

For a given parameter value $K = K^*$ the corresponding B_h values are obtained by iteration. (To a given $K = K^*$ value several approximate solutions may correspond to $h = 1, 2 \ldots$)

The improvement of the solution, is performed in the following way taking a finite number of higher harmonics into consideration.

If the harmonic linearization is applicable the improved and the approximate solutions are close to each other ($B_h \approx B_n^*$, $\omega_h = \omega_h^*$).

The coefficients Δb, Δa and ΔA_0 are determined on the basis of approximate solution B_h^*, ω_h^* according to formulae (21)—(23). Taking these concrete values Δb, Δa, ΔA_0 the outlined procedure for determining the periodic solu-

tion is repeated. This successive approximation procedure is continued till the required accuracy is reached.

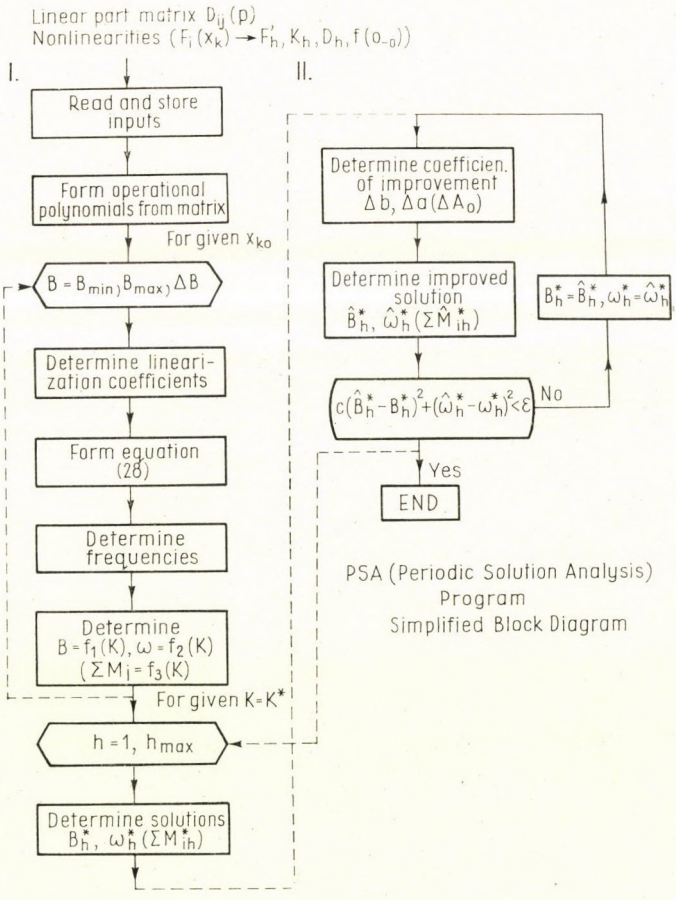
Figure 7 shows the computer block diagram of realization of this method.

The first experimental variant considered one nonlinear element. The program was developed in ALGOL.

It is an important fact that the describing function method, in its original form, is involved in the proposed procedure.

The results obtained in such a way may be used for further investigation of the system using the exact mathematical method proposed in [37].

It is noteworthy that the described computer program can find useful application in the synthesis of nonlinear compensating elements [29, 30] on the basis of E. P. Popov's exponential harmonic describing function method.



PSA (Periodic Solution Analysis) Program Simplified Block Diagram

Fig. 7

Example

To demonstrate the computer method of improvement in the approximate periodic solution, some of the results of the investigation of a simple problem will be given. Earlier, in [18 p. 619], this problem was analysed in a traditional way.

The system is shown in Fig. 8a. (The quantities are used without dimensions.)

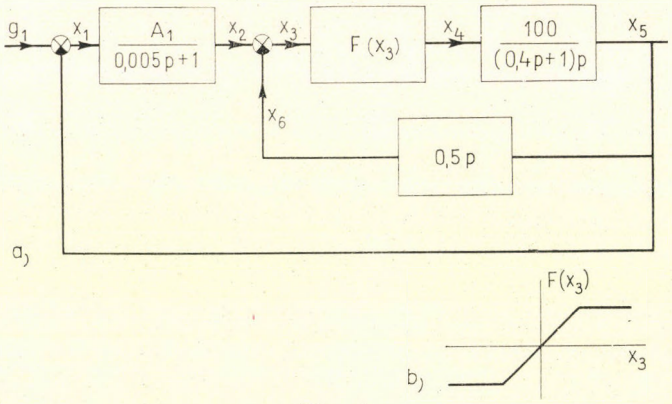


Fig. 8

The system differencial equations are

$$x_1 + x_5 = g_1$$

$$A_1 x_1 - (0.005\, p + 1) x_2 = 0$$

$$x_2 - x_3 - x_6 = 0$$

$$x_4 = F(x_3)$$

$$100 x_4 - (0.4\, p + 1) p x_5 = 0$$

$$0.5\, p x_5 - x_6 = 0.$$

This system of differential equations can be transformed into the following form

$$Q(p)\, x_3 = R_1(p)\, g - R_2(p)\, F(x_3)\,,$$

where

$$Q(p) = \begin{array}{|c|c|c|c|c|c|} \hline 1 & & & & 1 & \\ \hline A_1 & -0{,}005p-1 & & & & \\ \hline & 1 & -1 & & & -1 \\ \hline & & & 1 & & \\ \hline & & & 100 & -0{,}4p^2-p & \\ \hline & & & & 0{,}5p & -1 \\ \hline \end{array}$$

$R_1(p)$ — is the cofactor of (1, 3) element

$R_3(p)$ — is the cofactor of (4, 3) element of the $Q(p)$ determinant.

The subroutine for determining the polynomial determinants results:

$$Q(p) = 0.002\,p^3 + 0.406\,p^2 + p$$

$$R_1(p) = A\,(0.4\,p^2 + p)$$

$$R_2(p) = 100A + 0.25\,p^2 + 50\,p$$

(Of course, in this simple case these polynomials could be obtained easily in an elementary way.)

Let us suppose, that the nonlinearity is a simple saturation as shown in Fig. 8b.

This nonlinearity is characterized by the data:

$$n = 2 \text{ (number of linear sections)}$$

$$F_1 = 1,\ K_1 = 1,\ D_1 = 0,\ K_2 = 0,\ F(0_{-0}) = 0.$$

At the parameter value $A_1^* = 140$ the computer program results two approximate periodic solutions for describing the function method.

These are:

$$B_1 = 2.28 \qquad \omega = 118.32$$

$$B_2 = 21.21 \qquad \omega = 44.72.$$

It can easily be concluded that the first is unstable, the second is stable approximate solution.

The filter property of the linear part of the system is significant in this case. It seemed evident that the approximate solution would be close to he improved solution.

The simulation of the system by a digital analog simulation language resulted a periodic solution which significantly differs from the approximate solution.

The results of the improvement of the approximate solution justified the results of simulation. It was also detected that a significant number of higher harmonics effects the improved solution. This is demonstraated by Table 1.

Table 1

Number of higher harmonics taken into consideration	9		21	
Number of iterations	B	ω	B	ω
1	21.21	44.72	21.21	44.72
2	18.32	47.96	17.17	49.46
3	17.45	49.08	15.83	51.39
4	17.13	49.41	15.27	52.27
5	17.01	49.68	15.01	52.70
6	16.96	49.75	14,89	52.90
7	16.94	49.78	14.83	53.00
8			14.80	53.05

(A further increase of the number of higher harmonics did not effect the improved solution significantly. The last results $B_2 = 14.80$; $\omega = 53.05$ are in harmony with the results of simulation.)

In Table 2 the relative amplitudes (δ_ν) of higher harmonics are given at $B = 21.21$ and $\omega = 44.72$ in order to demonstrate the filter property.

Table 2

ν	3	5	7	9	11	13	15	17
δ_ν	0.0267	0.0047	0.0022	0.0015	0.0011	0.0008	0.0006	0.0005

The example seems to demonstrate that by the computer methods new effects can be detected, and, besides, the important investigation of accuracy and convergence problems, connected with the method of harmonic linearization [32, 33, 36, 37].

Further Perspectives

Recently, digital analog simulation languages with interactive computer graphics are used in several variants [34, 35]. It seems to us that implementations of theoretical methods of investigation could be used as options at such methods, of control system investigation.

A user-oriented design method should have a direct block diagram input form as well.

The digital analog simulation languages on the one hand, and the methods proposed in this paper on the other can be developed in this direction.

Conclusions

The idea of a new computer aided linear system design method was outlined.

As regards the nonlinear systems, only the investigation of a class of piecewise linear systems by harmonic linearization method was discussed in detail. But there is a great part of practical tasks belonging to such problems. In addition the developed subroutines can be used in the solution of many other problems.

The author believes that computer aided design methods together with simulation, being developed in interactive direction, will have a significant role in control system design in the future.

References

1. *Coffey, T. C.:* The application of modern computing technology to control system analysis and design problems. El Segundo Technical Operations Aerospace Corp., 1967.
2. *Saucedo, R.—Schiring, E. E.:* Introduction to continuous and digital control systems C. New York.
3. *Antstiss, B. M.:* Mdelta-A digital program for control system analysis. Proc. of Share-ACM-IEE D. A. Workshop. 1968.
4. *Singer, D.* et al.: Investigation of transient processes of technological system. (In Hungarian) MTA AKI Közlemények Budapest, 9, 1968.
5. *Kheir, N. A.:* Real-time digital simulation of dynamic systems. MTA AKI Közlemények Budapest, 3, 1969.
6. *Gyürki, J.:* A transformation method for the analysis and the synthesis of multivariable control system by digital computer. IFAC Congress, Warszawa 1969.
7. *Dixon, B. A.—Ironside, F. E.:* On-line computer-aided control system design using a graphical display. CAD IEE Conf. Publ. No. 51.
8. *Atkinson, P.—Davey, R. L.:* Computer-aided design of closed-loop control systems. CAD IEE Conf. Publ. No. 51.
9. *Aird, R. J.—Moseley, A. D.:* On line computer analysis for control system design. CAD IEE Conf. Publ. No. 51.
10. *Atkinson, P.—Dalvi, V. S.:* Computer-aided design of parallel compensated closed-loop control systems. Control. Summer 1969.
11. *Bass, R. W.—Gura, I.:* High order system design via state-space considerations. Joint Automatic Control Conference 1965.
12. *Sarkar, B.:* The correlation measure between the responses of two linear systems for transient input as a performance criterion for servomechanism. Indian J. Pure Appl. Phys. v. 6. Oct. 1968.
13. *Keviczky, L.—Kovács, T.:* Diagrams for optimal setting of regulars. (In Hungarian) 6th Hungarian Conference on Automation. Budapest, 1970.

14. *Mitsumaki, T.—Haruna, K.—Mizuno, K.:* Computer aided design of automatic control systems. IFAC Kyoto Symposium. Kyoto 1970.
15. *Chen, C. F.—Shieh, L. S.:* An algebraic method for control systems design. Int. J. Control, **11,** 1968.
16. *Cadena, F. E.—Thaler, G. F.:* Synthesis of systems with dominant complex roots. Int. J. Control, **12, 2,** 1969.
17. *Chua, L. O.—Brown, C. A.—Vizza, R. R.:* NAPP. A computer program for analysing nonlinear networks in the phase plane. Purdue University, 1969.
18. *Popov, E. P.—Paltov, I. P.:* Priblizhennye methody issledovania nelineinych avtomaticheskikh sistem. Moscow, 1960.
19. *Siljak, D.:* Nonlinear Systems. John Wiley, 1969.
20. *Popov, V. M.:* On the absolute stability of nonlinear automatic systems. Automation and Remote Control, 8, 1961.
21. *Aizerman, M. A.—Gantmacher, F. R.:* Absolute stability of regulator systems. AN SzSzSzR, Isdat 1963, In Russian Holden Day, 1964.
22. *Jakubovich, V. A.:* Method of matrix inequalities in the theory of absolute stability of nonlinear control systems. Automation and Remoty Control, 5, 1965.
23. *Nagy, J.:* Method of absolute stability and the Computer aided design of nonlinear control systems. (In Hungarian) 6th Hungarian Conference on Automation, Budapest, 1970.
24. *Magnus, K.:* Über ein Verfahren zur Untersuchung nichtlinearer Schwingung- und Regelungs-Systeme. VDI-Forschungscheft, 451, 1955.
25. *Gibson, J. E.:* Nonlinear automatic control. McGraw Hill, 1963.
26. *Grahm, D. —McRuer, D.:* Analysis of Nonlinear Control Systems. John Wiley, 1961.
27. *Hsu, J. C.—Mayer, A. U.:* Modern control principles and applications. McGraw Hill, 1968.
28 *Somló, J.:* Obobchenyj method dlja issledovania sistem kusochnolineinyimi characteristikami. Automation and Remote Control, 7, 10, 1966.
29. *Somló, J.:* Nekotorye obobchenyaja v oblasti chastotonykh metodov analyza i syntheza nelineinych avtomaticheskikh sistem. Automation Research Inst. Trans., Budapest, 12, 1967.
30. *Somló, J.:* On a computer aided method for nonlinear system design. IFAC Kyoto Symposium, Kyoto, 1970.
31 *Jarvis, R. A.:* Adaptive global search in a timevariant environment using a probabilistic automation (to be published).
32. *Garber, E. D.—Shifrin, E. Sh.:* Nelineinye zadachi avtomaticheskovo regulirovania sudovykh energeticheskikh ustanovok. Leningrad, 1967.
33. *Rosenwasser, E. N.:* Kolebania neleneinykh sistem. Nauka, Moszkva, 1969.
34. *Schroer, B. J.:* Expanded capabilities of CSMP graphics on the IBM 1130 digital computer. Simulation, May 1970.
35. *Cullen, P. C.:* OLDAS: On-line continuous system simulation language. Interactive Systems for Experimental Applied Mathematics. Academic Press, 1968.
36. *Fenyő, S.—Frey, T.:* Modern mathematical methods in technology. vol. 1, North-Holland Publishing Company, Amsterdam, 1969.
37. *Frey, T.—Somló, J.—Nguyen Van Quy:* The use of operators with degenerated kernel for nonlinear system investigation. 5th Congr. of IFAC, Paris, 1972.

Проектирование систем автоматического регулирования при помощи ЦВМ

Я. ШОМЛО

(Будапешт)

Разработка систем программ для проектирования систем автоматического регулирования является актуальной задачей современной теории управления.

Теоретические результаты в области оптимального и адаптивного управления дают возможность разрабатывать очень эффективные системы. Однако, в ряде практических вопросов постановка задачи существенно проще и проблемы могут быть решены применением классических, хорошо разработанных методов.

При проектировании с помощью ЦВМ эффективность проектирования определяется эффективностью выбранного метода и эффективностью связи человека с ЦВМ.

При разработке инженерных методов самыми важными факторами являются простота входных и содержательность выходных информаций.

Для иллюстрации вышесказанных, в настоящей статье предлагается метод для снитеза линейных систем автоматического регулирования с постоянными коэффициентми. В качестве примера эффективности расчетов на ЦВМ, дается метод для анализа периодических колебаний в кусочно-линейных системах. Упоминается программа для исследования абсолютной устойчивости нелинейных систем.

Для синтеза лниейных систем предлагается метод, который сочетает метод корневых годографов с определением временных характеристик. — Полюсы и нули передаточных функций, при различных значениях свободных параметров, дают сжатую информацию о поведении систем. Соответственно, по данному методу, проектирование проводится в комплексной плоскости. Точки пространства параметров, кандидаты в расчетные, проверюются с точки зрения временных характеристик (оцениваются также установившиеся ошибки). Предлагается применение нового метода для синтеза с доминирующими полюсами.

Разработана система подпрограмм, состоящая из следующих частей: 1. определение передаточных функций сложных систем (в которых символы тоже фигурируют); 2.отображение линий комплексной плоскости в плоскости параметров; 3. подпрограмма для при. менения корневого метода; 4. определение временных и 5. частотных характеристик, и т. д.

Главная программа дает возможность для гибкого содействия подпрограмм. Структура входных данных очень проста и не требует никакого преобразования уравнений системы.

В области проектирования нелинейных систем, в качестве примера, показывается эффективность применения ЦВМ для исследования периодических гармонической линеаризации с учтом высших гармоник. В этом случае также предлагаются стандартные программы для всех этапов исследования. Конкретные примеры показывали, что при расчетах на ЦВМ могут обнаружиться неожиданные эффекты, очень важные с точки зрения подевения систем (например, большая чувствительность к эффекту высших гармоник).

По мнению автора стати, в будущем наиболее используемым, мощным и быстрым орудием при проектировании систем автоматического регулирования будет широко разработанная система специальных программм, реализующая гибкую связь человека с машиной.

Dr. J. Somló
Research Institute for Automation
Hungarian Academy of Sciences
Budapest, XI, Kende u. 13—17, Hungary

Problems of Control and Information Theory, Vol. 2 (1), pp. 67—80 (1973)

DENSITY ESTIMATION AND PATTERN CLASSIFICATION

(Note to a paper of Wolverton and Wagner)

L. REJTŐ, P. RÉVÉSZ

(Budapest)

(Received April 15, 1972)

1. Density estimation

Let $\xi_1, \xi_2, \ldots, \xi_n$ be a sequence of independent identically distributed random variables with density function $f(x)$. M. ROSENBLATT [1] defined the empirical density function $f_n^{(r)}(x)$ (based on the sample) $\xi_1, \xi_2, \ldots, \xi_n$) as follows:

$$f_n^{(r)}(x) = \frac{1}{na_n} \sum_{i=1}^{n} h\left(\frac{x - \xi_i}{a_n}\right)$$

where $h(x)$ is an arbitrary density function and $\{a_n\}$ is a sequence of real numbers going to 0.

C. T. WOLWERTON and T. I. WAGNER [2] gave a little bit different definition of the empirical density function:

$$f_n^{(w)}(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{a_k} h\left(\frac{x - \xi_k}{a_k}\right) = \frac{n-1}{n} f_{n-1}^{(w)}(x) + \frac{1}{na_n} h\left(\frac{x - \xi_n}{n}\right)$$

where the meaning of $h(x)$ and $\{a_n\}$ is the same as above.

A great advantage of the second definition is that $f_n^{(w)}(x)$ can be evaluated making use of $f_{n-1}^{(w)}(x)$ and ξ_n only (not using of sample elements $\xi_1, \xi_2, \ldots, \xi_{n-1}$ directly).

It is easy to give a common generalization of the first and second definitions:

Definition:

$$f_n(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{a_{nk}} h\left(\frac{x - \xi_k}{a_{nk}}\right)$$

where $h(x)$ is an arbitrary (but given) density function and $a_{n1}, a_{n2}, \ldots, a_{nn}$ $(n = 1, 2, \ldots)$ is a monotonically decreasing sequence of real numbers.

Further we are interested in the more general case, when the independent random variables $\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_n$ (having density function $f(\vec{x})$) are elements of the r-dimensional Euclidean space E^r. We shall use the following definition of the empirical density function:

Definition:

$$f_n(\vec{x}) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{a_{nk}^r} h\left(\frac{\vec{x} - \xi_k}{a_{nk}}\right)$$

where $h(\vec{x})$ is an arbitrary (but given) density function on E^r and $a_{1n}, a_{2n}, \ldots a_{nn}$ $(n = 1, 2, \ldots)$ is a monotonically decreasing sequence of real numbers.

We shall use the following notations: $S(\vec{a};\varrho)$ denotes the sphere in E^r with centre $\vec{a}$ and radius ϱ.

$Fr\ A$ denotes the frontier of $A \subset E^r$. $||\vec{x}||$ is the norm of $\vec{x} \in E^r$.

$||\vec{x} - B||$ denotes the distance of $\vec{x}$ to B in E^r,

and $$A_\varepsilon = \{\vec{x}: \vec{} \in A,\ ||x - Fr A|| \leq \varepsilon\}.$$

In this part of the paper we are going to prove the following:

Theorem 1: Suppose that

a) *the density function* $f(\vec{x})$ $(\vec{x} \in E^r)$ *is obeying the following conditions:*

(i) $f(\vec{x})$ *is vanishing outside an open convex (not certainly bounded) domain* T*;*

(ii) *there exist the second order partial derivatives of* $f(\vec{x})$ *in* T *and*

$$\sup_{\vec{x} \in T} \max \left\{ |f(\vec{x})|; \left|\frac{\partial f}{\partial x_1}\right|; \ldots; \left|\frac{\partial f}{\partial x_r}\right|; \left|\frac{\partial^2 f}{\partial x_1^2}\right|; \left|\frac{\partial^2 f}{\partial x_1 \partial x_2}\right|; \ldots; \left|\frac{\partial^2 f}{\partial x_r^2}\right| \right\} \leq K$$

for some $0 < K < +\infty$

(iii) *there exists a* $\delta > 0$ *such that*

$$\int_{E^r} ||\vec{x}||^\delta f(\vec{x})\, d\vec{x} < +\infty$$

b) *The density function* $h(\vec{x})$ *is obeying the following conditions*

(i) $\int_{E^r} x_i h(\vec{x}) d\vec{x} = 0$ *for* $i = 1, 2, \ldots, r$,

$\int_{E^r} x_i x_j h(\vec{x}) d\vec{x}$ *exist for* $i = 1, \ldots, r,\ j = 1, \ldots, r$

and $\lim_{||\vec{x}|| \to \infty} ||\vec{x}||^{r+2} h(\vec{x}) = 0$

(ii) $|h(\vec{x})| \leq H_0$ *for all* x *in* E^r

(iii) $|h(\vec{x}) - h(\vec{y})| \leq D\, ||\vec{x} - \vec{y}||^\alpha$ *for all* $\vec{x}, \vec{y} \in E^r$ *where* D *is constant and* $0 < \alpha \leq 1$.

c) *the double array* $\{a_{nk}\}$ $(k = 1, 2, \ldots, n;\ n = 1, 2, \ldots)$ *is obeying the following conditions:*

(i) $1/2 \geq a_{n1} \geq a_{n2} \geq \ldots \geq a_{nn}$ $(n = 1, 2, \ldots)$

(ii) *there exists an* $\omega (0 < \omega < 1)$ *such that* $n^{1-\omega} a_{nn}^r \to \infty$

(iii) *for all* $\varepsilon > 0$ *there exist* K_0 *and* N_0 *such that if* $k > K_0$ *and* $n > N_0$ *then* $a_{nk} < \varepsilon$.

(iv) *there exists a constant* b $(b > 0)$ *for which*

$$\frac{1}{n} \sum_{k=1}^{n} a_{nk}^2 \leq b \cdot a_{nn}^2 \qquad (n = 1, 2, \ldots).$$

Then we have

(1) $$P\{\sup_{\vec{x} \in T - T_\varrho} |f_n(\vec{x}) - f(\vec{x})| \geq \varepsilon\} \leq e^{-c_0 n \cdot a_{nn}}$$

(2) $$P\{\int_{E^r} |f_n(\vec{x}) - f(\vec{x})|^p \, d\vec{x} \geq \varepsilon\} \leq e^{-c_1 \cdot a_{nn}}$$

(3) $$P\left\{\lim_{n \to \infty} \frac{\min\left(\sqrt{n a_{nn}^r}\,;\ \frac{1}{a_{nn}^2}\right)}{\log n} \sup_{T - T_\varrho} |f_n(\vec{x}) - f(\vec{x})| = 0\right\} = 1$$

and especially if $a_{nn} = n^{-\frac{1}{4+r}}$ *then*

(3*) $$P\left\{\lim_{n \to \infty} \frac{n^{\frac{2}{4+r}}}{\log n} \sup_{T - T_\varrho} |f_n(\vec{x}) - f(\vec{x})| = 0\right\} = 1$$

for any $\varrho > 0$, *where* $c_0 = c_0(\varrho, \varepsilon)$ *and* $c_1 = c_1(\varrho, \varepsilon)$ *are positive constants.*

Remark 1. The conditions c) (i)—(iv) hold if

$$a_{nk} = 1/k^\alpha \qquad (0 < \alpha < 1/r)$$

or if

$$a_{nk} = 1/n^\alpha \qquad (0 < \alpha < 1/r)$$

which are the really interesting cases.

The method of the proof of Theorem 1 is based on that of paper [3]. First we present several lemmas.

Lemma[1] *1. Let* $f(\vec{x})$ *and* $h(\vec{x})$ *be density functions on* E^r *obeying conditions*

[1] This Lemma is essentially due to E. PARZEN [5].

a) (i) – (ii) and b) (i) – (ii) of Theorem 1. Then for any $\vec{x} \in T - T_\varrho$

$$\left| \frac{1}{a^r} \int_{E^r} h\left(\frac{\vec{x} - \vec{y}}{a}\right) f(\vec{y})\, d\vec{y} - f(\vec{x}) \right| \leq d_1 a^2$$

where d_1 *and* ϱ *are positive constants.*

Proof

$$\left| \frac{1}{a^r} \int_{E^r} h\left(\frac{\vec{x} - \vec{y}}{a}\right) f(\vec{y})\, d\vec{y} - \int_{E^r} f(\vec{x})\, h(\vec{z})\, d\vec{z}) \right| = \left| \int_{E^r} h(\vec{z})\, [f(\vec{x} - a\vec{z}) - f(\vec{x})]\, d\vec{z} \right| \leq$$

$$\leq \left| \int_{\frac{x-T}{a}} h(\vec{z})\, [f(\vec{x} - a\vec{z}) - f(\vec{x})]\, d\vec{z} \right| + 2K \int_{E^r - \frac{\vec{x}-T}{a}} h(\vec{z})\, d\vec{z} \leq$$

$$\leq \left| \int_{\frac{\vec{x}-T}{a}} h(\vec{z}) \left[\sum_{i=1}^{r} \frac{\partial f}{\partial x_i} a z_i + K_0 \sum_{i,j} \frac{\partial^2 f}{\partial x_i\, \partial x_j} a^2 z_i z_j \right] d\vec{z} \right| +$$

$$+ 2\, K \mathsf{P}\left(\vec{\eta} \notin \frac{\vec{x} - T}{a}\right)$$

where $\vec{\eta}$ is a random variable having density function $h(\vec{x})$.

Since $\vec{x} \in T - T_s$, we have $S(\vec{0}, \varrho) \subset \vec{x} - T$, hence, by the Markov inequality:

$$\mathsf{P}\left(\vec{\eta} \notin \frac{\vec{x} - T}{a}\right) \leq \mathsf{P}\left(\vec{\eta} \notin S\left(\vec{0}\,; \frac{\varrho}{a}\right)\right) \leq K_1 a^2$$

then, using the Cauchy–Schwartz inequality:

$$\left| \int_{\frac{\vec{x}-T}{a}} z_i\, h(\vec{z})\, d\vec{z} \right| = \left| \int_{E^r - \frac{\vec{x}-T}{a}} z_i\, h(\vec{z})\, d\vec{z} \right| \leq \sqrt{\int_{E^r} z_i^2\, h(\vec{z})\, d\vec{z}} \sqrt{\int_{E^r - \frac{\vec{x}-T}{a}} h(\vec{z})\, d\vec{z}} \leq K_2\, a\,.$$

Lemma 2. Let $f(\vec{x})$ *and* $h(\vec{x})$ *be bounded density functions on* E^r. *Then for all* $\vec{x} \in E^r$

$$\mathsf{D}^2\left(\frac{1}{a^r} h\left(\frac{\vec{x} - \vec{\xi}}{a}\right)\right) \leq \frac{d_2}{d^r}$$

Proof

$$\mathsf{D}^2\left(\frac{1}{a^r}h\left(\frac{\vec{x}-\vec{\xi}}{a}\right)\right)\leq \mathsf{E}\left(\frac{1}{a^{2r}}h^2\left(\frac{\vec{x}-\vec{\xi}}{a}\right)\right)=\frac{1}{a^{2r}}\int\limits_{E^r}h^2\left(\frac{\vec{x}-\vec{y}}{a}\right)f(\vec{y})\,d\vec{y}=$$

$$=\frac{1}{a^r}\int\limits_{E^r}h^2(\vec{z})\,f(\vec{x}-a\vec{z})\,d\vec{z}\leq\frac{d_2}{a^r}.$$

Lemma 3. Let $f(\vec{x})$ and $h(\vec{x})$ be bounded density functions on $\mathsf{E}^r(h(\vec{x})\leq H_0)$ and let $a_{n1}\geq a_{n2},\ldots,a_{nn}$ be a double array obeying condition c) (ii).

Then for any $\vec{x}\in\mathsf{E}^r$

$$\mathsf{E}(e^{\lambda_n(f_n-\mathsf{E}f_n)}\leq e^{d_3\frac{\lambda_n^2}{n\cdot a_{nn}^r}}$$

provided that $0\leq\lambda_n\leq\dfrac{na_{nn}^r}{4H_0}$.

Proof

$$\mathsf{E}(e^{\lambda_n(f_n-\mathsf{E}f_n)})=\prod_{k=1}^{n}\mathsf{E}\left(e^{\frac{\lambda_n}{n\cdot a_{nk}^r}\left[h\left(\frac{\vec{x}-\vec{\xi}_k}{a_{nk}}\right)-\mathsf{E}h\left(\frac{\vec{x}-\vec{\xi}_k}{a_{nk}}\right)\right]}\right).$$

Using that $e^x\leq 1+x+x^2$ if $x<\dfrac{1}{2}$ we get

$$\mathsf{E}(e^{\lambda_n(f_n-\mathsf{E}f_n)})\leq\prod_{k=1}^{n}\left[1+\frac{\lambda_n^2}{n^2}\mathsf{D}^2\left(\frac{1}{a_{nk}^r}h\left(\frac{\vec{x}-\vec{\xi}_k}{a_{nk}}\right)\right)\right]\leq$$

applying Lemma 2

$$\leq\prod_{k=1}^{n}\left(1+\frac{\lambda_n^2}{n^2}\frac{d_2}{a_{nk}^r}\right)\leq e^{\frac{\lambda_n^2}{n_2}d_2\sum_{k=1}^{n}\frac{1}{a_{nk}^r}}\leq e^{d_3\frac{\lambda_n^2}{na_{nn}^r}}.$$

Lemma 4. Under the conditions of Lemma 3 for all $\vec{x}\in E^r$ if $n>n_0(\varepsilon)$ then

$$\mathsf{P}\{|f(\vec{x})-\mathsf{E}f_n(\vec{x})|\geq\varepsilon\}\leq e^{-d_4na_{nn}^r}$$

where d_4 is a positive constant.

Proof. Using Lemma 3 and the Markov inequality, we get

$$\mathsf{P}\{e^{\lambda_n(f_n-\mathsf{E}f_n)}\geq e^{\lambda_n\varepsilon}\}\leq e^{d_3\frac{\lambda_n^2}{na_{nn}^r}-\lambda_n\varepsilon}.$$

Choosing $\lambda_n=\dfrac{\varepsilon}{2d_3}\sqrt{n\,a_{nn}^r}$ and $n>n_0(\varepsilon)$ we have

$$\mathsf{P}\{f_n-\mathsf{E}f_n\geq\varepsilon\}\leq e^{-dna_{nn}^r}.$$

in a quite similar way one can get that

$$\mathsf{P}\{\mathsf{E}f_n - f_n \geq \varepsilon\} \leq e^{-dna_{nn}}.$$

Lemma 5. Under the conditions of Lemma 3, if $n > N_0(\varepsilon)$ then

$$\mathsf{P}\left\{\frac{\sqrt{na_{nn}^r}}{\log n}\,|f_n(\vec{x}) - \mathsf{E}f_n(\vec{x})| \geq \varepsilon\right\} \leq e^{-d_s(\log)^{s/2}}$$

for any $\vec{x} \in \mathsf{E}^r$ $(d_s > 0)$.

Proof. Using Lemma 3 and the Markov inequality we have

$$\mathsf{P}\left(\frac{\sqrt{n \cdot a_{nn}^r}}{\log n}\,(f_n - \mathsf{E}f_n) \geq \varepsilon\right) = \mathsf{P}\left(e^{\sqrt{n \cdot a_{nn}^r \log n}\,(f_n - \mathsf{E}f_n)} \geq e^{\varepsilon(\log n)^{s/2}}\right) \leq$$

$$\leq c^{(\log n)^{s/2}\left(\frac{d_3}{\sqrt{\log n}} - \varepsilon\right)}$$

which implies the lemma.

Lemma 6. Let $f(\vec{x})$ be a density function obeying condition a) (iii) of Theorem 1, and let $\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_n$ be a sequence of independent identically distributed random variables with density function $f(\vec{x})$. Further let B_n be the event that $\vec{\xi}_n$ is not belonging to $S(\vec{0}; n^{1/\delta})$ i.e.

$$B_n = \{\omega : \vec{\xi}_n(\omega) \notin S(\vec{O}; n^{1/\delta})\}.$$

Finally set

$$C_n = \{\omega : \max\{\|\vec{\xi}_1\|, \|\vec{\xi}_2\|, \ldots, \|\vec{\xi}_n\|\} \notin S(\vec{O}; n^{1/\delta}\}.$$

Then among the events B_n and among the events C_n only finitely many will occur with probability 1.

Proof

$$\sum_{n=1}^{\infty} \mathsf{P}(B_n) = \sum_{n=1}^{\infty} \mathsf{P}(\|\vec{\xi}_n\| \geq n^{1/\delta}) = \sum_{n=1}^{\infty} n\,\mathsf{P}\left(n^{1/\delta} \leq \|\vec{\xi}_n\| \leq (n+1)^{1/\delta}\right) \leq$$

$$\leq \int_{\mathsf{E}^r} \|\vec{x}\|^{\delta} f(\vec{x})\, d\vec{x}.$$

Thus, by the Borel–Cantelli lemma, among the events B_n only finitely many will occur (with probability 1).

This fact easily implies that among the events C_n finitely many will occur with probability 1 too.

Lemma 7. Under the conditions of Theorem 1

$$\mathsf{P}\{\sup_{\vec{x} \in \mathsf{E}^r} |f_n(\vec{x}) - \mathsf{E}f_n(\vec{x})| \geq \varepsilon\} \leq e^{-d_4 n a_{nn}^r}. \tag{4}$$

Proof. We introduce the following notations:

(i) Let C_n be the r-dimensional cube of edge $2n^T$:

$$C_n = \{\vec{x} = (x_1, x_2, \ldots, x_r) : -n^T \leq x_i < n^T, i = 1, \ldots, r\}.$$

(ii) Let $L_n = L_n(s)$ be the set of the lattice points:

$$L_n = \left\{ \vec{x} = (x_1, x_2, \ldots, x_r) . \quad -n^T \leq x_i < n^T, \; x_i = \frac{k_i}{n^s}, \; i = 1, 2, \ldots, r \right\}$$

where k_i is an integer and $s > \frac{1}{r} + \frac{1}{\alpha}$

(iii) $$D_n = L_n \cap C_n .$$

Then we have

$$\mathsf{P}\{\sup_{\vec{x} \in D_n} |f_n - \mathsf{E}f_n| \geq \varepsilon\} \leq \sum_{\vec{x} \in D} \mathsf{P}(|f_n(\vec{x}) - \mathsf{E}f_n(\vec{x})| \geq \varepsilon) \leq$$

$$\leq (2n^{T+S})^r \, e^{-d_5 n \cdot a_{nn}^r} \leq e^{-d_6 n a_{nn}^r}, \tag{5}$$

further if n is big enough then

$$\mathsf{P}\{\sup_{\vec{x} \notin C_n} |f_n - \mathsf{E}f_n| \geq \varepsilon\} \leq \mathsf{P}\{\sup_{\vec{x} \notin C_n} f_n \geq \varepsilon\} + \mathsf{P}(\sup_{\vec{x} \notin C_n} \mathsf{E}f_n \geq \varepsilon) =$$

$$= \mathsf{P}(\sup_{\vec{x} \notin C_n} f_n \geq \varepsilon) \leq \mathsf{P}\left\{ \sup_{\vec{x} \notin C_n} \frac{1}{n} \sum_{\left\{k : \frac{\vec{x} - \vec{\xi}_k}{a_{nk}} \in C_n\right\}} h\left(\frac{\vec{x} - \vec{\xi}_k}{a_{nk}}\right) \geq \frac{\varepsilon}{2} \right\} + \tag{6}$$

$$+ \mathsf{P}\left\{ \sup_{\vec{x} \notin C_n} \frac{1}{n} \sum_{\left\{k : \frac{\vec{x} - \vec{\xi}_k}{a_{nk}} \notin C_n\right\}} \frac{1}{a_{nk}^r} h\left(\frac{\vec{x} - \vec{\xi}_k}{a_{nk}}\right) \geq \frac{\varepsilon}{2} \right\}$$

and by condition b) (i)

$$\mathsf{P}\left\{ \sup_{\vec{x} \notin C_n} \frac{1}{n} \sum_{\left\{k : \frac{\vec{x} - \vec{\xi}_k}{a_{nk}} \notin C_n\right\}} \frac{1}{a_{nk}^r} h\left(\frac{\vec{x} - \vec{\xi}_k}{a_{nk}}\right) \geq \frac{\varepsilon}{2} \right\} \leq$$

$$\leq \mathsf{P}\left\{ \frac{1}{n a_{nn}^r} \frac{1}{n^{(r+2)r}} \geq \varepsilon/2 \right\} = 0 \tag{7}$$

and

$$\mathsf{P}\left\{ \sup_{\vec{x} \notin C_n} \frac{1}{n} \sum_{\left\{k : \frac{\vec{x} - \vec{\xi}_n}{a_{nk}} \in C_n\right\}} \frac{1}{a_{nk}^r} h\left(\frac{\vec{x} - \vec{\xi}_k}{a_{nk}}\right) \geq \frac{\varepsilon}{2} \right\} \leq$$

$$\leq \mathsf{P}\left(\frac{H_0}{n a_{nn}^r} \sum_{\{k : \vec{\xi}_n \notin C_{n/2}\}} 1 \geq \varepsilon/2 \right) \approx \left(\frac{n}{n^T} \right)^{a_{nn}^r \frac{\varepsilon}{2H_0}} \tag{8}$$

finally if

$$\vec{y}_0 = \left(\frac{k_1^0}{n^s}; \frac{k_2^0}{n^s}, \ldots, \frac{k_r^0}{n^s}\right)$$

and

$$\vec{x} = (x_1, x_2, \ldots, x_r)$$

such that

$$\frac{k_i^0}{n^s} \leq x_i < \frac{k_i^0 + 1}{n^s} \qquad (i = 1, 2, \ldots, r)$$

then

$$|f_n(\vec{x}) - f_n(\vec{y}_0)| \leq \frac{1}{n} \sum_{k=1}^{n} \left| \frac{1}{a_{nk}^r} \left[h\left(\frac{\vec{x} - \vec{\xi}_k}{a_{nk}}\right) - h\left(\frac{\vec{y}_0 - \vec{\xi}_k}{a_{nk}}\right) \right] \right| \leq$$

$$\leq \frac{1}{n} \sum_{k=1}^{n} \frac{1}{a_{nk}^r} \frac{D}{a_{nk}^\alpha} \|\vec{x} - \vec{y}_0\|^\alpha \leq O\left(\frac{1}{a_{nn}^{r+\alpha} n^{\alpha s}}\right) \tag{9}$$

and similarly

$$|\mathsf{E} f_n(\vec{x}) - \mathsf{E} f_n(\vec{y}_0)| \leq \frac{\varepsilon}{2} \tag{10}$$

if n is big enough.

Inequalities (5)—(10) clearly imply (4).

Lemma 8. Under the conditions of Theorem 1

$$\mathsf{P}\left\{\frac{\sqrt{n a_{nn}^r}}{\log n} \sup_{\vec{x}} |f_n(\vec{x}) - \mathsf{E} f_n(\vec{x})| \to O\right\} = 1 .$$

Proof. The proof of this lemma is similar to that of Lemma 8. As a first step we prove that

$$\mathsf{P}\left\{\frac{\sqrt{n a_{nn}^r}}{\log n} \sup_{\vec{x} \in D_n} |f_n(\vec{x}) - \mathsf{E} f_n(\vec{x})| \to O\right\} = 1 . \tag{11}$$

By Lemma 5 we have:

$$\mathsf{P}\left\{\frac{\sqrt{n a_{nn}^r}}{\log} \sup_{\vec{x} \in D_n} |f_n - \mathsf{E} f_n| \geq \varepsilon\right\} \leq (2n)^{(T+S)r} e^{-d_5 (\log n)^{3/2}}$$

which clearly implies (11).

The next part of the proof can be obtained by repeating the second part of Lemma 7.

Proof of (1)

(1) immediately follows from Lemmas 1 and 7.

Proof of (2). Clearly one can find a number $R > 0$ such that

$$\int \mathsf{E} f_n(\vec{x})\, d\vec{x} \geq 1 - \varepsilon/2 \qquad (n = 1, 2, \ldots)$$

and

$$|\, \mathsf{E} f_n(x)\, | \leq {}^1\!/_2 \quad \text{if} \quad x \notin S(\vec{O}\,; R)\,.$$

Then

$$\mathsf{P}\left\{\left[\int\limits_{\mathsf{E}^r} |f_n - \mathsf{E} f_n|^p\, d\vec{x}\right]^{1/p} \geq \varepsilon\right\} \leq \mathsf{P}\left\{\int\limits_{S(\vec{O}\,;\,R)} |f_n - \mathsf{E} f_n|^p\, d\vec{x} \geq \frac{\varepsilon^p}{2}\right\} +$$

$$+ \mathsf{P}\left\{\int\limits_{\vec{x} \notin S(O\,;\,R)} |f_n - \mathsf{E} f_n|^p\, d\vec{x} \geq \frac{\varepsilon^p}{2}\right\} \leq \mathsf{P}\left\{\sup_{\vec{x} \in S(O\,;\,R)} |f_n - \mathsf{E} f_n| \geq \frac{\varepsilon}{L}\right\} +$$

$$+ \mathsf{P}\left\{\int\limits_{\vec{x} \notin S(\vec{O}\,;\,R)} |f_n - \mathsf{E} f_n|^p\, d\vec{x} \geq \frac{\varepsilon^p}{2},\ \sup |f_n - \mathsf{E} f_n| \geq 1\right\} +$$

$$+ \mathsf{P}\left\{\int\limits_{\vec{x} \notin S(\vec{O}\,;\,R)} |f_n - \mathsf{E} f_n|^p\, d\vec{x} \geq \frac{\varepsilon^p}{2},\ \sup |f_n - \mathsf{E} f_n| < 1\right\} \leq$$

$$\leq e^{-dna_{nn}^r} + \mathsf{P}\left\{\int\limits_{\vec{x} \notin S(\vec{O}\,;\,R)} |f_n - \mathsf{E} f_n|\, d\vec{x} \geq \frac{\varepsilon^p}{2}\right\} \leq$$

$$\leq e^{-dna_{nn}^r} + \mathsf{P}\left\{\int\limits_{\vec{x} \notin S(\vec{O}\,;\,R)} f_n\, d\vec{x} \geq \frac{\varepsilon^p}{4}\right\} + \mathsf{P}\left\{\int\limits_{\vec{x} \notin S(\vec{O}\,;\,R)} \mathsf{E} f_n\, d\vec{x} \geq \frac{\varepsilon^p}{4}\right\} =$$

$$= e^{-dna_{nn}^r} + \mathsf{P}\left(\int\limits_{\vec{x} \notin S(\vec{O}\,;\,R)} f_n\, d\vec{x} \geq \frac{\varepsilon^p}{4}\right) = e^{-dna_{nn}^r} + \mathsf{P}\left\{1 - \int\limits_{\vec{x} \in S(\vec{O}\,;\,R)} f_n\, d\vec{x} \geq \frac{\varepsilon^p}{4}\right\} =$$

$$= e^{-dna_{nn}^r} + \mathsf{P}\left\{1 - \int\limits_{x \in S(\vec{O}\,;\,R)} (f_n - \mathsf{E} f_n)\, d\vec{x} - \int\limits_{\vec{x} \in S(\vec{O}\,;\,R)} \mathsf{E} f_n\, d\vec{x} \geq \frac{\varepsilon^p}{4}\right\} \leq$$

$$\leq e^{-dna_n^r} + \mathsf{P}\left\{\int\limits_{x \in S(\vec{O}\,;\,R)} |f_n - \mathsf{E} f_n|\, d\vec{x} \geq \frac{\varepsilon^p}{4}\right\} \leq e^{-dna_{nn}^r}.$$

Hence we have proved that

$$\mathsf{P}\left\{\left[\int\limits_{\mathsf{E}^r} |f_n - \mathsf{E} f_n|^p\, d\vec{x}\right]^{1/p} \geq \varepsilon\right\} \leq e^{-dna_{nn}^r}. \tag{12}$$

Making use of Lemma 1, it is easy to verify that

$$\int_{E^r} |f - \mathsf{E}f_n|^p \, d\vec{x} \to 0 \tag{13}$$

(12) and (13) clearly imply (2).

Proof of (3). (3) follows immediately from Lemmas 1 and 8. Making use of Lemma 6 statement (3) can be proved without the condition

$$\lim_{\|\vec{x}\| \to \infty} \|\vec{x}\|^{r+2} h(\vec{x}) = 0 .$$

Remark 2. Let us recall a result of [3] (Theorem 8) showing that (3) and (3*) are the best possible in some sense (at least in the one dimensional case).

Theorem ([3]). *If there exists an interval* [A, B] *such that* $0 < \alpha \leq f''(x) \leq \beta < +\infty$ (or $-\infty < -\beta \leq f''(x) \leq -\alpha < 0$) *when* $x \in [A, B]$ *and if* $h(x)$ *and* $a_{nk} = a_n$ *is satisfying conditions* b) *and* c) *of Theorem 1 respectively then there exist constants* $D > 0$ *and* $C > 0$ *such that*

$$\mathsf{P}\left\{\sup_x |f_n^{(r)}(x) - f(x)| \geq \frac{D}{\sqrt{n a_{nn}}} + C a_n^2\right\} \geq 0,1 .$$

Remark 3. Suppose that $a_{nn} = \frac{1}{n^\alpha}$ $(0 < \alpha < 1/r)$. In this case (1) and (2) suggest that we get the best rate of the convergence if α is a very small number, at the same time (3) and (3*) says that the best possible choosing is $\alpha = \frac{1}{4+r}$. This "contradiction" can be explained as follows: if we want to "minimize" the difference $|f_n(\vec{x}) - \mathsf{E}f_n(\vec{x})|$ then we have to choose α as a small number but if we want to minimize $|f_n - f|$ then the best choosing is $\alpha = \frac{1}{4+r}$.

In (1) and (2) the influence of the difference between $f_n - \mathsf{E}f_n$ and $f_n - f = (f_n - \mathsf{E}f_n) + (\mathsf{E}f_n - f)$ cannot be recognized because

$$\mathsf{P}\{\sup_{\vec{x}} |\mathsf{E}f_n(\vec{x}) - f(\vec{x})| \geq \varepsilon\} = 0$$

if n is big enough.

2. Pattern classification

In [2] the following model is treated:

"Let $X_1^1, X_2^1, \ldots$ $[X_1^2, X_2^2, \ldots]$ be a sequence of independent identically distributed random vectors in E^m having a common probability density $f_1(x)[f_2(x)]$. We then think of $\{X_i^1\}$ as a sequence of samples from Class 1 and

$\{X_i^2\}$ as a sequence of samples from Class 2. For each i either we independently observe X_i^1 with probability q_1, or we observe X_i^2 with probability $q_2 = 1 - q_1$. If we define $\varrho_i = 1$ when X_i^1 is observed and $\varrho_i = 0$ when X_i^2 is observed, then $\varrho_1, \varrho_2, \ldots$ is a sequence of independent identically distributed random variables such that $\mathsf{P}(\varrho_i = 1) = q_1$ and $\mathsf{P}(\varrho_i = 0) = q_2$. We will assume that X_i^1, X_j^2, ϱ_k are independent for all i, j, k. If we define $X_i = \varrho_i X_i^1 + (1 - \varrho_i) X_i^2$ then $\{X_i\}$ is the sequence of observed random vectors."

The problem is to find a decision procedure for classifying X_i $i > n$ that is based on the sequence $(X_1, \varrho_1), \ldots, (X_n, \varrho_n)$. When f_1, f_2, q_1, q_2 are known, the following procedure minimizes the probability of misclassification.

Let

$$\mathfrak{D}_0(x) = q_1 f_1(x) - q_2 f_2(x)$$

and

$$\mathfrak{D}_0 = \{x; \mathfrak{D}_0(x) \geq 0\}.$$

Then decide: X_i is from Class 1 if $X_i \in \mathfrak{D}_0$ and X_i is from Class 2 otherwise. In this case the probability of misclassification is

$$\mathsf{P}(\mathfrak{D}) = q_1 \int_{\mathfrak{D}_0} f_1(x)\, dx + q_2 \int_{\mathfrak{D}_0} f_2(x)\, dx\,.$$

In the case, when q_1, f_1, f_2 are unknown (instead of $\mathfrak{D}_0$) one can use the decision function[2]

$$\mathfrak{D}^{(n)}(x) = \frac{\mu_n}{n} \frac{1}{\mu_n} \sum_{k=1}^{n} \frac{\varrho_k}{a_{\mu_k}} h\left(\frac{x - X_k}{a_{\mu_k}}\right) - \frac{n - \mu_n}{n} \frac{1}{n - \mu_n} \sum_{k=1}^{n} \frac{1 - \varrho_k}{a_{k-\mu_k}} h\left(\frac{x - X_k}{a_{k-\mu_k}}\right)$$

where

$$\mu_k = \varrho_1 + \varrho_2 + \ldots + \varrho_k \qquad (k = 1, 2, \ldots)$$

$1 \geq a_1 > a_2 > \ldots$ is a monotonically decreasing sequence of postive numbers, $h(x)$ is a density function and

$$f^{(1)}_{\mu_n} = \frac{1}{\mu_n} \sum_{k=1}^{n} \frac{\varrho_k}{a_{\mu_k}} h\left(\frac{x - X}{a_{k-\mu_k}}\right)$$

resp.

$$f^{(2)}_{n-\mu_n} = \frac{1}{n - \mu_n} \sum_{k=1}^{n} \frac{1 - \varrho_k}{a_{k-\mu_k}} h\left(\frac{x - X_k}{a_{k-\mu_k}}\right)$$

[2] This definition is a little bit different from the definition used in [2], but the difference is minor.

is the estimation of f^1 resp. f^2 (based on the sample elements $X_1, X_2, \ldots, X_n$ belonging to Class 1, resp. Class 2) studied in Chapter 1, finally

$$q_1^{(n)} = \frac{\mu_n}{n}, \quad q_2^{(n)} = \frac{n - \mu_n}{n}$$

are the natural estimations of q_1 and q_2.

Clearly in this case we decide that X_i is from class 1 if $X_i \in \mathfrak{D}_n$ where

$$\mathfrak{D}_n = \{x : \mathfrak{D}^{(n)}(x) \geq 0\}$$

and the probability of misclassification is

$$\mathsf{P}(\mathfrak{D}_n) = q_1 \int_{\mathfrak{D}_n} f_1 + q_2 \int_{\mathfrak{D}_n} f_2 .$$

($P(\mathfrak{D}_n)$ is a random variable).

Theorem 2. Let us suppose that f_1, f_2, h and $a_k = a_{kn}$ are obeying the conditions of Theorem 1. Then

$$\mathsf{P}\{\sup_{x \in T - T_\varrho} |\mathfrak{D}^{(n)}(x) - \mathfrak{D}_0(x)| \geq \varepsilon\} \leq e^{-cna_n^m} \tag{14}$$

$$\mathsf{P}\{\int_{R^m} |\mathfrak{D}^{(n)}(x) - \mathfrak{D}_0(x)|^p \, dx \geq \varepsilon\} \leq e^{-c_1 na_n^m} \tag{15}$$

$$\mathsf{P}\left\{\frac{\min\left(\sqrt{na_n^m}, \dfrac{1}{a_n^2}\right)}{\log n} \sup_{T - T_s} |\mathfrak{D}^n(x) - \mathfrak{D}_0(x)| \to 0\right\} = 1 \tag{16}$$

and especially if $a_k = k^{-\frac{1}{4+m}}$ then

$$\mathsf{P}\left\{\frac{n^{\frac{2}{4+m}}}{\log x} \sup_{x \in T - T_s} |\mathfrak{D}^{(n)}(x) - \mathfrak{D}_0(x)| \to 0\right\} = 1 . \tag{16*}$$

Proof. This Theorem is a simple consequence of Theorem 1 and the obvious, inequality:

$$|\mathfrak{D}^{(n)}(x) - \mathfrak{D}_0(x)| \leq |q_1^{(n)} - q_1| f_{\mu_n}^{(1)} + q_1 |f_{\mu_n}^{(1)} - f_1| + \\ + |q_2^{(n)} - q_2| f_{n-\mu_n}^{(2)} + q_2 |f_{n-\mu_n}^{(2)} - f_2| .$$

As an example we give the proof of (14) in a more detailed form:

Clearly

$$\mathsf{P}\{\sup_{x \in T - T_s} |\mathfrak{D}^{(n)}(x) = \mathfrak{D}_0(x)| \geq \varepsilon\} \leq$$

$$\leq \mathsf{P}\{|q_1^{(n)} - q_1| \sup f_{\mu_n}^{(1)} \geq \varepsilon/4\} + \mathsf{P}\{q_1 \sup |f_{\mu_n}^{(1)} - f_1| \geq \varepsilon/4\} + \tag{17}$$

$$+ \mathsf{P}\{|q_2^{(n)} - q_2| \sup f_{n-\mu_n}^{(2)} \geq \varepsilon/4\} + \mathsf{P}\{q_2 \sup |f_{n-\mu_n}^{(2)} - f_2| \geq \varepsilon/4\}$$

and

$$\mathsf{P}\{|q_1^{(n)} - q_1| \sup_{x \in T - T_\varrho} f_{\mu_n}^{(1)} \geq \varepsilon/4\} \leq \mathsf{P}\{|q_1^{(n)} - q|\,(k+\varepsilon) \geq \varepsilon/4\} +$$

$$+ \mathsf{P}\{\sup_{T-T_\varrho} f_{\mu_n}^{(1)} > k + \varepsilon\} \leq e^{-cn} +$$

$$+ \mathsf{P}\{\sup f_{\mu_n}^{(1)} > k + \varepsilon,\ nq_1 - n^{3/4} \leq \mu_n \leq nq_1 + n^{3/4}\} +$$

$$+ \mathsf{P}\{\mu_n \notin (nq_1 - n^{3/4},\ nq_1 + n^{3/4})\} \leq e^{-cn} +$$

$$+ \sum_{k=nq_1-n^{3/4}}^{nq_1+n^{3/4}} \mathsf{P}\{\sup_{T-T_\varrho} f_k^{(1)} > k + \varepsilon \,|\, \mu_n = k\}\ \mathsf{P}\{\mu_n = k\} \leq e^{-cna_n^m}$$

further

$$\mathsf{P}\{q_1 \sup_{T-T_\varrho} |f_{\mu_n}^{(1)} - f_1| \geq \varepsilon/4\} \leq$$

$$\leq \sum_{k=nq_1-n^{3/4}}^{nq_1+n^{4/4}} \mathsf{P}\{\sup_{T-T_\varrho} |f_k^{(1)} - f_1| \geq \varphi/4 \,|\, \mu_n = k\}\ \mathsf{P}(\mu_n = k) +$$

$$+ \mathsf{P}\{\mu_n \notin (nq_1 - n^{3/4};\ nq_1 + n^{3/4}\} \leq e^{-cna_n^m}$$

The third and fourth members of (17) can be estimated in a quite similar way, which proves (14).

In the following theorem we investigate the difference between $\mathsf{P}(\mathfrak{D}_n)$ and $\mathsf{P}(\mathfrak{D})$. In fact we prove our

Theorem 3. Under the conditions of Theorem 1

$$\mathsf{P}(|\,\mathsf{P}(\mathfrak{D}_n) - \mathsf{P}(\mathfrak{D})\,| \geq \varepsilon) \leq e^{-\gamma n a_n^m}$$

This theorem immediately follows from Theorem 2. and the following result of L. Győrfi [4]

Lemma 5. We have

$$|\mathsf{P}(\mathfrak{D}_n) - \mathsf{P}(\mathfrak{D})| \leq \sqrt{\int_{E^m} [\mathfrak{D}_n(x) - \mathfrak{D}(x)]^2\, dx}$$

with probability 1.

In fact this Theorem can be generalized in the following way: Suppose that the sample $X_1, X_2, \ldots$, is coming from k classes having density functions $f_1, f_2, \ldots, f_k$ with probabilities $p_1, p_2, \ldots, p_k$.

If $f_1, f_2, \ldots, f_k$ and $p_1, p_2, \ldots, p_k$ are known then we decide: X_i is from class j $(j = 1, 2, \ldots, k)$ if

$$p_j f_j(X_i) \geq q_l f_l(X_i) \qquad (l = 1, 2, \ldots, j-1, j+1, \ldots, k)\ .$$

In the case when $p_1, p_2 \ldots, p_k; f_1, f_2, \ldots, f_k$ are unknown. we decide X is from class j if

$$p_j^{(n)} f_j^{(n)}(X_i) \geq p_e^{(n)} f_e^{(n)}(X_i) \qquad (l = 1, 2, \ldots, j-1, j+1, \ldots, n) .$$

Let $\mathsf{P}(\mathfrak{D}_n)$ and $\mathsf{P}(\mathfrak{D})$ denote the probability of misclassification used this decision procedure and the decision procedure described above respectively.

In this case it is easy to prove that

Theorem 3. Under the conditions of Theorem 1*

$$\mathsf{P}\{|\mathsf{P}(\mathfrak{D}_n) - \mathsf{P}(\mathfrak{D})| \geq \varepsilon\} \leq e^{-na_n^m} .$$

References

1. Rosenblatt, M.: *Remarks on some non-parametric estimates of density function, Ann. Math. Statist.*, **27** (1956), 832—837.
2. Wolverton, C. T. and Wagner, T. I.: Asymptotically optimal discriminant functions for pattern classification. *IEEE Transcation on Information Theory*, IT-**15** (1969), 258—266.
3. Révész, P.: On empirical density functions. *Periodica Mathematica, Hung.*, **2** (1972).
4. Győrfi, L.: Estimation of optimal decision functions. *Technical report.*
5. Parzen, E.: On estimation of a probability density function and mode. *Ann. Math. Statist*, **33** (1962), 832—837.

Оценка плотности вероятностей и классификация образов

Л. РЕЙТО, П. РЕВЕС

(Будапешт)

В работе дается обобщение определения эмпирической плотности вероятностей и на основе данного обобщения рассматривается проблема классификации образов. Результаты представлены в теоремах 1, 2, 3.

Lidia REJTŐ
P. RÉVÉSZ
Mathematical Research Institute
Hungarian Academy of Sciences
Budapest, V. Reáltanoda u. 13—15.

Problems of Control and Information Theory, Vol. 2 (1), pp. 81—93 (1973)

ON A PATTERN CLASSIFICATION ALGORITHM OF R. L. KASHYAP

J. FRITZ

(Budapest)

(Received May 22, 1972)

The convergence properties of a stochastic approximation type pattern recognition learning algorithm are investigated: some results of Aizerman—Braverman—Rozonoer and of R. L. Kashyap are developed further. Necessary and sufficient conditions for the convergence of the algorithm are given, and it is shown that under certain natural conditions the algorithm does not converge. An estimation for the probability of erroneous classification is also proved.

1. Introduction

Aizerman, Braverman and Rozonoer [2] have proposed a supervised learning procedure for the classification of patterns, which are coming from two nonoverlapping sets, and they have proved the convergence of the algorithm. The assumptions of this theorem are very restrictive, although the algorithm stops in a finite number of steps. R. L. Kashyap [1] has developed further this method for a more difficult case, when the pattern classes cannot be separated by a sign of any of the considered separating functions, and moreover, the teacher need not be quite reliable. Unfortunately, the proof presented by Kashyap is not complete, and some other problems remained also unsolved: neither the convergence properties of the algorithm, nor the asymptotic behaviour of the loss function are investigated in [1]. The aim of this report is to generalize the algorithm to the case of overlapping classes (alternative hypothesis testing), and to investigate the conditions of the convergence of the algorithm, and the asymptotic behaviour of some loss functions. I have to remark, however, that the algorithm works well only if the overlapping part of the classes is of small probability, as we shall see later. Besides the methods of the above-mentioned publications, a simple but useful observation of L. Győrfi [4] will be used in a little bit modified form.

2. Formulation of the Problem

Let $d(x)$, $-1 \leq d(x) \leq +1$ be a real valued measurable function on the probability space (S, Σ, Q), and suppose that we have to learn the sign of $d(x)$ in order to solve an alternative hypothesis testing problem: if the estimated

value of $d(x)$ is positive or negative, then we adopt hypothesis H_+ and H_- respectively. Of course, it may happen that we estimate $d(x)$ to be 0; in this case no decision can be made. The learning process itself is an iterative algorithm defined in terms of a double sequence $(X_1, Z_1), (X_2, Z_2), \ldots (X_n, Z_n), \ldots$ of random variables on an overall probability space (Ω, A, P), where X_n is a randomly selected point from S, while Z_n contains the information given by the teacher concerning the value of $d(X_n)$. The aim of the learning algorithm is to build up a function $f(x)$ on S from a given realization of this double sequence such that sign $f(x)$ should be a good estimation for sign $d(x)$. The sequence (X_n, Z_n) is specified as follows:

(i) Every X_n is measurable with respect to Σ and has distribution Q on (S, Σ).

(ii) The real sequence $Z_1, Z_2, \ldots, Z_n, \ldots$ is uniformly bounded by a constant $O < M < +\infty$, further $E(Z_n \mid X_n) = \mathrm{d}(X_n)$ for each nonnegative integer n.

(iii) The sequence of pairs $(X_1, Z_1), (X_2, Z_2), \ldots, (X_n, Z_n), \ldots$ is totally independent.

From the point of view of pattern recognition the following special cases are interesting:

A) H_+ and H_- are disjoint measurable subsets of S with $1 = Q(H_+) + Q(H_-)$; $d(x) = +1$ or -1 according as $x \in H_+$ or $x \in H_-$, and $Z_n = d(X_n)$. This is the simplest pattern classification problem, essentially solved by Aizerman, Braverman and Rozonoer [2].

B) The situation is the same as above, but Z_n is defined as $Z_n = \frac{1}{\lambda} W_n d(X_n)$, where W_n is uniformly bounded and independent of $X_1, X_2, \ldots, X_{n-1}$; further $E(W_n \mid X_n) = \lambda$ is a nonzero constant. An easy computation shows that then (ii) is satisfied. Especially, if W_n takes on two values only, then this means that the information $d(X_n)$ is coming from the teacher through a binary symmetric channel of nonzero capacity. If $P(W_n = 1) = 1$ then B) reduces to A). This problem has been raised by Kashyap [1], where another interpretation is given.

C) H_+ and H_- are alternative hypotheses with given a priori probabilities, and $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), \ldots$ is an independent and identically distributed sequence of pairs, where $Y_n = \pm 1$ according as H_+ or H_- has obtained at the n-th trial. If we can observe only the value of X_n, then the best decision (maximum a priori likelihood decision) is given by the sign of the standard decision function $d(x) = P(H_+ \mid x) - P(H_- \mid x)$, where $P(H_+ \mid x) = 1 - P(H_- \mid x) = P(Y_n = 1 \mid X_n = x)$ is independent of n.

Putting $Z_n = Y_n$, we see that (ii) is satisfied. This case reduces to A) if Y_n is a function of X_n.

D) The basic situation is the same as in C), but Z_n is defined as $Z_n = \frac{1}{\lambda} W_n Y_n$, where W_n is uniformly bounded and independent of $X_1, X_2, \ldots, X_{n-1}, Y_1, Y_2, \ldots, Y_{n-1}, W_1, W_2, \ldots, W_{n-1}$, further $E(W_n \mid X_n, Y_n) = \lambda$ is a nonzero constant. Then $E(W_n Y_n \mid X_n) = E(E(W_n Y_n \mid X_n, Y_n) \mid X_n) = E(Y_n E(W_n \mid X_n, Y_n) \mid X_n)) = \lambda d(X_n)$, thus (ii) is valid in this case, too. D) is a common generalization of the above situations.

Suppose that for the estimation of the sign of $d(x)$ we can use only functions of type

$$f(x, c) = t + \Sigma c^i \varphi_i(x), \qquad c = (c^1, c^2, \ldots),$$

where t is a fixed nonzero constant, $K(x) = \varphi_1(x), \varphi_2(x), \ldots)$ is a finite or infinite sequence of measurable functions on S, and c is the corresponding vector of coefficients. Let R denote the real linear normed space of the square summable ones among these coefficient sequences; that is $R = l_2$ if the sequence $\varphi_1, \varphi_2, \ldots, \varphi_i, \ldots$ of regression functions is infinite, and R is the r-dimensional Euclidean space if $K(x) = (\varphi_1(x), \ldots, \varphi_r(x))$. The norm and scalar product in R will be denoted as $|| c ||^2 = \Sigma \mid c_2 \mid^2$ and $(a, b) = \Sigma a^i b^i$. We see that $f(x, c) = (K(x), c) + t$, provided that the scalar product exists. If $c \in R$ and the following assumption holds, then this certainly will be true, at least Q — a.s.

(iv) $$\int || K(x) ||^2 Q(dx) = \Sigma \int \mid \varphi_i(x) \mid_2 Q(dx) < + \infty.$$

The importance of this assumption is demonstrated by the following lemma.

Lemma 1. Under (iv) we have $f(x, c) \in L_2(Q)$ if $c \in R$; further the weak convergence of a sequence c_n to c in R implies that $\lim f(x, c_n) = f(x, c)$ Q-a.s. and in the $L_2(Q)$ norm.

Proof. Since $(u + v)^2 \leq 2u^2 + 2v^2$, the inequality of Cauchy implies

(1) $$\mid f(x, c) \mid^2 \leq 2t^2 + 2(K(x), c)^2 \leq 2t^2 + 2 || c ||^2 || K(x) ||^2.$$

In view of (iv) this means that $f(x, c)$ is defined and finite Q-a.s., and $f(x, c) \varepsilon L_2(Q)$. Similarly,

(2) $$\mid f(x, c_n) - f(x, c) \mid^2 \leq || c_n - c ||^2 || K(x) ||^2,$$

which implies the second statement by the Dominated Convergence Theorem since the weak convergence of c_n to c means that $|| c_n ||^2$ is bounded, and $\lim (K(x), c_n) = (K(x), c)$ Q-a.s. <>

The main goal of this paper is to investigate a stochastic approximation algorithm that minimizes the average loss function

$$J(c) = \int f(x, c)\,(\operatorname{sign} f(x, c) - d(x))\,Q(dx) \tag{3}$$

on R, and we shall examine the results from the point of view of problems A)—D). Having a more or less good solution $c \in R$ of this problem, we can use the decision rule: adopt H_+ or H_- if $f(x, c)$ is positive or negative, but make no decision if $f(x, c) = 0$. The error probability of this decision is given by

$$P_e(c) = P_d + \frac{1}{2}\int d(x)\,(\operatorname{sign} d(x) - \operatorname{sign} f(x, c))\,Q(dx), \tag{4}$$

where P_d is the error probability of the standard decision, associated with $d(x)$; here the cases $f(x, c) = 0$ and $d(x) = 0$ are not considered as errors. Of course, the most desirable result would be a procedure which minimizes the second term on the right hand side of (4), but no such algorithm is known. Moreover, the fact that $J(c)$ is close to its minimum value, does not mean that $P_e(c)$ is close to its minimum: in the region where $| f(x, c) |$ is small, the decision rule sign $f(x, c)$ becomes unreliable. These questions will be discussed later.

Sometimes we shall assume that

$$Q(f(x, c) = 0) = 0 \quad \text{if} \quad c \in R, \tag{v}$$

which is a very strong form of the linear independence of the functions $1, \varphi_1, \varphi_2, \ldots, \varphi_i, \ldots$, and rules out the possibility that the evaluation of $f(x, c)$ does not lead to decision.

Observe that $J(c)$ is never negative, since $u(\operatorname{sign} u - d) = | u | - ud \geq$ ≥ 0 if $| d | \leq 1$. Two further important properties of J are formulated in Lemma 2.

Lemma 2. If (iv) holds, then $J(c)$ is a U-convex and weakly continuous functional on R.

Proof. Let $\lambda, \mu \geq 0$, $\lambda + \mu = 1$. Since $f(x, \mu a + \lambda b) = \mu f(x, a) + \lambda f(x\, b)$, the convexity of J follows by integration from the inequality

$$(\mu u + \lambda v)\,(\operatorname{sign}(\mu u + \lambda v) - d) = | \mu u + \lambda v | - d(\mu u + \lambda v) \leq \mu | u | +$$

$$+ \lambda | v | - d\,(\mu u + \lambda v) = \mu u(\operatorname{sign} u - d) + \lambda v(\operatorname{sign} v - d),$$

which expresses only the U-convesity of the function $| u |$.

On the other hand, if c_n converges weakly to c in R, then we know from Lemma 1 that

$$f(x, c_n)\,(\operatorname{sign} f(x, c_n) - d(x)) = |\, f(x, c_n)\,| - f(x, c_n)\, d(x)$$

converges Q-a.s. to

$$f(x, c)\,(\operatorname{sign} f(x, c) - d(x)) = |\, f(x, c)\,| - f(x, c)\, d(x),$$

and (1) shows that this is a dominated convergence. <>

Unfortunately, Lemma 2, does not mean that the functional J takes on its minimum value in R, though the closed spheres of R are weakly compact. For example, if (S, Σ, Q) is the interval $\left(-\frac{1}{2}, \frac{1}{2}\right)$ with the Lebesque measure on its Borel subsets, $d(x) = \operatorname{sign} x$, $t = 1$, and we have only one interpolation function $\varphi(x) = x$, then the considered functions are the linear functions $f(x, c) = 1 + cx$. We see, that $J(c) \geq \frac{1}{2}$ if $c \leq 2$, and $J(c) = \frac{1}{2c}$ if $c \geq 2$; that is $J(c) > 0 = \inf J(c)$ for each c. Similarily trivial examples show that the minimum of $J(c)$ and $P_e(c)$ does not coincide.

Therefore, we have to assume that $J(c)$ takes on its minimum:

(vi) There exists at least one $a \in R$ for which we have $J(a) = J_0 = \inf_{c \in R} J(c)$.

The set of these a-s will be denoted by R_0, it is a convex and weakly closed subset of R, as follows from Lemma 2.

In connection with problem B), R. L. Kashyap [1] has proposed the following stochastic gradient procedure to find the minimum place of $J(c)$:

$$C_{n+1} = C_n + \varrho_n K(X_n)\,(Z_n - \operatorname{sign} f(X_n, C_n))\,, \tag{5}$$

where C_1 is an arbitrarily chosen point of R, and the sequence $\varrho_1, \varrho_2, \ldots, \varrho_n, \ldots$ of given positive numbers satisfies the conditions

$$\sum_{n=1}^{\infty} \varrho_n = +\infty\,, \quad \sum_{n=1}^{\infty} \varrho_n^2 < +\infty\,. \tag{6}$$

$C_1, C_2, \ldots, C_n, \ldots$ is a sequence of abstract random variables taking values in R with probability one if (iv) is satisfied. Let $\mathfrak{A}_n$ denote the minimal σ-algebra on which $X_1, X_2, \ldots, X_{n-1}$ and $Z_1, Z_2, \ldots, Z_{n-1}$ are measurable; $\mathfrak{A}_1$ is the trivial σ-algebra. Then $\mathfrak{A}_1, \mathfrak{A}_2, \ldots, \mathfrak{A}_n, \ldots$ is an increasing sequence of σ-algebras, C_n is measurable on A_n, and (iii) means that (X_n, Z_n) is independent

of $\mathfrak{A}_n$. An easy computation shows that the conditional expectation of the correcting term

$$T_n = K(X_n)\left(Z_n - \operatorname{sign} f(X_n, C_n)\right), \tag{7}$$

given the σ-algebra $\mathfrak{A}_n$ (or C_n), really equals the negative gradient of $J(c)$ at C_n.

In the algorithm of Aizerman—Braverman—Rozonoer [2], ϱ_n is a small constant, and $Z_n = d(X_n) = \operatorname{sign} d(X_n)$, further, they suppose that there exists an $a \in R$ such that $\operatorname{sign} f(x, a) = d(x)$, and $0 < \varepsilon = \inf_x f(x, a)$ Q-a.s. In this case they prove that $d(x) = \operatorname{sign} f(x, C_n)$ Q-a.s. after a finite number of steps with probability one. The main assumption of Kashyap [1] is that R_0 is not empty, but his proof implies only that

$$\lim E \mid \operatorname{sign} f(x, C_n) - \operatorname{sign} f(x,) \mid = 0 \quad \text{if} \quad f(x, a) \neq 0, \quad a \in R_0.$$

Of course, this means only that

$$\lim \int \mid \operatorname{sign} f(x, C_n) - \operatorname{sign} f(x, a) \mid Q(dx) = 0$$

with positive probability.

3. The Convergence Properties of the Algortihm

Throughout we assume the validity of conditions (i)—(iv), and we shall use the above notations without any reference. First we investigate the role of (vi) and prove a negative result.

Theorem 1. If assumption (vi) does not hold, then algorithms (5)—(6) cannot converge in R, since in this case we have $\sup \| C_n \| = +\infty$ with probability one.

Proof. Let a denote an arbitrary element of R, and set $B_n = C_n - a$. From (5) we have

$$\| B_{n+1} \|^2 = \| B_n \|^2 + \varrho_n^2 \| T_n \|^2 + 2\varrho_n (T_n, B), \tag{8}$$

$$E_n \| B_{n+1} \|^2 = \| B_n \|^2 + \varrho_n^2 E_n \| T_n \|^2 + 2\varrho_n (E_n T_n, B_n), \tag{9}$$

where E_n denotes conditional expectation given by the σalgebra $\mathfrak{A}_n$. Observe, that

$$E_n T_n = K(x)\,(d(x) - \operatorname{sign} f(x, C))\, Q(dx), \tag{10}$$

since $E(Z_n \mid X_n) = d(X_n)$, and (X_n, Z_n) is independent of $\mathfrak{A}_n$, further

$$E_n \| T_n \|^2 \leq (M + 1)^2 \int \| K(x) \|^2 Q(dx) = M_1 < +\infty. \tag{11}$$

From (8) and (9) we obtain that

$$\sum_{=1}^{m} \| B_{n+1} \|^2 - E_n \| B_{n+1} \|^2 = \sum_{n=1}^{m} \varrho_n^2 (\| T_n \|^2 - E_n \| T_n \|^2) + \tag{12}$$

$$+ 2 \sum_{n=1}^{m} (T_n - E_n T_n, B_n) \varrho_n$$

where the first sum on the right-hand side certainly converges with probability one, as $\Sigma \varrho_n^2 < + \infty$. Observe that $(T_n - E_n T_n, B_n)$ is a centered sequence (adapted to $\mathfrak{A}_{n+1}$) as T_n is measurable on $\mathfrak{A}_{n+1}$, while B_n is $\mathfrak{A}_n$-measurable; further

$$E_n (T_n - E_n T_n, B_n)^2 \leq || B_n ||^2 E_n || T_n - E_n T_n ||^2 \leq 2 M_1 || B_n ||^2,$$

therefore Proposition IV. 6.2. of J. Neveue [5] yields that

$$\sum_{n=1}^{m} \varrho_n (T_n - E_n T_n, B_n)$$

converges with probability one to a finite limit on the set

$$\Omega_0 = \left(\sum_{n=1}^{\infty} \varrho_n^2 \| C_n \|^2 < + \infty \right).$$

as $|| B_n ||^2 \leq 2 || C_n ||^2 + 2 || a ||^2$. This means, however, that the left-hand side of (12) also converges to a finite limit with probability one on Ω_0. Since the sum of the equalities (9) is

$$- 2 \sum_{n=1}^{m} \varrho_n (E_n T_n, B_n) = \sum_{n=1}^{m} \| B_{n+1} \|^2 - E_n \| B_{n+1} \|^2 + \tag{13}$$

$$+ \| B_1 \|^2 - \| B_{n+1} \|^2 + \sum_{n=1}^{m} \varrho_n^2 \| T_n \|^2,$$

(11) implies that

$$\sup - \sum_{n=1}^{m} \varrho_n (E_n T_n, B_n) < + \infty \qquad \text{P-a. s. on } \Omega_0. \tag{14}$$

Taking into account (10), and the elementary inequality

$$(u - v) (\operatorname{sign} u - d) \geq u (\operatorname{sign} u - d) - v (\operatorname{sign} v - d),$$

we can conclude that

$$\sup \sum_{n=1}^{m} \varrho_n \big(J(C_n) - J(a) \big) < + \infty \qquad \text{P-a. s. on } \Omega_0. \tag{15}$$

Since $\Sigma \varrho_n = + \infty$, this is possible only if

$$\lim J(C_n) = J_0 \ P\text{-a.s. on } \Omega_0; \tag{16}$$

suppose now that $\sup \| C_n \| < + \infty$ on Ω_1. Obviously, Ω_1 is a subset of Ω_0. If $P(\Omega_1) > 0$, then we can select such an elementary event ω from Ω_1 that $\sup || c_n || < + \infty$, and $\underline{\lim} J(c_n) = J_0$, where $c_n = C_n(\omega)$. However, the closed spheres in R are weakly compact, therefore at least one subsequence $a_1, a_2, \ldots, a_k, \ldots$ of $c_1, c_2, \ldots c_n, \ldots$ converges weakly to an $a \in R$, for which we have from Lemma 2, that $\lim J(a_n) = J(a) = J_0$; that is (vi) holds. This proves the statement of the theorem. <>

The following theorem is mainly based on the achievements of this proof.

Theorem 2. Suppose that (vi) holds besides (i)—(iv). Then $\lim J(C_n) = J_0$ with probability one, and any weak limit point of the sequence C_n belongs to R_0. Moreover, if R is a finite dimensional space, then C_n converges with probability one.

Proof. Let a denote an element of R_0; that is $J(a) = J_0$. We use the notations of the preceeding proof. Observing that in this case

$$- (E_n T_n, B_n) \geq J(C_n) - J(a) \geq 0,$$

from (9) and (11) we have

$$E \| B_{n+1} \|^2 \leq E \| B_n \|^2 + M_1 \varrho_n^2 \leq \| B_1 \|^2 + M_1 \sum_{k=1}^{n} \varrho_k^2 ,$$

that is the sequence $E || B_{n+1} ||^2$ is bounded, and from the Fatou—Lebesque Theorem we obtain that $\Sigma \varrho_n^2 || B_n ||^2 < + \infty$ with probability one, that is $P(\Omega_0) = 1$ in the previous proof. Therefore, (13) implies that

$$\lim || B_n || \text{ exists and finite } P\text{-a.s.}, \tag{17}$$

further (15) becomes

$$\sum_{n=1}^{\infty} \varrho_n \big(J(C_n) - J(a)\big) < + \infty \qquad \text{P-a. s.}, \tag{18}$$

Moreover, the expectation of (13) yields

$$\sum_{n=1}^{\infty} \varrho_n E\big(J(C_n) - J(a)\big) < + \infty . \tag{19}$$

This implies the first statement by the following lemma, which is very similar to several lemmas of Aizerman—Braverman—Rozonoer [2].

Lemma 3. Let $U(c) \geq 0$ be strongly uniformly continuous on the bounded subsets of R. If $\sum_{1=n} \varrho_n E(U(C_n)) < + \infty$, then the assumptions of Theorem 1, imply that $\lim U(C_n) = 0$ with probability one.

Proof. Set $K_n = (M+1)\,||K(X_n)||$ and $I_n = 1$ if $U(C_n) \geq \varepsilon$, $I_n = 0$ otherwise, where $\varepsilon > 0$ is arbitrary. First we prove, that

$$\lim \varrho_n K_n = 0 \text{ and } \sum_{n=1}^{\infty} \varrho_n I_n K_n < +\infty \qquad \text{P–a. s.} \tag{20}$$

The first statement is obvious, since $E(\Sigma\, \varrho_n^2 K_n^2) < +\infty$. To prove the second one, we observe that I_n and K_n are independent, thus we have $E(I_n K_n) = E(I_n)\,E(K_n)$, and $E(K_n)$ is bounded. Further, as by definition, $\varepsilon I_n \leq U(C_n)$, we have $\Sigma\, \varrho_n E(I_n) < +\infty$, whence

$$E\left(\sum_{n=1}^{\infty} \varrho_n I_n K_n\right) = \sum_{n=1}^{\infty} \varrho_n E(I_n K_n) < +\infty\,,$$

which proves (20):

From the proof of Theorem 2, we know that the sequence C_n is bounded with probability one, since $||B_n||$ converges. Further, (5) implies

$$||C_m - C_n|| \leq \sum_{k=n}^{m} \varrho_k K_k\,, \tag{21}$$

therefore for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$U(C_m) < U(C_n) + \varepsilon \text{ if } \sum_{k=n}^{m-1} \varrho_k K_k < \delta\,. \tag{22}$$

On the other hand, $\Sigma\, \varrho_n U(C_n) < +\infty$ is possible only if $\underline{\lim}\, U(C_n) = 0$ P-a.s. Thus, if m is large enough, we have an N (which is a random variable, too), such that $m > N$ and

$$U(C_N) < \varepsilon\,,\ \varrho_n K_n < \frac{\delta}{2} \text{ if } n \geq N\,, \text{ and } \sum_{k=N}^{\infty} \varrho_k I_k K_k > \frac{\delta}{2}\,. \tag{23}$$

Let now $n \leq m$ be the greatest integer for which $U(C_n) < \varepsilon$ still holds; then $I_n = 1$ if $n < k < m$, and $\varrho_n K_n < \dfrac{\delta}{2}$, thus from (23) and (22) we have $U(C_m) + \varepsilon < 2\varepsilon$, which proves the lemma. ◇

Now we are in the position to prove Theorem 2. Let $U(c) = J(c) - J(a)$ in the lemma; in view of Lemma 2, $J(c)$ is weakly continuous, that is $J(c) - J(a)$ will be strongly uniformly continuous on the bounded subsets of R; thus Lemma 3 applies and yields that $J(a) = J_0 = \lim J(C_n)$ P-a.s. This implies even the second statement immediately as $J(c)$ is weakly continuous, since then $\lim\limits_{k} c_{n_k} = c$ in the weak sense, and $\lim J(c_n) = J_0$ is possible only if $J(c) = J_0$.

Finally, observe that (17) holds in the following stronger form: There exists a subset Ω_2 of Ω such that $P(\Omega_2) = 1$, and

(24) $\lim \| C_n - a \|$ exists and finite on Ω_2 for each $a \in R_0$.

Really, let $a_1, a_2, \ldots, a_m, \ldots$ be a countable dense set in R_0 in the strong topology. Since

$$\| C_n - a_m \| - \| a_m - a \| \leq \| C_n - a \| \leq \| C_n - a_m \| + \| a_m - a \| ;$$

we can choose Ω_2 such that $\lim \| C_n - a_m \|$ should exist for each a_m, and (24) is proved. On the other hand, if R is a finite dimensional space, then in view of the Axiom of Choice we can define such a function $A(\omega)$ on Ω_2 which takes values in R, and for each $\omega \in \Omega_2$ we have $\lim \| C_n(\omega) - A(\omega) \| = 0$. Thus, the second statement of this theorem implies that $A(\omega) \in R_0$ P-a.s., (24) means therefore that $\lim \| C_n(\omega) - A(\omega) \| = 0$ on a subset of Ω_2 having probability one. Consequently, $A(\omega)$ is measurable and $\lim C_n = A$ P-a.s. $\langle\rangle$

If $R = l_2$ then the weak convergence does not imply the strong one, thus the above proof does not work in this case. Fortunately, the proposed decision functions sign $f(x, C_n)$ do converge, and the $L_1(Q)$ limit is the common value (equivalence class) of the functions sign $f(x, a)$, $a \in R_0$.

Theorem 3. Under assumptions (i)—(vi) we have for each interpolation function $\varphi_i(x)$ that

$$\int \varphi_i(x) \left(\operatorname{sign} f(x, a) - d(x)\right) Q(dx) = 0 \quad \text{if} \quad a \in R_0 ,$$

and $\operatorname{sign} f(x, a) = \operatorname{sign} f(x, b)$ Q-a.s. if $a, b \in R_0$. Further, if $a \in R_0$, then

(26) $$\lim \int | \operatorname{sign} f(x, C_n) - \operatorname{sign} f(x, a) | \, Q(dx) = 0$$

with probability one.

Proof. We know that $J(c)$ is the integral of a U-convex function which is differentiable Q-a.s. if (v) holds. Thus the Monotone Convergence Theorem implies that $J(c)$ is also a differentiable function of each component of the vector c, and an easy computation shows that its i-th partial derivative is just the integral on the left-hand side of (25). Since J takes on its minimum at a, (25) must be true. The second statement is a simple consequence of (25), but it follows immediately from (26), too.

To prove (26), observe that $J(C_n) \geq J(a)$ implies

$$- (E_n T_n, B_n) = \int \left(f(x, C_n^{\tau}) - f(x, a)\right) \left(\operatorname{sign} f(x, C_n) - d(x)\right) Q(dx) \geq$$

$$\geq \int f(x, a) \left(\operatorname{sign} f(x, a) - \operatorname{sign} f(x, C_n)\right) Q(dx) = V(C_n) \geq 0 ;$$

thus from (13) we have

$$\sum_{n=1}^{\infty} \varrho_n E\big(V(C_n)\big) < +\infty . \tag{27}$$

However, if (v) holds then $V(c)$ is weakly continuous on R, therefore Lemma 3 applies and yields that $\lim V(C_n) = 0$ with probability one, which means that $f(x, a)$ sign $f(x, C_n)$converges to $f(x, a)$ sign $f(x, a)$ in the $L_1(Q)$ norm with probability one: But, for any $\varepsilon > \sigma > 0$ we have

$$Q(\,|\,\operatorname{sign} f(x, C_n) - \operatorname{sign} f(x, a)\,|\,> \varepsilon) \leq Q(|\,f(x, a)\,| < \sigma) +$$

$$+ \frac{2}{\sigma} \int f(x, a)\,(\operatorname{sign} f(x, a) - \operatorname{sign} f(x, C_n))\, Q(dx)\,,$$

where the first term on the right-handside will be arbitrarily small if σ is small enough, as follows from (v). Thus $\lim V(C_n) = 0$ P-a.s. implies that sign $f(x, C_n)$ converges in Q-probability to sign $f(x, a)$ with probability one, which proves (26) by the Dominated Convergence Theorem. <>

Remark. Let $L(\varphi)$ denote the subspace of $L_2(Q)$, spanned by the interpolation functions $\varphi_1, \varphi_2, \ldots, \varphi_i, \ldots$; then (25) expresses that $d(x) - \operatorname{sign} f(x, a)$ is orthogonal to $L(\varphi)$ for each $a \in R_0$. This shows that our assumptions (namely (iv) and (vi)) are very strong. Really, in the interesting special case, when the function 1 and the elements of $L(\varphi)$ jointly generate $L_2(Q)$, and $\int d(x)Q(dx) = 0$ (i.e. the a priori probabilities of H_+ and H_- are equal), then this is possible only if $d(x) = \operatorname{sign} d(x) = \operatorname{sign} f(x, a)$ Q-a.s. Therefore, we must be very careful with the application of the algorithm in such situations, when we are not sure that the classes are nonoverlapping; although in practical problems the number of interpolation functions is always limited.

Finally, let us turn to the question of error probability in cases A)—D). If the algorithm is stopped after the n-th step, then we shall use the decision rule: adopt H_+ or H_- according as the value of $f(x, C_n)$ is positive or negative. In view of (4), the error probability of this decision rule will be

$$P_e(C_n) = P + \frac{1}{2} \int d(x)\,(\operatorname{sign} d(x) - d(x))\, Q(dx) +$$

$$+ \frac{1}{2} \int d(x)\,(d(x) - \operatorname{sign} f(x, a))\, Q(dx) +$$

$$+ \frac{1}{2} \int d(x)\,(\operatorname{sign} f(x, a) - \operatorname{sign} f(x, C_n))\, Q(dx)\,,$$

where $d(x)$ denotes the standard decision function $d(x) = P(H_+ \mid x) - P(H_- \mid x)$, P_d is the corresponding minimal error probability, and $a \in R_0$. [We suppose here that (i)–(vi) hold.]

Observe that the first integral on the right hand side vanishes in cases A) and B), and it will be small if the overlapping part of the classes is of small probability; while the third integral goes to zero with probability one if n goes to infinity, as follows from (26). As regards the critical second integral, let $d(x) = d_1(x) + d_2(x)$ be the orthogonal decomposition of $d(x)$ with respect to $L(\varphi)$, that is $d_1 \in L(\varphi)$ and d_2 is orthogonal to $L(\varphi)$. Then from (25) we obtain that

$$I_2 = \frac{1}{2} \int d(x) \left(d(x) - \operatorname{sign} f(x, C_n)\right) Q(dx =$$

$$= \frac{1}{2} \int d_2(x) \left(d(x) - \left(\operatorname{sign} f(x, a)\right) Q(dx) \leq \sqrt{\int d_2^2(x)\, Q(dx)}\,.$$

Unfortunately, the $L_2(Q)$ norm of d_2 need not be small, even though the number of interpolation functions is large, since the constant function 1 — which does not belong to $L(\varphi)$ — may have an important role in the expansion of $d(x)$. However, if 1 is orthogonal to $L(\varphi)$, that is when assumption

(vii) $$\int \varphi_i(x)\, Q(dx) = 0 \quad \text{for each } i$$

holds, then $d_2(x) = p_+ - p_- + d_3(x)$, where d_3 is orthogonal to 1, and $p_+ - p_- \int = \int d(x) Q(dx)$ is the difference of the a priori probabilities. Therefore, as $t \int (d(x) - \operatorname{sign} f(x, a)) Q(dx) = -J_0$,

$$I_2 \leq -\frac{1}{t}(p_+ - p_-) J_0 + \sqrt{\int d_3^2(x)\, Q(dx)}\,,$$

which is a very good estimation, since the first term is negative if the sign of t is appropriately chosen, while the second term may be assumed to be small.

Of course, assumption (vii) is rather unnatural, since the distribution Q is not known. A modified algorithm to solve this problem will be investigated in a forthcoming paper.

References

1. *Kashyap, R. L.:* Algorithms for pattern classification. Adaptive, Learning, and Pattern Recognition Systems, Mendel-Fu Ed. Academic Press, New York 1970.
2. *Aizerman, M. A.—Braverman, E. M.—Rozonoer, L. I.:* The method of potential functions in the theory of learning machines. Moscow 1970.
3. *Csibi, S.:* Simple and compound processes in iterative machine learning. Read at the Summer School of the International Centre for Mechanical Sciences, Udine, 1971 (in print).
4. *Győrfi, L.:* A potential function type algorithm to estimate optimal decision functions. Research Report of the Telecommunication Institute, Budapest, 1971.
5. *Neveu, J.:* Mathematical Foundations of the Theory of Probability. Masson et Cie, Paris, 1964.

Об алгоритме распознавания образов по Р. Л. Кашиапу

Й. ФРИЦ

(Будапешт)

Представленный в работе алгоритм был рассмотрен М. А. Айзерманом, Э. М. Браверманом и Л. И. Розоноэром при условии существования разделяющей функции определенного типа. Р. Л. Кашиап исследовал более общую ситуацию, когда разделяющая функция определенного типа, не существует границы классов не пересекаются и учитель при этом является не полностью достоверным. Поскольку доказательство, данное Кашиапом, некорректно, цель настоящей работы заключается в некотором обобщении достигнутых результатов на пересекающиеся классы и в изучении необходимых и достаточных условий сходимости алгоритма Кашиапа. Показано, когда алгоритм может расходиться. Доказательства в работе основаны на применении теории центрированных случайных последовательностей также. Оценивается вероятность некорректной классификации образов. оценивается.

J. Fritz
Research Institute for Mathematics
Hungarian Academy of Sciences
Budapest, V. Reáltanoda u. 13—15.

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

PROBLEMS OF CONTROL- AND INFOMATION THEORY VOL. 2(1)
ENGLISH TRANSLATION OF COMMUNICATIONS IN RUSSIAN

ON OPTIMAL LINEAR ESTIMATES OF EXPECTATION AND REGRESSION COEFFICIENTS OF HOMOGENEOUS AND ISOTROPIC RANDOM FIELDS

M. I. YADRENKO

(Kiev)

Optimal (in the sense of minimal root-mean-square error) linear unbiased estimates of expectation and regression coefficients of homogeneous and isotropic random field observed on a circle, are considered. For the case where measurements are taken in the vertices of a regular polygon or on a system of concentric circles, the optimal estimates of the mean value of homogeneous and isotropic field are obtained.

1. Formulation of the problem

A variety of problems in the fields of automatic control, geology, meteorology, geophysics, pattern recognition, radiophysics, statistical optics and holography can be solved through statistical estimates of certain random field parameters (e.g., expectation, regression coefficients). Temperature evaluation in a rotating cement kiln, referred to in Ref. [1], may also be put on the list. It may be solved through the results obtained below.

Let $\eta(p)$ be a random function dependent on a point of a plane (i.e. a random field in a plane) with zero expectation and known correlation function. Assume that a random field $\xi(p) = m + \eta(p)$, or $\xi(p) = \sum_{s=1}^{N} a_s g_s(p) + \eta(p)$ is observed on a set D of the plane, where the parameters $m, a_1, \ldots, a_N$ are unknown, and function $g_s(p)$ $(s = \overline{1, N})$ are known. Estimates $\hat{m}, \hat{a}_s$ of the parameters m, a_s should be found which are linearly dependent on the measurement results, unbiased ($M\hat{m} = m$, $M\hat{a}_s = a_s$), and have the least r.m.s. error of all linear and unbiased estimates. In the further discussion such estimates will be referred to as optimal and linear estimates.

The form of optimal estimates depends, of course, on the set D over which the observations are made. Consideration will be given here to derivation of the optimal estimate of an unknown expectation m for the case where D is a system of concentrical circles or a set of vertices of a regular polygon,

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

and $\eta(p)$ is a homogeneous and isotropic random field such that

$$M\eta(p_1)\,\eta(p_2) = B(\varrho_{p_1,p_2}) \tag{1.1}$$

depends only on the distance ϱ_{p_1,p_2} between points p_1 and p_2.

In Section 4 optimal estimates of regression coefficients a_s are also given for the case where D is a circle.

Optimal estimates of expectation and regression coefficients of homogeneous and isotropic random fields have not yet been described in the literature. Among the results obtained below, two (Corollary 1 and Theorem 2) should be mentioned as being conceivably interesting for applications: if a homogeneous and isotropic random field is observed on a circle or in the vertices of a regular polygon, the mean of the set of observations is the optimal estimate of the unknown expectation.

2. Spectral expansion of a homogeneous and isotropic random field

The function $B(\varrho)$ for the homogeneous isotropic random field is as follows [2]:

$$B(\varrho) = \int_0^\infty J_0(\lambda\varrho)\,d\Phi(\lambda) \tag{2.1}$$

where $\Phi(\lambda)$ is a non-decreasing function over $(0, +\infty)$, J_0 is a Bessel function of the first kind and zero order [3]. The random field itself is representable as [4, 5].

$$\eta(p) = \sum_{n=-\infty}^{\infty} e^{in\varphi} \int_0^\infty J_n(\lambda r)\,d\,Z_n(\lambda) \tag{2.2}$$

where (r, φ) are polar coordinates of the point p, $J_n(x)$ is an n-order Bessel function, $Z_n(\lambda)$ is a sequence of stochastic processes with non-correlated increments such that

$$M d\,Z_n(\lambda)\,\overline{dZ_m}(\lambda) = \delta_n^m\,d\varphi(\lambda) \tag{2.3}$$

where the bar serves as a symbol of complex conjugation.

The representation (2.2) has not, unfortunately, become popular enough among specialists working in application of random field theory although it describes essentially the property of homogeneity and isotropy, and facilitates considerably solution of main statistical problems.

Functions

$$b_k(r) = \int_0^\infty I_k^2(\lambda r) = \frac{1}{T_1} \int_0^\pi e^{-2ikx} B(1r \sin x)\, dx \tag{2.4}$$

which may be called spectral coefficients of the random field $\xi(r, \varphi)$, will be essential below.

3. Estimate of expectation

Assume that the random field $\xi(p) = m + \eta(p)$ with $M\eta(p) = 0$ is observed on the point set D of the plane. Let H_D be a closed (in the sense of r.m.s. convergence) linear variety generated by the random variables $\xi(p)$, $p \in D$, i.e. a set of all possible linear combinations of the variables $\xi(p)$, $p \in D$, and their limits in r.m.s. Let Γ be a closed linear variety generated by random variables having the following form: $\sum_{P_k \in D} \xi(p_k) C_k$ where $\Sigma C_k = 1$. Each random variable $\tilde{m}$ of Γ may be regarded as a linear unbiased estimate ($M\tilde{m} = m$). Estimate with minimal variance $\hat{m}$ is known [6, 7] to be defined by the equation

$$M\eta(p)\, \hat{m} = C \tag{3.1}$$

where $C = D[\hat{m}]$.

Estimate of the unknown mean is considered for the case where D is a system of concentric circles.

Theorem 1. Let the random field $\xi(p) = m + \eta(p)$ be observed on concentric circles with the radii $R_1, \ldots, R_N$ and the center in the origin. Then the linear unbiased estimate m having minimal variance is as follows

$$\hat{m} = \sum_{k=1}^{N} \frac{\Delta_k}{\sum_{i=1}^{N} \Delta_i} \cdot \frac{1}{2\pi} \int_0^{2\pi} \xi(R_k, \varphi)\, d\varphi \tag{3.2}$$

where Δ is a matrix determinant with (s, k)-th entry determined by $\int_0^\infty J_0(\lambda R_k)\, J_0(\lambda R_s)\, d\varphi$ (λ and Δ_i is obtained from Δ) by substitution of 1 for all entries of the i-th column. Variance of $\hat{m}$ is as follows:

$$D\hat{m} = \frac{\Delta}{\sum_{i=1}^{N} \Delta_i} \tag{3.3}$$

Let

$$\xi_k(r) = \frac{1}{2\pi} \int_0^{2\pi} e^{-ik\varphi}\, \xi(r, \varphi)\, d\varphi$$

Owing to (2.2)

$$\xi_k(r) = \begin{cases} \int_0^{\infty} J_0(\lambda r)\, d\, Z_k(\lambda) \\ m + \int_0^{\infty} J_k(\lambda r)\, d\, Z_0(\lambda) \end{cases} \tag{3.4}$$

Each $\widehat{m}$ from H_0 has the following form:

$$\widetilde{m} = \sum_{i=1}^{N} \sum_{n=-\infty}^{\infty} C_{ni}\, \xi_n(R_i) \tag{3.5}$$

where $\xi_n(R_i)$ is defined by Eqs. (3.4). The basic equation (3.1) becomes

$$\sum_{S=1}^{N} \sum_{n=-\infty}^{\infty} C_{ns}\, e^{-in\varphi} \int_0^{\infty} J_n(\lambda R_s)\, J_n(\lambda R_k)\, d\Phi(\lambda) \tag{3.6}$$

for all $0 \leq \varphi \leq 2\pi$ and $k = 1, \ldots, N$. Hence the system of equations for C_{0s} is as follows

$$\sum_{S=1}^{N} C_{0s} \int_0^{\infty} J_0(\lambda R_s)\, J_0(\lambda R_k)\, d\varphi(\lambda) = C \tag{3.7}$$

With due regard to the condition $\sum_{k=1}^{N} C_{0n} = 1$ (unbiasedness), obtain

$$C_{0s} = \frac{C \Delta_s}{\Delta}, \qquad C = \frac{\Delta}{\sum_{S=1}^{N} \Delta_s} \tag{3.8}$$

In the particular case where observations are made on a single circle, Theorem 1 gives the following:

Corollary 1. Among all unbiased estimates m, the least variance is estimated as

$$\widehat{m} = \frac{1}{2\pi} \int_0^{2\pi} \xi(r, \varphi)\, d\varphi \tag{3.9}$$

which is mean $\xi(r, \varphi)$ over the circle of the radius r, and therewith

$$D\widehat{m} = b_0(r) = \int_0^\infty J_0^2(\lambda r)\, d\Phi(\lambda) \tag{3.10}$$

It is worth noting that in this case the form of the optimal estimate is independent of the random field correlation function $\eta(r, \varphi)$.

A problem naturally arises concerning the behaviour of (3.9) with increase of r. The answer may be found by the results of Ref. [8] where the behaviour of spherical mean variances is studied. It appears that if $\Phi'(\lambda) = \lambda^\alpha g(\lambda)$, where $0 \leq \alpha < 2$, $g(\lambda)$ is limited and continuous in zero, $g(0) \neq 0$, then

$$D\widehat{m} = \frac{1}{R^{1+\alpha}} g(0)\, O_\alpha + 0\left(\frac{1}{R^{1+\alpha}}\right.$$

Assume that random field correlation function $\eta(p)$ is known. Therefore, for practical applications of the results obtained, it is often necessary to start with choosing an appropriate *a priori* model for $\eta(p)$. Several examples of homogeneous and isotropic random fields are given below, accompanied in each case by expressions for spectral coefficients $b_k(r)$.

Example 1. Let

$$B(\varrho) = J_0(c\varrho) \tag{3.11}$$

Then

$$b_k(r) = J_k^2(cr) \tag{3.12}$$

Let λ_i be a sequence of the roots of the equation $J_0(x) = 0$. Then (3.10) and (3.12) imply that m is exactly restored through observations of $\xi(p)$ with the correlation function (3.11) on any of the circles with the radii $r_i \bar{c}$. Note that Gaussian random field with the correlation function (3.1) has the following Markovian property: for a sufficiently large class $\{K\}$ of closed Jordan contours, the values of $\eta(p)$ in the points p_1 and p_2 separated by the contour K, are independent if the values of $\eta(p)$ on the contour K are known [9].

Example 2. Let

$$B(\varrho) = c\varrho\, k_1(c\varrho) \tag{3.13}$$

where $K_\nu(x)$ is a modified Bessel function of the third kind [3]. Then

$$b_0(r) = 2c^2 r[I_1(cr)\, K_0(cr) - K_1(cr)\, I_0(cr)]$$

where $I_\nu(x)$ is a modified Bessel function of the first kind [3]. It is interesting to note that Gaussian random field with the correlation function (3.3) has the

following Markovian property: the values of $\eta(p_1)$ and $\eta(p_2)$ become independent if on the contour separating p_1 and p_2 values of the field and its normal derivative are known.

Example 3. Let

$$B(\varrho) = \exp\{-c\varrho^2\} \tag{3.14}$$

Then

$$b_k(r) = \exp\{-2cr^2\} I_k(2cr^2)$$

Consider now the case where D is a set of vertices of a regular N-corner polygon.

Theorem 2. Assume that a random field $\xi(r, \varphi) = m + \eta(r, \varphi)$ is observed in the points $\left(r, k\frac{2\pi}{N}\right)$ $k = 0, \ldots, N-1$, which are vertices of a regular N-corner polygon inscribed into a circle of the radius r. Then the linear estimate m with minimal variance has the following form:

$$\hat{m} = \frac{1}{N}\sum_{k=0}^{N-1} \xi\left(r, k\frac{2\pi}{N}\right) \tag{3.15}$$

with variance of this estimate being

$$D\hat{m} = \sum_{s=-\infty}^{\infty} b_{kN}(r) \tag{3.16}$$

Let us look for an estimate $\hat{m}$ having the form $\hat{m} = \sum_{k=0}^{N-1} c_k \xi(r, \varphi_k)$ where $\varphi_k = \frac{2\pi}{N}k$, $\Sigma c_k = 1$. The basic Eq. (3.1) gives a system of equations

$$\sum_{k=-\infty}^{\infty} e^{in\varphi_s}\left(\sum_{k=0}^{N-1} C_k \cdot e^{in\varphi_k}\right) b_n(r) = C \tag{3.17}$$

$$(s = 0, 1, \ldots, N-1)$$

By summing all equalities (3.17), dividing them by N, and using the relation

$$\frac{1}{N}\sum_{S=0}^{N-1} e^{in\varphi_s} = \begin{cases} 0 & n \neq Nl \\ 1 & n = Nl \end{cases} \tag{3.18}$$

we obtain

$$C = \sum_{l=-\infty}^{\infty} b_{nl}(r)$$

Now it is easy to verify that $C_k = \frac{1}{N}$ $(k = 0, \ldots, N-1)$ is the solution of the system (3.18). Since the solution of Eq. (3.1) is unique, Theorem 2 is proved.

4. On estimation of regression coefficients

Let a random field $\xi(r, \varphi) = ag(r, \varphi) + \eta(r, \varphi)$, where a is an unknown parameter and $g(r, \varphi)$ is a known function, be observed on a circle of radius r. Now we pose the problem of considering all unbiased (i.e. satisfying the condition $M\hat{a} = a$) estimates and finding one having the least variance. In particular, if $g(r, \varphi)1$, we obtain the problem of expectation estimate.

Let

$$g_n(r) = \frac{1}{2\pi} \int_0^{2\pi} e^{-in\varphi} g(r, \varphi)\, d\varphi \tag{4.1}$$

Assume that the following condition holds

$$\sum_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)} < +\infty \tag{4.2}$$

Theorem 3. Among all unbiased estimates of the parameter a, the following one has the least variance:

$$\hat{a} = \frac{1}{\sum\limits_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)}} \int_0^{2\pi} l(r, \varphi)\, \xi(r, \varphi)\, d\varphi \tag{4.3}$$

where

$$l(r, \varphi) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\varphi} \frac{g_n(r)}{b_n(r)} \tag{4.4}$$

therewith

$$D\hat{a} = M\,|\,\hat{a} - a\,|^2 = \frac{1}{\sum\limits_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)}} \tag{4.5}$$

In particular, if $g(r, \varphi) = g(r)$ is independent of φ, then

$$\hat{a} = \frac{1}{g(r)} \frac{1}{2\pi} \int_0^{2\pi} \xi(r, \varphi)\, d\varphi \tag{4.6}$$

Let us look for an estimate $\hat{a}$ having the form $\hat{a} = \sum\limits_{n=-\infty}^{\infty} Cn\, \xi_n(r)$ where $\xi_n(r) = ag_n(r) + \eta_n(r)$ and $\Sigma\, Cn\, g_n(r) = 1$ (i.e. condition of unbiasedness). Note that

$$D\hat{a} = \sum_{n=-\infty}^{\infty} C_n^2\, b_n(r) \tag{4.7}$$

From the inequality

$$1 = \left[\sum_{n=-\infty}^{\infty} C_n g_n(r)\right]^2 \leq \sum_{n=-\infty}^{\infty} C_n^2 b_n(r) \sum_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)} \tag{4.8}$$

we obtain

$$D\check{a} \geq \frac{1}{\sum\limits_{n=-\infty}^{\infty} \frac{g_n^2(r)}{b_n(r)}} \tag{4.9}$$

where the minimum in (4.9) is reached if

$$C_n = \frac{g_n(r)}{b_n(r)} \cdot \frac{1}{\sum\limits_{n=-\infty}^{\infty} \frac{g_k^2(r)}{b_k(r)}}$$

and this proves Theorem 3.

Note that if the condition (4.2) is broken, the parameter a may be restored exactly through observations on a circle.

Note also that [10] implies that in the Gaussian case the estimates (3.9) and (4.6) will be those of maximum likehood.

Assume that on a circle of the radius r a random field $\xi(p) = \sum\limits_{s=1}^{N} a_s g_s(p) + \eta(p)$ is observed, where $g_s(p)$ are linearly independent functions such that

$$\sum_{k=-\infty}^{\infty} \frac{g_{ks}^2(r)}{b_k(r)} < +\infty \qquad (s = 1, \ldots, N)$$

Denote

$$\gamma_{sJ} = \sum_{k=-\infty}^{\infty} \frac{g_{ks}(r)\, g_{kJ}(r)}{b_k(r)}$$

Then the following statement holds:

Theorem 4. The linear unbiased minimal-variance estimate of the parameter a_i has the following form

$$\hat{a}_i = \sum_{s=1}^{N} \alpha_{si}\, l_i(r, \varphi)$$

where α_{si} is the solution of the system of linear equations

$$\sum_{s=1}^{\infty} \alpha_{si}\, \gamma_{sJ} = \delta_i^J \qquad (J = 1, \ldots, N)$$

and $l_s(r, \varphi)$ is defined through (4.4) by substituting $g_{sn}(r)$ for $g_n r$.

The proof may be obtained through arguing similar to that of Theorem 3.

References

1. *Виленкин, С. Я., Дубенко, Т. И.:* Об оптимальных линейных оценках математического ожидания однородного случайного поля. Известия Академии наук СССР, Техническая кибернетика, 1971, № 1, 134—141.
2. *Гихман И. И., Скороход А. В.:* Введение в теорию случайных процессов. «Наука», 1965.
3. Higher Transcendental Functions, vol. 2, based on notes left by Harry Bateman and compiled by Arthur Erdélyi, New York, Toronto, London, McGrow-Hill Book Company, Inc. 1953.
4. *Яглом А. М.,* Спектральные представления для различных классов случайных функций. Труды IV Всесоюзного математического съезда, 1963, т. 1, 250—273.
5. *Yadrenko, M. I.* Spektralni rozklady isotropnykh vypadkovykh poliv, Dopovidi Akademii Nauk URSR, 1968, ¾ 10, 835—837 (in Ukrainian).
6. *Grenander, U.* Stochastic processes and statistical inference. Arkiv for Math., 1951, B. 1, H. 3, 195—277.
7. *Grenander, U.* On empirical spectral analysis of stochastic processes. Arkiv for Math., B. 1, H. 6, 1952, 503—531.
8. *Yadrenko, M. I.* Sferichni seredni izotropnykh vypadkovykh poliv, Dopovidi Akademii Nauk URSR, 1968, ¾ 12 (in Ukrainian).
9. *Ядренко М. И.:* Изотропные случайные поля марковского типа в эвклидовом и гильбертовом пространстве. Труды Всесоюзного совещания по теории вероятностей и математической статистике. Ереван, 19—25 сентября 1958, 263—279.
10. *Ядренко М. И.:* Об абсолютной непрерывности мер соответствующих гауссовским однородным случайным полям. Теория вероятностей и математическая статистика, 1972, в. 7, 212—225.

ANALYSIS TECHNIQUE FOR ASYNCHRONOUS AUTOMATA

B. LAGOVIER

(Moscow)

1. Introduction

Design of digital control devices involves analysis of automata given, as a rule, either as a circuit or as a list of excitation functions of storage cells. To solve this problem one has to obtain the operation algorithm of a given device, i.e. to construct its transfer and output functions. The automaton table defining its transfer δ and output λ functions is widely used for description of operation algorithms. The canonical automaton table has $2^m > k$ entries, where m and k are numbers of inputs and internal states, respectively. Each entry indicates, pair of input and internal states for the next internal state (transfer) and the output state. Ref. [1] contains the concept of the generalized input state and a technique for construction of transfer tables where each entry is defined by a pair "input function – internal state". In Refs. [2, 3] necessary and sufficient conditions for such a system of input functions are defined, the system is proved to be orthogonal, and methods of orthogonalization are worked up. A technique is also described for construction of generalized input transfer table through given excitation functions of storage cells. This technique, however, asks for rather complex calculations, and is acceptable only for circuits with a small number of storage cells. The present paper describes another technique for construction of a generalized input transfer table through a given circuit which is more obvious and, in some instances, more convenient in analysis of multi-cell circuits.

2. Description of the Technique

Let a set of cell excitation functions of a device be given:

$$\begin{aligned} y_1 &= \varphi_1(x_1, \ldots, x_m, y_1, \ldots, y_n) \\ y_2 &= \varphi_2(x_1, \ldots, x_m, y_1, \ldots, y_n) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= \varphi_n(x_1, \ldots, x_m, y_1, \ldots, y_n) \end{aligned} \tag{2.1}$$

where $x_1, x_2, \ldots, x_m$ are input variables, y_i is the input of a storage cell, and y_i is its output. The set of functions (2.1) describes a logical network with blocks having two internal states and implemented by one storage cell each. Consider the i-th block (see Fig. 1). The excitation function y_i is completely defined, and we will assume that it is equal to 1 over the set M_{1i} of operating states and to 0 over the set M of remaining states of input and internal variables. In doing so, each excitation function divides all the 2^{m+n} states of input and internal variables into two disjoint sets, with $M_{1i} \cap M_{0i} = \emptyset$. The transfer function of the given logic network may be found through multiplication of all block transfer functions.

Decomposing each excitation function of (2.1) by its own variable,

$$y_i \equiv g_{0i}\,\overline{y}_i + g_{1i} + g_{2i} \qquad (i = 1, 2, \ldots, n) \tag{2.2}$$

where g_{0i} is the factor of $\overline{y}$, g_{1i} is the factor of y_i, and g_{2i} is the free term. Using the decomposition factors of y_i, one may construct the transfer function of an i-th block consisting of a single storage cell. The transfer function of the block is represented by generalized input state table.

Table 1

y_i	f_{0i}	f_{1i}	f_{2i}	f_{3i}
0	0	1	0	1
1	1	1	0	0

where f_{0i} is called "retain" function of y_i; f_{1i} is "on" function, f_{2i} is "reset" function, and f_{3i} is "generation" function.

The generalized functions $(f_{0i}, f_{1i}, f_{2i}, f_{3i})$ may be obtained as follows. It is clear from (2.2) that for $y_i = 0$, $y_i^{(0)} = g_{0i} + g_{i2}$, and for $y_1 = 1$, $y_i^{(1)} = g_{1i} + g_{2i}$. From the first column of Table 1 one may see that for $y_i = 0$, $y_i = 0$, i.e. $f_{0i}^{(0)} = y_i^{(0)} = \overline{g_{0i} + g_{2i}}$, and for $y_1 = 1$, $yi = 1$, i.e. $f_{0i}^{(1)} = y_i^{(1)} = g_{1i} + g_{2i}$. Similarly, the remaining columns of Table 1 give:

$$f_{1i}^{(0)} = y_1^{(0)} = g_{0i} + g_{2i}\,; \qquad f_{1i}^{(1)} = y_i^{(1)} = g_{1i} + g_{2i}\,;$$

$$f_{2i}^{(0)} = \overline{y}_i^{(0)} = \overline{g_{0i} + g_{i2}}\,; \qquad f_{2i}^{(1)} = \overline{y}_i^{(1)} = \overline{g_{1i} + g_{2i}}\,;$$

$$f_{3i}^{(0)} = y_i^{(0)} = g_{0i} + g_{2i}\,; \qquad f_{3i}^{(1)} = \overline{y}_i^{(1)} = \overline{g_{1i} + g_{2i}}\,.$$

These expressions lead to

$$f_{0i} = f_{0i}^{(0)} f_{0i}^{(1)} = \overline{(g_{0i} + g_{2i})}\,(g_{1i} + g_{2i}) = \bar{g}_{0i}\, g_{1i}\, \bar{g}_{2i}$$
$$f_{1i} = f_{1i}^{(0)} f_{1i}^{(1)} = (g_{0i} + g_{2i})\,(g_{1i} + g_{2i}) = g_{0i}\, g_{1i}\, g_{2i}$$
$$f_{2i} = f_{2i}^{(0)} f_{2i}^{(1)} = \overline{(g_{0i} + g_{2i})}\,\overline{(g_{1i} + g_{2i})} = \bar{g}_{0i}\, \bar{g}_{1i}\, \bar{g}_{2i}$$
$$f_{3i} = f_{3i}^{(0)} f_{3i} = (g_{0i} + g_{2i})\,\overline{(g_{1i} + g_{2i})} = g_{0i}\, g_{1i}\, g_{2i} \qquad (2.3)$$

In a non-contradictorily defined asynchronous automaton, the generation function should be identically equal to zero ($f_{3i} = 0$), and the set of input function f_{0i}, f_{1i}, f_{2i} is complete and orthogonal.

Table 2

g_{0i}	g_{1i}	g_{2i}	f_{0i}	f_{1i}	f_{2i}	f_{3i}
0	0	0	0	0	1	0
0	0	1	0	1	0	0
0	1	0	1	0	0	0
0	1	1	0	1	0	0
1	0	0	0	0	0	1
1	0	1	0	1	0	0
1	1	0	0	1	0	0
1	1	1	0	1	0	0

Each function in the set (2.1) may be checked, at the very beginning of analysis with completeness and orthogonality conditions to be non-contradictorily defined. If this check is completed, each block may be represented by a transfer function (Table 3) where the input functions f_{0i}, f_{1i}, f_{2i} are obtained through the system (2.3).

Table 3

y_i	f_{0i}	f_{1i}	f_{2i}
0	0	1	0
1	1	1	0

Prior to using the union and multiplication operations over blocks with different external inputs, were proposed for canonical transfer tables [4], one has first to define the input state set. For canonical transfer tables, the cardinality

of this set equals the product of numbers of input states taken over all blocks. Input sets are formed as sums of input states of constituent blocks.

Let us verify whether these operations may be applied to the transfer tables with generalized input states. Let the system of input functions of each separate block be complete and orthogonal. For simplicity, a union of two blocks will be considered. All the proofs obtained may easily be extended to n blocks.

Let the systems of the first and second blocks consist of functions $f_0, f_1, \ldots, f_k$, and $h_0, h_1, \ldots, h_l$, respectively. According to [4], the system of input functions of the united block will comprise the following products:

$$f_0h_0, f_0h_1, \ldots, f_0h_l,$$
$$f_1h_0, f_1h_1, \ldots, f_1h_l,$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$f_kh_0, f_kh_1, \ldots, f_kh_l.$$

The system of functions thus obtained is complete and orthogonal. In fact, the union of all functions may be represented as

$$f_0h_0 + f_0h_1 + \ldots + f_kh_l = (f_0 + f_1 + \ldots + f_k)\,(h_0 + h_1 + \ldots + h_l).$$

but, by the completeness condition for the given function systems, each parenthesized term in the right-hand part of the equation is identically equal to 1, and therefore the function system is complete.

In order to prove orthogonality, the conjunction of two arbitrary products $f_jh_j \wedge f_ph_q$, $i \neq p$ or $j = q$ is considered. It is known that two functions are orthogonal if their conjunction is identically equal to zero. And $f_1h_jf_ph_q = f_if_ph_jh_q = 0$ since:

1) $h_ih_q = 0$, for $i = p$ and $j \neq q$,

2) $f_if_p = 0$ for $i \neq p$ and $j = q$, and

3) $f_if_if_qf_q = 0$ for $i \neq p$, $j \neq$ p, and $f_i = h_j$.

Multiplication of blocks defined by generalized input transfer tables, in the absence of additional external inputs does not differ from the multiplication of canonical transfer tables.

Thus, if there are transfer tables of individual storage cells one can obtain an orthogonal system of network input functions by multiplying the orthogonal systems of cell input functions.

Let us come now to the construction of the transfer table. Let the system (2.1) be given, and for each function y_i let the genralized input functions f_{0i}, f_{1i}, f_{i2} be obtained, next a generalized system of orthogonal functions is got:

$$\begin{aligned} F_1 &= f_{01}f_{02}\cdots f_{0n} \\ F_2 &= f_{01}f_{02}\cdots f_{0(n-1)}f_{1n} \\ &\cdots\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots\cdots \\ F_{3n} &= f_2 f_{22}\cdots f_{2n}. \end{aligned} \tag{2.4}$$

The system has at most 3^n functions. In effect, some of the functions in the system (2.4) may be identically equal to 0, thus reducing the number of functions in the system.

Now derivation of the transfer table through the system (2.4) will be described. Two cases are considered. First, if all y_i's in the system (2.1) depend only on input and proper internal variables; second, if a y_i depending on improper internal variables may be met in the system (2.1).

The first case corresponds to parallel connections of blocks. In this case, all F_i's in (2.4) are independent of internal variables by construction. The transfer table has 2^n rows and N columns where N is the number of $F_i \not\equiv 0$ in (2.4). The rows are then coded by products of internal variables $y_1^{\sigma_1}y_2^{\sigma_2}\ldots y_n^{\sigma_n}$, where $\sigma \in \{0, 1\}$.

Let $F_1 = f_{\alpha,1}f_{\alpha_2 2}\ldots f_{\alpha_n n}$, where $\alpha_i \in \{0, 1, 2\}$. The intersection of a row $y_1^{\sigma}y_2^{\sigma_2}\ldots y_n^{\sigma_n}$ and a column F_1 contains the next state $y_1^{\tau_1}y_2^{\tau_2}\ldots y_n^{\tau_n}$. The value of r_j is found in accordance with the following rule: if $\sigma_j = 0$ and $\alpha_j = 0$ or 2, $\tau_j = 0$; if $\tau_j = 0$ and $\alpha_j = 1$, $\tau_j = 1$; if $\sigma_j = 1$ and $\alpha_j = 0$ or 1, $r_j = 1$; and if $\tau_j = 1$ and $\alpha_j = 2$, $\tau_j = 0$. It is easy to see that this rule indeed enables one to determine the next state. If for all j, $1 \leq j \leq n$, $\tau_j = \tau_j$, then the state is stable. Note that if in the function $F_i = f_{\alpha_1 1}f_{\alpha_2 2}\ldots f_{\alpha_n n}$ none of α'_{js} is 0, the column F_i contains just one stable state, namely the state $y_1^{\tau_1}y_2^{\tau_2}\ldots y_n^{\tau_n}$, where

$$\tau_j = \begin{cases} 1 \text{ if } \alpha_j = 2 \\ 0 \text{ if } \alpha_j = 1. \end{cases}$$

Similarly, if a function F_i contains one $\alpha_j = 0$, the column of F_j contains two stable states ($\tau_j = 0$ in one of them, $\tau i = 1$ in the other. Other values are determined as before), and so on.

To construct a transfer table allowing for all of the possible internal states and transfers among them, one does not have to check each of the 2^n

internal states for stability. The system of generalized input functions (2.4) is taken. They may be united into groups, containing $C^r 2^{n-r}$ functions each, where r is the number of functions $f_{\alpha_i i}$ having $\alpha_i = 0$ in the product of F_i.

If in an individually taken group none of $F_i \not\equiv 0$, it defines all the 2^n internal states. Let, for instance, $r = 0$, i.e. we take a group F_i where all $\alpha \in \{1, 2\}$. We assume that this group does not have any $F_i = 0$. The number of functions in the group is $C_n^0 2^{n-0} = 2^n$. According to the above, each function in the group defines one stable state of storage cells, and all transfers to one (proper) stable state are allowable in each of transfer table columns. Since $F_i F_j \equiv 0 (i \neq j)$, stable states in different columns are different, and their total number is 2^n. If for each group $F_i = 0$, then each group is considered individually, and the set of automaton internal states is determined as a sum of different internal states defined by each group.

Example 1. Let $F_1 = f_{01} f_{02}$, $F_2 = f_{01} f_{12}$, $F_3 = f_{01} f_{22}$, $F_4 = f_{11} f_{02}$, $F_5 = f_{11} f_{12}$, $F_6 = f_{11} f_{22}$, $f_7 = f_{21} f_{02}$, $F_8 = f_{21} f_{12}$, $F_9 = f_{21} f_{22}$
then Table 4 is transfer table of united blocks.

Table 4

$y_1 y_2$	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
00	⟨00⟩	01	⟨00⟩	10	11	10	⟨00⟩	01	⟨00⟩
01	⟨01⟩	⟨01⟩	00	11	11	10	⟨01⟩	⟨01⟩	00
10	⟨10⟩	11	⟨10⟩	⟨10⟩	11	⟨10⟩	00	01	00
11	⟨11⟩	⟨11⟩	10	⟨11⟩	⟨11⟩	10	01	01	00

Consider now the second case of the system (2.1) where several or all of y_i's depend on improper internal variables. In this case, some of F_i's in the system (2.3) depend on y_i's $(i \neq j)$. $F_i = \psi_i(x_1, \ldots, x_m, y_1, \ldots, y_n)$ may be represented as a sum $F_i = F_i' + F_i''$, where $F_1^i = \varphi_i(x_1, \ldots, x_m)$, and $F_i'' = \psi_i(x_1, \ldots, x_m, y_1, \ldots, y_n)$. $F_i F_j = 0 (i \neq j)$ implies that $F_i'' F_j = 0$, and $F_i'' F_j = 0$. One on determine the next states for all 2^n internal states in corresponding columns of transfer table through the functions F_i', as well as through the function F_i in the above case. In order to determine the states defined by the functions F_i'', they are represented in the sum-of-products form. $F_i'' = k_{1i}^x k_{1i}^y + k_{2i}^x k_{2i}^y + \ldots + k_{s(i)i}^x k_{s(i)i}^x$, where k_{li}^x is l-th product of input variables, and k_{li}^y is l-th product of internal variables $(1 \leq l \leq s(i))$. Denote a column by k_{li}^x. $F_i F_j = 0$ implies that $k_{li}^x k_{li}^y F_j = 0$. In particular, for functions such as F_i'' $k_{li}^x k_{li}^y k_{pj}^x k_{pj}^y \equiv 0$. A pair of products k_{li}^x and k_{li}^x defines next states of the transfer function only for the rows where $k_{li}^y = 1$. It may turn out that for

somes $l's$ and $p's$ $(1 \leq l \leq s(i);\ 1 \leq p \leq s(j)ti, j = 1, 2, \ldots, 3^n)$ $k_{li}^x k_{pj}^x \neq 0$. Obviously $k_{li}^y k_{pj}^y = 0$. In this case, a column is introduced into the transfer table for each product $k_{li}^x k_{pj}^x \neq 0$ where each next state for the entry $k_{li}^x k_{pj}^x k_{li}^x$ is determined through F_i, and for the entry $k_{li}^x k_{pj}^x k_{pi}^z$ is determined through F_j as above. Thus for the case at hand, the transfer table is obtained through the steps described below.

At the first step a transfer table is constructed consisting as it were of two groups of columns. The first group comprises columns denoted by functions such as F_i, and F_i', and the second one comprises columns denoted by products k^x. At the second step the columns dehnoted by te conjunctions k^x are considered only, and all the possible intersections of these conjunctions are determined. The first conjunction in the sequence is chosen at the same time and is decided whether it has pairwise intersections with all of the others. If not, then the column, denoted by this conjunction, is taken in the first part of the table. Otherwise, if such intersections exist, then all these columns, as new ones, are taken in the same part of the table. The column, denoted by the controlled conjunction, will be omitted further on. The second conjunction is considered thereafter, and its intersections are examined with all of the other ones, including those, obtained because of the intersections of the first conjunction with the others. If the second conjunction does not possesss any intersection with the others, the corresponding column is immediately taken in the first part of the table, just as before. At the other hand, if such conjunctions do exist, they are taken in the second part of the table (to be controlled). The conjunction, having been controlled already, will not be controlled ever more.

The composition of the transfer table is finished if the last column is taken in the first part of the table too.

Practically those automation are of interest, where the transfer graph is strongly connected. Such automations are usually applied to control cyclical processes of the production.

Certain internal states cannot be reached among the functions of the transfers of such automations. This states are not used in general. This is the case if a part of the input functions is zero. E.g., if $F_3 - F_4 - F_6 = 0$ in the above example, then the state 10 $(y_1\bar{y}_2)$ has been proved to be unattainable and it is possible that this state does not occur in the transfertable obtained.

A simplified algorithm is suggested below to fill the transfer-table of such automations.

Unattainable states are neglected.

1. Write down the titles of columns.
2. Put stable states in columns having one such state.

3. Put transfers (unstable states) and corresponding stable states in columns having two stable states, and so on until columns with maximal number of stable states are reached.

4. Starting from a column with the maximal number of stable states, return to the preceding column and insert missing transfers, and so on until columns with one stable state are reached.

This technique is illustrated by two examples. The first one is simpler, it shows all steps of writing the transfer table. The second example is borrowed from Ref. [2] and is solved by means of the technique proposed, thus the two techniques can be compared.

Example 2. A system of functions is given

$$y_1 = x_1\bar{y}_3(x_2 + y_1) = x_1\bar{y}_3y_1 + x_1x_2\bar{y}_3$$

$$y_2 = x_1(\bar{x}_2y_1 + y_2) = x_1y_2 + \bar{x}_2y_1$$

$$y_3 = x_1(x_2y_2 + y_3) = x_1y_3 + x_1x_2y_2$$

Compute functions of individual blocks and find block transfer tables.

block I

$f_{11} = x_1\bar{y}_3x_2$

$f_{21} = \bar{x}_1 + y_3$

$f_{01} = x_1\bar{y}_3\bar{x}_2$

Table 5

y_1	f_{01}	f_{11}	f_{21}
0	⟨0⟩	1	⟨0⟩
1	⟨1⟩	⟨1⟩	0

block II

$f_{12} = x_1\bar{x}_2y_1$

$f_{22} = \bar{x}_1$

$f_{02} = x_1(\bar{x}_2 + \bar{y}_1)$

$f_{32} = 0$

Table 6

y_2	f_{02}	f_{12}	f_{22}
0	⟨0⟩	1	⟨0⟩
1	⟨1⟩	⟨1⟩	0

block III

$f_{13} = x_1\bar{x}_2y_2$

$f_{23} = \bar{x}_1$

$f_{03} = x_1(\bar{x}_2 + \bar{y}_2)$

$f_{33} = 0$

Table 7

y_2	f_{02}	f_{12}	f_{22}
0	⟨0⟩	1	⟨0⟩
1	⟨1⟩	⟨1⟩	0

Let us define overall system of logic network input functions. To this end, one has to compute at most 3^3 products of block input functions, but it is easy to verify that only the following products are non-zero:

$$F_1 = f_{01}f_{02}f_{03} = x_1\bar{x}_2y_1y_3, \qquad F_2 = f_{11}f_{02}f_{03} = x_1x_2y_2\bar{y}_3,$$

$$F_3 = f_{21}f_{02}f_{03} = x_1x_2\bar{y}_2y_3 + x_1\bar{x}_2\bar{y}_2y_3, \qquad F_4 = f_{01}f_{12}f_{03} = x_1\bar{x}_2y_1y_3,$$

$$F_5 = f_{21}f_{12}f_{03} = x_1\bar{x}_2y_1y_3, \qquad F_6 = f_{11}f_{02}f_{15} = x_1x_2y_2\bar{y}_3,$$

$$F_7 = f_{21}f_{02}f_{19} = x_1x_2y_2y_3, \qquad F_8 = \bar{x}_1.$$

Write down a transfer table where F_8 belongs to the first type (is independent of internal variables), and the remaining seven functions are of the second type. At first let us write a column for F_8 in the table, and denote it by $\bar{x}_1$. The other seven functions are decomposed by products and ordered as follows:

$$k_{11} = x_1\bar{x}_2,\ k_{21} = x_1x_2,\ k_{31} = x_1\bar{x}_2,\ k_{31} = x_1\bar{x}_2,\ k_{11} = x_1\bar{x}_2,$$

$$k_{51} = x_1\bar{x}_2,\ x_{61} = x_1x_2,\ k_{71} = x_1x_2.$$

Let us take the first product k_{11} and check it for mutual intersection with others. The intersections are written into the table, and k_{11} is rejected.

$$k_{21} = x_1x_2,\ k^{*\circ} = x_1x_2, \quad k_{31} = x_1\bar{x}_2, \quad k_{41} = x_1\bar{x}_2,\ x_{51} = x_1\bar{x}_2,$$

$$k_{61} = x_1x_2,\ k_{71} = x_1x_2,\ k_{11}k_{32} = x_1\bar{x}_2,\ k_{11}k_{41} = x_1\bar{x}_2,\ k_{11} = k_{51} = x_1x_2.$$

k_{21} is rejected at the second step; k_{31} at the third, etc. As a result we have:

$$k_{11}k_{32}k_{41}k_{51} = x_1\bar{x}_2 \quad \text{and} \quad k_{21}k_{32}k_{61}k_{71} = x_1x_2.$$

These products will denote columns in the transfer table (see Table 8).

To check for orthogonality, we verify whether the corresponding k^y's intersect:

$$k^y_{11}k^y_{32} = 0,\ k^y_{11}k^y_{41} = 0,\ k^y_{11}k^y_{51} = 0,\ k^y_{32}h^y_{41} = 0,\ k^y_{32}k^y_{51} = 0,$$

$$k^y_{41}k^y_{51} = 0,\ k^y_{21}k^y_{51} = 0,\ k^y_{21}k^y_{61} = 0,\ k^y_{21}k^y_{71} = 0,\ k^y_{31}k^y_{61} = 0,$$

$$h^y_{31}k^y_{71} = 0,\ k^y_{61}k^y_{71} = 0.$$

Thus we get only three generalized input functions $x_1\bar{x}_2$, x_1x_2, $\bar{x}_1$, and the following internal states $\bar{y}_1\bar{y}_3$, $\bar{y}_2\bar{y}_3$, $\bar{y}_2y_3$, $\bar{y}_1y_3$, $y_1\bar{y}_3$, $y_1\bar{y}_3$, $y_2\bar{y}_3$, y_2y_3.

A transfer table is written containing these states (Table 8).

We start filling the first row of the transfer table from the column $\bar{x}_1$ containing one stable state $\langle 000 \rangle$ which may be introduced in either of the entries $\bar{y}_1\bar{y}_3$ and $\bar{y}_2\bar{y}_3$. The former is considered, as a result for this state the value of y_2 is defined. Since under the state $\bar{y}_1\bar{y}_2\bar{y}_3$ there may bi that $F = 1$ (for $x_1\bar{x}_2 = 1$) and $F_2 = 1$ (for $x_1x_2 = 1$), and $F_1 = f_{01}f_{02}f_{03}$ and $F_2 = f_{11}f_{02}f_{03}$, stable state $\langle 000 \rangle$ is inserted in the column $x_1\bar{x}_2$ (row 1), and unstable state 100 (row 1) and stable state $\langle 100 \rangle$ (row 5) are inserted in the column x_1x_2.

Table 8

y_1	y_2	y_3	$x_1\bar{x}_2$	x_1x_2	$\bar{x}_1$	
→0		0	$\langle 000 \rangle$	100	$\langle 000 \rangle$	1
0	0	0	—	—	—	2
	0	1	—	—	—	3
0		1	$\langle 011 \rangle$	$\langle 011 \rangle$	000	4
1		0	110	$\langle 100 \rangle$	000	5
→1		1	011	011	000	6
	1	9	$\langle 110 \rangle$	111	000	7
1	1		—	—	—	8

The second row may be defined to be equivalent to row 1 or 5. Since it makes no difference, the choice may be arbitrary, so let row 2 be equivalent to row 5. Row 5 is considered. In the column $x_1\bar{x}_2$ we introduce the next state 110 because $F_4 = f_{01}f_{12}f_{03} = 1$ for $y_1\bar{y}_2\bar{y}_3 = 1$ and $x_1x_2 = 1$ implies that y_1 and y_3 should retain their former states, and y_2 should go to $y_2 = 1$. In row 7 stable state $\langle 110 \rangle$ is put thus defining the value of y_1. We proceed now to column x_1x_2 in the same row. Since $F_6 = f_{11}f_{02}f_{12} = 1$ for $y_1y_2\bar{y}_3 = 1$ and $x_1x_2 = 1$ the next state is 111. This transfer may lead either to row 6, or 8. We join them, and, since in this case $F_7 = f_{11}f_{02}f_{13} = 1$, put in row 6 of column x_1x_2 the next state 011. This unstable state leads to row 4, where in the same column the stable state $\langle 011 \rangle$ is put. Go now to the column $x_1\bar{x}_2$ of the same row, and put there the same stable state $\langle 011 \rangle$. Finally the unstable state 011 is put in row 6 of the same column.

Note that the internal state $\bar{y}_2y_3$ in row 3 has proved to be unattainable, and may be disregarded further on. Indeed, for this state to be attainable, it is necessary that at least one of the products $f_{22}f_{13}$, $f_{22}f_{03}y_3$ or for $f_{13}\bar{y}_2$ is included in one of the functions F_i, but this is not the case.

Now one may proceed to column $\bar{x}_1$ and insert there all lacking unstable states 000 leading to row 1.

Table 9 is the final one.

Table 9

$y_1y_2y_2$	$x, \bar{x}_2$	x_1x_2	$\bar{x}_1$	
000	⟨000⟩	1000	⟨000⟩	1
011	⟨011⟩	⟨011⟩	000	2
100	110	⟨100⟩	000	3
111	011	011	000	4
110	⟨110⟩	111	000	5

Example 3. Using Ref. [2], two original functions are considered as

$$Y_1 = (x_1 + x_1\bar{x}_5)x_2, \quad Y_2 = [(x_1 + y_1\bar{x}_5)]x_2x_6 + x_4 + x_{12} + x_7y_2]x_3\bar{x}_5.$$

We define the block transfer functions:

block I

$$f_{01} = \bar{x}_1x_2\bar{x}_5,$$
$$f_{11} = x_1x_2$$
$$f_{21} = \bar{x}_2 + \bar{x}_1x_5$$
$$f_{31} = 0$$

Table 10

f_{01}	f_{11}	f_{21}
⟨0⟩	1	⟨0⟩
⟨1⟩	⟨1⟩	0

block II

$$f_{02} = x_3\bar{x}_4\bar{x}_5x_7x_{12}(\bar{x}_2\bar{x}_6 + \bar{x}_1\bar{y}_1),$$
$$f_{12} = x_3\bar{x}_5/x_4 + x_{12} + x_1x_2x_6 + x_2x_6y_1),$$
$$f_{22} = \bar{x}_3 + x_5 + \bar{x}_4\bar{x}_7x_{12}(\bar{x}_2 + x_6 + \bar{x}_1\bar{y}_1 + \bar{x}_1x_5),$$
$$f_{32} = 0$$

Table 11

f_{02}	f_{12}	f_{22}
⟨0⟩	1	⟨0⟩
⟨1⟩	⟨1⟩	0

Compute overall system of input functions consisting of 9 products:

$$F_1 = f_{01}f_{02} = \bar{x}_1x_2x_3\bar{x}_4\bar{x}_5x_7\bar{x}_{12}(\bar{x}_6 + \bar{y}_1) = F^3 + F^3,$$
$$F_1' = f_{01}f_{02} = \bar{x}_1x_2x_3\bar{x}_4x_5\bar{x}_6x_7\bar{x}_{12}, \quad k_{21}^x = \bar{x}_1x_2x_3\bar{x}_4x_5x_7\bar{x}_{12},$$
$$F_2 = f_{01}f_{12} = \bar{x}_1x_2x_3\bar{x}_5(x_4 + x_{12} + x_6y_1) = F_2' + F_2'',$$
$$F_2' = f_{01}f_{12} = \bar{x}_1x_2x_3\bar{x}_5(x_4 + x_{12}), \quad k_{32}' = \bar{x}_1x_2x_3\bar{x}_5x_6;$$
$$F_3 = f_{01}f_{22} = \bar{x}_1x_2\bar{x}_5(\bar{x}_3 + \bar{x}_4x_6\bar{x}_7\bar{x}_{12} + \bar{x}_4\bar{x}_7\bar{x}_{12}y_1) = F_3' + F'',$$

$$F_3' = f_{01}f_{22} = \bar{x}_1x_2\bar{x}_5(\bar{x}_3 + \bar{y}_4\bar{x}_6\bar{x}_7\bar{x}_{12}),\ k_{33} = \bar{x}_1x_2\bar{x}_5x_4\bar{x}_7\bar{x}_{12};$$

$$F_4 = f_{11}f_{02} = x_1x_2x_3\bar{x}_4\bar{x}_5\bar{x}_6x_7\bar{x}_{12};$$

$$F_5 = f_{11}f_{02} = x_1x_2x_3\bar{x}_5(x_4 + x_6 + x_{12});$$

$$F_6 = f_{11}f_{22} = x_1x_2(\bar{x}_3 + x_5 + \bar{x}_4\bar{x}_6\bar{x}_7x_{12});$$

$$F_7 = f_{12}f_{02} = \bar{x}_2x_3\bar{x}_5 \cdot \bar{x}_4x_7\bar{x}_{12};$$

$$F_8 = f_{21}f_{12} = \bar{x}_2x_3\bar{x}_5(y_4 + x_{12});$$

$$F_8 = f_{21}f_{12} = \bar{x}_2x_3\bar{x}_5(y_4 + x_{12});$$

$$F_9 = f_{21}f_{22} = \bar{x}_2(\bar{x}_3 + x_5 + \bar{x}_4\bar{x}_7\bar{x}_{12}) + \bar{x}_1x_5;$$

$$k_{21}^x\, k_{32}^x \neq 0\,,\ k_{21}^x\, {}^x_{33} = 0\,,\ k_{32}^v\, 0_{33}^x \neq 0\,,\ k_{21}^y \cdot k_{32}^y = 0\,,$$

$$k_{21}^y\, k_{32}^y \neq 0\,,\ k_{32}^y\, k_{33}^y = 0\,.$$

and make the following transfer table based on this system.

Table 12

k_{21}^x	k_{32}^x	k_{33}^x	F_1'	F_2	F_3'	F_4	F_5	F_6	F_7	F_8	F_9
⟨00⟩	—	⟨00⟩	⟨00⟩	01	⟨00⟩	10	11	10	⟨00⟩	01	⟨00⟩
⟨01⟩	—	00	⟨01⟩	⟨01⟩	00	11	11	10	⟨01⟩	⟨01⟩	00
—	11	—	⟨10⟩	11	⟨10⟩	⟨10⟩	11	⟨10⟩	00	01	00
—	⟨11⟩	—	⟨11⟩	⟨11⟩	10	⟨11⟩	⟨11⟩	10	01	01	00

Define products $k_{21}^x\, k_{32}^x$ and $k_{32}^x k_{33}^x$:

$$k_{21}^x\, k_{32}^x = \bar{x}_1x_2x_3\bar{x}_4\bar{x}_5x_6x_7\bar{x}_{12},\ k_{32}^x\, k_{33}^x = \bar{x}_1x_2x_3\bar{x}_4\bar{x}_5x_6\bar{x}_7x$$

and obtain Table 13.

Table 13

$k_{21}^x\ k_{32}^x$	$k_{32}^x\ k_{33}^x$	F_1'	F_2'	F_3'	F_4	F_5	F_6	F_7	F_8	F_9
⟨00⟩	⟨00⟩	⟨00⟩	01	⟨00⟩	10	11	10	⟨00⟩	01	⟨00⟩
⟨01⟩	00	⟨01⟩	⟨01⟩	00	11	11	10	⟨01⟩	⟨01⟩	00
11	11	⟨10⟩	11	⟨10⟩	⟨10⟩	11	⟨10⟩	00	01	00
⟨11⟩	⟨11⟩	⟨11⟩	⟨11⟩	10	⟨11⟩	⟨11⟩	10	01	01	00

It is easy to see that the set of functions defined here —

$$F_1'\,F_2',\ F_3'\,F_4,\ F_5,\ F_6,\ F_7,\ F_8,\ F_9,\ k_{21}^x k_{32}^x\ k_{32}^x k_{33}^x \quad \text{and} \quad k_{32}^x k_{23}^x$$

coincides with the set of Ref. [2].

References

1. *Gavrilov, M. A.:* Synthesis of Transfer Tables by Generalized Input State Technique, "Automatics and Remote Control", No. 1, 1967.
2. *Gavrilov M. A. et al.:* Description and Analysis of Switching Circuits with Large Numbers of Input Variables, "Automatics and Remote Control", No. 10, 1969.
3. *Kuznetsov, O. P.:* Orthogonal Systems of Boolean Functions and Their Application to the Analysis and Synthesis of Logic Circuits: "Automatics and Remote Control", No. 10, 1970.
4. *Gavrilov, M. A.:* "Synthesis of state tables" Text-book for students by correspondance, 1964.

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

Budapest XI., Goldman György tér 3, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it.

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4—5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction, (outlining the problem), Methods and results, Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary — possibly in Russian if the paper is in English and *vice-versa* — should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10—15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которых должен содержать рисунки, таблицы и литературу) направляются

Е. Д. Теряеву — Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский Проспект, 14, Москва В-71, СССР

или

Я. Кощишу — Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета, Будапешт XI, площадь
Голдманн Д., 3, Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре.

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4—5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру гранок.

Статье должна предшествовать аннотация объемом до 50—100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме — на русском языке, если статья написана на английском, и наоборот — должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем разюме не должен превышать 10—15 % объема статьи.

Авторам высылаются гранки статьи, которые они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Petrov, B. N.—Popov, E. P.: Fundamental problems of technical plants control science (*Петров Б. Н.—Попов Е. П.:* Основные проблемы науки об управлении техническими объектами) 3

Ядренко М. И.: Об оптимальных линейных оценках математического ожидания и коэффициентов регрессии однородных и изотропных случайных полей (*Yadrenko, M. I.:* On optimal linear estimates of mean value and regression coefficients of homogeneous and isotropic random field) 17

Лаговиер Б. А.: Метод анализа асинхронных автоматов (*Lagovier, B.:* An analysis technique for asynchronous automata) 27

Somló, J.: On the computer aided design of control systems (*Шомло Я.:* Проектирование систем автоматического регулирования при помощи ЦВМ) 41

Rejtő, L.—Révész P.: Density estimation and pattern classification (*Рейто Л.—Ревес П.:* Оценка плотности вероятностей и классификация образов) 67

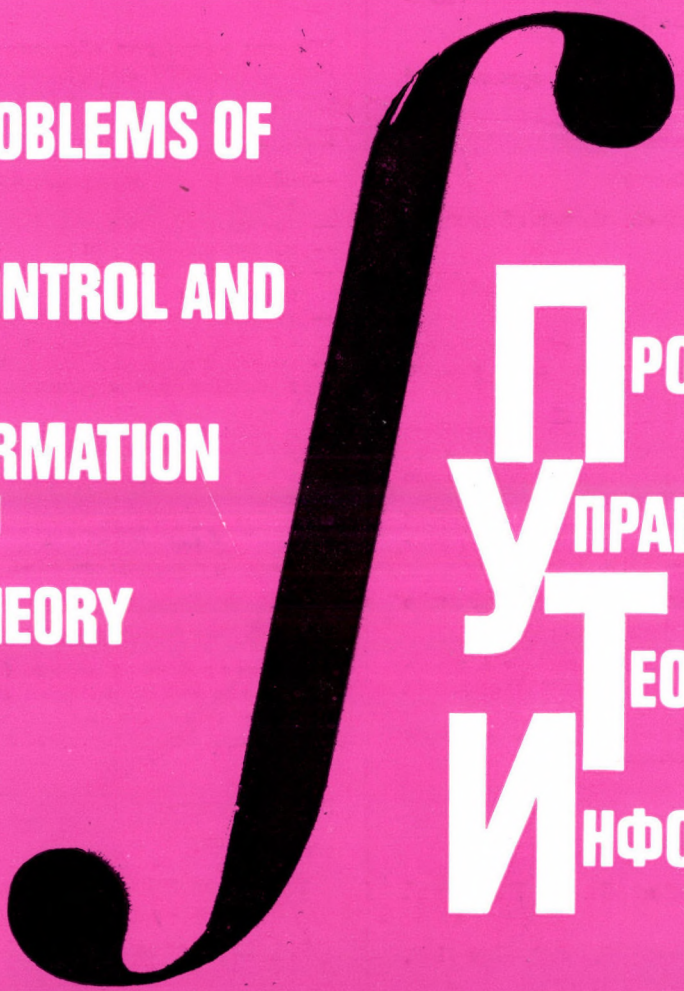
Fritz, J.: On a pattern classification algorithm of R. L. Kashyap (*Фриц Й.:* Об алгоритме распознавания образов по Р. Л. Кашиап) 81

316.920

VOL. 2 * NUMBER 2
TOM ТОМ НОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY



ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1973

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

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

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

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

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

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

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

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

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

H-1363 Budapest, P.O. Box 24, H. ngary

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал создан для обеспечения быстрейшей публикации материалов о новейших результатах научных исследований, разработок и кратких сообщений о достижениях в следующих областях науки:

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

Журнал издается четыре раза в год общим объёмом около 320 печатных страниц.

Журнал в первую очередь будет публиковать научные достижения обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

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

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии Наук Венгрии

H-1363 Будапешт, п. о. 24, Венгрия

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

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

DEPUTY EDITORS

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

CO-ORDINATING EDITORS

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

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

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

B. S. SOTSKOV (Moscow)
L. VARGA (Budapest)

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

РЕДАКТОРЫ ЖУРНАЛА

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

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

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

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

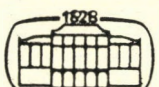
С. В. ЕМЕЛЬЯНОВ (Москва)
Л. КАЛМАР (Будапешт)

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

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

Б. С. СОТСКОВ (Москва)
Л. ВАРГА (Будапешт)

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



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI KIADÓ, BUDAPEST

Problems of Control and Information Theory, Vol. 2 (2), pp. 95–105 (1973)

THE OPTIMIZATION OF MULTIVARIABLE SYSTEMS AND MARKOVIAN PROGRAMMING

M. V. MEEROV, B. L. LITVAK

(Moscow)

(Received October 5, 1972)

1. Introduction

Refs [2–4] have shown that the problems of statical optimization for multivariable systems [1] are reducible to a special form of linear programming problems. The specifics of these problems is in that the constraint matrix has a special structure so that each column and each row contains at most one positive element, all other elements are non-positive. Such matrices are referred to as matrices of the required form or *RF* matrices. In a number of cases when multivariable systems are optimized [2] linear programming problems are also used whose constraint matrix consists of a series of submatrices of a specified form.

The mathematical tool of Markovian processes with rewards [5, 6] (the so-called Markovian programming) can be effectively used to solve certain *dynamic* problems in the optimization of multivariable systems. In this case the optimization problem is reducible to a linear programming problem. The constraint matrix of the linear programming problem thus obtained has the same specifics of signs as in the case of statical problems in the optimization of multivariable systems. The Markovian programming problems are of great dimensionality. The constraint matrix is as a rule filled with nonzero elements. Therefore it is very important to use the specific features of the problem.

1. The statement of the problem

Let us express the equations of a stationary ergodic uniform Markovian process with discrete time

$$\pi_j = \sum_{i=1}^{N} \pi_i P_{ij}, \qquad j = 1, 2, \ldots, N \tag{1}$$

$$\sum_{i=1}^{N} \pi_i = 1 \tag{2}$$

where π_j is the limit probability of reaching the j-th state; N is the total number of states (N is a finite number); P_{ij} is the probability of transition from the i-th state into the j-th state within one period of time.

For a controlled Markovian process with the possibility of using mixed strategies Eqs (1), (2) will be

$$\sum_{k \in K_j} \pi_j^k - \sum_{k \in K_i} \sum_{i=1}^{N} \pi_i^k P_{ij}^K = 0 \qquad j = 1, \ldots, N \tag{3}$$

$$\sum_{k \in K_i} \sum_{i=1}^{N} \pi_i^k = 1 \tag{4}$$

$$\pi_i^k \geqq 0, \qquad i = 1, 2, \ldots, N, \qquad k \in K_i$$

where k is the index of strategy;

K_i is a set of indices of possible behaviour strategies in the i-th state (a set of strategies with the indices $k \in K$ can be specified by a discrete number of points or by a convex polyhedron).

The maximization condition of average rewards within one period of an infinite duration process can be expressed as

$$\sum_{k \in K_j} \sum_{j=1}^{N} q_j^k \pi_j^k \to \max \tag{6}$$

where q_j^k is the reward gained due to the transition from the j-th state into any other one when choosing to K-strategy. The problem of optimizing a system described by the Markovian process under consideration consists [5, 6], in selecting for each state of the process, those behaviour strategies which permit to obtain the maximal average rewards at a sufficiently long (theoretically infinite) process and is formulated as a linear programming problem (3)—(6).

The problem can also be in considering the reevaluation coefficients β for future rewards relative to the present. The coefficient β is also considered as the probability that the process will be continued. It is easily shown that the problem of optimizing the Markovian processes with rewards in case of reevaluation is in finding the optimal feasible solution to the following linear programming problem

$$\sum_{k \in K_j} \sum_{j=1}^{N} q_j^k \pi_j^k \to \max$$

at

$$\sum_{k \in K_j}^{N} \pi_j^k - \beta \sum_{k \in K_i} \sum_{i=1}^{N} \pi_i^k P_{ij}^k = \varepsilon_j \tag{7}$$

$$\pi_i^k \geqq 0, \qquad i = 1, 2, \ldots, N, \qquad k \in K_i$$

where ε_j are arbitrary positive (or nonnegative) numbers.

The indices k of the variables π_i^k included in the optimal feasible solution of the problem (7) correspond to the ordinal numbers of optimal strategies for each i-th state. The values of the variables π_i^k themselves in the problem (7) have no physical sense.

The equations of the process (1) show that the matrix $(E - P)$ used in describing a Markovian process belongs to RF matrices[1] (here ε is a unitary matrix, $P = [P_{ij}]$ is a stochastic matrix). The constraint matrix of problem (7) consists of series of RF matrices. Apart from a series of matrices, problem (3)—(6) has one more row (4). The characteristic form of the right hand part of the conditions (3) permits to use all useful properties of problem (7) to solve problem (3)—(6). The properties of problem (7) are maintained also in the case if it is required to minimize, not maximize, the objective function, in other words the objective function coefficients signs can have any signs. The signs of constraints are not fixed "$=$", "$\geq$", "$\leq$" since the conditions "$\leq$" or "$\geq$" are reduced to $=$ by introducing a matrix of additional variables which is a particular case of RF matrices. The only assumption additional to the specifics of signs in the matrix is the nonnegativeness of the vector in the right-hand part of the constraints. The number of rows "n" in RF matrices is not necessarily equal to that of columns "m".

We have described in some details certain modifications of the linear-programming problem (7) for the sole reason that such problems whose constraint matrix consists of a series of RF matrices do not arise in Markovian programming alone. Such problems arise, for instance, in statical optimization of nonlinear multivariable systems, in using a generalized Leontieff model, etc.

For simplicity let us combine the stages of finding the initial feasible plan with finding the optimal plan. For this purpose it is sufficient to introduce artificial variables x_j^p with objective function coefficients of sufficiently large magnitude into the problem and take the initial feasible solution $x_j^0 > 0$. Let us assume that we have completed these operations and obtained the following linear programming problem

$$\sum_{i=1}^{m} \sum_{k \in K_i} c_i^k x_i^k \to \max$$

at

$$\sum_{k \in K_j} a_{ji}^k x_k^j - \sum_{i=1}^{m} \sum_{k \in K_i} a_{ij}^k x_i^k + x_j^0 = b_{0j} \tag{8}$$

where $x_i^k \geq 0, \quad x_j^0 \geq 0, \quad i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n,$

$$b_j \geq 0, \qquad k \in K_i.$$

[1] A special matrix termed here a RF matrix is sometimes referred to as a Metzner matrix (or M matrix) if it is square as in our case.

2. Solution of the problem

Before proceeding to describe the solution procedures for the problem we will record several of its properties which are easy to be proved.

Property 1. A qualitative solution of problem (8) does not depend on the magnitude of the components in the right hand part of the constraints.

A qualitative solution means in this case the determination of the vectors which are included into the optimal feasible solution.

Property 2. Any initial feasible to problem (8) contains just one positive variable x_i^k for each i, where $k = 0$ or $k \in K_i$.

Note that these properties can be updated for different versions of the problem in different ways. Ref. [4] describes such properties for problems which arise in optimizing nonlinear multivariable systems. Ref. [6] cites Property 2 for the problems (3)—(6). Property 2 for Markovian programming problems is also known as the property of optimality of pure strategies.

The advantages of the specifies of the problem are as follows. Property 1 permits the decomposition of the problem into two stages, qualitative solution (finding the optimal feasible solution) and quantitative solution (finding the values of variables in the optimal point). Property 2 permits the decomposition of the problem into "m" parts individual treatment of each i-th block of the matrix columns ($i = 1, 2, \ldots, m$). Each i-th group of columns can be given implicitly as a certain set determined by a convex polyhedron. Property 2 permits a simplification of the finite procedures used in solving the problem, and the construction of finite and iterational solution algorithms not involving a transformation of the initial matrix.

A simplex method procedure with using special properties of the problem at hand differs in that the selection of the vector introduced into the feasible solution can be performed for each i-th group of the matrix columns separately. This decomposition as applied to problems (3)—(6) is described in [6].

In solving the above problems the number of iterations is effectively reduced by simultaneous introduction of vector blocks into the feasible solution. This is convenient in our case because there is no need in finding vectors taken out of the feasible solution when a block of vectors is introduced there. A vector of the i-th group can oust only vectors of the same i-th group. In case of simultaneous introduction of the vectors A_i^l into the feasible solution the simplex method algorithm can change only in that the directing element will be, instead of the element a_{li}^l, the matrix composed of the l-th elements of the vectors A_i^l in transition to a new feasible solution.

The disadvantage of these algorithms is the need in transforming the matrix at each iteration which increases the number of nonzero elements, accumulation of computation errors, etc. Using the main algorithm of the simplex method, the entire constraint matrix is transformed, while using a modified simplex method the submatrix of feasible solution vectors is transformed and thus at each iteration the inverse matrix of the basis is determined. The inverse matrix is needed in these cases only to determine the rows of dual estimates λ

$$\lambda = C_\delta A_\delta^{-1}$$

where C_δ is the vector of the objective function coefficients which correspond to feasible solution variables; A_δ is the feasible solution matrix.

The vector λ can, however, be determined without computing the inverse matrix by solving the system

$$\lambda A_\delta = C_\delta. \tag{9}$$

Note that the method of Howard [5] who used quite a different approach to optimize Markovian processes with rewards in terms of linear programming is a modification of the simplex method as applied to this kind of problems with using by blocks introduction of vectors into the feasible solution and system (9) is solved instead of finding the inverse matrix.

All these procedures are finite in terms of linear programming. Iterational algorithms are also applicable.

The idea of the iterational algorithm is as follows. In the finite procedure described above a system of linear algebraic equations (9) was to be solved at each step. This system is solved in the final analysis in order to determine whether there are vectors which are preferable for the introduction into the feasible solution (in terms of the objective function increment) as compared with those which are the components of the feasible solution matrix A_δ of the given iteration. Consequently, system (9) can be solved only till the point of finding the preferable vectors. As soon as they appear, these vectors A_i^l should replace the feasible solution vectors of the system (9) with the same indices i and continue the solution of the system for the matrix A_δ thus modified.

In this way the replacement of feasible solution vectors will take place while solving the system (9). The process ends when this system is solved completely and there are no "preferable vectors".

The algorithm for finding the optimal feasible solution of the problem (8) is as follows:

1) A system, (9) is constructed of columns A_i^l with the estimates $\Delta_i^l = \min \Delta_i^k < 0$ where

$$\Delta_i^k = -\sum_{\substack{j=1 \\ j\neq 1}}^{n} a_{ij}^k \lambda_i + a_{ii}^k \lambda_i - c_i^k \quad i = 1, 2, \ldots, m \quad k \in K_i \tag{10}$$

$$\Delta_i^k = \lambda_i \qquad i = j, \; k = 0. \tag{11}$$

At the first iteration $\lambda = 0$ and $\Delta_i^k = -C_i^k$ for $k \in K_i$. Note also that if a set of columns with the indices $k \in K_i$ is not finite while A is a convex polyhedron, then at subsequent iterations the linear programming subproblems are to be used if Δ_i^l is to be found, the subproblems consisting in minimizing the objective function (10) for the set K_i. The subproblems are solved separately.
System (9) has the form

$$-\sum_{\substack{j=1 \\ j=i}}^{n} a_{ij}^l \lambda + a_{ii}^l \lambda_i - c_i^l = 0, \qquad i \in s \tag{12}$$

$$\lambda_i = 0, \qquad i \in s, \; l = 0$$

where s is the set of indices i for which there are negative estimates of Δ_i^k, $k \in K_i$.

2) Complete several steps of solving system (12). If the method of simple iteration is used, the following values of the variables will be obtained at each step

$$\lambda_i^{(p)} = \frac{1}{a_{ii}^l}\left(\sum_{j=1}^{n} a_{ij}^l \lambda_j^{(p-1)} + c_i^l\right), \qquad i \in s \tag{13}$$

$$\lambda_i^{(p)} = 0 \qquad i \notin s$$

where $\lambda_0^{(p)} = 0$

3) Compute the values $\Delta_i^m = \min \Delta_i^k$ where Δ_i^k is found from Eqs (10) and (11) at $\lambda = \lambda(p)$.

If the elements $\Delta_i^m < \Delta_i^l < 0$ appear then we return to $n - a$ and replace the i-th rows in the system (12), otherwise we continue solving the system by the iterational formula (13) from the value of 0 until a solution of the required accuracy is obtained.

Note that if a row containing the element $a_{ii}^l < 0$ appears in the system (12), this would mean that the objective function is not constrained.

Consequently, the above algorithm for the problem (8) is reducible to the iterational procedure (13) with the replacement of rows while solving.

This iterational algorithm is an extension of the existing iterational methods of solving systems of linear algebraic equations to the case of linear programming problems of the class what we are concerned with and preserves

all the advantages of these iterational methods (the matrix is not transformed in solution, the algorithm is simple, the random computation errors are smoothed, etc.). On the other hand, the iterational algorithm for the problem at hand preserves the disadvantages of common iterational procedures, the most important one of which is the narrow field of application (the convergence conditions should be met and the race of convergence should be reasonable). Note, however, that in comparison with the iterational methods used nowadays to solve the general problem of linear programming (for instance, by methods involving infinite iterational processes of solving matrix games) these iterational methods are beyond any doubt more developped and effective in the problems where they are applicable.

The conditions for the convergence of the iterational method in the problems at hand can e.g. be as follows: for columns of problem (8) having $a_{ii}^k > 0$

$$|a_{ii}^k| > \left| \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}^k \right| \qquad \begin{matrix} i = 1, 2, \ldots, m \\ k = 1, 2, \ldots, k_i . \end{matrix}$$

Let us now proceed to the justification of the algorithm described.

In order to solve problem (8) the subsystem of conditions (12) should be isolated from conditions (10), (11); this should be done such that the vector λ^* obtained by solving the system (12) should satisfy all the other conditions

$$\Delta_i^k \geqq 0; \qquad i = 1, 2, \ldots, m, \qquad k \in K_i, \qquad k = 0 . \tag{14}$$

In the suggested algorithm system (12) consists at each p-th step of the conditions (10), (11) having a minimal value $\Delta_i^k(\lambda^{(p)})$ for each i.

Therefore if this system has been solved completely (the value $\lambda^{(p)} = \lambda^*$ has been obtained at which $\Delta_i^l(\lambda^*) = 0$) then all other conditions (14) are observed

$$\Delta_i^{k_i}(\lambda^*) \geq \Delta_i^l(\lambda^*) = 0 .$$

Consequently, if the exact solution λ^* of the system (12) is obtained, the initial problem has been solved.

It is required to show that if the convergence conditions are met, then the suggested procedure converges, following all replacements of rows in system (12), to an exact solution. For this purpose it is sufficient to verify the monotonic increase of the vector $\lambda^{(p)}$ at each step of the iteration.[2]

[2] Monotonic increase in this case signifies that at each step of the procedure at least one component of the vector $\lambda^{(p)}$ increases while others do not decrease.

Indeed, the values of the components of the vector $\lambda^{(p)}$ cannot grow indefinitely, since the observance of the convergence conditions makes the process converge to the unique solution of system (12) composed of any rows of Eqs (10), (11). The rule for the termination also means that the iterational process (13) ends as soon as at a certain step no component of the vector $\lambda^{(p)}$ changes its values within the specified computational accuracy.

Consequently, in order to justify the algorithm, it is necessary to make sure that the vector $\lambda^{(p)}$ increases monotonically ot the steps of the solution. The procedure for determining the indefiniteness of the objective function also needs a justification.

Theorem 1. For the vector $\lambda^{(p)}$ to increase monotonically at the steps of the iterational process (13), it is sufficient to select the initial point $\lambda^{(0)}$ to satisfy the conditions

$$-\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}^{l}\,\lambda_j^{(0)} + a_{ii}^{l}\,\lambda_i^{(0)} - c_i^{l} \leq 0. \tag{15}$$

The proof is given in the Appendix.

Note that the vector of the initial point $\lambda^{(0)}$ satisfying the conditions (15) is in this case evidently $\lambda^{(0)} = 0$. The conditions (15) at this $\lambda^{(0)}$ are satisfied since, by virtue of its structure, system (12) contains all $c_i^{l} > 0$.

Theorem 2. The objective function of problem (8) is infinite on the set in discussion, if at a certain step of the iterational process (15) the element $\Delta_i^k(\lambda^{(p)}) < 0$ appears for a row of system (10) having $a_{ii}^{k} < 0$.

In order to prove Theorem 2 it is sufficient to make sure that if $\Delta_i^k < 0$ at the given $\lambda^{(p)}$, then the value Δ_i^k will remain negative even at an exact solution of system (12) which follows immediately from (10) for rows having $a_{ii}^{k} < 0$ at monotonic increase of λ.

To conclude let us note that the iterational algorithm of problem (8) can be obtained not only through the procedure (13) which is similar to that of solving a system of linear equations by the method of simple iteration but also through any other method to solve systems of linear algebraic equations provided that this method ensures a monotonic increase of the vector λ in the process of solution.

Appendix

Proof of Theorem 1

Assume that we have the vector $\lambda^{(0)}$ satisfying the conditions (15) or

$$\lambda_i^{(0)} \leq \frac{1}{a_{ii}^{l}}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}^{l}\,\lambda_j^{(0)} + c_i^{l}\right); \qquad i \in s.$$

At the same time from Eq. (13)

$$\lambda_i^{(1)} = \frac{1}{a_{ii}^l}\left(\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}^l \lambda_j^{(0)} + c_i^l\right); \qquad i \in s.$$

Comparing these two expressions we have

$$\lambda_i^{(1)} \geq \lambda_i^{(0)} \tag{16}$$

The conditions (15) are not only sufficient for Eq. (16) to hold but necessary as well. Indeed, considering all the above relations in the reverse order we can see that if conditions (16) are observed, inequalities (15) are valid.

Consequently, we have thus far only confirmed the fact that the conditions (15) are necessary and sufficient for the vector $\lambda^{(p)}$ to grow monotonically at the first step of the iteration process. Note that this is true in solving, by the iteration method, any system of linear algebraic equations irrespective of its specific properties. The maintenance of the monotonicity on the increase of the vector $\lambda^{(p)}$ at the subsequent steps of the iteration process is ensured by the specifics of the system in question.

Let us describe the relation of the vectors $\lambda^{(p+1)}$ and $\lambda^{(p)}$ for any step of the iterational process (13). Let us consider separately the case where the coefficients a_{ii}^l, c_i^l in (13) remain the same at the p-th and the $(p+1)$-th step, and the case where the coefficients at the $(p+1)$-th step are replaced by others.

Let us start with the case where the coefficients a_{ii}^l, c_i^l are not changed at the $(p+1)$-th step as compared with the p-th step.

Then from Eq. (13)

$$\lambda_i^{(p+1)} - \lambda_i^{(p)} = \frac{1}{a_{ii}^l}\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}^l(\lambda_j^{(p)} - \lambda_j^{(p-1)}); \qquad i \in s.$$

Thence, if $\lambda_i^{(p)} \geq \lambda^{(p-1)}$ then $\lambda_i^{(p+1)} \geq \lambda_i^{(p)}$ since $a_{ij}^l \geq 0$ and $a_{ii}^l \geq 0$.

Consequently, in this case for monotonic growth of the vector $\lambda^{(p)}$ at the steps of the iterational process (13) it is sufficient that $\lambda^{(p)}$ grows at least at one preceding step. On the other hand, the increase of the vector $\lambda^{(p)}$ is ensured at the very first step of the process if the initial point $\lambda^{(p)}$ satisfies the condition (15).

Let us proceed now to the case where the coefficients a_{ii}^l and c_i^l are replaced at the (p + 1)-th step by the algorithm suggested.

Two remarks are required now. Firstly, by analogy with Eq. (16) it is easy to show that if $\lambda^{(p)}$ satisfies the inequalities

$$\Delta_i^l(\lambda^{(p)}) = -\sum_{\substack{j=1 \\ j\neq}}^{n} a_{ij}^l \lambda_{ij}^{(p)} + a_{ii}^l \lambda_i^{(p)} - c_i^l \leqq 0 \qquad i \in s. \tag{17}$$

then

$$\lambda_i^{(p+1)} \geq \lambda_i^{(p)} \tag{18}$$

and conversely, if conditions (18) are satisfied, then $\lambda^{(p)}$ satisfies the inequalities (17). Thence the second remark follows: if the coefficients a_{ii}^l, c_i^l do not change at the steps of the iterational process (see the case considered above), then

$$\Delta_i^l(\lambda)^{(p)} \leqq 0; \quad i \in s \quad \text{for any} \quad p. \tag{19}$$

Assume that by the above algorithm we replace at the $(p+1)$-th step the i-th row of system (12) by a row for which

$$\Delta_i^k(\lambda^{(p)}) < \Delta_i^l(\lambda^{(p)}); \qquad i \in s$$

or

$$\Delta_i^k(\lambda^{(p)}) < \Delta^{(0)}(\lambda^{(p)}); \quad i \notin s.$$

Then considering (19) and the fact that $\Delta_i^{(0)} = 0$ we will have $\Delta_i^k(\lambda^{(p)}) < 0$; $i \in s$ where the set s is expanded at the expense of i-th rows having $\Delta_i^k(\lambda^{(p)}) < \Delta^{(0)}(\lambda^{(p)})$. Thence, by virtue of Eqs (17)—(18), $\lambda^{(p+1)} \geq \lambda^{(p)}$ in this case too. The monotonic growth of the vector $\lambda^{(p)}$ has been proved for all cases.

References

1. *Мееров М. В.* Системы многосвязного регулирования. Наука, М., 1965.
2. *Литвак Б. Л.* К вопросу оптимизациионого класса объектов многосвязного регулирования. Доклад на 1У Всесоюзном совещании по автоматическому управлению (технической кибернетике), М., 1968.
3. *Мееров М. В., Литвак Б. Л.* К вопросу оптимизации сложных многосвязных систем большой размерности. Доклады АН СССР, **187**, 5, 1969.
4. *Мееров М. В., Литвак Б. Л.* Статическая опимизация систем многосвязного регулирования большой размерности I, II, III. Автоматика и телемеханика, 3, 4, 5, 1970.
5. *Ховард Р. А.* Динамическое программирование и марковские процессы. "Советское радио", М., 1964.
6. *Вольф Ф., Данциг Дж.* Марковские цепи и линейное программирование. Кибернетический сборник (новая серия), 4, "Мир", М., 1967.

Оптимизация многосвязных систем и марковское программирование

М. В. МЕЕРОВ—Б. Л. ЛИТВАК

(Москва)

Как известно, задачи статической оптимизации многосвязных систем могут быть сведены к задачам линейного программирования специального вида. Для решения некоторых динамических задач оптимизации многосвязных систем в ряде случаев представляется целесообразным использование аппарата теории марковских процессов с доходами (марковского программирования). Весьма интересно при этом, что задачи марковского

программирования сводятся к задачам линейного программирования, обладающим такой же особенностью, как и задачи статической оптимизации многосвязных систем. Особенность эта заключается в том, что матрица ограничений обладает специальной структурой знаков элементов, а именно: каждый столбец матрицы содержит не более одного положительного элемента, все остальные элементы неположительные. Использование специфики задачи позволяет предложить некоторые эффективные алгоритмы ее решения.

В работе прежде всего приведены различные постановки задач марковского программирования в терминах линейного программирования. Далее рассматриваются свойства задачи, которые позволяют, во-первых, проводить качественное решение (определение оптимального базиса) независимо от величины неотрицательных компонент правой части ограничений и, во-вторых, разбить задачу на части по столбцам и рассматривать при решении отдельно блоки столбцов. Указывается, как использовать данные свойства при применении для решения задачи конечных методов линейного программирования.

Основное внимание в алгоритмической части работы уделено рассмотрению итерационных алгоритмов решения задачи. Идея предлагаемых итерационных алгоритмов заключается в следующем. В конечных процедурах на каждом шаге по сути дела решается система линейных алгебраических уравнений, которая позволяет ответить на вопрос, имеются ли вектора, более предположительные для введения в базис (в отношении величины приращения целевой функции) чем те, из которых составлена базисная матрица данной итерации. В рассматриваемых задачах систему можно решать не до конца, а только до выявления более предположительных для введения в базис векторов.

Производя замену базисных векторов непосредственно в процессе решения системы, одновременно с получением решения системы, будет получено решение всей задачи. В работе приводятся итерационные алгоритмы, являющиеся обобщением известных итерационных схем решения систем линейных алгебраических уравнений. Для обоснования предлагаемых алгоритмов доказываются соответствующие теоремы.

М. В. Мееров
Институт проблем управления
СССР, Москва В—485,
Профсоюзная ул., 81

Problems of Control and Information Theory, Vol. 2 (2), pp. 107—122 (1973)

АСИМПТОТИЧЕСКИЕ ВЫРАЖЕНИЯ ДЛЯ ИНФОРМАЦИИ В КАНАЛЕ С РАЗРЫВНОЙ ПЛОТНОСТЬЮ ШУМА

М. В. БУРНАШЕВ, Р. З. ХАСЬМИНСКИЙ

(Москва)

(Поступила в редакцию 15 февраля 1972 г.)

Для канала с разрывной плотностью шума исследовано асимптотическое поведение величины информации при малой мощности входного сигнала.

1. Введение

В настоящей работе изучаются асимптотические выражения для средней взаимной информации между входом и выходом аддитивного канала без памяти, в котором на каждый входной сигнал S_i накладывается «шум» ξ_i так, что сигнал на выходе η_i находится по формуле

$$\eta_i = S_i + \xi_i . \tag{1.1}$$

Здесь $\xi_1, \ldots, \xi_n, \ldots$ — независимые одинаково распределенные случайные величины с плотностью $p(x)$ относительно лебеговой меры на прямой. Случай, когда $p(x)$ является абсолютно непрерывной функцией и для нее существует информационное количество Фишера, т. е.

$$\int_{-\infty}^{\infty} p(x) \left| \frac{p'(x)}{p(x)} \right|^2 dx = I < \infty , \tag{1.2}$$

рассматривался в [1, 2]. Например, если входной сигнал удовлетворяет следующему «пиковому» ограничению

$$|S - MS| < K \sqrt{DS} ,$$

где K — произвольная, но не зависящая от S постоянная, то для информации $J(S, \eta)$ справедлива асимптотическая формула [2]

$$J(S, \eta) = \frac{1}{2} J \cdot DS + o(DS), \qquad (DS \to 0). \tag{1.3}$$

Интересно изучить асимптотическое поведение $J(S, \eta)$, если для плотности $p(x)$ нарушается одно из вышеприведенных условий. В работе [3] рассматривался частный случай, когда $p(x)$ является абсолютно-непрерыв-

ной функцией, но имеет конечное число «нерегулярных» точек $x_1, \ldots, x_n$, в окрестности которых

$$p(x)\left|\frac{p'(x)}{p(x)}\right|^2 \sim C_k |x - x_k|^{-1}.$$

В данной статье рассматривается случай, когда плотность $p(x)$ имеет конечное число разрывов первого рода. При этом оказывается, что главный член величины информации в выходном сигнале относительно слабого входного представляет собой довольно простой функционал от распределения входного сигнала, зависящий от предельных значений плотности $p(x)$ в окрестности точек разрыва как от параметров. Из этого выражения для информации нетрудно получить (2) и асимптотические выражения для пропускной способности данного канала при различных ограничениях.

Некоторые оценки пропускной способности для разрывных плотностей частного вида были доказаны в [1]. Основной результат (1.4—1.6) для частного случая приведен без доказательства в [4].

Чтобы описать полученный результат более точно, для произвольных $p \geq 0$, $q \geq 0$ и случайной величины S с функцией распределения $F_S(x)$ введем функционал $B_{p,q}(S)$ формулой

$$B_{p,q}(S) = \int \Phi_{p,q}(F_S(x))\, dx,$$

где

$$\Phi_{p,q}(y) = p \ln p \cdot y + q \ln q(1-y) - [py + q(1-y)] \ln [py + q(1-y)] \quad (1.4)$$

и считается, что $0 \cdot \infty = 0$.

Так как $\Phi_{p,q}(y) \geq 0$ при $0 \leq y \leq 1$, то $B_{p,q}(S) \geq 0$, причем $B_{p,q}(S) = 0$, если S — детерминированная величина или $p = q$. Если плотность $p(x)$ имеет r разрывов 1-го рода, причем q_k и p_k суть предельные значения плотности в точках разрыва слева и справа соответственно, то при некоторых ограничениях на класс входных распределений будет доказана асимптотическая формула

$$J(S, \eta) = B(S) + 0((M|S|^\gamma)^{1/\gamma}) \quad \text{при} \quad M|S|^\gamma \to 0, \quad (1.5)$$

где

$$B(S) = \sum_{k=1}^{r} B_{p_k, q_k}(S), \quad \gamma > 1. \quad (1.6)$$

Таким образом, выражения главного члена информации для гладких и разрывных плотностей существенно различны. Так, например, очевидно, что для входного сигнала вида $S_\varepsilon = \varepsilon S$, где S — фиксированная случайная величина с конечной дисперсией

$$B(S_\varepsilon) = \varepsilon \cdot B(S)$$

и, таким образом, главный член информации пропорционален не дисперсии, как в гладком случае, а корню из нее. (Первым этот факт для частного случая разрывной плотности отметил Прелов в [1].)

Р. Л. Добрушин поставил вопрос об устойчивости приведенного результата по отношению к слабым дополнительным шумам в канале. В связи с этим доказываем, что формулы (1.5), (1.6) сохраняются, если мощность дополнительных шумов по порядку величины меньше входного сигнала (лемма 2.1).

2. Основная теорема

Сформулируем теперь предположения, при которых будет доказан приведенный выше результат.

Обозначим через $\mathfrak{K}$ канал без памяти с независимым аддитивным шумом, плотность которого $p(x)$ удовлетворяет следующим условиям:

1) $p(x)$ непрерывна всюду, за исключением конечного числа точек $x_1, \ldots, x_r$, и существуют односторонние конечные пределы

$$p(x_k - 0) = q_k, \quad p(x_k + 0) = p_k; \quad k = 1, \ldots, r. \tag{2.1}$$

2) для некоторого $b > 1$ существуют интегралы

$$\int_{x_k}^{x_{k+1}} p(x) \left| \frac{p'(x)}{p(x)} \right|^b dx < \infty, \quad k = 0, \ldots, r; \; x_0 = -\infty, \; x_{r+1} = +\infty. \tag{2.2}$$

Функционал $B(S)$, определенный формулами (1.4), (1.6), играет в дальнейшем основную роль. Изучим в связи с этим некоторые его свойства.

1. Для любого h справедливо равенство

$$B(S) = B(S + h). \tag{2.3}$$

2. Для любого $\gamma > 1$ и любой постоянной h справедливо неравенство

$$B(S) \leq C(p_1, \ldots, p_r, q_1, \ldots, q_r) \cdot C(\gamma) \cdot (M\,|S - h|^{\gamma})^{1/\gamma}. \tag{2.4}$$

3. Если $\max\limits_{i=1,\ldots,r} |p_i - q_i| > 0$, то для некоторой положительной постоянной $C = C(p_1, \ldots, p_r, q_1, \ldots, q_r)$

$$B(S) \geq C \cdot M\,|S - \operatorname{med} S|, \tag{2.5}$$

где $\operatorname{med} S$ — медиана случайной величины S.

Первое свойство очевидно. Перейдем к доказательству второго и третьего свойств. Обозначив $C_k = \dfrac{(p_k - q_k)^2}{\max(p_k, q_k)}$, получим из (1.4) неравенства

$$\frac{1}{2} C_k y(1-y) \leq \Phi_{p_k, q_k}(y) \leq C_k \left[y \ln \frac{1}{y} + (1-y) \ln \frac{1}{1-y} \right].$$

Поэтому

$$B(S) \leq \sum_{k=1}^{r} C_k \left\{ \int_{-\infty}^{\infty} F(x) \ln \frac{1}{F(x)} dx + \int_{-\infty}^{\infty} [1 - F(x)] \ln \frac{1}{1-F(x)} dx \right\}. \quad (2.6)$$

Пусть теперь для некоторого $\gamma > 1$

$$\int |x|^{\gamma} dF(x) < \infty.$$

Тогда, как известно, при $x \to \infty$

$$x^{\gamma_1} [F(-x) + 1 - F(x)] \to 0 \quad (2.7)$$

для любого $\gamma_1 < \gamma$.

Ограничиваясь для определенности первым интегралом в (2.6), получим, учитывая (2.7),

$$\int F(x) \ln \frac{1}{F(x)} dx = x F(x) \ln \frac{1}{F(x)} \Big|_{-\infty}^{\infty} - \int x \ln \frac{1}{F(x)} dF(x) + \int x \, dF(x) \leq$$

$$\leq \int |x| \ln \frac{1}{F(x)} dF(x) + \int |x| \, dF(x).$$

Применяя теперь неравенство Гельдера и обозначая $\Gamma(x)$ — гамма-функцию Эйлера, находим

$$\int |x| \, dF(x) + \int |x| \ln \frac{1}{F(x)} dF(x) \leq \left[\int |x|^{\gamma} dF(x) \right]^{1/\gamma} \cdot$$

$$\cdot \left\{ \left[\int \left(\ln \frac{1}{F(x)} \right)^{\gamma/\gamma - 1} dF(x) \right]^{\gamma - 1/\gamma} + 1 \right\} = \left[\int |x|^{\gamma} dF(x) \right]^{1/\gamma} \cdot$$

$$\cdot \left\{ \left[\int_0^1 \left(\ln \frac{1}{y} \right)^{\gamma/\gamma - 1} dy \right]^{\gamma - 1/\gamma} + 1 \right\} = \left[\int |x|^{\gamma} dF(x) \right]^{1/\gamma} \cdot \left[\Gamma^{\gamma - 1/\gamma} \left(\frac{\gamma}{\gamma - 1} + 1 \right) + 1 \right].$$

Соотношение (2.4) доказано. Неравенство (2.5) следует из оценки

$$B(S) \geq \frac{1}{2} \sum_{k=1}^{r} C_k \int_{-\infty}^{\infty} F(x) [1 - F(x)] \, dx \geq \frac{1}{4} \sum_{k=1}^{r} C_k \cdot M |S - \operatorname{med} S|.$$

Обозначим $\|S\|_\gamma = (M|S|^\gamma)^{1/\gamma}$ и условимся писать $\psi = 0(\|S\|_\gamma)$, если $\sup_{\|S\|_\gamma \le \varepsilon} \frac{\psi}{\|S\|_\gamma} \to 0$ при $\varepsilon \to 0$.

Теорема 2.1. Для канала $\mathcal{A}$ с плотностью шума ξ, удовлетворяющей условиям (2.1), (2.2), при любом $\gamma > 1$ справедливо соотношение для информации между входным S и выходным η сигналами:

$$J(S, \eta) = B(S) + o(\|S\|_\gamma), \quad (\|S\|_\gamma \to 0) \tag{2.8}$$

В оставшейся части этого параграфа докажем некоторые следствия из этой теоремы, касающиеся пропускной способности рассматриваемого канала при различных ограничениях на входной сигнал, и исследуем «устойчивость» выражения (2.8) по отношению к малым дополнительным шумам в канале. В 3 и приложении приводится доказательство теоремы 1.

Замечание. В формуле (2.8) нельзя написать $o(\|S\|_1)$ или $o(B(S))$ вместо $o(\|S\|_\gamma)$, как показывает следующий простой пример. Пусть случайная величина S может принимать три значения $-\frac{1}{\varepsilon}$, 0, $\frac{1}{\varepsilon}$ с вероятностями

$$p\left(S = -\frac{1}{\varepsilon}\right) = p\left(S = \frac{1}{\varepsilon}\right) = \frac{\varepsilon^2}{2}, \quad p(S = 0) = 1 - \varepsilon^2.$$

Тогда, как легко проверить, справедливы следующие оценки ($\varepsilon \to 0$):

$$H(S) \sim \varepsilon^2 \ln \frac{1}{\varepsilon}; \quad B(S) \sim \varepsilon \ln \frac{1}{\varepsilon}; \quad M|S| = \varepsilon; \quad M|S|^\gamma = \varepsilon^{2-\gamma},$$

где $H(S)$ — энтропия дискретной случайной величины S. Учитывая, что $J(S, \eta) \le H(S)$, получаем $J(S, \eta) = o(B(S))$. Формула (2.8) позволяет в данном случае получить при любом $\delta > 0$ довольно грубую оценку $J(S, \eta) = o(\varepsilon^{1-\delta})$, а «уточненная» формула дала бы неверный результат.

Ясно, однако, что оценка (2.8) точна в главном члене, если рассматриваемое семейство случайных величин S при некотором $\gamma > 1$ удовлетворяет условию

$$\|S\|_\gamma = o(B(S)), \qquad (B(S) \to 0).$$

Из свойства (2.5) видно, например, что для семейства $\mathfrak{L}$ случайных величин S с $M|S| \to 0$, удовлетворяющих при некоторых $\gamma > 1$, $K > 0$ условию

$$\|S\|_\gamma < K \cdot \|S - \operatorname{med} S\|_1,$$

оценка

$$J(S, \eta) = B(S) + o(B(S))$$

имеет место равномерно для $S \in \mathfrak{L}$.

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Теорема 2.2. Пропускная способность канала $\mathfrak{N}$ при ограничении

$$|S| \leq \varepsilon$$

имеет при $\varepsilon \to 0$ представление

$$C(\varepsilon) = 2\Phi(x^*) \cdot \varepsilon + o(\varepsilon), \tag{2.9}$$

где $\Phi(x) = \sum_{k=1}^{r} \varphi_{p_k, q_k}(x)$, функции $\Phi_{p_k, q_k}(x)$ определены в (1.4), а x^* — единственное в интервале (0.1) решение уравнения

$$\Phi'(x) = 0. \tag{2.10}$$

При этом входным распределением, на котором с точностью до $o(\varepsilon)$ реализуется пропускная способность, является распределение:

$$p\{S = -\varepsilon\} = x^*, \quad p\{S = \varepsilon\} = 1 - x^*. \tag{2.11}$$

Доказательство. Очевидно $\Phi(0) = \Phi(1) = 0$. Поэтому

$$J(S, \eta) = \int_{-\varepsilon}^{\varepsilon} \Phi(F(u))du + o(\varepsilon). \tag{2.12}$$

Используя очевидное при $xy > 0$, $x \neq y$ неравенство $x \ln \frac{x}{y} > x - y$, получим

$$\Phi'(0) = \sum_{k=1}^{r} \left[p_k \ln \frac{p_k}{q_k} - (p_k - q_k) \right] > 0,$$

$$\Phi'(1) = \sum_{k=1}^{r} \left[-q_k \ln \frac{q_k}{p_k} + (q_k - p_k) \right] < 0.$$

Далее

$$\Phi''(x) = -\sum_{k=1}^{r} \frac{(p_k - q_k)^2}{q_k(1 - x) + p_k \cdot x} < 0.$$

Следовательно, уравнение $\Phi'(x) = 0$ имеет единственное решение x^* в интервале (0,1) и

$$\max_{0<x<1} \Phi(x) = \Phi(x^*). \tag{2.13}$$

Теперь из (2.12) и (2.13) легко следуют утверждения теоремы.

Замечание. Если $p(x)$ — симметричная функция (т. е. $p(x) = p(-x)$), то нетрудно заметить, что $x^* = \frac{1}{2}$ и оптимальным распределением является следующее

$$p\{S = -\varepsilon\} = p\{S = \varepsilon\} = \frac{1}{2}.$$

Теорема 2.3. Пропускная способность канала $\mathfrak{A}$ при ограничении

$$(M\,|S|^{\gamma})^{1/\gamma} \leq \varepsilon\,, \quad (\gamma > 1) \tag{2.14}$$

имеет при $\varepsilon \to 0$ представление

$$C(\varepsilon) = \varepsilon \left(\int_0^1 |\Phi'(x)|^{\beta}\, dx \right)^{1/\beta} + o(\varepsilon), \qquad \left(\frac{1}{\gamma} + \frac{1}{\beta} = 1 \right) \tag{2.15}$$

Такое значение информации $J(S, \eta)$ достигается на функции распределения $F^*(u)$, обратной к функции

$$g^*(x) = -\frac{\varepsilon\,|\Phi'(x)|^{1/\gamma - 1}}{\left(\int_0^1 |\Phi'(x)|^{\beta} dx \right)^{1/\gamma}} \operatorname{sgn} \Phi'(x)\,. \tag{2.16}$$

Доказательство. Достаточно проверить, что верхняя грань функционала $\int_{-\infty}^{\infty} \Phi\big(F(u)\big)\, du$ при ограничении

$$\int_{-\infty}^{\infty} |u|^{\gamma}\, dF(u) \leq \varepsilon^{\gamma}$$

достигается на распределении $F^*(u)$, определяемом формулой (2.16), и дается выражением (2.15). Для доказательства этого факта нам удобно сделать замену переменной $F(u) = x$. Обозначим $u = g(x)$ обратную к $F(u)$ функцию, которую будем считать для определенности непрерывной слева. В этих обозначениях наша задача сводится к нахождению

$$\sup \int_0^1 \Phi(x)\, dg(x) \tag{2.17}$$

при условии

$$\int_0^1 |g(x)|^{\gamma}\, dx \leq \varepsilon^{\gamma}.$$

Функция $g(x)$ монотонна, но, вообще говоря, неограничена в окрестности точек $x = 0$ и $x = 1$. Однако нетрудно проверить, что $\Phi(x)\, g(x) \to 0$ при $x \to 0$ или $x \to 1$. Поэтому, интегрируя в (2.17) по частям и используя неравенство Гельдера, найдем

$$\int_0^1 \Phi(x)\, dg(x) = -\int_0^1 g(x)\, \Phi'(x)\, dx \leq \left(\int_0^1 |g(x)|^{\gamma}\, dx \right)^{1/\gamma} \left(\int_0^1 |\Phi'(x)|^{\beta}\, dx \right)^{1/\beta} \leq$$

$$\leq \varepsilon \left(\int_0^1 |\Phi'(x)|^{\beta}\, dx \right)^{1/\beta},$$

причем последнее неравенство превращается в равенство при $g(x) = g^*(x)$, где функция $g^*(x)$ определяется равенством (2.16). Отсюда вытекают утверждения теоремы.

Устойчивость теоремы 2.1 по отношению к достаточно слабым дополнительным шумам в канале вытекает из следующей леммы.

Лемма 2.1. Если канал связи описывается соотношениями

$$\eta = S + \xi + \zeta,$$

где случайная величина ξ имеет плотность $p(x)$, удовлетворяющую условиям (2.1), (2.2), то при любых $\gamma > 1$, $\gamma_1 > 1$

$$J(S, \eta) = B(S) + o(\| S \|_\gamma) + o(\| \zeta \|_{\gamma_1}) + o(B(\zeta)).$$

Доказательство. Из тождества (см. [5], формула (3,8,7) на стр. 35)

$$J(S, \eta - \zeta) + J(\zeta, \eta) = J(S, \eta) + J(\zeta, \zeta + \xi)$$

вытекают оценки

$$J(S, \eta - \zeta) - J(\zeta, \zeta + \xi) \leq J(S, \eta) \leq J(S, \eta - \zeta).$$

Так как, в силу теоремы 2.1,

$$J(S, \eta - \zeta) = B(S) + o(\| S \|_\gamma) \quad \text{и}$$

$$J(\zeta, \zeta + \xi) = B(\zeta) + o(\| \zeta \|_{\gamma_1}),$$

то лемма доказана.

Ясно, что выражения в правых частях (2.8) и (2.18) совпадают, если, например, при некоторых $\gamma > 1$, $\gamma_1 > 1$

$$\| \zeta \|_{\gamma_1} = o(\| S \|_\gamma).$$

3. Доказательство теоремы 2.1

Для сокращения записи введем обозначения

$$\varepsilon = \| S \|_\gamma, \quad U = \sum_{k=1}^{r} [x_k - \beta(\varepsilon), x_k + \beta(\varepsilon)], \qquad R = (-\infty, \infty),$$

где функция $\beta(\varepsilon) \to 0$ при $\varepsilon \to 0$ и будет выбрана позже. Обозначим также

$$\psi(x, y) = p(x - y) \ln \frac{p(x - y)}{\int p(x - z)\, dF(z)}.$$

Тогда выражение для информации $J(S, \eta)$ можно представить в следующем виде

$$J(S,\eta) = \iint \psi(x,y)\,dF(y)\,dx = \sum_{k=1}^{r} \int_{x_k-\beta(\varepsilon)}^{x_k-\beta(\varepsilon)} \int \psi(x,y)\,dF(y)\,dx + \\ + \int_{R/U} \int \psi(x,y)\,dF(y)\,dx = \sum_{k=1}^{r} I_k + I_0 . \tag{3.1}$$

В приложении будет доказана следующая

Лемма 3.1. Если $\beta(\varepsilon) = \varepsilon^{\frac{\gamma-1}{2\gamma}}$ и выполнены предположения теоремы 2.1, то

$$I_0 = o(\varepsilon) . \tag{3.2}$$

Из леммы 3.1 и (3.1) ясно, что доказательство теоремы 2.1 сводится к доказательству следующей леммы.

Лемма 3.2. Если выполнены предположения теоремы 2.1, а $\beta(\varepsilon) = \varepsilon^{\frac{\gamma-1}{2\gamma}}$, то

$$|I_k - B_{p_k,q_k}(S)| = \left| \int_{x_k-\beta(\varepsilon)}^{x_k+\beta(\varepsilon)} \int \psi(x,y)\,dF(y)\,dx - B_{p_k,q_k}(S) \right| = o(\varepsilon) . \tag{3.3}$$

Доказательство. 1) Из неравенства Чебышева получим следующие оценки

$$P\{p(x-S) > e^{\lambda k} \cdot Mp(x-S)\} = \int_{\left\{\frac{p(x-y)}{Mp(x-S)} > e^{\lambda k}\right\}} dF(y) \le e^{-\lambda k}, \tag{3.4}$$

$$P\{|S| > \beta(\varepsilon)\} = \int_{|y|>\beta(\varepsilon)} dF(y) \le \frac{\varepsilon^{\gamma}}{\beta^{\gamma}(\varepsilon)} = \varepsilon^{1+\frac{\gamma-1}{2}} = o(\varepsilon) . \tag{3.5}$$

С помощью (3.4) получим, обозначая $\varphi(x,y) = \dfrac{p(x-y)}{\int p(x-z)\,dF(z)}$, оценку

$$\iint_{\{\ln\varphi(x,y)\ge\lambda\}} \psi(x,y)\,dF(y)\,dx = \sum_{k=1}^{\infty} \int_{\{\lambda(k+1)>\ln\varphi(x,y)\ge\lambda k\}} \int \psi(x,y)\,dF(y)\,dx \le \\ \le \sum_{k=1}^{\infty} \lambda(k+1)\,e^{-\lambda k} < \frac{2\lambda\,e^{-\lambda}}{(1-e^{-\lambda})^2} .$$

Кроме того, очевидно

$$\left| \iint\limits_{\{\ln \varphi(x,y) \le -\lambda\}} \psi(x,y)\, dF(y)\, dx \right| \le$$

$$\le \sum_{k=1}^{\infty} \int\limits_{\{-\lambda(k+1) < \ln \varphi(x,y) \le -\lambda k\}} \int \lambda(k+1)\, p(x-y)\, dx\, dF(y) \le$$

$$\le \sum_{k=1}^{\infty} \lambda(k+1)\, e^{-\lambda k} < \frac{2\lambda\, e^{-\lambda}}{(1-e^{-\lambda})^2}\,.$$

Полагая $\lambda = 2 \ln \dfrac{1}{\varepsilon}$, получим

$$\iint\limits_{\{|\ln \varphi(x,y)| > \lambda\}} |\psi(x,y)|\, dF(y)\, dx = o(\varepsilon)\,. \tag{3.6}$$

Далее из (3.5) и (3.3) находим

$$\left| \iint\limits_{|y| > \beta(\varepsilon)} \psi(x,y)\, dF(y)\, dx \right| = o(\varepsilon) + \iint\limits_{\{|y| > \beta(\varepsilon),\, |\ln \varphi(x,y)| \le \lambda\}} \psi(x,y)\, dF(y)\, dx \le$$

$$\le o(\varepsilon) + \lambda \cdot P\{|S| > \beta(\varepsilon)\} = o(\varepsilon)\,. \tag{3.7}$$

2) Используя простое неравенство $\ln x \le x - 1$ и (3.5), получим при $\varepsilon \le \varepsilon_0$

$$\left| \iint\limits_{|y| < \beta(\varepsilon)} \psi(x,y)\, dF(y)\, dx - \iint\limits_{|y| < \beta(\varepsilon)} p(x-y) \ln \frac{p(x-y)}{\int\limits_{|z| < \beta(\varepsilon)} p(x-z)\, dF(z)}\, dF(y)\, dx \right| =$$

$$= \iint\limits_{|y| < \beta(\varepsilon)} p(x-y) \ln \left[1 + \frac{\int\limits_{|z| > \beta(\varepsilon)} p(x-z)\, dF(z)}{\int\limits_{|z| < \beta(\varepsilon)} p(x-z)\, dF(z)} \right] dF(y)\, dx \le$$

$$\le \iint\limits_{|z| > \beta(\varepsilon)} p(x-z)\, dF(z)\, dx \le \int\limits_{|z| > \beta(\varepsilon)} dF(z) = o(\varepsilon)\,. \tag{3.8}$$

Приведенные оценки (3.7), (3.8) позволяют сразу написать для I_k, полагая без потери общности $x_k = 0$:

$$I_k = \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{|y| < \beta(\varepsilon)} \psi(x,y)\, dF(y)\, dx + o(\varepsilon) =$$

$$= \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{-\beta(\varepsilon)}^{x} p(x-y) \ln \frac{p(x-y)}{\int\limits_{|z|<\beta(\varepsilon)} p(x-z)\, dF(z)}\, dF(y)\, dx +$$

$$+ \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon)} p(x-y) \ln \frac{p(x-y)}{\int\limits_{|z|<\beta(\varepsilon)} p(x-z)\, dF(z)}\, dF(y)\, dx + o(\varepsilon) =$$

$$= J_k^{(1)} + J_k^{(2)} + o(\varepsilon)\,. \tag{3.9}$$

3) Таким образом, лемма 3.2, а вместе с ней и теорема 2.1 будут доказаны, если доказать следующие утверждения, которые можно сформулировать в виде отдельной леммы.

Лемма 3.3. Если выполнены предположения теоремы 2.1, то

$$J_k^{(1)} = \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{-\beta(\varepsilon)}^{x} p(x-y) \ln \frac{p(x-y)}{\int\limits_{|z_1|<\beta(\varepsilon)} p(x-z)\, dF(z)}\, dF(y)\, dx =$$

$$= p_k \int\limits_{-\infty}^{\infty} F(x) \ln \frac{p_k}{p_k F(x) + q_k [1 - F(x)]}\, dx + o(\varepsilon)\,, \tag{3.10}$$

$$J_k^{(2)} = \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon)} p(x-y) \ln \frac{p(x-y)}{\int\limits_{|z|<\beta(\varepsilon)} p(x-z)\, dF(z)}\, dF(y)\, dx =$$

$$= q_k \int\limits_{-\infty}^{\infty} [1 - F(x)] \ln \frac{q_k}{p_k F(x) + q_k [1 - F(x)]}\, dx + o(\varepsilon)\,, \tag{3.11}$$

где оценки $o(\varepsilon)$ равномерны по F при $\varepsilon \to 0$.

Доказательство. Так как оба утверждения леммы доказываются одинаково, остановимся лишь на втором. Замечая, что функция $p(x)$ равномерно непрерывна на каждом интервале (x_k, x_{k+1}), $k = 0, \ldots, r$; введем

$$\sigma(\varepsilon) = \sup_{\substack{|z| \le \varepsilon \\ x,\, x-z \in (x_k, x_{k+1})}} |p(x) - p(x-z)|\,. \tag{3.12}$$

Кроме того, предположим для определенности, что $p_k \neq 0$. С помощью оценки

$$\ln(1+x) < \frac{x^\alpha}{\alpha}$$

при $\alpha < \dfrac{\gamma-1}{\gamma}$ находим

$$\left| J_k^{(2)} - \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon)} p(x-y) \ln \frac{p(x-y)}{p_k \int\limits_{-\beta(\varepsilon)}^{x} dF(z) + \int\limits_{x}^{\beta(\varepsilon)} p(x-z)\, dF(z)} dF(y)\, dx \right| \leq$$

$$\leq \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon)} p(x-y) \ln \left[1 + \frac{2\sigma(2\beta) \cdot \int\limits_{-\beta(\varepsilon)}^{x} dF(z)}{\int\limits_{x}^{\beta(\varepsilon)} p(x-z)\, dF(z)} \right] dF(y)\, dx \leq \tag{3.13}$$

$$\leq 2\sigma(2\beta) \cdot \int\limits_{-\beta(\varepsilon)}^{\varepsilon} \int\limits_{-\beta(\varepsilon)}^{x} dF(z)\, dx + \frac{2\sigma^{\alpha}(2\beta)}{\alpha} \int\limits_{\varepsilon}^{\beta(\varepsilon)} \left[\int\limits_{x}^{\beta(\varepsilon)} dF(y) \right]^{1-\alpha} dx \sim \frac{\varepsilon \cdot \sigma^{\alpha}(2\beta)}{\alpha} = o(\varepsilon).$$

Совершенно аналогично получим

$$\left| \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon)} p(x-y) \left[\ln \frac{p(x-y)}{p_k \int\limits_{-\beta(\varepsilon)}^{x} dF(z) + \int\limits_{x}^{\beta(\varepsilon)} p(x-z)\, dF(z)} - \right.\right.$$

$$\left.\left. - \ln \frac{p(x-y)}{p_k \int\limits_{-\beta(\varepsilon)}^{x} dF(z) + p(x-y) \int\limits_{x}^{\beta(\varepsilon)} dF(z)} \right] dF(y)\, dx \right| = o(\varepsilon); \tag{3.14}$$

$$\left| \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon)} \left[q_k \ln \frac{q_k}{p_k \int\limits_{-\beta(\varepsilon)}^{x} dF(z) + q_k \int\limits_{x}^{\beta(\varepsilon)} dF(z)} - \right.\right.$$

$$\left.\left. - p(x-y) \ln \frac{p(x-y)}{p_k \int\limits_{-\beta(\varepsilon)}^{x} dF(z) + p(x-y) \int\limits_{x}^{\beta(\varepsilon)} dF(z)} \right] dF(y)\, dx \right| = o(\varepsilon); \tag{3.15}$$

$$\left| \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon} q_k \ln \frac{q_k}{p_k \int\limits_{-\beta(\varepsilon)}^{x} dF(z) + q_k \int\limits_{x}^{\beta(\varepsilon)} dF(z)} dF(y)\, dx - \right.$$

$$\left. - \int\limits_{-\infty}^{\infty} \int\limits_{x}^{\infty} q_k \ln \frac{q_k}{p_k \int\limits_{-\infty}^{x} dF(x) + q_k \int\limits_{x}^{\infty} dF(x)} dF(y)\, dx \right| = o(\varepsilon). \tag{3.16}$$

Теперь из оценок (3.13)—(3.16) легко следует утверждение леммы.

Теорема 2.1 доказана.

Приложение

Доказательство леммы 3.1.

Так как для любого x

$$\int p(x-y)\ln\frac{p(x-y)}{\int p(x-z)\,dF(z)}\,dF(y)\geq 0,$$

в силу неравенства $\ln z\geq 1-\frac{1}{z}$, то из (3.1) следует

$$I_0\geq 0. \tag{п.1}$$

Оценим теперь интеграл I_0 сверху. Очевидно

$$I_0=\int\limits_{R\setminus U}\int\psi(x,y)\,dF(y)\,dx\leq \iint\limits_{\{R\setminus U\}\cap\{|\ln\varphi(x,y)|<\delta\}}+$$

$$+\iint\limits_{\{R\setminus U\}\cap\{\delta\leq\ln\varphi(x,y)<\lambda\}}+\iint\limits_{\{\ln\varphi(x,y)\leq\lambda\}}=Z_1+Z_2+Z_3, \tag{п. 2}$$

где $\lambda=2\ln\frac{1}{\varepsilon}$, а $\delta\to 0$ при $\varepsilon\to 0$ и будет выбрано позже.

Интеграл Z_3 уже оценивался в (3.6)

$$Z_3=o(\varepsilon). \tag{п.3}$$

Рассмотрим теперь интеграл Z_2. Заметим, что при достаточно малом $\delta>0$ можно написать при $\ln\varphi(x,y)\geq\delta$ неравенство

$$1+\delta<\frac{(1-2\delta^3)\,p(x-y)}{\int p(x-z)\,dF(z)}\leq\frac{(1-2\delta^3)}{\int\limits_{|z|<\frac{\varepsilon}{\delta^3}}dF(z)}\cdot\sup_{|z|\leq\frac{\varepsilon}{\delta^3}}\frac{p(x-y)}{p(x-z)}\leq\sup_{|z|\leq\frac{\varepsilon}{\delta^3}}\frac{p(x-y)}{p(x-z)}.$$

Другими словами, если $\dfrac{p(x-y)}{\int p(x-z)\,dF(z)}>\varepsilon^{\delta}$, то найдется такое $|z|\leq\dfrac{\varepsilon}{\delta^3}$, что $\dfrac{p(x-y)}{p(x-z)}>1+\delta$. Вспоминая теперь оценку (3.7) и используя нера-

венства Гельдера и Чебышева, нетрудно получить [2, 3]:

$$Z_2 \leq \lambda \iint\limits_{\{R\setminus U\}\cap\{\delta<\ln\varphi(x,y)\}\cap\{|y|<\beta(\varepsilon)\}} p(x-y)\,dF(y)\,dx + \lambda \iint\limits_{\{|y|\geq\beta(\varepsilon)\}} p(x-y)\times$$

$$\times\, dx\, dF(y) \leq \frac{\lambda}{\delta^b} \iint\limits_{\{R\setminus U\}\cap\{|y|<\beta(\varepsilon)\}} |p^{1/b}(x-y) - p^{1/b}(x-z)|^b\, dF(y)\,dx + o(\varepsilon) \leq$$

$$\leq \frac{\lambda}{\delta^b} \iint\limits_{\{R\setminus U\}\cap\{|y|<\beta(\varepsilon)\}} \left| \int\limits_y^z \frac{p'(x-t)}{p^{1-1/b}(x-t)}\,dt \right|^b dF(y)\,dx + o(\varepsilon) \leq$$

$$\leq \frac{\lambda}{\delta^b} \iint\limits_{\{R\setminus U\}\cap\{|y|<\beta(\varepsilon)\}} |z-y|^{b-1} \left| \int\limits_y^z p(x-t) \left| \frac{p'(x-t)}{p(x-t)} \right|^b dt \right| dx\, dF(y) + o(\varepsilon) \leq$$

$$\leq K\cdot\frac{\lambda}{\delta^b} \int\limits_{|y|<\beta(\varepsilon)} |z-y|^b\, dF(y) + o(\varepsilon) \leq K\cdot\frac{\lambda\cdot\varepsilon^b}{\delta^{4b}} + o(\varepsilon) = o(\varepsilon), \qquad \text{(п. 4)}$$

если $\dfrac{\varepsilon^b\cdot\ln\dfrac{1}{\varepsilon}}{\delta^{4b}} \to 0$ при $\varepsilon\to 0$.

Для оценки оставшегося интеграла Z_1 воспользуемся неравенством

$$-\ln x \leq x-1+\delta|x-1|,$$

справедливым для $|\ln x|<\delta$. Тогда получим

$$Z_1 \leq \iint\limits_{\{R\setminus U\}\cap\{|\ln\varphi(x,y)|<\delta\}} \left[p(x-y) - \int p(x-z)\,dF(z)\right] dF(y)\,dx +$$

$$+\delta\cdot \iint\limits_{\{R\setminus U\}\cap\{|\ln\varphi(x,y)|<\delta\}} \left| p(x-y) - \int p(x-z)\,dF(z)\right| dF(y)\,dx \leq$$

$$\leq \iiint\limits_{\{R\setminus U\}\cap\left\{\frac{\int p(x-z)\,dF(z)}{p(x-y)}>e^{\delta}\right\}\cap\{|y|,\,|z|<\beta(\varepsilon)\}} |p(x-z) -$$

$$- p(x-y)|\,dF(z)\,dF(y)\,dx + \delta\cdot K\varepsilon + o(\varepsilon).$$

Для $x \in R \setminus U$ можно написать

$$\int p(x-z)\,dF(z) = p(x) + \int\limits_{|z|<\beta(\varepsilon)} [p(x-z)-p(x)]\,dF(z) +$$

$$+ \int\limits_{|z|\geq\beta(\varepsilon)} [p(x-z)-p(x)]\,dF(z) \leq p(x) + \sigma(\beta) + o(\varepsilon).$$

Поэтому справедливо включение

$$\{R\setminus U\} \cap \left\{\frac{\int p(x-z)\,dF(z)}{p(x-y)} > e^{\delta}\right\} \cap \{|y|, |z| < \beta(\varepsilon)\} \subset$$

$$\subset \{R\setminus U\} \cap \left\{p(x) < \frac{2\sigma(\beta)}{\delta}\right\} \cap \{|y|, |z| < \beta(\varepsilon)\},$$

из которого легко получим, меняя порядок интегрирования

$$Z_1 \leq \iiint\limits_{\left\{p(x)<\frac{2\sigma(\beta)}{\delta}\right\}\cap\{R\setminus U\}\cap\{|y|,\,|z|<\beta(\varepsilon)\}} \left|\int\limits_y^z p'(x-t)\,dt\right| dF(z)\,dF(y)\,dx + o(\varepsilon) \leq$$

$$\leq \iint |z-y|\,dF(y)\,dF(z) \cdot \int\limits_{\left\{p(x)<\frac{2\sigma(\beta)}{\delta}\right\}} |p'(x)|\,dx + o(\varepsilon) = o(\varepsilon), \qquad \text{(п. 5)}$$

если $\dfrac{\sigma(\beta)}{\delta} \to 0$ при $\varepsilon \to 0$

Из (п.4) и (п.5) ясно, что в качестве δ можно взять, например,

$$\delta = \frac{1}{\ln\dfrac{1}{\varepsilon}} + \sqrt{\sigma(\beta(\varepsilon))}.$$

Лемма 3.1 доказана.

Литература

1. *Прелов В. В.* Асимптотика пропускной способности непрерывного канала с большим шумом. Проблемы передачи информации, 1970, **6**, 2.
2. *Ибрагимов И. А., Хасьминский Р. З.* Передача слабого сигнала по каналу без памяти. Проблемы передачи информации, 1972, **8**, 4.
3. *Бурнашев М. В.* Блочный метод передачи слабого сигнала по каналу без памяти. Проблемы передачи информации, 1973, **9**, 1.
4. *Хасьминский Р. З.* Каналы с малой пропускной способностью и посимвольная передача. Проблемы передачи информации, 1970, **6**, 2, 58—67.
5. *Пинскер М. С.* Информация и информационная устойчивость случайных величин и процессов. Проблемы передачи информации, вып. 7. Изд.-во АН СССР, 1960.

Asymptotical expressions for the information in the channel with discontinuous noise density

M. V. BURNASHEV, R. Z. HAS'MINSKY

(Moscow)

Let $\mathcal{A}$ be a memoryless channel with additive noise and the following conditions for the noise density be satisfied:

1) $p(x)$ is continuous everywhere except a finite number of points $x_1, \ldots, x_r$ and there exist finite one-sided limits

$$p(x_k - 0) = q_k, \quad p(x_k + 0) = p_k; \qquad k = 1, \ldots, r \tag{1}$$

2) for some $b > 1$ there exist integrals

$$\int_{x_k}^{x_{k+1}} p(x) \left| \frac{p(x')}{p(x)} \right|^b dx < \infty, \qquad k = 0, \ldots, r, \tag{1}$$

$$x_0 = -\infty, \qquad x_{r+1} = +\infty.$$

Let us denote S — the input signal, ξ — independent additive noise, η — the output signal. We shall introduce also for arbitrary $p \geq 0$, $q \geq 0$ and random value S with the distribution function $F_S(x)$ the functional $B_{p,q}(S)$ by the formula

$$B_{p,q}(S) = \int \Phi_{p,q}(F_S(x))\, dx,$$

where

$$\Phi_{p,p}(y) = yp \ln p + (1 - y)q \ln q - [py + q(1 - y)] \ln [py + q(1 - y)] \tag{4}$$

and $0 \cdot \infty = 0$ is supposed. Then the following theorem is valid

Theorem. For the channel $\mathcal{A}$ with arbitrary $\gamma > 1$ the following asymptotical expression for the information between input signal S and output one η holds

$$J(S, \eta) = B(S) + \overline{o}((M \mid S \mid^{\gamma})^{1/\gamma}), \qquad M \mid S \mid^{\gamma} \to 0, \tag{5}$$

where

$$B(S) = \sum_{k=1}^{r} B_{p_k, q_k}(S). \tag{6}$$

Some exact formulas for the capacity of such channel under various restrictions on the input signal are obtained, and "stability" of the expression (5) to little additional noise is investigated.

Р. З. Хасьминский

Институт проблем передачи информации

СССР, Москва Е—24, Авиамоторная 8а, корп. 2

Problems of Control and Information Theory, Vol. 2 (2), pp. 123—132 (1973)

НЕСУЩЕСТВОВАНИЕ СОВЕРШЕННЫХ КОДОВ НАД ПОЛЯМИ ГАЛУА*

В. А. ЗИНОВЬЕВ, В. К. ЛЕОНТЬЕВ

(Москва)

(Поступила в редакцию 25 апреля 1972 г.)

Решена проблема существования совершенных кодов над полями Галуа. Доказано, что единственными нетривиальными совершенными кодами над конечными полями Галуа являются известные коды Хэмминга и Голея.

Пусть E^n — n-мерное векторное пространство над полем Галуа $GF(q)$, где $q = p^m$ (p — простое число). Код $A \subseteq E^n$ с минимальным расстоянием Хэмминга $d = 2t + 1$ называется совершенным, если шары радиуса t с центрами в кодовых точках покрывают все пространство E^n.

Тривиальным примером совершенного кода с любым $t \geq 1$ является один кодовый вектор длины $n = t$, а в случае $q = 2$ — два вектора длины $n = 2t + 1$ с расстоянием $d = n$. Известными нетривиальными совершенными кодами являются коды Хэмминга [1] с параметрами $t = 1, n = (q^m - 1)/(q-1)$, где $m \geq 2$, и коды Голея [2] с параметрами $q = 2$, $t = 3$, $n = 23$ и $q = 3$, $t = 2$, $n = 11$.

Целью настоящей работы является доказательство следующей теоремы.

Теорема 1. Единственными нетривиальными совершенными кодами над полями Галуа $GF(q)$ при всех q являются указанные выше коды Хэмминга и Голея.

Различные частные случаи этой теоремы были доказаны в работах [3—10]. При доказательстве ее будем существенно использовать результаты предыдущей работы [10].

Основой доказательства теоремы являются два необходимых условия существования совершенных кодов. Первое из них принадлежит Хэммингу и состоит в следующем. Пусть $m(A)$ — число всех кодовых точек совершенного кода с расстоянием $d = 2t + 1$. По определению совершенного кода

$$m(A) = q^n \Big/ \sum_{i=0}^{t} (q-1)^i C_n^i .$$

* Предварительное изложение этого результата содержится в работе Зиновьева В. А., Леонтьева В. К. "Теорема о несуществовании совершенных кодов над полями Галуа". Препринт ИППИ АН СССР, Москва, 1972.

Нетрудно показать [5], что из этого равенства следует, что

$$\sum_{i=0}^{t} (q-1)^i C_n^i = q^{n-k}, \tag{1}$$

где $q^k = m(A)$. Разрешимость в целых числах уравнения (1) является необходимым условием существования совершенного кода с параметрами t, q, n.

Второе необходимое условие существования совершенного кода принадлежит Ллойду [3] и состоит в следующем.

Теорема Ллойда.[1] Если существует совершенный код длины n над полем $GF(q)$, исправляющий t ошибок, то полином $P_t(n, \xi)$ степени t по переменной ξ, определяемый равенством

$$P_t(n, \xi) = \sum_{i=0}^{t} (-1)^i C_{n-\xi}^{t-i} C_{\xi-1}^{i} (q-1)^{t-i},$$

где $C_a^i = a(a-1)\ldots(a-i+1)/i$, для любых a имеет t различных целых корней среди чисел $1, 2, \ldots, n-1$.

Основная идея получения теоремы 1 из этих необходимых условий состоит в следующем.

Пусть существует совершенный код длины n над полем $GF(q)$. Тогда полином $P_t(n, \xi)$ имеет t различных целых корней $\xi_1, \xi_2, \ldots, \xi_t$, которые можно упорядочить следующим образом: $\xi_1 < \xi_2 < \ldots < \xi_t$. Нетрудно показать [8], что

$$\xi_1 \xi_2 \ldots \xi_t = \frac{t!}{q^t} \sum_{i=0}^{t} (q-1)^i C_n^i = t!\, q^{n-k-t}, \tag{2}$$

$$\xi_1 + \xi_2 + \ldots + \xi_t = t \frac{q-1}{q} \left(n + 1 - \frac{(t+1)(q-2)}{2(q-1)} \right). \tag{3}$$

В дальнейшем будем пользоваться обозначением

$$\xi_1 + \xi_2 + \ldots + \xi_t = a(q, t).$$

В равенстве (2) участвует необходимое условие Хэмминга существования совершенного кода, и таким образом это равенство является синтезом условий Хэмминга и Ллойда. Сравнивая с помощью равенств (2) и (3) сумму и произведение корней $\xi_1, \ldots, \xi_t$, нетрудно убедиться, что при достаточно больших n среднее арифметическое чисел $\xi_1, \ldots, \xi_t$ примерно равно их среднему геометрическому, т. е. эти числа должны быть примерно одинаковы. Однако условие целочисленности корней $\xi_1, \ldots, \xi_t$, записанное в виде (2),

[1] Эта формулировка теоремы Ллойда принадлежит ван Линту [5]

показывает, что на самом деле эти корни должны существенно различаться. Исходя из этих соображений, можно получить верхнюю границу для длины n возможного совершенного кода [10]. Используя связь совершенного кода с тактическими конфигурациями, можно получить нижнюю оценку длины возможного совершенного кода [10]. Сравнение этих оценок и приводит к теореме 1.

Доказательство теоремы. Так как для $t \leq 7$ [5—7] и $t < p$ [8] теорема 1 доказана, то будем рассматривать лишь случай $t \geq 8$ и $t \geq p$, хотя тот же метод позволяет доказать эту теорему в полном объеме.[2] Пусть существует совершенный код A с параметрами n, t, q. Тогда, как показано в [10], длина n удовлетворяет следующему неравенству:

$$n \geq \max\left\{\frac{(t+1)(t+2)}{q-1} + t,\ qt+1\right\}. \tag{4}$$

Нащей целью теперь является получение верхней оценки для n, которая была бы меньше (4). В силу теоремы Ллойда полином $P_t(n, \xi)$ имеет t различных корней $\xi_1, \ldots, \xi_t$, и поэтому любые два корня ξ_j и ξ'_j, $\xi_j < \xi'_j$, связаны соотношением $\xi'_j = \xi_j \mu_j$, где $\mu_j > 1$. Возьмем $h \leq t/2$ таких непересекающихся пар корней $\xi_j, \xi_{j'}$. Тогда с учетом условия (3), как это сделано в [10], можно получить следующее неравенство:

$$\xi_1 \xi_2 \ldots \xi_t \leq \left(\frac{a(q,t)}{t}\right)^t \prod_{j=1}^{h} \frac{4\mu_j}{(1+\mu_j)^2}. \tag{5}$$

Оценим снизу сумму $\sum_{i=0}^{t}(q-1)^i C_n^i$. Для этого воспользуемся неравенством $n \geq 4t$, которое является тривиальным следствием (4). Так как при $n \geq 4t$

$$n(n-t+1) > (n-\alpha t)^2,$$

где $\alpha = 2(2-\sqrt{3}) < 0{,}536$ — корень уравнения $\alpha^2 - 8\alpha + 4 = 0$, то

$$n(n-1)\ldots(n-t+1) > (n-\alpha t)^t > (n-0{,}536t)^t,$$

и значит

$$\frac{t!}{q^t}\sum_{i=0}^{t}(q-1)^i C_n^i > \left(\frac{q-1}{q}\right)^t t!\, C_n^t > \left(\frac{q-1}{q}\right)^t (n-0{,}536t)^t. \tag{6}$$

[2] На самом деле теорема 1 доказана также для $q = 2$ [9, 10] и $q = 3, 5, 7$ [10]. Здесь мы несколько изменили доказательство (уточнили верхнюю оценку), а потому повторяем его и для указанных случаев.

Обозначив через

$$\Psi(\mu_1, \ldots, \mu_h) = \prod_{j=1}^{h} \frac{4\mu_j}{(1+\mu_j)^2},$$

получим из (2), (5) и (6)

$$\left(1 - \frac{t(1+0{,}072(q-1))+q}{2(n+1)(q-1)-(t+1)(q-2)}\right)^t < \Psi(\mu_1, \ldots, \mu_h). \tag{7}$$

(Довольно точные численные значения встречающихся в этой работе постоянных необходимы для того, чтобы доказать несуществование совершенных кодов при начальных значениях параметров n, t, q.)

Верхняя оценка для n, которую легко получить из (7), зависит от чисел $\mu_1, \ldots, \mu_h$ и монотонно уменьшается с их возрастанием. Поэтому нам важно получить хорошие нижние оценки для этих чисел. Из равенства (2) следует, что каждый корень ξ_i можно представить в следующем виде: $\xi_j = \alpha_j p^{r_j}$, где $(\alpha_j, p) = 1$, и поэтому

$$\alpha_1 \alpha_2 \ldots \alpha_t = t!/p^{[t/p]+[t/p^2]} + \ldots,$$

где $[x]$ — целая часть x.

Разобьем все корни $\xi_1, \ldots, \xi_t$ на классы, относя в один класс корни, имеющие одинаковые α_i. Пусть эти классы будут $L_1, \ldots, L_g$ и x_i — число классов мощности i. Тогда

$$\sum_{i\geq 1} x_i = g, \quad \sum_{i\geq 1} i x_i = t. \tag{8}$$

Отметим, что в классе мощности $i (i \geq 2)$ отношение максимального корня к минимальному не менее чем p^{i-1}. Поэтому из корней $\xi_1, \ldots, \xi_t$ можно заведомо выделить $h = \sum_{i\geq 2} x_i$ непересекающихся пар корней с отношением $\mu_i \geq p$. Тогда из оценки (5), учитывая, что при $i \geq 2$

$$\frac{4p^{i-1}}{(1+p^{i-1})^2} \leq \left(\frac{4p}{(1+p)^2}\right)^{i-1}$$

следует, что

$$\xi_1 \xi_2 \ldots \xi_t \leq \left(\frac{a(q,t)}{t}\right)^t \prod_{i\geq 2} \left(\frac{4p^{i-1}}{(1+p^{i-1})^2}\right)^{x_i} \leq$$

$$\leq \left(\frac{a(q,t)}{t}\right)^t \left(\frac{4p}{(1+p)^2}\right)^{\sum_{i\geq 2}(i-1)x_i} = \left(\frac{a(q,t)}{t}\right)^t \left(\frac{4p}{(1+p)^2}\right)^{t-g}. \tag{9}$$

Остается оценить снизу разность $t - g$. Эта оценка по существу получена в [10]. Хотя там было дано несколько иное определение классов, сам вывод нижней оценки для $t - g$ остается справедливым, если в лемме 1 работы [10] число $q = p^m$ заменить на p. Поэтому приводим только окончательный результат. Пусть

$$\beta(p) = \begin{cases} 5 \text{ для } p = 2,3, \\ 4 \text{ для } p = 5, \\ \text{максимальному целому решению неравенства } (\beta - 1)'! \leq \beta p \text{ для } p \geq 7. \end{cases}$$

Тогда

$$t - g \geq t \cdot \mu(p, t), \tag{10}$$

$$\mu(p,t) = \begin{cases} \left[\frac{t}{p}\right] \Big/ t, & \text{если } p \leq t < p \cdot \beta(p), \quad (11) \\ \dfrac{p \log p - (p-1)\log(p-1) - \dfrac{p-1}{t} \cdot r(p,t)}{(p-1)\log\left(\dfrac{p}{p-1} \cdot t\right)}, & \text{если } t \geq p \cdot \beta(p), \quad (12) \end{cases}$$

где

$$r(p,t) = \left(\frac{3}{2} + \frac{p-2}{p}\right)\log t - \frac{1}{2}\log\frac{p}{4,2} + \frac{0,36}{p}\left(1 + \frac{4(p-2)}{p-1}\right).$$

Здесь и далее все логарифмы берутся по основанию два.

Применение (9) и (10) в (7) и использование неравенства $(1 - \gamma)^a \geq 1 - a\gamma$, где $a \geq 1$ и $0 < \gamma < 1$, приводит к следующей верхней оценке длины n:

$$n + 1 < \left(\frac{p+1}{p-1}\right)^2 \frac{t(1 + 0,072(q-1)) + q}{2(q-1)\mu(p,t)} + \frac{(t+1)(q-2)}{2(q-1)}. \tag{13}$$

Так как при всех $t \geq 8$ и $q \geq 2$ справедливо неравенство

$$t(1 + 0,072(q-1)) + q \leq \left(\frac{1,125}{q-1} + 0,197\right) t(q-1), \tag{14}$$

то (13) можно переписать следующим образом:

$$n + 1 < \left(\frac{p+1}{p-1}\right)^2 \left(\frac{0,5625}{q-1} + 0,0985\right) \frac{t}{\mu(p,t)} + \frac{(t+1)(q-2)}{2(q-1)}. \tag{15}$$

Сравним теперь полученную верхнюю оценку (15) с нижней оценкой (4).

Случай $max\{8, p\} \leq t < p \cdot \beta(p)$. Для этого случая верхняя оценка имеет вид

$$n + 1 < \left(\frac{p+1}{p-1}\right)^2 \left(\frac{0{,}5625}{q-1} + 0{,}0985\right) \frac{t^2}{[t/p]} + \frac{(t+1)(q-2)}{2(q-1)}. \qquad (16)$$

1. $p = 2$. Так как $\beta(2) = 5$, то при $p = 2$ возможны значения $t = 8, 9$, так что $[t/2] = 4$ и оценка (16) перепишется следующим образом:

$$n + 1 < \left(\frac{1{,}266}{q-1} + 0{,}224\right) t^2 + \frac{(t+1)(q-2)}{2(q-1)}. \qquad (17)$$

При $q = 2$ $n + 1 < 1{,}49t^2$,

что меньше нижней оценки $n + 1 \geq (t + 1)(t + 3)$ для значений $t = 8{,}9$.

При $q = 4$ возможные длины совершенных кодов заключены в пределах

$$\text{при } t = 8 \qquad 38 \leq n \leq 43,$$

$$\text{при } t = 9 \qquad 46 \leq n \leq 52. \qquad (18)$$

Для значений $q = 2^s \geq 8$

$$n + 1 < 0{,}405t^2 + (t + 1)/2,$$

что меньше нижней оценки $n + 1 \geq qt + 2$ при $t = 8{,}9$.

2. $p = 3$. В этом случае t заключено в пределах $8 \leq t < 15$. Оценка (16) при этом перепишется так:

$$n + 1 < \left(\frac{2{,}25}{q-1} + 0{,}394\right) \frac{t^2}{[t/3]} + \frac{(t+1)(q-2)}{2(q-1)}. \qquad (19)$$

В случае $q = 3$ справедливо неравенство

$$n + 1 \geq \frac{(t+1)(t+2)}{2} + t + 1 > \frac{1{,}52\,t^2}{[t/3]} + \frac{t+1}{4},$$

а в случае $q = 3^s \geq 9$ — неравенство

$$n + 1 \geq qt + 2 > \frac{0{,}676\,t^2}{[t/3]} + \frac{t+1}{2}.$$

Последнее неравенство очевидно, если учесть, что всегда $q > \frac{t}{2}$, а $[t/3] \geq 2$.

3. $p = 5$. Этот случай также требует индивидуального рассмотрения. При всех $8 \leq t < 20$ и $q \geq 5$ верхняя оценка

$$n + 1 < \left(\frac{1{,}266}{q-1} + 0{,}224\right) \frac{t^2}{[t/5]} + \frac{(t+1)(q-2)}{2(q-1)}$$

меньше нижней оценки $n + 1 \geq qt + 2$.

4. $p \geq 7$. В этом случае оценку (16) можно переписать следующим образом:

$$n + 1 < 0{,}343 \frac{t^2}{[t/p]} + \frac{t+1}{2}. \tag{20}$$

Следующая серия очевидных неравенств

$$n + 1 \geq qt + 2 > pt > \frac{0{,}5t^2}{[t/p]} > 0{,}343 \frac{t^2}{[t/p]} + \frac{t+1}{2}$$

говорит о том, что верхняя оценка (20) всюду меньше нижней оценки. Таким образом, в рассмотренном диапазоне $\max\{8, p\} \leq t < p\beta(p)$ единственные возможные длины совершенных кодов при $q = 4$ и $t = 8{,}9$ заключены в пределах (18). Но для этих значений длин условие Хэмминга не выполняется [8].

Случай $t \geq p\beta(p)$. Для значений $p = 2, 3, 5, 7$ функция $\mu(p, t)$ может быть оценена следующим образом, если учесть, что $t \geq p\beta(p)$:

$$\mu(2, t) > 1{,}43/\log 2t, \tag{21}$$

$$\mu(3, t) > 0{,}86/\log \frac{3}{2}t, \tag{22}$$

$$\mu(5, t) > 0{,}42/\log \frac{5}{4}t, \tag{23}$$

$$\mu(7, t) > 0{,}37/\log \frac{7}{6}t. \tag{24}$$

1. $p = 2$. Верхняя оценка (15) имеет вид

$$n + 1 < \left(\frac{3{,}4}{q-1} + 0{,}62\right) t \log 2t + \frac{(t+1)(q-2)}{2(q-1)}. \tag{25}$$

При $q = 2$ эта оценка меньше нижней оценки $n + 1 \geq (t + 1)(t + 3)$ для всех $t \geq 18$. При $t \leq 18$ длина $n < 380$.

При $q = 4$ оценка (25) меньше нижней оценки $n + 1 \geq \frac{(t+1)(t+2)}{3} + t + 1$ для всех $t \geq 24$.

При $t \leq 24$ длина $n < 240$.

При $q \geq 8$ верхняя оценка

$$n + 1 < 1{,}12\, t \log 2t + \frac{t+1}{2}$$

для всех $t \geq 2 \cdot \beta(2) = 10$ меньше нижней оценки (4).

2. $p = 3$. Верхняя оценка (15) имеет вид

$$n + 1 < \left(\frac{2{,}62}{q-1} + 0{,}46\right) t \log \frac{3}{2} t + \frac{(t+1)(q-2)}{2(q-1)}. \tag{26}$$

Эта оценка для всех $t \geq 3\beta(3) = 15$ меньше при $q = 3$ нижней оценки $n \geq \frac{(t+1)(t+2)}{2} + t$, а при $q \geq 9$ нижней оценки (4).

3. $p = 5$. Оценка (15) имеет вид

$$n + 1 < \left(\frac{3{,}15}{q-1} + 0{,}54\right) t \log \frac{5}{4} t + \frac{(t+1)(q-2)}{2(q-1)}. \tag{27}$$

При $q = 5$

$$n + 1 < 1{,}34 t \log \frac{5}{4} t + \frac{3}{8}(t+1),$$

что меньше при всех $t \geq 5 \cdot \beta(5) = 20$ нижней оценки $n \geq \frac{(t+1)(t+2)}{4} + t$.

При $q \geq 25$

$$n + 1 < 0{,}67 t \log \frac{5}{4} t + \frac{t+1}{2},$$

что меньше оценки (4) при всех $t \geq 20$.

4. $p = 7$. В этом случае

$$n + 1 < 0{,}93 t \log \frac{7}{6} t + \frac{t+1}{2}, \tag{28}$$

что при всех $t \geq 7 \cdot \beta(7) = 35$ меньше оценки (4).

5. $p \geq 11$. При $p \geq 11$ функцию $\mu(p, t)$ можно оценить следующим образом

$$\mu(p, t) \geq \frac{0{,}6 \log p}{p \log t}. \tag{29}$$

Это дает следующую верхнюю оценку для всех $p \geq 11$ и $t \geq p \cdot \beta(p)$:

$$n < 0{,}4 tp \frac{\log t}{\log p}. \tag{30}$$

Покажем, что для всех $p \geq 11$ и $q \geq p$ и всех $t \geq p \cdot \beta(p)$ оценка (30) меньше нижней оценки (4). В диапазоне $t < q^2$

$$n > qt > 0{,}4 tp \log t / \log p,$$

так как

$$0{,}4tp\frac{2\log q}{\log p} > 0{,}4tp\frac{\log t}{\log p}$$

и

$$q/\log q > 0{,}8\,p/\log p \tag{31}$$

при всех $q \geq p$. Пусть теперь $t \geq q^2$. Тогда

$$n \geq \frac{(t+1)(t+2)}{q-1} + t > \frac{t^2}{q} > 0{,}4tp\frac{\log t}{\log p}.$$

Последнее неравенство

$$t/\log t > 0{,}4qp/\log p$$

выполняется, так как при всех $t \geq q^2$ выполняется неравенство

$$t/\log t \geq q^2/2\log q,$$

а также неравенство (31).

Таким образом, в рассмотренном случае $t \geq p \cdot \beta(p)$ — возможные длины совершенных кодов ($q = 2$, $10 \leq t \leq 18$, $n < 380$ и $q = 4$, $10 \leq t \leq 24$, $n < 240$) лежат в диапазоне, когда условие Хэмминга не выполняется [8].

Таким образом, в статье показано, что полученная верхняя оценка длины (13), кроме указанных случаев, всегда лежит ниже нижней оценки (4). Но это и означает, что для всех этих значений t не могут существовать нетривиальные совершенные коды.

Теорема 1 полностью доказана.

Литература

1. *Hamming, R. W.* Error Detecting and Error Correcting Codes. Bell Syst. Techn. J., 1950, *29*, 2, 147—160. (Русск. перев.: Хэмминг Р., в сб. «Коды с обнаружением и исправлением ошибок», М., Изд-во иностр. лит., 1956, 7—22).
2. *Golay, M. J. E.* Notes on Digital Coding. Proc. IRE, 1949, *37*, 6, 657.
3. *Lloyd, S. P.* Binary Block Coding. Bell Syst. Techn. J., 1957, *36*, 2, 517—535. (Русск. перев.: Ллойд С. П. Бинарное блочное кодирование. Кибернетический сборник, I, М., Изд-во иностр. лит., 1960, 206—226.)
4. *Shapiro, H. S., Slotnick, D. L.* On the Mathematical Theory of Error Correcting Codes. IBM J. Res. Develop., 1959, *3*, 1, 25—34. (Русск. перев.: Шапиро Г. С., Злотник Д. Л. К математической теории кодов с исправлением ошибок. Кибернетический сборник, 5, М., Изд-во иностр. лит., 1962, 7—32.)
5. *Van Lint, J. H.* On the Nonexistence of Perfect 2- and 3-Hamming-Error-Correcting Codes. Inform. and Control, 1970, *16*, 4, 396—401.
6. *Tietäväinen, A.* On the Nonexistence of Perfect 4-Error-Correcting Codes. "Suomal. tiedeak. toimit.", 1970, AI, 485.

7. *Van Lint, J. H.* On the Nonexistence of Perfect 5-, 6- and 7-Hamming-Error-Correcting Codes over CG (q). Report 70—WSK—06, Technological University Eindhoven (1970).
8. *Van Lint, J. H.* Error Correcting Codes. Lecture Notes Math., 201, 1971.
9. *Tietäväinen, A., Perko, A.* There are no Unknown Perfect Binary Codes. Annales Universitatis Turkuensis, 1971, 148, ser. A.
10. *Зиновьев В. А., Леонтьев В. К.* О совершенных кодах. Проблемы передачи информации, 1972, *8*, 1, 26—25.

The Nonexistence of Perfect Codes over Galois Fields

V. A. ZINOVJEV, V. K. LEONTJEV

(Moscow)

Let E^n be the n-dimensional vector space over the Galois Field $GF(q)$, where q is a power of a prime p. The Hamming distance $d(x, y)$ of vectors $x, y \in E^n$ is defined to be the weight (that is the number of nonzero components) of $x - y$. If t is a positive integer, we define the sphere $B(a, t)$ a centre a and radius t as the set

$$B(a, t) = \{x \in E^n \mid d(a, x) \leq t\}.$$

A subset A of E^n is called a perfect t-error-correcting code of block length n if

(1) $x, y \in A$, $x \neq y$, implies $B(x, t) \cap B(y, t) = \varnothing$,

(2) $\bigcup_{\alpha \in A} B(a, t) = E^n$.

The following perfect codes are known:

1) Trivial perfect codes in cases $n = t$ (one-word code) and $n = 2t + 1$ for $q = 2$ (repetition code of two words).

2) Hamming codes in case $t = 1$, $n = (q^m - 1)/(q - 1)$, $m \geq 2$ [1].

3) Golay codes in cases $t = 2$, $q = 3$, $n = 11$ and $t = 3$, $q = 2$, $n = 23$ [2].

The purpose of this paper is to prove the nonexistence of unknown nontrivial perfect codes over $GF(q)$ for all q. Hence we state

Theorem. For all q(q is a prime pover) there are no nontrivial perfect t-error-correcting codes over $GF(q)$ except known Hamming and Golay codes.

The different particular cases of this theorem were proved earlier in papers [3—10].

Proof of this theorem is essentially based on Hamming condition (1), Lloyd's Theorem and methods and results of our preceding work [10].

В. А. Зиновьев, В. К. Леонтьев

Институт проблем передачи информации

СССР, Москва Е—24, Авиамоторная 8а, корп. 2

Problems of Control and Information Theory, Vol. 2(2), pp. 133—137 (1973)

ОБОБЩЕНИЕ ТЕОРЕМЫ ЛЛОЙДА НА СЛУЧАЙ ПРОИЗВОЛЬНОГО АЛФАВИТА

Л. А. БАССАЛЫГО

(Москва)

(Поступила в редакцию 3 мая 1972 г.)

Пусть $A(n, 2t+1, q)$ — код длины n, заданный над алфавитом из q символов, с минимальным кодовым расстоянием Хэмминга $2t+1$. Данный код называется совершенным, если его мощность $|A(n, 2t+1, q)|$ удовлетворяет следующему условию:

$$|A(n, 2t+1, q)| = \frac{q^n}{\sum_{i=0}^{t} (q-1)^i C_n^i}. \tag{1}$$

В последнее время появилось много работ, посвященных проблеме существования совершенных кодов над полем Галуа $GF(p^m)$. Окончательный результат здесь принадлежит В. А. Зиновьеву и В. К. Леонтьеву [1], которые доказали, что над полем Галуа $GF(p^m)$ не существует нетривиальных совершенных кодов, кроме известных кодов Хэмминга и Голея. Доказательство этого факта, как и доказательства предварительных частных результатов, основаны на использовании теоремы Ллойда [2], доказанной самим Ллойдом для случая $q = 2$, а затем обобщенной рядом авторов на случай $q = p^m$ (p — простое). В случае же, когда q не является степенью простого числа и $\frac{q^n}{\sum_{i=0}^{t} (q-1)^i C_n^i}$ является целым числом, известен лишь следующий результат: при $q = 6$, $n = 7$, $t = 1$ не существует совершенного кода. Естественным поэтому представляется желание обобщить теорему Ллойда на случай произвольного алфавита, тем более что это обобщение достигается лишь усложнением выкладок, проделанных Ллойдом [2].

Теорема Ллойда. Пусть $\varphi_s(n, \zeta)$ — полином по ξ степени s, задаваемый равенством

$$\varphi_s(n, \xi) = \sum_{r=0}^{s} (-1)^r (q-1)^{s-r} C_\xi^{s-r} C_{n-\xi}^r, \tag{2}$$

где $C_u^i = \dfrac{u(u-1)\ldots(u-i+1)}{i!}$. Если существует совершенный код длины n с исправлением t ошибок над алфавитом из q символов, то $\varphi_t(n-1, \xi)$ имеет t различных целых корней среди чисел $0, 1, 2, \ldots, n-1$.

Доказательство. Так как данное доказательство отличается от доказательства Ллойда [2] лишь более сложными выкладками, то и отметим лишь эти различия, не проводя полного доказательства. Все наши обозначения совпадают с обозначениями Ллойда: $v_{j,s}$ $(j = 0, 1, \ldots, n)$ — число последовательностей веса s, находящихся на расстоянии j от какого-либо кодового слова.[1] $G_j(x) = \sum_{s=0}^{n} v_{j,s} x^s$ — производящая функция чисел $v_{j,s}$, где x — комплексное переменное.[2]

Число последовательностей, находящихся на расстоянии j от фиксированного кодового слова веса s, равно

$$\sum_{m=0}^{\infty} C_{n-s}^{j-m}(q-1)^{j-m} C_s^m \sum_{i=0}^{\infty} C_m^i (q-2)^{m-i}$$ [3]

и, следовательно, $G_j(x) = \sum_{s=0}^{\infty} v_s g_s(j, x)$, где

$$g_s(j, x) = \sum_{m=0}^{\infty} C_{n-s}^{j-m}(q-1)^{j-m} C_s^m \sum_{i=0}^{\infty} C_m^i (q-2)^{m-i} x^{s+j-m-i}.$$

Так как

$$\sum_{j=0}^{\infty} y^j G_j(x) = \sum_{j=0}^{\infty} y^j \sum_{s=0}^{\infty} v_s g_s(j, x) =$$

$$= \sum_{s=0}^{\infty} v_s \sum_{m=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} C_s^m C_m^i (q-2)^{m-i} x^{s-i} y^m C_{n-s}^{j-m}(q-1)^{j-m} x^{j-m} y^{j-m} =$$

$$= \sum_{s=0}^{\infty} v_s \sum_{m=0}^{\infty} \sum_{i=0}^{\infty} C_s^m x^{s-m} y^m [1 + (q-1)xy]^{n-s} C_m^i (q-2)^{m-i} x^{m-i} =$$

$$= \sum_{s=0}^{\infty} v_s [1 + (q-1)xy]^{n-s} \sum_{m=0}^{\infty} C_s^m x^{s-m} y^m [1 + (q-2)x]^m =$$

$$= \sum_{s=0}^{\infty} v_s [1 + (q-1)xy]^{n-s} [x + y + (q-2)xy]^s =$$

[1] Каждая последовательность учитывается столько раз, сколько имеется кодовых слов, находящихся от нее на расстоянии j.

[2] В дальнейшем опустим индекс 0 у $v_{0,s}$ и $G_0(x)$.

[3] Примем соглашение, что $C_k^l = 0$ при $l > \kappa$ и $l < 0$.

$$= [1 + (q-1)xy]^n \sum_{s=0}^{\infty} v_s \left[\frac{x + y + (q-2)xy}{1 + (q-1)xy}\right]^s =$$

$$= [1 + (q-1)xy]^n G\left(\frac{x + y + (q-2)xy}{1 + (q-1)xy}\right),$$ [4]

то

$$G_j(x) = \frac{1}{2\pi i} \int_{C_0} \frac{[1 + (q-1)xy]^n G\left(\frac{x + y + (q-2)xy}{1 + (q-1)xy}\right)}{y^{j+1}} dy = L_j G(x), \quad (3)$$

где контур C_0 представляет собой маленькую окружность с центром в начале координат, которая обходится в положительном направлении. Положив $Z = \frac{x + y + (q-2)xs}{1 + (q-1)xy}$, получаем

$$G_j(x) = L_j G(x) = \frac{1}{2\pi i} \int_{C_x} \frac{[1 + (q-2)x - (q-1)x^2]^{n+1} G(z)}{(z-x)^{j+1}[1 + (q-2)x - (q-1)xz]^{n-j+1}} dz =$$

$$= \frac{1}{j!}[1 + (q-2)x - (q-1)x^2]^{n+1} \lim_{z \to x} \frac{d^j \frac{G(z)}{[1 + (q-2)x - (q-1)xz]^{n-j+1}}}{dz^j} =$$

$$= \frac{1}{j!}[1 + (q-2)x - (q-1)x^2]^{u+1} \sum_{k=0}^{j} C_j^k (j-k)!\, C_{n-k}^{j-k} \times$$

$$\times \frac{[(q-1)x]^{j-k}}{[1 + (q-2)x - (q-1)x^2]^{n+1-k}} \frac{d^k G(x)}{dx^k} =$$

$$= \sum_{k=0}^{j} C_{n-k}^{j-k} \frac{[(q-1)x]^{j-k}[1 + (q-2)x - (q-1)x^2]^k}{k!} \frac{d^k G(x)}{dx^k}. \quad (4)$$

Здесь C_x — малая окружность, охватывающая x, но не $\frac{1 + (q-2)x}{(q-1)x}$. Так как

$$\sum_{j=0}^{t} G_j(x) = [1 + (q-1)x]^n, \quad (5)$$

то получаем, что $G(x)$ удовлетворяет неоднородному линейному дифферен-

[4] Чтобы не иметь дело с особенностями, будем рассматривать достаточно малые x и y; например, $|x| < \frac{1}{q-1}$, $|y| < 1$.

циальному уравнению t-ого порядка:

$$MG(x) = \sum_{k=0}^{t} \frac{[1+(q-2)x-(q-1)x^2]^k}{k!} \left(\sum_{r=0}^{t-k} C_{n-k}^{r} [(q-1)x]^r \right) \frac{d^k G(x)}{dx^k} = \\ = [1+(q-1)x]^n . \tag{6}$$

Положим $f_{n,\xi}(x) = [1+(q-1)x]^{\xi}[1-x]^{n-\xi} = \sum_{s=0}^{\infty} \varphi_s(n,\xi)\, x^s$, где $\varphi_s(n,\xi) = \\ = \sum_{r=0}^{s} (-1)^r (q-1)^{s-r} C_{\xi}^{s-r} C_{n-\xi}^{r}$. Из равенства $(1-x)^{-1}[1+(q-1)x]^{\xi} \times \\ \times [1-x]^{n-\xi} = [1+(q-1)x]^{\xi}[1-x]^{n-1-\xi}$ следует, что $\sum_{k=0}^{s} \varphi_k(n,\xi) = \\ = \varphi_s(n-1,\xi)$. Так как

$$[1+(q-1)xy]^n f_{n,\xi}\left(\frac{x+y+(q-2)xy}{1+(q-1)xy}\right) = [1+(q-1)xy + \\ + (q-1)x + (q-1)y + (q-1)(q-2)xy]^{\xi}[1+(q-1)xy - x - y - (q-2)xy]^{n-\xi} = \\ = [1+(q-1)x]^{\xi}[1-x]^{n-\xi}[1+(q-1)y][1-y]^{n-\xi} = f_{n,\xi}(x) f_{n,\xi}(y),$$

то

$$L_j f_{n,\xi}(x) = \varphi_j(n,\xi) f_{n,\xi}(x)$$

и, следовательно,

$$Mf_{n,\xi}(x) = \sum_{j=0}^{t} \varphi_j(n,\xi) f_{n,\xi}(x) = \varphi_t(n-1,\xi) f_{n.\xi}(x) .$$

Дальнейшие рассуждения опускаем, так как они практически дословно совпадают с рассуждениями Ллойда [2].

Следствие 1. Теорема Ллойда заменой переменной ξ на $n-x$ может быть переформулирована следующим образом: если существует совершенный код длины n над алфавитом из q символов, исправляющий t ошибок, то полином $P_t(n,q,x)$ от x степени t, определяемый равенством

$$P_t(n,q,x) = \sum_{r=0}^{t} (-1)^r (q-1)^{t-r} C_{n-x}^{t-r} C_{x-1}^{r} \tag{7}$$

имеет t различных корней среди чисел $1, 2, \ldots, n$.

Литература

1. *Зиновьев В. А., Леонтьев В. К.* Несуществование совершенных кодов над полями Галуа. Проблемы управления и теории информации, 1973, **2**, 2, 123–132.
2. *Lloyd, S. P.* Binary Block Coding. Bell Syst. Techn. J., 1957, **36**, 2, 517–535.

The Generalizations of Lloyd's Theorem on an Arbitrary Alphabet

L. A. BASSALYGO

(Moscow)

In 1957 S. P. Lloyd proved a theorem which gave a necessary condition for the existence of a binary ($q = 2$) perfect t-error-correcting code:

If a perfect t-error-correcting code of block length n over q symbol alphabet exists then the polynomial $P_t(x) = \sum_{i=0}^{t} (-1)^i C_{n-x}^{t-i} C_{x-1}^{i} (q-1)^{t-i}$ where

$$C_a^i = \frac{a(a-1)\ldots(a-i+1)}{i!},$$

has t distinct integer zeros among $1, 2, \ldots, n$. For $q = p^m$ (p-prime) this theorem was proved by F. J. MacWilliams.

In the present paper this theorem is proved for arbitrary q.

Л. А. Бассалыго

Институт проблем передачи информации

СССР, Москва Е—24, Авиамоторная 8а, корп. 2

Problems of Control and Information Theory, Vol. 2 (2), pp. 139—147 (1973)

О СВОЙСТВАХ ЭЛЕМЕНТОВ, УДОБНЫХ ДЛЯ СИНТЕЗА d-БЕЗОТКАЗНЫХ СТРУКТУР

М. А. ГАВРИЛОВ, В. М. ОСТИАНУ, Л. Г. БИВОЛ

(Москва)

(Поступила в редакцию 2 февраля 1972 г.)

Рассмотрен метод построения дискретных структур без памяти, который обобщается на случай построения таких структур, обладающих свойством d-безотказности.

Под d-безотказным понимается устройство, выполняющее заданный алгоритм функционирования при отказах любых $t \leq d$ его элементов, причем найдется такое сочетание $d + 1$ элементов, одновременные отказы которых нарушают выполнение устройством заданного алгоритма функционирования.

Определяются необходимые и достаточные условия, которым должна удовлетворять функция, реализуемая абсолютно надежным выходным элементом для того, чтобы последний был способен исправлять ошибки на его входах.

В общем виде задача обеспечения надежности формулируется обычно в терминах теории вероятностей, т. е. как задача обеспечения определенных заданных вероятностных характеристик надежности, таких как вероятность безотказной работы, среднее врема безотказной работы и др. [1]. Однако для дискретных управляющих устройств подавляющее большинство методов введения избыточности основано на обеспечении так называемой (d_0, d_1)-безотказности.[1] Поскольку величины d_0 и d_1 связаны определенными соотношениями с вероятностью безотказной работы устройства (различными по виду для различных методов введения избыточности), то в дальнейшем будем говорить о построении (d_0, d_1)-безотказных устройств.

Очевидно, что сложность реализации (d_0, d_1)-безотказного устройства зависит как от метода введения избыточности, так и от приспособленности самих элементов, из которых реализуется устройство, к реализации функций (d_0, d_1)-безотказных устройств.

Работы, в которых исследовались бы требования к внутренней структуре элементов, позволяющих синтезировать d-безотказные устройства с наи-

[1] Под (d_0, d_1)-безотказным дискретным устройством понимается такое устройство, которое при отказах типа $1 \to 0$ любых $t_0 \leq d_0$ элементов и отказах типа $0 \to 1$ любых $t_1 \leq d_1$ элементов правильно выполняет заданный алгоритм функционирования, но найдется хотя бы одно сочетание $d_1 + 1$ отказов типа $0 \to 1$ при $t_0 \leq d_0$ отказах типа $1 \to 0$ и хотя бы одно сочетание $d_0 + 1$ отказов типа $1 \to 0$ при $t_1 \leq d_1$ отказах типа $0 \to 1$, приводящие к невыполнению заданного алгоритма функционирования. Если $d_0 = d_1$, то устройство будем называть d-безотказным.

меньшим числом используемых элементов, насколько известно авторам, почти полностью отсутствуют.

В настоящей статье задача получения (d_0, d_1)-безотказности рассматривается лишь по отношению к комбинационным устройствам (в том числе и по отношению к комбинационной части последовательностных устройств). Рассматривается общая постановка этой задачи и формулируются некоторые принципиальные требования, которые должны быть предъявлены к внутренней структуре логических элементов.

Представим условия работы устройства в виде системы $\Phi = \{F_i\}$ булевых функций F_i $(i = 1, 2, \ldots, m)$. Пусть условия работы заданы в виде таблиц состояний функций F_i, причём F_i могут принимать значения 0, 1 и $\frac{1}{2}$ (ложное, истинное и безразличное). Набор значений входных переменных $x_1, x_2, \ldots, x_t$, для которых j-ый выход реализует значение 1, 0 или $\frac{1}{2}$ будем называть соответственно рабочим, нерабочим или безразличным состоянием рассматриваемого выхода.

Рассмотрим некоторую функцию $F \in \{F_i\}$. Обозначим множество её рабочих состояний через M_1, а множество нерабочих состояний через M_0. Эта функция может быть реализована как по рабочим (из M_1), так и по нерабочим (из M_0) состояниям. Будем называть в соответствии с этим «реализуемой» ту часть таблицы состояний M_1, или M_0, по состояниям которой реализуется функция F, и обозначать её через M_ε, и «нереализуемой» — вторую часть таблицы состояний, и обозначать её через $M_{\bar{\varepsilon}}$.

При реализации структуры для каждого из m выходов структуры должен быть выбран из набора логических элементов, имеющихся в распоряжении проектировщика, выходной элемент, оптимально реализующий структуру дискретного устройства совместно с выходными элементами,выбранными для других выходов.[2]

Определение 1. Будем говорить, что функция F реализована (d_0, d_1)-безотказно элементом φ, если:

а) функция F реализована элементом φ (на входы которого поданы входные переменные, или функции входных переменных или те и другие вместе), т. е. существует некоторая структура A с выходным элементом φ, которая для любого состояния из таблицы M_1 функции F выдает на своём

[2] Метод синтеза избыточных структур дискретных устройств в такой постановке рассмотрен в [2], где описана некоторая процедура направленного поиска структур, близких к оптимальным.

выходе воздействие, равное 1, а для любого состояния из M_0 — воздействие, равное 0;

б) эти условия реализации сохраняются также и при появлении на любых входах элемента φ не более d_0-отказов типа $1 \to 0$ и не более d_1-отказов типа $0 \to 1$. [3]

Очевидно, для того, чтобы устройство, реализующее систему функций Φ, было (d_0, d_1)-безотказным, достаточно реализовать (d_0, d_1)-безотказно каждую функцию $F \in \{F_i\}$.

Пусть выходной элемент φ имеет n входов $q_1, q_2, \ldots, q_n$ и условия его работы заданы таблицей состояний μ, причём множества рабочих и нерабочих состояний таблицы обозначены соответственно через μ_1 и μ_0. Тогда функция F может быть реализована на его выходе, если будет выбрана такая совокупность входных переменных $x_1, x_2, \ldots, x_t$ или функций от них, подаваемая на его входы, что будет выполнено условие а) из определения 1.

Назовём «приведенной» таблицу, полученную из таблицы состояний элемента φ путём подстановки вместо значений входов $q_1, q_2, \ldots, q_n$ значений указанных выше совокупностей переменных $x_1, x_2, \ldots, x_t$ или функций, подаваемых на входы элемента φ и сопоставленных состояниям s из таблицы состояний $M_1 \cup M_0$ функции F.

Метод, описанный в [2], распространен в [3] для (d_0, d_1)-безотказной реализации функции F элементом φ из класса элементов с симметричными входами. Однако этот метод может быть обобщен и на случай произвольных элементов. Его идея заключается в следующем.

Пусть известна дизъюнктивная нормальная форма (ДНФ) функции, реализуемой элементом φ, и её отрицание. Обозначим i-ый конъюнктивный член этой ДНФ переменной g_i, а j-ый конъюнктивный член её отрицания через h_i.

Построим таблицу λ_ε, строки которой соответствуют строкам таблицы M_ε, а столбцы — переменным g_i, и таблицу $\lambda_{\bar{\varepsilon}}$, строки которой соответствуют строкам таблицы $M_{\bar{\varepsilon}}$, а столбцы — переменным h_j. Реализация заданной функции F элементом φ будет непротиворечивой, если в каждой строке таблиц λ_ε и $\lambda_{\bar{\varepsilon}}$ будет содержаться хотя бы по одной единице. Реализация функции F будет d-безотказной, если в каждой строке этих таблиц будет хотя бы $d + 1$ единиц. Из таблиц λ_ε и $\lambda_{\bar{\varepsilon}}$ можно получить систему условий, которым должны удовлетворять исходные переменные или доопределяющие функции для непротиворечивой и d-безотказной реализации функции F элементом φ.

[3] Выходной элемент φ предполагается вполне надежным.

Рассмотрим некоторые требования к внутренней структуре выходного элемента, т. е. к таблице μ состояний, характеризующих его логические свойства.

В [7] уже были исследованы эти требования в частном случае. Теперь рассмотрим эту задачу в общем виде.

Следуя [4], будем говорить, что состояния $s' \in \mu_\varepsilon$ и $s'' \in \mu_{\bar{\varepsilon}}$ имеют обязательную переменную $\tilde{x}_\nu$, если они отличаются друг от друга только значением переменной $\tilde{x}_\nu$.

Определение 2. Будем называть (d_0, d_1)-соседствующим с s состоянием всякое состояние s^*, которое можно получить из s, заменяя в последнем $t_0 (0 \leq t_0 \leq d_0)$ единиц нулями и $t_1 (0 \leq t_1 \leq d_1)$ нулей единицами, причём хотя бы одно из чисел t_0, t_1 должно быть больше нуля.

Множество всех (d_0, d_1)-соседствующих с s состояний будем обозначать через $\{s(d_0, d_1)\}$.

Теорема. Для того чтобы элемент φ реализовал (d_0, d_1)-безотказно функцию F, необходимо и достаточно выполнение следующих условий.

1) В каждом из множеств μ_ε и $\mu_{\bar{\varepsilon}}$ таблицы истинности элемента φ должно иметься хотя бы по одному состоянию $s' \in \mu_\varepsilon$ и $s'' \in \mu_{\bar{\varepsilon}}$, не содержащему ни одной обязательной переменной.

2) Вместе с s' в μ_ε должно содержаться все множество $\{s'(d_0, d_1)\}$ и вместе с s' в $\mu_{\bar{\varepsilon}}$ должно содержаться все множество $\{s''(d_0, d_1)\}$.

Доказательство

Достаточность. Пусть в таблице состояний выходного элемента φ содержатся два состояния $s' \in \mu_\varepsilon$ и $s'' \in \mu_{\bar{\varepsilon}}$, не имеющие обязательных переменных, и вместе с s' в μ_ε содержатся все (d_0, d_1)-соседствующие с s' состояния, а вместе с s'' в $\mu_{\bar{\varepsilon}}$ содержатся все (d_0, d_1)-соседствующие с s'' состояния. Тогда при реализации заданной функции F состояниям из её таблицы M будем ставить в соответствие только состояния s' и s''. В результате отказов состояние $s' \in \mu_\varepsilon$ и состояние $s'' \in \mu_{\bar{\varepsilon}}$ могут перейти в некоторые другие s'^* и s''^*. Если число отказов типа $0 \to 1$ не больше, чем d_1, и число отказов типа $1 \to 0$ не больше, чем d_0, то $s'^* \in \{s'(d_0, d_1)\} \subset \mu_\varepsilon$, а $s''^* \in \{s''(d_0, d_1)\} \subset \mu_{\bar{\varepsilon}}$. Следовательно, при состояниях s'^* на выходе будет реализовано значение ε, а при s''^* — значение $\bar{\varepsilon}$. Тем самым выход будет реализован правильно, чем и завершается доказательство достаточности.

Необходимость. Пусть на выходе элемента φ реализуется (d_0, d_1)-безотказно функция F. Это значит, что состояниям $S \in M_\varepsilon$ сопоставлены такие

состояния s' из μ_ε, что замена в последних не более чем d_0 значений 1 на 0 и не более чем d_1 значений 0 на 1 переводит их в состояния $s'^* \in \mu_\varepsilon$. Следовательно, $\{s'(d_0, d_1)\} \subset \mu_\varepsilon$ и в s' нет ни одной обязательной переменной. Аналогичная картина имеет место для s'' из $\mu_{\bar{\varepsilon}}$. Этим доказательство теоремы завершено.

Наличие в таблице μ более чем двух состояний, удовлетворяющих условиям теоремы, позволяет получить менее сложную структуру при осуществлении надежностного синтеза.

Заметим, что можно говорить о (d_0, d_1)-безотказной реализации не всей функции F в целом, а лишь одного из подмножеств её значений: равных ε, или равных $\bar{\varepsilon}$. В таком случае требования теоремы должны формулироваться лишь по отношению к одному из множеств μ_ε или $\mu_{\bar{\varepsilon}}$ соответственно.

Покажем, что произвольный элемент φ, реализующий (d_0, d_1)-безотказно произвольную (но заданную) функцию F, реализует функцию выхода корректора либо реализует сумму двух функций, из которых одна является функцией выхода корректора, а вторая (обозначаемая буквой ψ) не может быть использована при (d_0, d_1)-безотказной реализации функции F.

Если строится d-безотказное устройство методом корректоров, то, как известно из [5], любой выход корректора реализует дизъюнкции обобщенных симметрических функций

$$\vee S_{t,t-1,\ldots,t-d}(x_1, x_2, \ldots, x_t) \qquad (1)$$

с рабочими числами t, $t-1, \ldots, t-d$, зависящую от t входных переменных $x_1, x_2, \ldots, x_t$. В силу доказанной теоремы, таблица t-входового элемента φ, реализующего заданную функцию d-безотказно, вместе с каждым состоянием $s' \in \mu_\varepsilon$ должна содержать в μ_ε все множество $\{s'(d_0, d_1)\}$ и вместе с каждым $s'' \in \mu_{\bar{\varepsilon}}$ должна содержать в $\mu_{\bar{\varepsilon}}$ все множество $\{s''(d_0, d_1)\}$. Следовательно, при наличии нескольких порождающих состояний[4] $s' \in \mu_\varepsilon$ и $s'' \in \mu_{\bar{\varepsilon}}$ элемент φ реализует функцию (1).

Но поскольку в таблице истинности элемента φ помимо состояний s', s'', $\{s'(d_0, d_1)\}$ и $\{s''(d_0, d_1)\}$ могут присутствовать и другие, то кроме (1) элемент φ может реализовать ещё некоторую функцию $\psi(x_1, x_2, \ldots, x_t)$. Следовательно

$$\varphi(x_1, x_2, \ldots, x_j) = \vee S_{t,t-1,\ldots,t-d}(x_1, \ldots, x_t) + \psi(x_1, \ldots, x_t).$$

С другой стороны, поскольку элемент φ d-безотказно реализует функцию F, то состояниям из таблицы M сопоставлены только те состояния из таблицы μ, все d-соседствующие состояния $\{s(d)\}$ которых находятся в той же таблице μ_ε или $\mu_{\bar{\varepsilon}}$, которой принадлежит s. Таким образом, из всех конъюнктивных

[4] Определение порождающего состояния приведено в [5].

членов функции $\varphi(x_1, \ldots, x_t)$ использованы только те, которые образуют функцию (1).

Пусть теперь строится (d_0, d_1)-безотказное устройство методом корректоров. Вид функций, реализуемых выходами корректора (d_0, d_1)-безотказного дискретного устройства, построенного из функциональных элементов, определен в [6]. При построении таких устройств используется корректирующий код Кима-Фреймана [8], характерной чертой которого является расстояние $D(\alpha, \beta)$ между любой парой его слов α и β, удовлетворяющее соотношению:

$$\mathfrak{D}(\alpha, \beta) \geq 2(d_0 + d_1 + 1) - (|\alpha| - |\beta|),$$

где $|x|$ есть норма кодового слова x.

Лусть на входы корректора подается слово γ из кода Кима-Фрейма. При наличии $t_0 \leq d_0$ отказов типа $1 \to 0$ в этом слове сохраняются без искажения все символы, равные 0, а любые t_0 символов, равные 1, перейдут в нули. Таким образом, дизъюнкция состояний на входе корректора, получающихся в результате воздействий t_0 отказов типа $1 \to 0$, может быть представлена в виде конъюнкций неизменных букв и элементарной обобщенной симметрической функции с рабочим числом t_0, порожденной остальными символами слова. Аналогичная картина имеет место при наличии $t_1 \leq d_1$ отказов типа $0 \to 1$.

Поскольку отказы типа $1 \to 0$ и типа $0 \to 1$ могут иметь место одновременно (в количестве не более, чем d_0 и d_1 соответственно), то совокупность состояний на входах корректора при этом определяется произведением двух элементарных обобщенных симметрических функций.

$$\sigma_{t_0}(x_{i_1}, x_{i_2}, \ldots, x_k) \cdot \sigma_{t_1}(x_{i_{k+1}}, \ldots, x_{i_t}), \tag{2}$$

где k — число символов со значением 1 в рассматриваемом слове кода. Отсюда следует, что на любом выходе корректора реализуется частично симметрическая функция, представляющая собой булеву сумму по всем γ произведений двух элементарных обобщенных симметрических функций вида (2):

$$L(x_1, \ldots, x_t) = \sum_{k=1}^{\nu} \sum_{\substack{0 \leq t_0 \leq d_0 \\ 0 \leq t_1 \leq d_1}} \sigma_{t_0}(\alpha_k)\, \sigma_{t_1}(\alpha_{t-k}) = \sum_{k=0}^{t} \sigma_{0,\ldots,d_0}(\alpha_k)\, \sigma_{0,\ldots,d_1}(\alpha_{t-k}), \tag{3}$$

где α_k и α_{t-k} — наборы переменных, от которых зависят соответственно σ_{t_0} и σ_{t_1} из (2). Число и индексы переменных, от которых зависят функции $\sigma_{0,\ldots,d_0}$ и $\sigma_{0,\ldots,d_1}$, соответственно, определяется значениями символов слова γ.

Рассмотрим теперь функцию элемента φ, реализующего (d_0, d_1)-безотказно заданную функцию F. В силу доказанной теоремы, в таблице μ элемента φ вместе с каждым $s' \in \mu_\varepsilon$ в μ_ε должно содержаться все множество

$\{s'(d_0, d_1)\}$ и вместе с каждым $s'' \in \mu_{\bar{\varepsilon}}$ в $\mu_{\bar{\varepsilon}}$ должно содержаться и все множество $\{s''(d_0, d_1)\}$. Очевидно, что дизъюнкция всех этих множеств состояний представляет собой не что иное, как частично-симметрическую функцию (3). Следовательно, функция элемента φ имеет вид $\alpha(x_1, \ldots, x + \psi(x_1, \ldots, x_l)$. С другой стороны, если элемент φ реализует (d_0, d_1)-безотказно функцию F, то при реализации функции F состояниям из таблицы M были сопоставлены только те состояния из таблицы μ, все (d_0, d_1)-соседствующие состояния $\{s(d_0, d_1)\}$ которых находились в той же таблице μ_ε или $\mu_{\bar{\varepsilon}}$, которой принадлежит s. Таким образом, из всех конъюнктивных членов функции φ использованы только те, которые образуют функцию (3).

Итак, логическая функция элемента φ, способного реализовать (d_0, d_1)-безотказно функцию F (и, следовательно, обладающего указанными выше свойствами), либо тождественно равна функции (3), реализуемой любым выходом корректора (d_0, d_1)-безотказного устройства, либо представляет собой сумму функции (3) и некоторой другой $\psi(x_1, x_2, \ldots x_l)$, которая не используется при надежностном синтезе.

Таким образом, наиболее выгодным элементом для использования в качестве выходного при надежностном синтезе является элемент, реализующий частично-симметрическую функцию.

Литература

1. *Гаврилов М. А., Остиану В. М., Потехин А. И.* Надежность дискретных систем. Итоги науки. Теория вероятностей. Математическая статистика. Теоретическая кибернетика, 1969, Москва, 1970, 104 стр.
2. *Гаврилов М. А., Копыленко В. М.* Синтез структуры бесконтактных релейных устройств. Bull. Math. de la Soc. Math. de la RSR 1966, 10 (58), 3.
3. *Гаврилов М. А., Копыленко В. М.* Инженерные методы синтеза комбинаторных структур на произвольных элементах. Изв. на математически институт БАН, 1971, том 12, 211—230.
4. *Гаврилов М. А.* Структурная теория релейных устройств. Лекция по курсу «Основы теории автоматических и телемеханических систем» ч. II, ВЗЭИ, Москва, 1960.
5. *Гаврилов М. А.* Сигнализация и прогнозирование отказов в дискретных управляющих устройствах со структурной избыточностью. Доклад на II Международном конгрессе по автоматическому управлению, Базель, 1963, Москва, «Наука», 1965, 211—220.
6. *Гаврилов М. А., Кришталь В. З., Остиану В. М.* Синтез надежных релейных устройств, (d_0, d_1)-безотказных относительно внутренних элементов и входов. Сб. Трудов III Конгресса ИФАК (Лондон, июнь 1966), том «Технические средства автоматики», Москва, «Наука», 1971, 362—371.
7. *Бивол Л. Г.* О надежности схем, построенных из многофункциональных элементов. Материалы докладов VII Научно-технической конференции, Кишиневский политехнический институт, Кишинев, 1971.
8. *Kim, W. N.—Freiman, G. W.* Multi-Error Correcting Code for a Binary Asymmetric Channel. IEEE Trans. v. **CT—5**. Special Supplement, 1959.

On properties of elements suitable for the synthesis of d-faultless structures

M. A. GAVRILOV, V. M. OSTIANU, L. G. BIVOL

(Moscow)

In the present paper the problem of obtaining the (d_s, d_1)-faultlessness is considered only with respect to combinational devices (including also combinational parts of sequential devices). The general statement of this problem is discussed and certain principal requirements which must be produced to the internal structure of logical elements are formulated.

Let us represent operating conditions of a device in the form of the system of Boolean functions F_i $(i = 1, 2, \ldots, m)$.

Let the operating conditions be given in the form of state tables of functions $F_i \cdot F_i$ may assume the values 0, 1 and the indifferent one designated by $\frac{1}{2}$.

The set of values of input variables $x_1, x_2, \ldots, x_t$, for which the j-th input has the value 1, 0 or $\frac{1}{2}$, let respectively be called operating, non-operating and indifferent state of the output under discussion.

Let us consider a certain function $F \in \{F_i\}$. Designate the set of its operating states by M_1, and the set of non-operating states by M_0.

This function can be realized both by operating (from M_1) or by non-operating (from M_0) states. According to this let us call "realized" the part of the state table M_1 or M_0 by which the function F is realized and designate it by M_g, and "non-realized" — the second part of the state table and designate it by $M_{\bar{g}}$.

Definition 1. Assume that the function F is realized (d_0, d_1)-faultlessly by the element φ if:

a) the function F is realized by the element φ, that is there exists a certain structure A with the output element φ, which for any state from M_1 for the function F generates at its input the action equal to 1 and for any state from M_0 — the action equal to 0, and

b) these realization conditions remain also in appearing at any inputs of the element φ, not more than d_0-faults of $0 \to 1$ type; and not more the d_1 faults of $0 \to 1$ type.[1]

Obviously, if the device realizing the set of functions $\{F_i\}$ could be (d_0, d_1)-faultless, it is sufficient to realize each function $F_1(d_0, d_1)$-faultlessly.

Let the output element φ have n inputs $q_1, q_2, \ldots, q_n$ and conditions of its operation are set up by truth table μ, the sets of operating and non-operating states of the truth table being designated by μ_1 and μ_0, respectively. Then the function F will be realized at its output if such a set of input variables $x_1, x_2, \ldots, x_t$ is chosen (or functions of them) and fed to its inputs that the condition a) from Definition 1 is fulfilled.

Certain requirements to the internal structure of the output elements is considered that is to the truth table μ which characterizes its logical properties.

Following [7] let us assume that the states $S' \in \mu_g$ and $S'' \in \mu_{\bar{g}}$ have obligatory variable $\tilde{x}_\nu$ if they differ from each other by the value of the variable $\tilde{x}_\nu$ only.

Definition 2. Let us call (d_0, d_1)-neighbouring with states any state S^* which can be obtained from S by replacing in the latter $t_s (0 \leq t_s \leq d_0)$ ones by zeroes and $t_1 (0 \leq \leq t_1 \leq d_1)$.

zeroes by ones and in this case at least one of the numbers t_s, t_1 must be greater than zero.

The set of all (d_0, d_1)-neighbouring with S states is denoted by $\{s(d_0, d_1)\}$.

Theorem. In order to realize the function F (d_0, d_1)-faultlessly it is necessary and sufficient that the element φ fulfill the following conditions.

[1] The output element φ is assumed to be quite reliable.

1) Each set μ_ε and $\mu_{\bar{\varepsilon}}$ of states of the element φ must have at least one state $s' \in \mu_\varepsilon$ and one state $s'' \subset \mu_{\bar{\varepsilon}}$ containing no obligatory variable.

2) Together with S', μ_ε must contain the total set $\{s'(d_0, d_1)\}$ and, together with S'', $\mu_{\bar{\varepsilon}}$ must contain the total set $\{s''(d_0, d_1)\}$.

It is shown that a random element φ which realizes function F (d_0, d_1)-faultlessly, realizes only the corrector output function.

М. А. Гаврилов

Институт проблем управления

СССР, Москва В—485

Профсоюзная 81

Problems of Control and Information Theory, Vol. 2 (2), pp. 149—162 (1973)

COMMON INFORMATION IS FAR LESS THAN MUTUAL INFORMATION

P. GÁCS and J. KÖRNER

(Budapest)

(Received February 5, 1972)

1. Introduction

Up to now mutual information has no immediate interpretation. It appears nowhere as the result of a coding problem. Imre Csiszár raised the problem whether there is any connection between the mutual information $E\left(\log_2 \frac{P(\xi = x, \eta = y)}{P(\xi = x) \cdot P(\eta = y)}\right)$ of two random variables and such a coding of them which allows to discern all their common information. More precisely: given the random variables ξ and η we are looking for a code consisting of three parts: A, B and C where ξ can be decoded from A and C, while η can be decoded if B and C are known. We shall call such a code a Y-scheme in the sequel.

Our results are of negative character. Prescribing for the total length of the codes A, B and C that it should not exceed too much the optimum length of a common code of the pair (ξ, η), the optimum taken without any structural condition on the encoding, we shall prove in the case of memoryless sources that common information always corresponds to some deterministic interdependence of the two sources. Thus our aim is to show that common information has nothing to do with mutual information.

One can formulate the same question also in terms of Kolmogorov's "complexity" and "mutual information" of individual sequences. Our results remain here valid. (See § 6 of the present paper).

2. Formulation of results

First of all let us introduce some notations and definitions.

Definition 1. A discrete memoryless stationary information source (DMSS) with finite alphabet X is a sequence $\{\xi_n\}_{n=1}^{\infty}$ of independent, identically distributed random variables ξ_n with values in the finite set X. We denote the source by $\mathcal{X}$.

We shall often use the following notations:

$P(\xi = x)$: the probability of the event $[\xi = x]$

$H(\xi)$: the entropy of the random variable ξ.

$\xi^n = \xi_1 \xi_1 \dots \xi_n$.

X^n: the n-th Cartesian power of the set X.

$I = \{0, 1\}$: the set consisting of 0 and 1.

$I^* = \bigcup_{k=0}^{\infty} I^k$

$l(f)$: the length of the binary word f.

Definition 2. A mapping $\varphi: X^n \to I^k$ is called a (binary) block-code. φ is an ε-code $(0 < \varepsilon < 1)$ of the DMSS $\mathfrak{X} = \{\xi_n\}_{n=1}^{\infty}$ if there exists a mapping $\varphi': I^k \to X^n$ such that the inequality $P[\varphi'(\varphi(x)) \neq x]$ holds. $P[\varphi'(\varphi(x)) \neq x]$ is the probability of error of the code described above.

All the log's in this paper are to the base 2.

After these generally adopted definitions let us turn to the concepts of Y-scheme and common information.

Definition 3. Let $\mathfrak{X} = \{\xi_n\}_{n=1}^{\infty}$ and $\mathfrak{Y} = \{\eta_n\}_{n=1}^{\infty}$ be two DMSS's with finite alphabets X and Y. The joint distribution of the pair (ξ_1, ξ_1) — where the pairs (ξ_i, η_i) are supposed to be independent for different i's, — and $\varepsilon > 0$, $\delta > 0$ are given. We call a $Y_{\varepsilon,\delta}^{(n)}$-scheme (or simply Y-scheme) the following coding scheme:

(i) f and f' are binary block codes of the set X^n,
g and g' are binary block codes of Y^n with

$$P[f(\xi^n) = g(\eta^n)] > 1 - \varepsilon. \tag{1}$$

(ii) the juxtaposition ff' of f and f' is an ε-code of $\mathfrak{X}$ and similarly gg' is an ε-code of $\mathfrak{Y}$.

(iii) $l(ff') = l(f) + l(f') < (1 + \delta) H(\eta^n)$

$$l(gg') = l(g) + l(g') < (1 + \delta) H(\eta^n). \tag{2}$$

Definition 4. Let us write $J_n(\varepsilon, \delta) = \max l(f)$ where the maximum is taken over all the $Y_{\varepsilon,\delta}$-schemes coding the sources $\mathfrak{X}$ and $\mathfrak{Y}$. $J_n(\varepsilon, \delta)$ is the *common information* of ξ^n and η^n.

One could think that the quantity $J_n(\varepsilon, \delta)$ and the mutual information $I(\xi^n \wedge \eta^n)$ are closely related but this is not true. We shall show that in general cases of dependence $J_n(\varepsilon, \delta)$ is of smaller order of magnitude.

In order to formulate our result let us introduce a notion of accessibility with respect to the joint distribution of ξ_1 and η_1, which is closely related to the one introduced in the theory of Markov-chains (see [5]). We can assume without loss of generality that for every $x \in X$ and $y \in Y$ $P[\xi = x] > 0$ and $P[\eta = y] > 0$. We shall say that $x \in X$ and $x' \in X$ are *communicating* (write $x \sim x'$) if there exists some sequence $x_0 y_1 x_1 \ldots x_N$ with $x_i \in X$, $y_i \in Y$ and $P[\xi = x_{i-1}, \eta = y_i] > 0$, $P[\xi = x_i, \eta = y_i] > 0$, where $x_1 = x$ and $x_N = x'$. It is obvious that $x' \sim x$ is an equivalence relation on X. Let the corresponding equivalence classes be X_i for $i = 1, 2, \ldots, r$. Similarly we define a relation $y \sim y'$ on Y equivalence classes Y_i, $i = 1, 2, \ldots, s$. It is easy to see that one can pass from a given X_i only to a single Y_j with positive probability and vice versa. Hence one has $r = s$. If we give the same index to these "communicating" classes, we obtain

$$P(\eta \in Y_j \mid \xi \in X_i) = P(\xi \in X_i \mid \eta \in Y_i) = \delta_{ij} \tag{3}$$

where δ_{ij} denotes the Kronecker symbol.

Definition 5. We shall call the above unique partition of X and Y an *ergodic decomposition.*

This terminology will be motivated later on. The introduction of a random variable ζ taking its values from the set of the indices of the equivalence classes and defined by

$$\zeta = i \quad \text{if} \quad \xi \in X_i$$

proves to be useful. One has obviously:

$$P[\zeta = i] = P[\xi \in X_i, \quad \eta \in Y_i]. \tag{4}$$

Passing to the product space the sequence $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ defines a sequence of random variables $\zeta_1 \ldots \zeta_n$ in a similar way. It is evident that the ζ_i's are independent and identically distributed. The different classes of the ergodic decomposition of X^n and Y^n correspond to the values of ζ^n in a one-to-one way.

It is clear that ζ^n represents some "common information" of ξ^n and η^n. The amount of this information is $H(\zeta^n) = nH(\zeta_1)$. The following theorem states that there is no more common information in ξ^n and η^n.

Theorem 1. Two DMSS's $\mathfrak{X}$ and $\mathfrak{Y}$ and ε $(0 < \varepsilon < 1)$ are given. If δ_n tends to 0, we have

$$\limsup_{n \to \infty} \frac{1}{n} J_n(\varepsilon, \delta_n) \leq H(\zeta_1)\,, \tag{5a}$$

Furthermore there exists a sequence $\delta_n \to 0$ such that the equality

$$\lim_{n\to\infty} \frac{1}{n} J_n(\varepsilon, \delta_n) = H(\zeta_1) \tag{5b}$$

holds. We remark that the above limit does not depend on ε.

Originally we were interested to know in which cases is equality between the limit of (5b) and the mutual information $\lim_{n\to\infty} \frac{1}{n} I(\xi^n \wedge \eta^n) = I(\xi_1 \wedge \eta_1)$. The inequality $H(\zeta_1) \leq I(\xi_1 \leq \eta_1)$ is obvious.

Corollary 1. The limit $H(\zeta_1)$ of the common information equals $I(\xi_1 \wedge \eta_1)$ only in the trivial case when ξ_1 and η_1 are "conditionally independent", i.e. if and only if ξ_1 and η_1 are independent for any fixed value of ζ_1.

Proof of the corollary. The statement is an immediate consequence of the following trivial identity:

$$I(\xi \wedge \eta) = H(\zeta) + \sum_z P(\zeta = z) \cdot I(\xi \wedge \eta \mid \zeta = z).$$

Hence follows that $I(\xi \wedge \eta) = H(\zeta)$ holds if and only if $I(\xi \wedge \eta \mid \zeta = z) = 0$ for any fixed z.

Proving Theorem 1 we shall answer also the following question which we believe to be interesting in itself.

Definition 6. Given $\lambda (0 < \lambda \leq 1)$ we call a λ-block a pair of sets (A, B) with $A \subset X^n$ and $B \subset Y^n$ when

$$P(\xi^n \in A \mid \eta^n \in B) \geq \lambda \qquad P(\eta^n \in B \mid \xi^n \in A) \geq \lambda \tag{6}$$

holds.

Comparing (3) and (6) it is evident that our ergodic classes are λ-blocks for every λ. The question is whether a λ-block may be essentially smaller than the smallest ergodic class. According to Theorem 2 the answer is negative.

Theorem 2. Putting

$$Q(n, \lambda) = \min P(\xi^n \in A,\ \eta^n \in B)$$

$$P(\xi^n \in A \mid \eta^n \in B) \geq \lambda$$

$$P(\eta^n \in B \mid \xi^n \in A) \geq \lambda$$

for any sequence $\lambda_n = 2^{-n\varepsilon_n}$ with $\varepsilon_n \to 0$ we have

$$\lim_{n\to\infty} \frac{1}{n} \cdot \log Q(n, \lambda_n) = \min_z P(\zeta = z)\,.$$

In the following paragraphs we shall reduce the coding problem to a problem concerning blocks and prove Theorems 1 and 2 by means of a pivotal lemma. We shall often use a very elementary Markov-type inequality for bounded random variables. Throughout the paper it is referred to as the reverse Markov inequality. A proof — if needed — is to be found in Loeve [5]. For the terminology of Markov-chains see [1] and [5].

3. Blocks

The consequences of inequality (1) are presented.

Proposition 1. The binary block codes f and g of ξ^n and η^n respectively are supposed to have the same length $l = l(f(\xi^n))$ and satisfy (1). Denoting by $C(\varepsilon)$ the set of those c's in I^l for which

$$P(f(\xi^n) = g(\eta^n) \mid f(\xi^n) = c) \geq 1 - \sqrt{\varepsilon} \qquad \text{and}$$

$$P(f(\xi^n) = g(\eta^n) \mid g(\eta^n) = c) \geq 1 - \sqrt{\varepsilon} \tag{7}$$

we have $P[f(\xi^n) = g(\eta^n) \in C(\varepsilon)] \geq 1 - 4\sqrt{\varepsilon}$.

Proof. Let us introduce the sets

$$A(\varepsilon') = \{c \in I^l; \quad P(g(\eta^n) = c \mid f(\xi^n) = c) > 1 - \varepsilon'\} \qquad \text{and}$$

$$B(\varepsilon') = \{c \in I^l; \quad P(f(\xi^n) = c \mid g(\eta^n) = c) > 1 - \varepsilon'\}.$$

A reverse Markov inequality combined with (1) implies that $P[f(\xi) \in A(\varepsilon')] \geq$ $\geq 1 - \frac{\varepsilon}{\varepsilon'}$ and $P[g(\eta) \in B(\varepsilon') \geq 1 - \frac{\varepsilon}{\varepsilon'}$. Choosing $\varepsilon' = \sqrt{\varepsilon}$ we write $C(\varepsilon) =$ $= A(\sqrt{\varepsilon}) \cap B(\sqrt{\varepsilon})$. Obviously for $c \subset C(\varepsilon)$ one has (7).

On the other hand we obtain

$$\sum_{c \in A(\sqrt{\varepsilon})} P[f(\xi^n) = g(\eta^n) = c] \geq (1 - \sqrt{\varepsilon}) \cdot P[f(\xi^n) \in A(\sqrt{\varepsilon})] \geq (1 - \sqrt{\varepsilon})^2 \geq 1 - 2\sqrt{\varepsilon}$$

and similarly: $P[f(\xi^n) = g(\eta^n),\ g(\eta^n) \in B(\sqrt{\varepsilon})] \geq 1 - 2\sqrt{\varepsilon}$.

For the intersection of $A(\sqrt{\varepsilon})$ and $B(\sqrt{\varepsilon})$ thus follows: $P[f(\xi^n) = g(\eta^n) \in$ $\in C(\varepsilon)] \geq 1 - 4\sqrt{\varepsilon}$. Proposition 1 says that a large part of (X^n, Y^n) consists of $(1 - \sqrt{\varepsilon})$-block. In this way one can see that the number $l(f)$ — i.e. the length of the "stem" of the Y-scheme is related to the "number of blocks in (X^n, Y^n)". We shall show that every block is essentially an ergodic class or the union of several ones. Before doing this we need one more proposition. We use the notations of Definition 3. Let us be given a sequence of Y-schemes $Y^{(n)}_{\varepsilon, \delta_n}$ with codes

f_n, g_n, f'_n, g'_n, and δ_n tending to 0. Write $l_n = l(f_n)$. We want to obtain an upper bound for $\limsup_{n\to\infty} \frac{1}{n} l_n$. Doing this it will be useful to replace f_n by a fuitable code of it.

Proposition 2. If $\tilde{f}_n = U_n(f_n)$ is any code of f_n having error probability less than λ (where $\lambda < 1 - \varepsilon$), of some fixed length, the inequality

$$\limsup_{n\to\infty} \frac{1}{n} l(\tilde{f}_n) \geqq \limsup_{n\to\infty} \frac{1}{n} l_n \qquad \text{follows.}$$

Proof. It is obvious that juxtaposing $\tilde{f}_n$ and f_n we obtain a code $\tilde{f}_n f'_n$ of ξ^n with error probability less than $\varepsilon + \lambda$. Thus Shannon's classical source coding theorem implies that $\limsup \frac{1}{n} l(\tilde{f}_n f'_n) \geq H(\xi_1)$. Since $\delta_n \to 0$ implies that $\lim_{n\to\infty} \frac{1}{n} l(\tilde{f}_n f'_n) = H(\xi_1)$, we have $\limsup_{n\to\infty} \frac{1}{n} l(\tilde{f}_n) f'_n) \geq \lim \frac{1}{n} l(f_n f_n)$. Hence our assertion becomes trivial. Put

$$\varphi_n(\xi^n) = \big(f_n(\xi^n), \zeta^n\big) \quad \Psi_n(\eta^n) = \big(g_n(\eta^n), \zeta^n\big) \tag{8}$$

Here ζ_n is the random variable defined in § 1. φ_n and ψ_n are such refinements of f_n and g_n which take different values in different ergodic classes. Obviously:

$$P\big(\varphi_n(\xi^n) = \Psi_n(n^n)\big) = P\big(f_n(\xi_n) = g_n(\eta^n)\big) > 1 - \varepsilon\,. \tag{9}$$

Thus the codes φ_n and ψ_n satisfy the conditions of Proposition 1 and so the assertion remains valid for them. From now on, $C(\varepsilon)$ corresponds to φ_n and ψ_n. Since φ_n is a refinement of f_n, from Proposition 2 we obtain for any code $\tilde{\varphi}_n = U_n(\varphi_n)$ having error probability less than λ that

$$\limsup_{n\to\infty} \frac{1}{n} \cdot l(\tilde{\varphi}_n) \geq \limsup_{n\to\infty} \frac{1}{n} \cdot l(f_n)\,.$$

In the following we shall give an upper bound of $\limsup_{n\to\infty} \frac{1}{n} \cdot l(\tilde{\varphi}_n)$ for a suitable series of codes U_n.

4. A lemma

First of all we prove that φ_n (defined by (8)) does not take too many different values in a single ergodic class. The proof is based upon our main

Lemma. Suppose for the sets $A_n \subset X^n$ and $B_n \subset Y^n$ that

$$P(\xi^n \in A_n \,|\, \eta^n \in B_n) \geq 2^{n-\varepsilon_n} \text{ and } P(\eta^n \in B_n \,|\, \xi^n \in A_n) \geq 2^{n-\varepsilon_n} \tag{10}$$

where $\varepsilon_n \to 0$. For the sake of simplicity let $\varepsilon_n > \frac{1}{n}$. If $A_n \subset \tilde{A}_n$ and $B_n \subset \tilde{B}_n$ where $\tilde{A}_n$ and $\tilde{B}_n$ are ergodic classes, it follows that

$$\frac{P(\xi^n \in A_n, \eta^n \in B_n)}{P(\xi^n \in \tilde{A}_n, \eta^n \in \tilde{B}_n)} \geq 2^{-nC \cdot \varepsilon_n \cdot \log^2 \varepsilon_n}$$

where C is a constant depending only on the distribution of the pair (ξ_1, η_1).

Proof. The simple main idea of the proof is that if we get from A_n to B_n with a large probability and vice versa then applying formally both transitions one after the other we get from A_n back to A_n with quite a large probability. On the other hand iterating this formal transition many times, the set A_n "spreads" over its whole ergodic class. Passing over to the demonstration, we write

$$W(x' \mid x) = \sum_y P(\xi_1 = x' \mid \eta_1 = y) \cdot P(\eta_1 = y \mid \xi_1 = x) \quad \text{if} \quad x, x' \in X.$$

Now we define Markov chains $\{\xi_{nk}\}_{k=1}^{\infty}$. Let $\xi_{n0} = \xi_n$ and $P(\xi_{n(k+1)} = x' \mid \xi_{nk} = x) = W(x' \mid x)$. For different n's we suppose the variables ξ_{nk} to be independent. Let us write $W^{(k)}(x' \mid x) = P(\xi_{1k} = x' \mid \xi_{10} = x)$. It is easy to see that the ergodic classes of the transition matrix $W(x' \mid x)$ just coincide with the X_i's of (3). That is why they were called ergodic classes. Especially the matrix $W(x' \mid x)$ has no transient element. This follows also from the fact that $P[\xi_1 = x]$ — evidently an invariant distribution of W — is positive for every $x \in X$. Moreover $W(x \mid x)$ is positive for every $x \in X$, hence no ergodic class is cyclic. Thus we can apply the well-known theorem on convergence of the transition probabilities to the case of $W(x' \mid x)$. It states that there exist probability distributions $\Pi_i(x)$ for which $\Pi_i(x) > 0$ if and only if $x \in X_i$, and also a δ $(0 \leq \delta < 1)$ depending only on the matrix W such that $| W^{(k)}(x' \mid x) - \Pi_i(x') | < \delta^k$ for every x and x' in X_i. From the last inequality we obtain

$$W^{(k)}(x' \mid x) \leq 2^{c_1 \delta^k} \Pi_i(x') \tag{11}$$

where C_1 is a constant depending only on the matrix W. Let us write $\xi^{nk} = \xi_{1k}\, \xi_{2k} \ldots \xi_{nk}$ and especially $\xi^{n0} = \xi^n$. We define

$$W_n^{(k)}(\bar{x} \mid x^n) = P(\xi^{nk} = \bar{x}^n \mid \xi^{n0} = x^n)$$

and especially $W_1^{(1)}(\bar{x} \mid x) = W(\bar{x} \mid x)$. Similarly $W_n^{(k)}(A' \mid A) = P(\xi^{nk} \in A' \mid \xi^{n0} \in A)$. $\{\xi^{ni}\}_{i=0}^{\infty}$ is a Markov chain with k-step transition probabilities $W_n^{(k)}(\bar{x}^n \mid x^n)$. The identity

$$W_n^{(1)}(\bar{x}_n \mid x^n) = \sum_{y^n \in Y^n} P(\xi^n = \bar{x}^n \mid \eta^n = y^n) \cdot P(\eta^n = y^n \mid \xi^n = x^n)$$

remains valid, hence $W_n^{(1)}$ plays the role of the matrix $W(x' \mid x)$ on the product space. On the other hand

$$W_n^{(k)}(\bar{x}^n \mid x^n) = \prod_{j=1}^{n} W^{(k)}(\bar{x}_j \mid x_j) . \tag{12}$$

Every sequence $t = i_1 i_2 \ldots i_n$ of the indices correspond to the ergodic class $X^{(t)} = X_{i_1} \times X_{i_2} \ldots \times X_{i_n} \subset X^n$. We write $\Pi_t(\bar{x}^n) = \prod_{j=1}^{n} \Pi_{ij}(\bar{x}_j)$. The invariant distribution Π_t plays the role of Π_i in the product space. From (11) and (12) we obtain for every $t = i_1 i_2 \ldots i_n$ and $\bar{x}^n$, $x^n \in X^{(t)}$ that

$$W_n^{(k)}(\bar{x}^n \mid x^n) \leqq 2^{C_1 \cdot n \delta^k} \Pi_t(\bar{x}^n) .$$

Now $\left\{\dfrac{P(\xi^n = x^n)}{P(\xi^n \in X^{(t)})}\right\}_{x^n \in X^{(t)}}$ is an invariant distribution on $X^{(t)}$. However every ergodic class has a unique invariant distribution and therefore $\dfrac{P(\xi^n = x^n)}{P(\xi^n \in X^t)} = \Pi_t(x^n)$, i.e.

$$W_n^{(k)}(\bar{x}_n \mid x^n) \leqq 2^{C_1 \cdot n \cdot \delta^k} \frac{P(\xi^n = \bar{x}^n)}{P(\xi^n \in X^{(t)})} . \tag{13}$$

Let us now consider the set A_n of the lemma. $A_n \subset \tilde{A}_n$ with $\tilde{A}_n$ being an ergodic class. Replacing $X^{(t)}$ by $\tilde{A}_n$ (12) results in the inequality

$$W_n^{(k)}(A_n \mid A_n) = \frac{1}{P(\xi^n \in A_n)} \cdot \sum_{x^n \in A_n} W_n^{(k)}(A_n \mid x^n) \cdot P(\xi^n = x^n) \leqq$$

$$\leqq \frac{P(\xi^n \in A_n)}{P(\xi^n \in \tilde{A}_n)} \cdot 2^{C_1 \cdot n \cdot \delta^k} .$$

We want to obtain a lower bound for $W_n^{(k)}(A_n \mid A_n)$. Let us write $\lambda_n = 2^{-n\varepsilon_n}$. Starting with the estimstion of $W_n^{(1)}(\mathrm{A}_n \mid A_n)$ we have

$$W_n^{(1)}(A_n \mid A_n) = \sum_{y^n \in Y^n} P(\xi^n \in A_n \mid \eta^n = y^n) \cdot P(\eta^n = y^n \mid \xi^n \in A_n) =$$

$$= \frac{1}{P(\xi^n \in A_n)} \cdot \sum_{y^n \in Y^n} P^2(\xi^n \in A_n \mid \eta^n = y^n) \geq \tag{14}$$

$$\geq \frac{1}{P(\xi^n \in A_n)} \cdot \sum_{y^n \in B_n} P^2(\xi^n \in A_n \mid \eta^n = y^n) \cdot P(\eta^n = y^n) .$$

From the conditions of the lemma we have

$$\lambda_n \leq \sum_{y^n \in B_n} P(\xi^n \in A_n \mid \eta^n = y^n) \cdot \frac{P(\eta^n = y^n)}{P(\eta^n \in B^n)} .$$

Applying a reverse Markov inequality we obtain

$$\sum_{y^n;\, P(\xi^n \in A_n \mid \eta^n = y^n) > \frac{1}{2}\lambda_n} \frac{P(\eta^n = y^n)}{P(\eta^n \in B_n)} \geq \frac{1}{2}\lambda_n .$$

Multiplying both sides by $P(\eta^n \in B_n)$ this gives

$$\sum_{y^n;\, P(\xi^n \in A_n \mid \eta_n) > \frac{1}{2}\lambda_n} P(\eta^n = y^n) \geq \frac{1}{2}\lambda_n \cdot P(\eta^n \in B_n) .$$

Combining the last inequality with (14) and (10) we obtain

$$W_n^{(1)}(A_n \mid A_n) \geq \frac{P(\eta^n \in B_n)}{P(\xi^n \in A_n)} \cdot \frac{1}{8}\lambda_n^3 \geq \frac{P(\xi^n \in A_n, \eta^n \in B_n)}{P(\xi^n \in A_n)} \cdot \left(\frac{\lambda_n}{2}\right)^3 \geq 2 \cdot \left(\frac{\lambda_n}{2}\right)^4 .$$

Now we iterate the above estimates. By induction

$$W_n^{(2^r)}(A_n \mid A_n) \geq 2 \cdot \left(\frac{\lambda_n}{2}\right)^{4^{r+1}} .$$

Putting $k = 2^r$ and $\lambda_n = 2^{-n\varepsilon_n}$

$$W_n^{(k)}(A_n \mid A_n) \geq 2^{-4k^2(n \cdot \varepsilon_n + 1)} .$$

From the assumption that $\varepsilon_n > \frac{1}{n}$ we thus obtain

$$W_n^{(k)}(A_n \mid A_n) \geq 2^{-8k^2 \cdot n \cdot \varepsilon_n} .$$

Comparing this inequality with (13) we have

$$2^{-8k^3 n \varepsilon_n} \leq W_n^{(k)}(A_n \mid A_n) \leq \frac{P(\xi^n \in A_n)}{P(\xi^n \in \widetilde{A}_n)} \cdot 2^{c_1 n \delta^k},$$

i.e.

$$\frac{P(\xi^n \in A_n)}{P(\xi^n \in \widetilde{A}_n)} \geq 2^{-n[8k^2 \varepsilon_n + c_1 \cdot \delta^k]} .$$

Finally let us choose k in the exponent of the right-hand side in a suitable manner. Let $\delta^k = 2^{-\delta_1 k}$. We prescribe for k that $\frac{2}{\delta_1} \log \frac{1}{\varepsilon_n} \leq k \leq \frac{2}{\delta_1} \log \frac{1}{\varepsilon_n}$. It is possible to choose $k = 2^r$ within these bounds. Now

$$\frac{P(\xi^n \in A_n)}{P(\xi^n \in \widetilde{A}_n)} \geq 2^{-n(C_2 \cdot \varepsilon_n \cdot \log^2 \varepsilon_n)}$$

Using (10) this yields

$$\frac{P(\xi^n \in A_n, \eta^n \in B_n)}{P(\xi^n \in \tilde{A}_n, \eta^n \in \tilde{B}_n)} \geq 2^{-n\varepsilon_n} \cdot \frac{P(\xi^n \in A_n)}{P(\xi^n \in \tilde{A}_n)} \geq 2^{-n \cdot C \cdot \varepsilon_n \cdot \log^2 \varepsilon_n}$$

what we wanted to prove.

5. Proof of Theorems 1 and 2. Random common length

We have all the tools to finish the proofs of Theorems 1 and 2.

Proof of Theorem 2. This theorem is an immediate consequence of the lemma. Given any series $\lambda_n = 2^{-n\varepsilon_n}$ with ε_n tending to 0 one has the inequality:

$$\frac{1}{n} \cdot \log Q(n, \lambda_n) \geq -C \cdot \varepsilon_n \cdot \log^2 \varepsilon_n + \min_t \left(\frac{1}{n} \cdot \log P(\xi^n \in X^{(t)}, \eta^n \in Y^{(t)}) \right).$$

Since $\varepsilon_n \cdot \log^2 \varepsilon_n$ tends to 0 when $\varepsilon_n \to 0$, evidently

$$\liminf_{n \to \infty} \frac{1}{n} \cdot \log Q(n_1 \lambda_n) \geq \min_z P(\zeta = z).$$

On the other hand the choice $(A_n, B_n) = X_{z_0}, Y_{z_0})^n$ where z_0 is the index of the ergodic class having minimum probability, gives that $\lim_{n \to \infty} \frac{1}{n} \log Q(n_1 \lambda_n) = \min_z P(\zeta = z]$.

End of the proof of Theorem 1. Now we turn to the proof of the inequality (5a). Let us consider the function φ_n defined in (8). Apart from a set of probability $4\sqrt{\varepsilon}$ this function is defined on $(1 - \sqrt{\varepsilon})$-blocks (see Proposition 1). Let us choose $\lambda > 4\sqrt{\varepsilon}$. We shall prove that there exists a code $\tilde{\varphi}_n = U_n(\varphi_n)$ of $\varphi_n(\xi_n)$ with some fixed length and error probability less than λ, which performs the inequality $\limsup_{n \to \infty} \frac{1}{n} \cdot l(\varphi_n) \leq H(\zeta_1)$.

The construction of the code U_n is the following: let U_n take a constant value on the complementary set of $C(\varepsilon)$ defined in Proposition 1. On $C(\varepsilon)$ we code the values of $\varphi_n(\xi^n) = (f_n(\xi^n), \zeta^n)$ in two steps. First we construct a $(\lambda - 4\sqrt{\varepsilon})$ — code (see Definition 2) of ζ^n with codewords' length less than $nH(\zeta_1) + k_1\sqrt{n}$. According to the lemma φ_n takes less than $2^{K_2 \log^2 n}$ different values in an ergodic class, i.e. for any fixed value of ζ^n. This is so because the different values of the restriction $\varphi_n \mid C(\varepsilon)$ are taken of different $(1 - \sqrt{\varepsilon})$-blocks. Thus the whole code U_n may have length less than $nH(\zeta_1) + K_1\sqrt{n} + K_2 \log^2 n$.

So we obtain that $\limsup_{n\to\infty} \frac{1}{n} l(\tilde{\varphi}_n) \leq H(\zeta_1)$. This proves (5a). The second part of the theorem is rather trivial. It is enough to prove at this point that any ε_1-code of the sequence ζ^n with $\varepsilon_1 < \varepsilon$ can be completed by addition of a complementary code of length $nH(\xi_1 \mid \zeta_1) + K\sqrt{n}$ to an $\varepsilon = \varepsilon_1 + \varepsilon_2$-code of the sequence ξ^n. Although this is very simple to show we prefer to give a reference where the proof is done (see [4]). Now our demonstration is complete. We have shown that it is not possible in general to code two DMSS's so that the resulting codes have some fixed common length of order n. All the same problem concerning the length of a possible random coincidence of codes has remained open. Which is the behaviour of the maximum expected length of random concidence? We shall show that this length is asymptotically equal to $nH(\zeta_1)$. Let us formulate first an exact form of the question.

Definition 6. Let $F_n(\xi^n)$ and $G_n(\eta^n)$ be ε-codes of the DMSS's $\mathfrak{X} = \{\xi_n\}$ and $\mathfrak{Y} = \{\eta^n\}$ respectively. Suppose that

$$l\big(F_n(\xi^n)\big) < (1 + \delta_n) \cdot H(\xi^n) \text{ and } l\big(G_n(\eta^n)\big) < (1 + \delta_n) \cdot H(\eta^n) .$$

Let the integer valued random variable ν_n be the length of the maximum common beginning segment of the binary sequences $F_n(\xi^n)$ and $G_n(\eta^n)$. The random variable ν_n is called the *random common length* of the codes F_n and G_n.

Theorem 3. Let $E(\nu_n)$ denote the expected value of ν_n. Supposing that $\delta_n \to 0$ we have

$$\limsup_{n\to\infty} E(\nu_n) \leq H(\zeta_1) .$$

Proof. Let F_n and G_n be given. From a reverse Markov inequality we obtain for $0 \leq x < 1$ that

$$P[\nu_n > (1 - \alpha) \cdot E(\nu_n)] \geq \frac{\alpha \cdot E(\nu_n)}{[(1 + \delta_n) \cdot H(\xi^n)] - (1 - \alpha) \cdot E(\nu_n)} \geq$$

$$\geq \frac{\alpha}{2H(\xi_1)} \cdot \frac{E(\nu_n)}{n} . \tag{15}$$

Let us fix α independently of n. Put $l_n = (1 - \alpha) \cdot E(\nu_n)$. If $f_n(\xi^n)$ and $g_n(\eta^n)$ denote the first l_n letters of the binary sequences $F_n(\xi^n)$ and $G_n(\eta^n)$ respectively, (15) yields

$$P[f_n(\xi^n) = g_n(\eta^n)] > \frac{\alpha}{2H(\xi_1)} \cdot \frac{E(\nu_n)}{n} .$$

Choosing for f'_n and g'_n the remaining segments of F_n and G_n we obtain Y-schemes having error probabilities equal to $\max\left(\varepsilon, 1 - \frac{\alpha}{2H(\xi_1)} \cdot \frac{E(\nu_n)}{n}\right)$. The common length is these Y-schemes is $l_n = (1 - \alpha) \cdot E(\nu_n)$. Thus Theorem 1 implies that

$$\limsup_{n\to\infty} \frac{E(\nu_n)}{n} \leq H(\zeta_1) .$$

The complexity interpretation of the result

This section will be understandable only for readers acquainted with the notion of complexity of a finite word given by Kolmogorov [1, 2] (see also [3], [4]). Let X be a finite alphabet containing 0 and 1, $| X | = S$. Let T be a universal Turing machine the operation of which results in a partial function $T(p, x)$, $T : I^* \times X^* \to X^*$. We define

$$K_T(y \mid x) = \min_{T(p,x)=y} l(p) .$$

It is known that if T' is another universal machine then $| K_T - K_{T'} | \leq C_{TT'}$ where the constant $C_{TT'}$ depends only on the two machines. Let us fix T once for ever and denote $K(y \mid x) = K_T(y \mid x)$ the "conditional compexity" of y with respect to x, $K(x) = K(x \mid \Lambda)$ the "complexity" of x, $K(x, y)$ the "common complexity" of x and y. The last can be defined for example as the length of the shortest program $p \in I^*$ for which $T(0p, \Lambda) = x$, $T(1p, \Lambda) = y$. We define further $I(x : y) = K(x) - K(x \mid y)$, the "mutual information" of x and y. We cite from [3] the following facts among which only (16) is not very easy to see. Let $\leq$ resp. $\asymp$ denote the inequality resp. equality up to an additive constant. By $\approx$ we mean equality up to an additive term $0(\log K(x, y))$. Then we have

$$K(x \mid y) \leq K(x) \leq l(x) \log s$$

$$K(z \mid y) + K(y \mid x) \geq K(z \mid x)$$

hence denoting $\Delta(x, y) = K(x \mid y) + K(y \mid x)$ Δ has approximately the properties of a metrics. Further

$$K(x, y) \approx K(x) + K(y \mid x) \tag{16}$$

hence $I(x : y) \approx I(y : x)$.

If $(\mathcal{X}, \mathcal{Y}) = \{\xi_n, \eta_n\}$ is a stationary source which is ergodic then for almost all $x = x_1x_2 \ldots$, $y = y_1y_2 \ldots$

$$\lim_{n\to\infty} \frac{1}{n} K(x^n) = H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(\xi^n), \quad \lim_{n\to\infty} \frac{1}{n} K(x^n, y^n) = \lim_{n\to\infty} \frac{1}{n} H(\xi^n, \eta^n)$$

$$\lim_{n\to\infty} \frac{1}{n} K(x^n \mid y^n) = \lim_{n\to\infty} \frac{1}{n} H(\xi^n \mid \eta^n), \quad \lim_{n\to\infty} \frac{1}{n} I(x^n : y^n) = \lim_{n\to\infty} \frac{1}{n} I(\xi^n, \eta^n).$$

The approximate symmetry of $I(x : y)$ suggested here the question: can we give a symmetric definition of it? For example is there a triple $a, b, c \in I^*$ of programs with

$$T(ca, \Lambda) = x, \; T(cb, \Lambda) = y$$

$$l(a) \approx K(x \mid y), \; l(b) \approx K(y \mid x), \; l(c) = I(x : y).$$

More generally let $x, y \in X^n$, we shall say that z represents some common information of x and y if $K(z \mid x) \approx 0$, $K(z \mid y) \approx 0$. Let us define

$$J(x, y) = \max_{z\,:\,K(z\mid x)\approx 0,\; K(z\mid y)\approx 0} K(z).$$

Here $\approx$ must be given in a bit more exact sense but we do not wish to be quite precise here. If $K(z) \approx J(x, y)$ then z is almost determined in the sense that it is in a given "sphere" of radius ≈ 0 in the metrics Δ. In this case we shall say that z *is* the common information of x and y. Our question was whether $J(x, y) \approx I(x : y)$. The answer is negative in a rather general case. Let ξ^n and η^n denote the random variables defined in § 2. (Suppose that $X = Y$.) Then in the product space with a probability tending to 1 we have by Theorem 1 that ζ^n *is* the common information in ξ^n and η^n.

We are indebted to Imre Csiszár sr. for his stating of the problem and constant aid.

REFERENCES

1. *Doob, J. L.:* Stochastic processes. Wiley, New York, 1953, pp. 170—185.
2. *Kolmogorov, A. N.:* Logical basis for information theory and probability theory. IEEE Transactions IT—14 (1968) pp. 662—664.
3. *Kolmogorov, A. N.:* Three approaches to the concept of "information quantity" (in Russian). Problemi peredachi informatsii 1 : 1 (1965) pp. 3—7.
4. *Körner, J.:* On a property of conditional entropy. Stud. Sci. Math. Hung. (1971) (to appear).
5. *Loeve, M.:* Probability Theory. Van Nostrand. New York, 1955. pp. 157 and 28—42.
6. *Martin-Löf, P.:* The definition of random sequences. Inf. and Control 9 (1966) pp. 602—619.
7. *Zvonkin, A. K.—Levin, L. A.:* Complexity of finite objects etc. (in Russian) Uspekhi Mat. Nauk 25 : 6 (1970) pp. 85—127.

Общая информация существенно больше, чем взаимная информация

П. ГААЧ—Й. КОРНЕР

(Будапешт)

С помощью трех представленных теорем уточняется понятие взаимной информации и доказывается, что термин «общая информация» является более общим, чем термин «взаимная информация».

P. Gács, J. Körner

Mathematical Research Institute

Budapest, V. Reáltanoda u. 13—15. Hungary

Problems of Control and Information Theory, Vol. 2 (2), pp. 163–171 (1973)

RELATIONS ENTRE LES MORPHISMES ET L'ÉQUIVALENCE DES AUTOMATES INDÉTERMINÉS

SYLWESTER PRZYBYLO

(Pologne)

(Received February 5, 1972)

Cette note a pour but d'établir des relations entre l'homomorphisme, l'épimorphisme et l'isomorphisme des automates abstraits indéterminés et la transformation d'équivalence (transformation d'équivalence faible) l'un des leurs dans l'autre et leurs d'équivalence (d'équivalence faible).

La note présente est une transposition des considérations sur des automates déterminés dans [1] pour des automates indéterminés et elle est á la fois un élargissement de la note mentionnée á des théorèmes qui ne sont caractéristiques que pour des automates indéterminés.

Comme on peut identifier les automates déterminés avec les automates indéterminés qui sont déterminés donc cette note comprend les résultats de la note [2].

Considérons les deux automates abstraits indéterminés ([2], p. 152)

$$S = [X, Y, Z, h] \quad \text{et} \quad S'[X', Y', Z', h']$$

et rappelons quelques notions nécessaires.

Nous considérons les trois transformations ξ, η, ϱ telles que:

$$\xi : X \to X', \quad \eta : Y \to Y', \quad \varrho : Z \to Z'.$$

Pour simplifier la notation designons pour $z \in Z$, $y \in Y$, $M \subseteq Y \times Z$

$$\eta\varrho(y, z) = [\eta(y), \varrho(z)], \quad \eta\varrho(M) = \{[\eta(y), \varrho(z)] / [y, z] \in M\}.$$

Définition 1

a) $[\xi, \eta, \varrho]$ sont dites l'homomorphisme de l'automate indéterminé S dans l'automate indéterminé S' si:

$$\bigwedge_{z \in Z} \bigwedge_{x \in X} h'[\varrho(z), \xi(x)] =$$

$$= \{[y', z'] / \bigvee_{y \in Y} \bigvee_{z_1 \in Z} ([y, z_1] \in h(z, x) \wedge \eta(y) = y' \wedge \varrho(z_1) = z')\}$$

c'est-á-dire si:

$$\bigwedge_{z \in Z} \bigwedge_{x \in X} h'(\varrho(z), \xi(x)) = \eta\varrho(h(z, x));$$

b) $[\xi, \eta, \varrho]$ sont dites l'épimorphisme s'il est l'homomorphisme et, en outre, les fonctions ξ, η, ϱ transforment respectivement X sur X', Y sur Y', Z sur Z';

c) $[\xi, \eta, \varrho]$ sont dites l'isomorphisme s'il est l'épimorphisme et, en outre, les fonctions ξ, η, ϱ sont univoques.

A present nous allons introduire les notions de morphismes forts.

Définition 2

a) $[\xi, \eta, \varrho]$ sont dites l'homomorphisme fort si:

$$\bigwedge_{z,z^* \in Z} \bigwedge_{x \in X} \bigwedge_{y \in Y} [y, z^*] \in h(z, x) \Leftrightarrow [\eta(y), \varrho(z^*)] \in h' [\varrho(z), \xi(x)];$$

b) $[\xi, \eta, \varrho]$ sont dites l'épimorphisme fort s'il est l'homomorphisme fort et, en outre, les fonctions ξ, η, ϱ transforment respectivement X sur X', Y sur Y', Z sur Z'.

c) $[\xi, \eta, \varrho]$ sont dites l'isomorphisme fort s'il est l'épimorphisme fort et, en outre, les fonctions ξ, η, ϱ sont univoques.

Les corollaires suivants s'ensuivent directement des définitions 1 et 2.

Corollaire 1

1. Chaque homomorphisme fort de S dans S' est l'homomorphisme de S dans S'.

2. Chaque épimorphisme fort de S sur S' est l'épimorphisme de S sur S'.

3. $[\xi, \eta, \varrho]$ est l'isomorphisme fort si et seulement s'il est l'isomorphisme de S sur S'.

Maintenant désignons par v_s une fonction de sortie définie sur des mots ([2], p. 123), et par e le mot vide et prénons:

Définition 3

a) $$v_s(z_1, p) \overset{df}{=} \begin{cases} \{e\} & \text{si} \quad p = e \\ \left\{ y_1, \ldots, y_n \middle| \bigvee_{z_2} \ldots \bigvee_{z_{n+1}} \left(\bigwedge_{i=1}^{n} [y_i, z_{i+1}] \in h(z_i, x_i) \right) \right\} & \\ & \text{si } p = x_1 \ldots x_n, n \geq 1 \end{cases}$$

b) $$v_s(M, p) \overset{df}{=} \bigcup_{z' \in Z} h_{z'}(M, p).$$

Définition 4

a) L'état $z \in Z$ est équivalent á l'état $z' \in Z'$ (nous écrivons $z \sim z'$) si

$$\bigwedge_{p \in W(X)} v_s(z, p) = v_{s'}(z', p),$$

où $W(X)$ est l'ensemble des mots construits sur l'alphabet X.

b) L'ensemble $M \subset Z$ est équivalent á l'ensemble $M' \subset Z'$ (nous écrivons $M \sim M'$) si

$$\bigwedge_{p \in W(X)} v_S(M, p) = v_{S'}(M', p) .$$

Définition 5

a) L'automate indéterminé S est transformé équivalentement dans l'automate indéterminé S', ce que nous écrivons $S \subsetneqq S'$, si

$$\bigwedge_{z \in Z} \bigvee_{z' \in Z'} z \sim z' .$$

Nous disons que l'automate indéterminé S est équivalent á l'automate indéterminé S', en symbole $S \sim S'$, si $S \subsetneqq S'$ et $S' \subsetneqq S$;

b) L'automate indéterminé S est transformé équivalentement faiblement dans l'automate indéterminé S', ce que nous écrivons $S \underset{\approx}{\subsetneqq} S'$, si

$$\bigwedge_{M \subset Z} \bigvee_{M' \subset Z'} M \sim M' .$$

Nous disons que l'automate indéterminé S est équivalent faiblement á l'automate indéterminé S', en symbole $S \approx S'$, si $S \underset{\approx}{\subsetneqq} S'$ et $S' \underset{\approx}{\subsetneqq} S$.

Corollaire 2

1. Si $S \subsetneqq S'$ alors $S \underset{\approx}{\subsetneqq} S'$.
2. Si $S \sim S'$ alors $S \approx S'$.

Définition 6

a) L'automate indéterminé $S = [X, Y, Z, h]$ est dit réduit si

$$\bigwedge_{z_1, z_2 \in Z} [(z_1 \sim z_2) \Rightarrow (z_1 = z_2)] :$$

b) Nous disons que l'automate indéterminé S' est le réducteur de l'automate indéterminé S si S' est reduit équivalent á S.

c) L'automate indéterminé S est dit minimal si:

$$\bigwedge_{z \in Z} \bigwedge_{M \subset Z} [(\{z\} \sim M) \Rightarrow (\{z\} = M)] ;$$

d) Nous disons que l'automate indéterminé est le minimal de l'automate S si S' est minimal et équivalent faiblement á S;

e) L'automate indéterminé S est dit fort réduit si:

$$\bigwedge_{M, N \subset Z} [(M \sim N) \Rightarrow (M = N)] ;$$

f) Nous disons que l'automate indéterminé S' est le réducteur fort de l'automate indéterminé S si S' est fort réduit et équivalent faiblement á S.

Corollaire 3

1. Chaque automate indéterminé fort réduit est minimal, chaque automate indéterminé minimal est réduit.

2. Chaque réducteur fort de l'automate indéterminé S est le minimal de S.

Définition 7

Nous disons que les trois fonctions $[\xi, \eta, \varrho]$ étant l'homomorphisme (l'homomorphisme fort) de l'automate indéterminé S' sont l'homomorphisme de Z (l'homomorphisme fort de Z) si ξ et η transforment idéntiquement respectivement X sur X' et Y sur Y' (alors évidemment $X = X'$, $Y = Y'$).

Désignons par $[I_X, I_Y, \varrho]$ l'homomorphisme de Z defini ci-dessus. Nous disons que $[I_X, I_Y, \varrho]$ est l'épimorphisme de Z (l'épimorphisme fort de Z) si ϱ transforme Z sur Z', et qu'il est l'isomorphisme de Z (l'isomorphisme fort de Z) si en outre la fonction ϱ est univoque.

Dans la suite de la note nous allons écrive simplement ϱ au lieu de $[I_X, I_Y, \varrho]$ et nous omettrons "de Z" dans ce cas.

Théorème 1

Si la fonction ϱ établit l'homomorphisme de l'automate S dans l'automate indéterminé S', alors l'automate S est transformé équivalentement dans S'.

Il suffit de prouver que

$$\bigwedge_{z \in Z} \bigwedge_{p \in W(X)} v_s(z, p) = v_{s'}\big(\varrho(z), p\big). \tag{1}$$

Le raisonnement sera ici le même que celui de la démonstration du théorème 3.5 (qui concerne l'épimorphisme) dans [2], p. 155.

On voit facilement que la thèse du théorème 1 ne peut pas être remplacée par l'équivalence entre l'automate indéterminé S et S'.

Nous voyons par l'example 1 que le théorème inverse en théorème 1 n'est pas valable.

Exemple 1

Considérons les deux automates indéterminés S et S' aux graphes suivants:

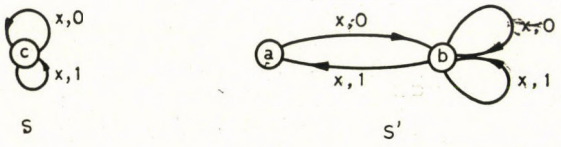


Fig. 1

S est transformé équivalentement dans S' car $c \sim b$. Nous allons prouver que l'homomorphisme de S dans S' n'existe pas ici.

Comme la fonction ϱ transforme Z dans Z' nous n'avons que les deux cas:

a) $\varrho(c) = a$ et alors

$$h'(\varrho(c), x) = h'(a, x) = \{(0, b)\} \neq \{(0, a), (1, a)\} = I_Y \varrho(h(c, x)),$$

ou bien

b) $\varrho(c) = b$ et alors

$$h'(\varrho(c), x) = h'(b, x) = \{(0, b), (1, b), (1, a)\} \neq \{(0, b), (1, b)\} = I\varrho_Y(h(c, x)).$$

Théorème 2

Si ϱ est l'épimorphisme de l'automate indéterminé S sur S' alors $S \sim S'$.

Démonstration. Vu le théorème 1 il n'y a qu'à prouver que $S' \subsetneq S$. Pour la fonction ϱ il existe la fonction (partiellement inverse) $\gamma : Z' \to Z$ telle que

$$\varrho(\gamma'(z')) = z'$$

pour chaque z' de Z'.

Prenons en considération un état z' de Z', donc

(2) $$\varrho(\gamma(z_2)) = z_2,$$

et posons

(3) $$z_1 = \gamma(z_1).$$

Il s'ensuit de (1) que

(4) $$z_1 \sim \varrho(z_1).$$

En tenant compte de (2), (3), (4) nous avons

$$z_2 = \varrho(\gamma(z_2)) = \varrho(z_1) \sim z_1$$

Pour le z_2 nous avons trouvé l'état z_1 tel que

$$z_2 \sim z_1, \quad \text{donc} \quad S' \subsetneq S.$$

Il est évident que ϱ étant l'isomorphisme de l'automate S sur S' entraîne $S \sim S'$.

On déduit le corollaire suivant dans la note [1] pour des automates déterminés S et S':

"Si $S \subsetneq S'$ et l'automate S' est réduit alors il existe l'homomorphisme de S dans S'."

Cette affirmation n'est plus valable pour des automates indéterminés.

En effet, dans l'exemple $1S \subsetneq S'$ et l'automate S' est réduit (parce que a n'est pas équivalent á b), et pourtant, ce qui a été déjá démontré, l'homomorphisme de S dans S' n'existe pas.

Nous avons toutefois le théorème suivant

Théorème 3

Si $S \subsetneq S'$ et l'automate S^3 est fort réduit alors il existe l'homomorphisme de S dans S'.

Démonstration. Comme S' est fort réduit, donc réduit, alors pour chaque état $z \in Z$ il existe un état $z' \in Z'$ et un seul tel que $z \sim z'$. Ainsi nous pouvons définir la fonction ϱ

$$\varrho : z \to \varrho(z) = z' \quad \text{tel que} \quad z \sim z',$$

La fonction ϱ transforme Z dans Z'.

Nous avons

$$\bigwedge_{z \in Z} z \sim \varrho(z), \tag{5}$$

De (5) et du corollaire 2.6.1 dans [2], p. 136 nous recevons pour tout $x \in X$ et $y \in Y$:

$$h_y(z, x) \sim h'_y(\varrho(z), x). \tag{6}$$

Comme

$$h_y(z, x) \sim \varrho(h_y(z, x)), \tag{7}$$

il s'ensuit de (6) et (7)

$$\varrho(h_y(z, x)) \sim h'_y(\varrho(z), x).$$

Vu que l'automate S' est fort réduit on a:

$$\zeta(h_y(z, x)) = h'_y(\varrho(z), x),$$

de la

$$\zeta(h_y(z, x)) \times \{y\} = h'_y(\varrho(z), x) \times \{y\},$$

alors

$$\bigcup_{y \in Y} \varrho\big(h_y(z, x)\big) \times \{y\} = \bigcup_{y \in Y} h'_y\big(\varrho(z), x\big) \times \{y\},$$

donc

$$I_Y \varrho(h(z, x)) = h'(\varrho(z), x),$$

et cela finit la démonstration.

Si en outre nous admettons que l'automate S est réduit alors la fonction ϱ (définie dans la démonstration ci-dessus) sera univoque.

La relation

$$\varrho(h_y(z, x)) = h'_y(\varrho(z), x)$$

est pour tout $z \in Z$ et $x \in X$ équivalent á ceci:

$$z^* \in h_y(z, x) \Leftrightarrow (\varrho z^*) \in h'_y(\varrho(z), x)$$

donc á

$$[y, z^*] \in h(z, x) \Leftrightarrow [y, \varrho(z^*)] \in h'(\varrho(z), x),$$

ce qui montre qu'il est exact le théorème suivant

Théorème 4

Si $S \underset{\sim}{\subseteq} S'$, S est réduit et S' est fort réduit, alors il existe l'homomorphisme fort de S dans S'.

L'hypothese $S \underset{\sim}{\subseteq} S'$ est essentielle dans les théorèmes 3 et 4 et elle ne peut pas être remplacée par $S \underset{\approx}{\subseteq} S'$.

Exemple 2

Considérons les deux automates indéterminés S et S' aux graphes suivants:

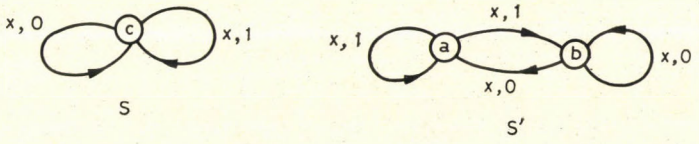


Fig. 2

L'automate S est transformé équivalentement faiblement dans S' car $\{c\} \sim \{a, b\}$ et S est réduit, l'automate S' est fort réduit. Pourtant l'homomorphisme de S dans S' n'existe pas. Comme la fonction ϱ transforme Z dans Z' nous n'avons que les deux cas:

a) $\varrho(c) = a$ et alors

$$h'(\varrho(c), x) = h'(a, x) = \{(1, a), (1, b)\} \neq \{1, a), (0, a)\} = I_Y \varrho(h(c, x))$$

ou bien

b) $\varrho(c) = b$ et alors

$$h'(\varrho(c), x) = h'(b, x) = \{(0, a), (0, b)\} \neq \{(0, b), (1, b)\} = I_Y \varrho(h(c, x)).$$

Dans tous les deux cas, comme nous le voyons, la fonction ϱ n'est pas l'homomorphisme de S dans S'.

Théorème 5

Si $S \sim S'$ et l'automate indéterminé S' est fort réduit, alors il existe l'épimorphisme (de Z) de l'automate S sur S'.

Démonstration. Comme $S \sim S'$ alors il existe une fonction $\varrho : Z \to Z'$ et une fonction $g : Z' \to Z$ telles que

$$z_1 \sim \varrho(z_1) \qquad \text{pour chaque} \qquad z_1 \in Z \tag{8}$$

et

$$z_2 \sim g(z_2) \qquad \text{pour chaque} \qquad z_2 \in Z^3. \tag{9}$$

Il s'ensuit de (8) et (9) que

$$z_2 \sim g(z_2) \sim \varrho(g(z_2)).$$

Puisque l'automate indéterminé S' est fort réduit, donc réduit, on a $z_2 = \varrho(g(z_2))$. Cela signifie que la fonction ϱ transforme Z sur Z', ce qui montre vu le théorème 3 que ϱ établit l'épimorphisme de S sur S'.

L'hypothèse que l'automate indéterminé S' est fort réduit ne peut pas être éaffaiblie par la condition que l'automate S' soit minimal. Cela résulte du théorème suivant:

Théorème 6

Soit S un automate réduit, ϱ l'épimorphisme de l'automate S sur S', $\overline{\varrho}$ l'épimorphisme de S sur S''. Alors S' et S'' sont isomorphes.

Démonstration. Comme l'automate S est réduit, les fonctions ϱ et $\overline{\varrho}$ sont univoques, donc $\overline{\varrho}$ est l'isomorphisme de S sur S'' et ϱ est l'isomorphisme de S sur S'', et par conséquent la fonction $\overline{\varrho}\varrho^1$ établit l'isomorphisme de S' sur S''.

Dans la monographie [2], p. 157, il y a les exemples des automates indéterminés équivalents S_7 et S_7^* tels que S_7 et S_7^* sont les minimals de l'automate S_7 (qui est réduit puisqu'il est minimal).

Comme il n'y a pas d'isomorphisme entre eux alors l'épimorphisme de S_7 sur S_7^* n'exists pas.

Donc il n'y a pas d'épimorphisme d'un automate quelconque S sur un automate minimal S' équivalent á S (cet exemple nous le montre et ceci même avec la condition complémentaire que l'automate S est aussi minimal).

Théorème 7

Si $S \sim S'$, l'automate indéterminé S est réduit et l'automate indéterminé S' est fort réduit, alors il existe l'isomorphisme de l'automate S sur S'.

La démonstration du théorème 7 ressemble á celle du théorème 5. La fonction ϱ, definie comme dans la démonstration mentionnée, est univoque parce que l'automate S est réduit et ϱ transforme Z sur Z', donc elle établit l'isomorphisme de S sur S'.

Comme les deux quelconques réducteurs forts du même automate indéterminé sont équivalents (théorème 2.17 dans [2], p. 146), donc il s'ensuit directement du théorème 7.

Corollaire 4

Il existe l'isomorphisme entre deux quelconques réducteurs forts du même automate indéterminé.

Vu le corollaire 1.3 toutes les considérations concernant l'isomorphisme s'appliquent aussi en cas de l'isomorphisme fort.

Travaux cités

1. *Moszner, Z., Tabor, J.:* Relations entre les morphismes et l'équivalence des automates. Demonstratio Math. (sous presse).
2. *Starke, P. H.:* Abstracte Automaten. Berlin, 1969.

О связях между морфизмом и эквивалентностью недетерминированных автоматов

С. ПРШИБИЛО

(Польша)

Определяются связи между гомоморфизмом, эпиморфизмом и изоморфизмом недетерминированных абстрактных автоматов. Представлены эквивалентные (слабо-эквивалентные) преобразования между морфизмами и их эквивалентность (слабая эквивалентность).

Дается обощение результатов, представленных в [1] для детерминированных автоматов, и представлены новые теоремы, действительные к недетерминированным автоматам.

Problems of Control and Information Theory, Vol. 2 (2), pp. 173—178 (1973)

THIRD INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY

(Tallinn, Estonia, June 18—23, 1973)

V. I. SIFOROV

(Moscow)

The Third International Symposium on Information Theory held in Tallinn (Estonian SSR) on June 18—23, 1973 was dedicated to a number of urgent problems in the information theory and its applications. It was sponsored by the Scientific Council for Cybernetics, the Institute of Information Transmssion Problems (both of the USSR Academy of Sciences) and the Institute of Cybernetics of Estonian Academy of Sciences. The Symposium was supported by the Sixth Scientific Commission and XVIIth General Assembly of the International Union of Radio Science (Union Radioscientifique Internationale — URSI) held in Warsaw in August, 1972.

The symposium was attended by scientist from Bulgaria, Hungary, German Democratic Republic, Poland, Romania, USSR, Czechoslovakia, West Berlin, Italy, Canada, USA, Federal Republic of Germany, Sweden, Finland, Japan and other countries, with American, Soviet and Hungarian participants constituing the majority. The total number of participants ran into 250 persons. At 5 plenary and 36 sectional meetings 130 reports were presented and discussed.

The first joint meeting was opened by an address of the Organizing Committee Chairman Dr. V. I. Siforov, by speeches of welcome delivered by the Vice-Chairman of the Council of Ministers of Estonian SSR A. G. Vendelin and the Vice-President of the Academy of Sciences of Estonian SSR, academician N. A. Alumöe, and by salutatory addresses of symposium members Dr. S. Csibi (Hungary) and Dr. E. Berlekamp (USA). The opening ceremony was followed by a report of L. A. Bassalygo, G. V. Zaitzev, V. A. Zinov'ev and V. K. Leont'ev "Uniform packages in the spaces with Hamming-Lee metrics".

In the opening address the Chairman of the Organizing Committee pointed out that the scientific aims of the Tallinn Symposium had been defined as early as at the First International Symposium on Information Theory held in Dubna (USSR) in the summer of 1969, and reaffirmed at the Second Symposium in Tsahkadsor (Armenian SSR). He emphasized that, first of all, the Tallinn Symposium was devoted to the general mathematical problems in the

information theory, to the study of coding and decoding methods, and to application of information-theoretical methods to the study of stochastic processes, to the mathematical statistics and to the theory of detection. Attention was also drawn to the fact that the program of meetings contains all key problems arising in practical application of main results of the information theory. Among most difficult and important ones, the following were mentioned:

- estimation of capacity of main channel types and of message generation rates of close-to-real sources;
- estimation of coding and decoding complexity and finding of practical circuits with close-to-optimal parameters;
- working out and study of physical channel and communication system mathematical models.

It was also noted that there are good reasons to believe that the information theory is entering a new era where it would not only develop its principles but might see practical realization of its ideas. It was also marked that the date of the Symposium coincided with the 25th anniversary of C. Shannon's "A mathematical theory of communication" where the main points of the information theory had been brilliantly formulated.

In his speech at the first plenary session of the Tallinn symposium the Chairman of the IEEE Professional group on information theory Dr. Elwyn Berlekamp announced that Soviet scientist V. D. Goppa for his papers published in the journal "Problems of information transmission" was awarded the IEEE International prize.

Great consideration was given to the results of studies in the coding theory, including error-correcting codes, and to the development of the information source coding theory.

E. L. Blokh and V. V. Zyablov (USSR) introduced in their report the notion of the generalized concatenated code, involving as particular cases all previously known concatenated and error-localizing codes, and discussed the main features of these generalized codes which were shown to have better correcting properties. The authors have also suggested an algorithm realizing the lower bound for the code distance.

Dr. E. R. Berlekamp (USA) and E. Justensen (Denmark) discussed construction of cyclic over GF (2) concatenated codes and studied their correcting properties. V. I. Levenshtein (USSR) suggested upper bounds for the number of vectors of binary error-correcting code and for the number of non-overlapping spherical segments on the Euclidean sphere that are improved as compared with those previously found by P. Elice and Csapoti—Rankin. M. J. Fergusson (Canada) discussed several questions concerning convolutional

codes and showed that specific kinds of them do not change the form of signal spectrum.

D. Slepian and J. K. Wolf (USA) developed the coding theory for multiple access channels with correlated sources and proved their coding theorem.

Papers of R. M. Gray and L. D. Davisson (USA), V. N. Koshelev (USSR), R. E. Krichevskii (USSR), A. Perez (Czechoslovakia), Yu. N. Lin'kov (USSR), V. K. Trofimov (USSR), V. F. Babkin (USSR) and A. B. Kryukov (USSR) were devoted to various aspects of information source coding theory.

A. Perez discussed probability of ambiguous coding of discrete stationary source vs. coding rate. Yu. N. Lin'kov presented an analysis of the rate-distortion function for discrete-parameter sources. V. N. Koshelev studied linear compressing encoding of binary messages.

Information redundancy reduction is known to be one of burning problems of the present-day communication theory. In many applications (in particular, in information measuring systems using overloaded communication channels) a need arises to reduce the transmitted information at the expense of redundant readings. Transmission of analog signals over digital communication lines using data compression units requires the influence of signal form and other waiting unit parameters on reconstruction error to be established. V. P. Evdokimov (USSR) presented results of an experimental verification of compression effectiveness and reconstruction accuracy of compressed signals, obtained through computer simulation of signals and compressors.

V. Cappellini (Italy) discussed application of certain data compression methods to signal processing. Both known compression algorithms and their original modifications were analyzed. Consideration was given also to digital filtration techniques. Conclusion was drawn that these methods may be applied successfully to a number of signal transmission and processing problems. T. Nemetz (Hungary) spoke about asymptotically optimal quantization. He proved a theorem establishing the form of optimal m-valued transformation and suggested a suboptimal iterative procedure determining a "good" transformation which is, in many cases, the optimal one.

Several papers were devoted to the theory of quantum information transmission channels. A. S. Kholevo (USSR) treated the question of optimal measurements of quantum signal parameters and formulated the general optimality conditions of quantum measurements. Based on these conditions, the optimal Bayes measurement of Gaussian state parameters was found. V. P. Belavkin and R. L. Stratonovich (USSR) studied optimization of quantum observables by an information criterion. An equation for Gaussian boson multi-dimensional case was derived and discussed. Means were indicated to use the found maximal amount of information.

Papers of R. Gallagher (USA), A. K. Gorbunov and M. S. Pinsker (USSR) T. T. Kadota (USA), and V. V. Prelov (USSR) were devoted to the capacity theory of various information transmission channels and to definition of the source epsilon-entropy.

R. Gallagher (USA) treated capacity of degenerate channels which he referred to as "broadcasting" since information here is transmitted from one sources to several users. The work of A. K. Gorbunov and M. S. Pinsker (USSR) dealt with epsilon-entropy of channels without anticipation and with prediction of Gaussian processes. T. T. Kadota (USA) solved the problem of continuous-time channel capacity using variable-length codes. The capacity of stationary finite memory using variable-length codes was shown to be equal to that of channels using fixed-length codes. V. V. Prelov (USSR) examined channel capacity under high additive noise and established its asymptotic behavior for output signal power tending to zero. I. Csiszár (Hungary) reviewed capacity of noise channels with arbitrary signal costs.

A number of reports were devoted to the noise immunity of information stores. In the paper of A. V. Kuznetsov and B. S. Tsybakov (USSR) "Noise-immune information storage in memory with errors", for instance, an existence theorem was proved for memory consisting of non-reliable elements and having a finite number of elements for a bit of information where storage error probability tends exponentially to zero with increase of code word length.

Report of D. J. Sacrison (USA) was devoted to urgent questions of the choice of a distortion measure for image transmission. The author had analyzed various transmitted image distortion estimation techniques. Report of D. S. Lebedev and L. I. Mirkin (USSR) dealt with a burning problem of transmitted noise-distorted image reconstruction. Here, the smoothing algorithm for noisy images is based on the composite source model having several states corresponding to both isotropic and anisotropic structure of image segment. An optimal, in the terms of mean-square criterion, algorithm was suggested for smoothing images distorted with an additive Gaussian noise.

Great attention at the symposium was paid to such important issues as the further development of the signal detection theory, and various aspects related to improvement and application of statistical methods of study. The paper of R. Esposito and D. Middleton (USA) was concerned with the theory of detection of local and distributed signal sources in space-time noise fields. Signal detection for quantized measurements in the presence of noise was discussed by B. R. Levin and V. M. Baronkin (USSR). T. Kailath (USA) showed validity of high-speed algorithms for recurrent estimation in linear systems and gave their analysis. Yu. M. Shtar'kov and M. M. Lange (USSR) estimated maximum likelihood reconstruction for a sequence of distorted sam-

ples with Markov-type restrictions. G. P. Tartakovksii and V. G. Repin (USSR) discussed adaptive pattern recognition based on the observation of coupled random signals. G. O. Katona (Hungary) gave consideration to a search model enabling one to find unknown elements of a given finite sequence of elements. I. A. Ibragimov and R. Z. Khas'minskii (USSR) estimated parameters of a signal in the Gaussian white noise. An interesting problem of learning under computational constraints from weakly dependent samples was examined by S. Csibi (Hungary). The report of R. L. Dobrushin (USSR) was devoted to stochastic Markov fields which are an important, yet poorly studied area.

A number of reports were concerned with various aspects of the communication theory, with the theory of channels having various kinds of feedback and with improvement of information flows in large management systems. J. Seidler (Poland) discussed several properties of multiple access communication system. Measurement of stochastic channel space-time response in optimum reception problems was treated by D. D. Klovskii and V. A. Soifer (USSR).

Several lectures dealt with various aspects of organization and processing of information in programming systems. Among them the following reports may be mentioned: E. H. Tóugu (USSR) "Data schemes in programming", S. I. Khmelnik (USSR) "Quick-acting procedures", L. Vyhandu (USSR) "Quick methods of data analysis in multipleattribute systems of objects", T. Atse, M. Tombak, M. Räbovyitra (USSR) "On a data processing method", M. H. Sarv, M. O. Tombak "Increasing text processing effectiveness during the parsing", J. Pruuden, B. Tamm (USSR) "Principles of the integrated programming system".

Along with the reports presented at sections and plenary meetings, several invited papers were read. Among them were those of G. D. Forney (USA) on an information-theoretical inference principle, C. W. Helstrom (USA) on resolution, estimation and other quantum detection theory problems, A. Niemi (Finland) on employment distribution expressed through a new variable, and A. N. Kolmogorov (USSR) on algorithmic foundations of the information and probability theories. All the lectures attracted great attention of the audience. In Kolmogorov's talk, algorithmic construction of the information and probability theories was related to the complexity estimation of mathematical models under consideration, and, in particular, to the complexity estimation of sequences. Such fundamental concepts as regularity, chance and probability were discussed in the light of the recent progress of mathematics, and areas of further studies required to extend and expand these and other fundamental concepts were outlined.

The Symposium has demonstrated that studies in the field of the information theory and its applications have advanced significantly in the USSR, USA, Hungary and other countries. The detailed discussions of the presented papers and contacts between scientists of various countries will, undoubtedly, enhance development of information sciences and application of attained results in a wide variety of fields.

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

PROBLEMS OF CONTROL AND INFORMATION THEORY VOL. 2(2)
ENGLISH TRANSLATION OF THE COMMUNICATIONS IN RUSSIAN

ASYMPTOTIC EXPRESSIONS FOR INFORMATION IN DISCONTINUOUS-NOISE-DENSITY CHANNEL

M. V. BURNASHEV, R. Z. KHASMINSKY

(Moscow)

Asymptotic behaviour of information value under low power of inpu signal is studied for the channel with discontinuous noise density.

1. Introduction

This paper is concerned with asymptotic expressions for mean mutual information of input and output in additive memoryless channel where the "noise" ξ_i is superimposed over each input signal S_i such that the output signal η_i is as follows:

$$\eta = S_i + \xi_i. \tag{1.1}$$

Here $\xi_1, \ldots, \xi_n, \ldots$ are independent similarly distributed random variables of density $p(x)$ relative to the Lebesgue measure on the straigt line. The case where $p(x)$ is an absolutely continuous function having Fisher information quantity, i.e.

$$\int_{-\infty}^{\infty} p(x) \left| \frac{p'(x)}{p(x)} \right|^2 dx = I < \infty, \tag{1.2}$$

was discussed in [1, 2]. If, for example, input signal complies with the following "peak" restriction

$$|S - MS| < K \cdot \sqrt{DS}$$

where K is an arbitrary S-independent constant, the asymptotic formula from [2] holds for information $J(S, \eta)$:

$$J(S, \eta) = \frac{1}{2} J \cdot DS + o(DS), \ (DS \to 0). \tag{1.3}$$

It is of interest to study the asymptotic behaviour of $J(S, \eta)$ for the case where one of the above conditions for $p(x)$ is broken. A particular case where $p(x)$ is absolutely continuous but has a finite number of "irregular" points $x_1, \ldots, x_n$ in whose vicinity

$$p(x) \left| \frac{p'(x)}{p(x)} \right|^2 \sim C_k |x - x_k|^{-1}$$

was discussed in [3].

The present paper discusses the case where $p(x)$ has a finite number of first-kind discontinuities. It turns out in this case that the main term of the value of information contained in the output signal is, for low input signal, a rather simple functional of input signal distribution depending on the limiting values of $p(x)$ in the vicinity of discontinuity points taken as parameters. This expression easily yields asymptotic expressions of channel capacity under various restrictions (see Section 2).

Several channel capacity estimates for particular forms of discontinuous densities were obtained in [1]. Reference [4] gives without proof the main result [(1.4), (1.5), (1.6)] for a particular case.

In order to obtain more precise expression of this result for arbitrary $p \geq 0$ and $q \geq 0$, and for the random variable S with distribution function $F_S(x)$, we introduce the functional

$$B_{p,q}(S) = \int \Phi(F_S(x))\, dX,$$

where

$$\Phi_{p,q}(y) = p \ln p \cdot y + q \ln q\,(1-y) - [py + q(1-y)] \ln [py + q(1-y)] \quad (1.4)$$

and by assumption $0 \cdot \infty = 0$.

Since $\Phi_{p,q}(y) \geq 0$ for $0 \leq y \leq 1$, we obtain $B_{p,q}(S) \geq 0$ with $B_{p,q}(S) = 0$ if S is a determinate value or $p = q$. If $p(x)$ has r first-order discontinuities, with q_k and p_k being limiting values of density at discontinuity points on the left and right, respectively, the following asymptotic formula may be proved to hold under certain restrictions on the class of input densities:

$$J(S, \eta) = B(S) + 0((M \mid S \mid^{\gamma}))^{1/\gamma} \quad \text{for} \quad M \mid S \mid^{\gamma} \to 0 \quad (1.5)$$

where

$$B(S) = \sum_{k=1}^{r} B_{p_k, q_k}(S), \ \gamma > 1. \quad (1.6)$$

Thus expressions for the main term of smooth and discontinuous densities differ significantly. To take an example, for input signal $S_\varepsilon = \varepsilon S$, S being a fixed random variable of finite variance, the following evidently holds:

$$B(S_\varepsilon) = \varepsilon \cdot B(S)$$

Thus the main term of information is in proportion to the root of the variance, rather than to the variance itself as in the case of smooth density. (V. V. Prelov [1] was the first to mention this for a particular case of discontinuous density.)

R. L. Dobrushin raised the question of validity of this result for low additional noise in channel. In this connection we are going to show that (1.5) and (1.6) hold if additional noise power is of a lower order of magnitude than the input signal (see Lemma 2.1).

2. Main theorem

Now the assumptions used in the proof of the above result are stated.

Let $\mathcal{K}$ denote a memoryless channel with independent additive noise whose density $p(x)$ satisfies the following conditions:

1) $p(x)$ is continuous everywhere except a finite number of points $x_1, \ldots, x_r$ and the one-sided finite limits

$$p(x_k - 0) = q_k,\ p(x_k + 0) = p_k, \qquad k = 1, \ldots, r \tag{2.1}$$

exist, and

2) for a certain $b > 1$ there exist the integrals

$$\int_{x_k}^{x_{k+1}} p(x) \left| \frac{p'(x)}{p(x)} \right|^b dx < \infty, \quad k = 0, \ldots, r; \quad x_0 = -\infty, \quad x_{r+1} = +\infty. \tag{2.2}$$

Let us consider now the functional $B(S)$ defined by (1.4) and (1.6) which will be of prime importance in further discussion.

1. For any h the following equality holds

$$B(S) = B(S + h). \tag{2.3}$$

2. For any $\gamma > 1$ and any constant h the following inequality holds:

$$B(S) \leq C(p_1, \ldots, p_r, q_1, \ldots, q_r) \cdot C(\gamma) \cdot (M \mid S - h \mid^{\gamma})^{1/\gamma}. \tag{2.4}$$

3. If $\max \mid p_i - q_i \mid > 0$, $i = 1, \ldots, k$, then for a certain positive constant $C = C(p_1, \ldots, p_r, q_1, \ldots, q_r)$

$$B(S) \geq C \cdot M \mid S - \operatorname{med} S \mid \tag{2.5}$$

where med S is the median of a random variable S.

The first property is evident, let us prove the second and the third. By denoting $C_k = \dfrac{(p_k - q_k)^2}{\max(p_k, q_k)}$ (1.4) implies the following inequalities:

$$\frac{1}{2} C_k y(1 - y) \leq \Phi_{p_k, q_k}(y) \leq C_k \left[y \ln \frac{1}{y} + (1 - y) \ln \frac{1}{1 - y} \right].$$

Therefore

$$B(S) \leq \sum_{k=1}^{r} C_k \left\{ \int_{-\infty}^{\infty} F(x) \ln \frac{1}{F(x)} dx + \int_{-\infty}^{\infty} \left[1 - F(x) \ln \frac{1}{1 - F(x)} dx \right] \right\}. \tag{2.6}$$

Now assume for a certain $\gamma > 1$ that

$$\int \mid x \mid^{\gamma} dF(x) < \infty.$$

Then for $x \to \infty$ and any $\gamma_1 < \gamma$

$$x^{\gamma_1}[F(-x) + 1 - F(x)] \to 0. \tag{2.7}$$

More specifically, confining ourselves to the first integral in (2.6) and allowing for (2.7), we obtain

$$\int F(x) \ln \frac{1}{F(x)}\, dx = xF(x) \ln \frac{1}{F(x)} \Big|_{-\infty}^{\infty} - \int x \ln \frac{1}{F(x)}\, dF + \\ + \int x d\, F(x) \leq \int |x|\, dF(x) + \int |x| \ln \frac{1}{F(x)}\, dF(x).$$

Using the Hölder inequality and denoting the Euler gamma function by $\Gamma(x)$, we come to

$$\int |x|\, dF(x) + \int |x| \ln \frac{1}{F(x)}\, dF(x) \leq$$

$$\leq \left[\int |x|^{\gamma}\, dF(x)\right]^{1/\gamma} \left\{\left[\int \left(\ln \frac{1}{F(x)}\right)^{\gamma/\gamma-1} dF(x)\right]^{\gamma-1/\gamma} + 1\right\} =$$

$$= \left[\int |x|^{\gamma}\, dF(x) \cdot \left\{\left[\int_0^1 \left(\ln \frac{1}{y}\right)^{\gamma/\gamma-1} dy\right]^{\gamma-1/\gamma} + 1\right\}\right. =$$

$$= \left[\int |x|^{\gamma}\, dF(x)\right]^{1/\gamma} \cdot \left[\Gamma^{\gamma-1/\gamma}\left(\frac{\gamma}{\gamma-1} + 1\right) + 1\right].$$

The relation (2.4) is thus proved. Inequality (2.5) follows from the estimate

$$B(S) \geq \frac{1}{2} \sum_{k=1}^{r} C_k \int_{-\infty}^{\infty} F(x)\left[1 - F(x)\right] dx \geq \frac{1}{4} \sum_{k=1}^{r} C_k \cdot M\,|S - \operatorname{med} S|\,.$$

We denote $||S||_{\gamma} = (M\,|\,S\,|\,\gamma)^{1/\gamma}$ and agree to write $\psi = 0(||S||_{\gamma})$ if $\sup_{||S||_{\gamma}} \frac{\psi}{||S||_{\gamma}} \to 0$ for $\varepsilon \to 0$.

Theorem 2.1. For a channel $\mathfrak{A}$ with the noise density ξ satisfying (2.1) and (2.2) for any $\gamma > 1$, the following relation between input (S) and output (η) signals holds:

$$J(S, \eta) = B(S) + o(||S||\gamma),\ (||S||\gamma \to 0)\,. \tag{2.8}$$

The proof of Theorem 1 is given in Section 3 and in the Appendix. The remainder of this section contains the proof of several results concerning

channel capacity under various restrictions on input signal, and study of the "validity" of expression (2.8) for low additional noise in the channel.

Remark. One cannot write $0(||S||_1)$ or $o(B(S))$ in (2.8) instead of $o(||S||\gamma)$ as the following simple example shows. Let the random variable S have three value $-\frac{1}{\varepsilon}$, 0, $\frac{1}{\varepsilon}$ with respective probabilities

$$p\left(S=-\frac{1}{\varepsilon}\right)=p\left(S=\frac{1}{\varepsilon}\right)=\frac{\varepsilon^2}{2}, \qquad p(S=0)=1-\varepsilon^2.$$

It is easy to verify that the following estimates are then valid ($\varepsilon \to 0$):

$$H(S)\sim\varepsilon^2\ln\frac{1}{\varepsilon};\ B(S)\sim\varepsilon\ln\frac{1}{\varepsilon};\ M|S|=\varepsilon;\ M|S|^\gamma=\varepsilon^2=\varepsilon^{2-\gamma},$$

where $H(S)$ is the entropy of the discrete random variable S. Since $J(S,\eta)\leq$ $\leq H(S)$, we obtain $J(S,\eta)=o(B(S))$. In this case (2.8) gives a rather rough estimate of $J(S,\eta)=o(\varepsilon^{1-\delta})$ for any $\delta>0$, and the "refined" formula would give an incorrect result.

It is yet clear that if the family of random variables S under consideration satisfies

$$||S||_\gamma = o(B(S)),\ (B(S)\to 0)$$

under a certain $\gamma>1$, then the main term of (2.8) is correct. From (2.5) one can for instance, see that for a family $\mathfrak{L}$ of random variables S having $M|S|\to 0$ and satisfying for certain $\gamma>1$ and $K>0$

$$||S||_\gamma < K\cdot||S-\operatorname{med} S||_1,$$

the estimate

$$J(S,\eta)=B(S)+o(B(S))$$

takes place uniformly for $S\in\mathfrak{L}$.

Theorem 2.2. Capacity of a channel $\mathfrak{A}$ under the restriction $|S|\leq\varepsilon$ and for $\varepsilon\to 0$ may be represented as

$$C(\varepsilon)=2\Phi(x^*)\cdot\varepsilon+o(\varepsilon) \tag{2.9}$$

where $\Phi(x)=\sum_{k=1}^{r}\Phi_{p_k,q_k}(x)$ and the functions $\Phi_{p_k,q_k}(x)$ are as defined in (1.4), and x^* is the only solution in the interval (0, 1) of the equation

$$\Phi'(x)=0. \tag{2.10}$$

Input distribution realizing channel capacity to within $o(\varepsilon)$ is as follows:

$$p\{S=-\varepsilon\}=x^*,\ p\{S=\varepsilon\}=1-x^*. \tag{2.11}$$

Proof. It is apparent that $\Phi(0) = \Phi(1) = 0$, therefore

$$J(S, \eta) = \int_{-\varepsilon}^{\varepsilon} \Phi(F(u))\, du + o(\varepsilon)\,. \tag{2.12}$$

By the inequality $x \ln \frac{x}{y} > x - y$ which is evident for $xy > 0$ and $x \neq y$, we obtain

$$\Phi'(0) = \sum_{k=1}^{r} \left[q_k \ln \frac{p_k}{q_k} - (p_k - q_k) \right] > 0\,,$$

$$\Phi'(1) = \sum_{k=1}^{r} \left[-q_k \ln \frac{q_k}{p_k} + (q_k - p_k) \right] < 0$$

and

$$\Phi''(x) = -\sum_{k=1}^{r} \frac{(p_k - q_k)^2}{q_k(1-x) + p_k \cdot x} < 0\,.$$

The equation $\Phi'(x) = 0$ has therefore the only solution x^* inthe interval $(0, 1)$, and

$$\max_{0<x<1} \Phi(x) = \Phi(x^*)\,. \tag{2.13}$$

The theorem easily follows from (2.12) and (2.13).

Remark. If $p(x)$ is symmetric (i.e. $p(x) = p(-x)$), it is readily seen that $x^* = \frac{1}{2}$ and that the optimal distribution is as follows:

$$p\{S = -\varepsilon\} = p\{S = \varepsilon\} = \frac{1}{2}\,.$$

Theorem 2.3. The capacity of a channel $\mathcal{K}$ under restriction

$$(M \mid S \mid^{\gamma})^{1/\gamma} \leq \varepsilon, \quad (\gamma > 1) \tag{2.14}$$

for $\varepsilon \to 0$ may be represented as follows:

$$C(\varepsilon) = \varepsilon \left(\int_0^1 |\Phi'(x)|^{\beta}\, dx \right)^{1/\beta} + o(\varepsilon), \quad \left(\frac{1}{\gamma} + \frac{1}{\beta} = 1 \right). \tag{2.15}$$

This value of information $J(S, \eta)$ results from distribution $F^*(u)$ which is an inverse function of

$$g^*(x) = -\frac{\varepsilon\, |\Phi'(x)|^{1/\gamma - 1}}{\left(\int_0^1 |\Phi'(x)|^{\beta}\, dx \right)^{1/\gamma}} \operatorname{sgn} \Phi'(x)\,. \tag{2.16}$$

Proof. It is sufficient to verify that under the restriction

$$\int_{-\infty}^{\infty} |u|^{\gamma} dF(u) \leq \varepsilon^{\gamma}$$

the upper bound of the functional $\int_{-\infty}^{\infty} \Phi(F(u))\, du$ may be reached through the distribution $F^*(x)$ (2.16) and is defined by (2.15). To this end it would be convenient to make the substitution $F(u) = x$. We denote the inverse function of $F(u)$ as $u = g(x)$ and assume it to be continuous to the left. With this notation our task boils down to finding

$$\sup \int_0^1 \Phi(x)\, dg(x) \tag{2.17}$$

under the condition

$$\int_0^1 |g(x)|^{\gamma} dx \leq \varepsilon^{\gamma}.$$

The function $g(x)$ is monotonic but, generally speaking, is unlimited in the vicinity of the points $x = 0$ and $x = 1$. It is yet easy to verify tha $\Phi(x)g(x) \to 0$ for $x \to 0$ or $x \to 1$. Therefore through integration by parts of (2.17) and use of the Hölder inequality, we obtain

$$\int_0^1 \Phi(x)\, dg(x) = -\int_0^1 g(x)\, \Phi'(x)\, dx \leq$$

$$\leq \Big(\int_0^1 |g(x)|^{\gamma} dx\Big)^{1/\gamma} \Big(\int_0^1 |\Phi'(x)|^{\beta} dx\Big)^{1/\beta} \leq \Big(\int_0^1 |\Phi'(x)|^{\beta} dx\Big)^{1/\beta}$$

with the last inequality becoming an equality for $g(x) = g^*(x)$ where $g^*(x)$ is defined by (2.16). This reasoning implies the theorem.

The validity of Theorem 2.1 for low enough additional noises stems from the following lemma.

Lemma 2.1. If a communication channel is described by

$$\eta = S + \xi + \zeta$$

where the random variable ξ has density $p(x)$ satisfying (2.1), (2.2), then

$$J(S, \eta) = B(S) + o(||S||_{\gamma}) + o(||\zeta||_{\gamma_1}) + o(B(\zeta))$$

for any $\gamma > 1$ and $\gamma_1 > 1$.

Proof. The identity

$$J(S, \eta - \zeta) + J(\zeta, \eta) = J(S, \eta) + J(\zeta, \zeta + \xi)$$

(see [5], page 35, formula (3, 8.7)) leads to the estimates

$$J(S, \eta - \zeta) - J(\zeta, \zeta + \xi) \leq J(S, \eta) \leq J(S, \eta - \zeta).$$

By virtue of Theorem 2.1

$$J(S, \eta - \zeta) = B(S) + o(|| S ||_\gamma)$$

and

$$J(\zeta, \zeta + \xi) = B(\zeta) + o(|| \zeta ||_{\gamma_1})$$

Lemma 2.1 is thus proved.

The right parts of (2.8) and (2.18) coincide apparently if, for instance,

$$|| \zeta ||_{\gamma_1} = o(|| S ||_\gamma)$$

for certain $\gamma > 1$ and $\gamma_1 > 1$.

3. *Proof of Theorem 2.1.* We denote for the sake of simplicity

$$\varepsilon = \|S\|_\gamma, \; U = \sum_{k=1}^{r} [x_k - \beta(\varepsilon), x_k + \beta(\varepsilon)], \; R = (-\infty, \infty)$$

where the function $\beta(\varepsilon) \to 0$ for $\varepsilon \to 0$ will be chosen later. Let us denote also

$$\psi(x, y) = p(x - y) \ln \frac{p(x - y)}{\int p(x - z)\, dF(z)}.$$

The expression for information $J(S, \eta)$ may be written as

$$J(S, \eta) = \int \int \psi(x, y)\, dF(y)\, dx =$$

$$= \sum_{k=1}^{r} \int_{x_k - \beta(\varepsilon)}^{x_k + \beta(\varepsilon)} \int \psi(x, y)\, dF(y)\, dx + \int_{R \setminus V} \int_{U} \psi(x, y)\, dF(y)\, dx = \sum_{k=1}^{r} I_k + I_0. \quad (3.1)$$

The appendix contains the proof of the following:

Lemma 3.1. If $\beta(\varepsilon) = \varepsilon^{\frac{\gamma - 1}{2\gamma}}$ and assumptions of Theorem 2.1 are satisfied, then

$$I_0 = o(\varepsilon). \quad (3.2)$$

Lemma 3.1 and (3.1) imply that the proof of Theorem 2.1 reduces to the proof of the following lemma.

Lemma 3.2. If assumptions of Theorem 2.1 are satisfied and $\beta(\varepsilon) = \varepsilon^{\frac{\gamma-1}{2\gamma}}$, then

$$|I_k - B_{p_k,q_k}(S)| = \left| \int_{x_k-\beta(\varepsilon)}^{x_k+\beta(\varepsilon)} \int \psi(x,y)\, dF(y)\, dx - B_{p_k,q_k}(S) \right| = o(\varepsilon). \qquad (3.3)$$

Proof. 1) We obtain first the following estimates through Chebyshev's inequality:

$$P\{p(x-S) > e^{\lambda k} \cdot Mp(x-S\} = \int dF(y) \le e^{-\lambda k}, \quad \left\{\frac{p(x-y)}{Mp(x-s)} > e^{\lambda k}\right\} \quad (\cdot)$$

$$p\{S > \beta(\varepsilon)\} = \int_{|y|>\beta(\varepsilon)} dF(y) \le \frac{\varepsilon^{\lambda}}{\beta^{\gamma}(\varepsilon)} = \varepsilon^{1+\frac{\gamma-1}{2}} = 0(\varepsilon). \qquad (\cdot\cdot)$$

Denoting $f(x,y) = \dfrac{p(x-y)}{\int p(x-z)\, dF(z)}$, we easily obtain through $(\cdot)$, the following estimate:

$$\iint_{\{\ln \varphi(x,y) \ge \lambda\}} \psi(x,y)\, dF(y)\, dx = \sum_{k=1}^{\infty} \iint_{\{\lambda(k+1) < \varphi(x,y) \ge \lambda_k\}} \psi(x,y)\, dF(y)\, dx \le$$

$$\le \sum_{k+1}^{\infty} \lambda(k+1)\, e^{-\lambda k} < \frac{2\lambda e^{-\lambda}}{(1-e^{-\lambda})^2}.$$

It is also evident that

$$\left| \iint_{\{\ln \varphi(x,y) \le -\lambda\}} \psi(x,y)\, dF(y)\, dx \right| \le \sum_{k=1}^{\infty} \iint_{\{-\lambda(k+1) < \ln \varphi(x,y) \le -\lambda k\}} \lambda(k+1)\, p(x-y) \times$$

$$\times\, dx\, dF(y) \le \sum_{k+1}^{\infty} \lambda(k+1)\, e^{-\lambda k} < \frac{2\lambda e^{-\lambda}}{(1-e^{-\lambda})^2}.$$

By taking $\lambda = 2 \ln \dfrac{1}{\varepsilon}$ we obtain

$$\iint_{\{|\ln \varphi(x,y)| > \lambda\}} |\psi(x,y)|\, dF(y)\, dx = 0\,|\varepsilon| \qquad (3.4)$$

and (. .) and (3.3) imply that

$$\left| \iint\limits_{|y|>\beta(\varepsilon)} \psi(x,y)\,dF(y)\,dx \right| = 0(\varepsilon) + \tag{3.5}$$

$$+ \iint\limits_{\{|y|>\beta(\varepsilon),\ |\ln\varphi(x,y)\le\lambda\}} \psi(x,y)\,dF(y)\,dx \le 0(\varepsilon) + \lambda\cdot p\,\{|S|>\beta(\varepsilon)\} = 0(\varepsilon)\,.$$

2) For $\varepsilon \le \varepsilon_0$ we obtain through the simple inequality $\ln x \le x - 1$ and (. .)

$$\left| \iint\limits_{|y|<\beta(\varepsilon)} \psi(x,y)\,dF(y)\,dx - \iint\limits_{|y|<\beta(\varepsilon)} p(x-y)\ln\frac{p(x-y)}{\int p(x-z)\,dF(z)}\,dF(y)\,dx \right| =$$

$$= \iint\limits_{|y|<\beta(\varepsilon)} p(x-y)\ln\left[1 + \frac{\int\limits_{|z|<\beta(c)} p(x-z)\,dF(z)}{\int\limits_{|z|<\beta(\varepsilon)} p(x-z)\,dF(z)}\right] dF(y)\,dx \le \tag{3.6}$$

$$= \iint\limits_{|z|>\beta(\varepsilon)} p(x-z)\,dF(z)\,dx \le \int\limits_{|z|>\beta(\varepsilon)} dF(z) = 0(\varepsilon)\,.$$

Estimates (3.5) and (3.6) enable one to write immediately I_k (assuming without loss of generality that $x_k = 0$):

$$I_k = \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{|y|<\beta(\varepsilon)} \psi(x,y)\,dF(y)\,dx + 0(\varepsilon) =$$

$$= \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{-\beta(\varepsilon)}^{x} p(x-y)\ln\frac{p(x-y)}{\int\limits_{|z|<\beta(\varepsilon)} p(x-z)\,dF(z)}\,dF(y)\,dx + \tag{3.7}$$

$$+ \int\limits_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int\limits_{x}^{\beta(\varepsilon)} p(x-y)\ln\frac{p(x-y)}{\int\limits_{|z|<\beta(\varepsilon)} p(x-z)\,dF(z)}\,dF(y)\,dx + o(\varepsilon) = J_k^{(1)} + J_k^{(2)} + o(\varepsilon)\,.$$

3) Thus Lemma 3.2 and, consequently, Theorem 2.1 will be proved if the following is true.

Lemma 3.3. If the assumptions of Theorem 2.1 are satisfied, then

$$J_k^{(1)} = \int_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int_{-\beta(\varepsilon)}^{x} p(x-y) \ln \frac{p(x-y)}{\int_{|z|<\beta(\varepsilon)} p(x-z)\, dF(z)} dF(y)\, dx =$$

$$= p_k \int_{-\infty}^{\infty} F(x) \ln \frac{p_k}{p_k F(x) + q_k [1 - F(x)]} dx + o(\varepsilon) \tag{3.8}$$

$$J_k^{(2)} = \int_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int_{x}^{\beta(\varepsilon)} \frac{p(x-y)}{\int_{|z|>\beta(\varepsilon)} p(x-z)\, dF(z)} dF(y)\, dx =$$

$$= q_k \int_{-\infty}^{\infty} [1 - F(x)] \ln \frac{q_k}{p_k F(x) + q_k [1 - F(x)]} dx + o(\varepsilon) \tag{3.9}$$

where the estimates $o(\varepsilon)$ are uniform over F for $\varepsilon \to 0$.

Proof. Reasoning for both assertions being identical, we shall dwell only on the second one. Note that $p(x)$ is uniformly continuous over each $[x_k, x_{k+1}]$ $k = 0, \ldots, r$, and introduce

$$\sigma(\varepsilon) = \sup_{|z| \le \varepsilon} [p(x) - p(x - z)] \tag{3.10}$$

$$x, x - z \in [x_k, x_{k+1}].$$

Moreover, for certainty $p_k \neq 0$ is assumed.

Through the estimate $\ln (1 + x) < \frac{x^\alpha}{\alpha}$ we find for $\alpha < \frac{\gamma - 1}{\gamma}$

$$\left| J_k^{(2)} - \int_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int_{x}^{\beta(\varepsilon)} p(x-y) \ln \frac{p(x-y)}{p_k \int_{-\beta(\varepsilon)}^{x} dF(z) + \int_{x}^{\beta(\varepsilon)} p(x-z)\, dF(z)} dF(y)\, dx \right| \le$$

$$\le \int_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int_{x}^{\beta(\varepsilon)} p(x-y) \ln \left[1 + \frac{2\sigma(2\beta) \cdot \int_{-\beta(\varepsilon)}^{x} dF(z)}{\qquad} \right] dF(y)\, dx \le \tag{3.11}$$

$$\le 2\sigma(2\beta) \int_{-\beta(\varepsilon)}^{\varepsilon} \int_{-\beta(\varepsilon)}^{x} dF(z)\, dx + \frac{2\sigma^\alpha(2\beta)}{\alpha} \int_{\varepsilon}^{\beta(\varepsilon)} \left[\int_{x}^{\beta(\varepsilon)} dF(y) \right]^{1-\alpha} dx \sim \frac{\varepsilon \cdot \sigma^\alpha(2\beta)}{\alpha} = o(\varepsilon).$$

Similarly we obtain

$$\left| \int_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int_{x}^{\beta(\varepsilon)} p(x-y) \left[\ln \frac{p(x-y)}{p_k \int_{-\beta(\varepsilon)}^{x} dF(z) + \int_{x}^{\beta(\varepsilon)} p(x-z)\, dF(z)} - \right. \right.$$

$$\left. \left. - \ln \frac{p(x-y)}{p_k \int_{-\beta(\varepsilon)}^{x} dF(z) + p(x-y) \int_{x}^{\beta(\varepsilon)} dF(z)} \right] dF(y)\, dx \right| = o(\varepsilon) \quad (3.12)$$

$$\left| \int_{-\beta(\varepsilon)}^{(\beta)} \int_{x}^{\beta(\varepsilon)} \left[q_k \ln \frac{q_k}{p_k \int_{-\beta(\varepsilon)}^{x} dF(z) + q_k \int_{x}^{\beta(\varepsilon)} dF(z)} - \right. \right.$$

$$\left. \left. - p(x-y) \ln \frac{p(x-y)}{p_k \int_{-\beta(\varepsilon)}^{x} dF(z) + p(x-y) \int_{x}^{\beta(\varepsilon)} dF(z)} \right] dF(y)\, dz \right| = o(\varepsilon) \quad (3.13)$$

$$\left| \int_{-\beta(\varepsilon)}^{\beta(\varepsilon)} \int_{x}^{\beta(\varepsilon)} q_k \ln \frac{q_k}{p_k \int_{-\beta(\varepsilon)}^{x} dF(z) + q_k \int_{x}^{\beta(\varepsilon)} dF(z)}\, dF(y)\, dx - \right.$$

$$\left. - \int_{-\infty}^{\infty} \int_{x}^{\infty} q_k \ln \frac{q_k}{p_k \int_{-\infty}^{x} dF(z) + g_k \int_{x}^{\infty} dF(z)}\, dF(y)\, dx \right| = o(\varepsilon)\,. \quad (3.14)$$

The lemma now readily follows from the estimates (3.11)–(3.14). Theorem 2.1 is therefore proved.

Appendix

Proof of Lemma 3.1. Since, by virtue of $\ln z \geq 1 - \frac{1}{z}$, the following holds for any x:

$$\int p(x-y) \ln \frac{p(x-y)}{\int p(n-z)\, dF(z)}\, dF(y) \geq 0\,.$$

(3.1) implies that

$$I_0 \geq 0. \quad (a.1)$$

Let us find now the upper bound of the integral I_0. It is clear that

$$I_0 = \int\int_{R/U} \psi(x,y)\,dF(y)\,dx \leq \iint_{\{R/U\}\cap\{|\ln\varphi(x,y)|<\delta\}} + \iint_{\{R\setminus V\}\cap\{\delta\leq\ln\varphi(x,y)<\lambda\}} + \iint_{\{\ln\varphi(x,y)\geq\lambda\}} = Z_1 + Z_2 + Z_3 \qquad \text{(a. 2)}$$

where $\lambda = 2\ln\frac{1}{\varepsilon}$, and $\delta \to 0$ for $\varepsilon \to 0$ will be chosen later.

The integral Z_3 is already evaluated in (3.4)

$$Z_3 = o(\varepsilon) \qquad \text{(a.3)}$$

Z_2 is considered now. We note that for $\delta > 0$ small enough one can write

$$1+\delta < \frac{(1-2\delta^3)\,p(x-y)}{\int p(x-z)\,dF(z)} \leq \frac{(1-2\delta^3)}{\int\limits_{|z|<\frac{\varepsilon}{\delta^3}} dF(z)} \cdot \sup\frac{p(x-y)}{p(x-z)} \leq \sup_{|z|\leq\frac{\varepsilon}{\delta^3}} \frac{p(x-y)}{p(x-z)}.$$

To put it differently, if $\dfrac{p(x-y)}{\int p(x-z)\,dF(z)} > e^{\delta}$ there will be $|z| \leq \dfrac{\varepsilon}{\delta^3}$ such that $\dfrac{p(x-y)}{p(x-z)}$. Reminding now (3.5) and using the Hölder's and Chebyshev's inequalities, one readily obtains (see [2, 3]):

$$z_2 \leq \lambda \iint_{\{R\setminus U\}\cap\{\delta<\ln\varphi(x,y)\}\cap\{|y|<\beta(\varepsilon)\}} p(x-y)\,dF(y)\,dx + \lambda \iint_{\{|y|>\beta(\varepsilon)\}} p(x-y)\,dx\,dF(y) \leq$$

$$\leq \frac{\lambda}{\delta^b} \iint_{\{R\setminus U\}\cap\{|y|<\beta(\varepsilon)\}} |p^{1/b}(x-y) - p^{1/b}(x-z)|^b\,dF(y)\,dx + 0(\varepsilon) =$$

$$= \frac{\lambda}{\delta^b} \iint_{\{R\setminus U\}\cap\{|y|<\beta(\varepsilon)\}} \left|\int_y^z \frac{p'(x-t)}{p^{1-1/b}}\,dt\right|^b dF(y)\,dx + 0(\varepsilon) \leq$$

$$\leq \frac{\lambda}{\delta^b} \iint_{\{R\setminus U\}\cap\{|y|<\beta(\varepsilon)\}} |z-y|^{b-1} \left|\int_y^z p(x-t)\left|\frac{p'(x-t)}{p(x-t)}\right|^b dt\right| dx\,dF(y) + 0(\varepsilon) \leq$$

$$\leq K\cdot\frac{\lambda}{\delta^b} \int_{|y|<\beta(\varepsilon)} |z-y|^b\,dF(y) + 0(\varepsilon) \leq K\cdot\frac{\lambda\cdot\varepsilon^b}{\delta^{4b}} + 0(\varepsilon) = 0(\varepsilon) \qquad \text{(a.4)}$$

if $\dfrac{\varepsilon^b\cdot\ln\frac{1}{\varepsilon}}{\delta^{4b}} \to 0$ for $\varepsilon \to 0$.

In order to evaluate the last integral Z_1, we make use of the inequality

$$-\ln x \leq x - 1 + \delta \mid x - 1 \mid,$$

which holds for $\mid \ln x \mid < \delta$, and obtain then

$$Z_1 = \iint\limits_{\{R \setminus U\} \cap \{|\ln \varphi(x,y)| < \delta\}} \left[p(x-y) - \int p(x-z)\, dF(z) \right] dF(y)\, dx +$$

$$+ \delta \iint\limits_{\{R \setminus U\} \cap \{|\ln \varphi(x,y)| < \delta\}} \left| p(x-y) - \int p(x-z)\, dF(z) \right|$$

$$dF(y)\, dx \leq \iiint\limits_{\{R \setminus U\} \cap \left\{ \frac{\int p(\lambda - z)\, dF(z)}{p(x-y)} > e^{\delta} \right\} \cap \{|y|,\, |z| < \beta(\varepsilon)\}} |p(x-z) - p(x-y)| \times$$

$$\times\, dF(z)\, dF(y)\, dx\, \delta \cdot KE + 0(\varepsilon).$$

For $x \in R \setminus U$ the following may be written

$$\int p(x-z)\, dF(z) = p(x) + \int\limits_{|z| < \beta(\varepsilon)} [p(x-z) - p(x)]\, dF(z) +$$

$$+ \int\limits_{|z| \geq \beta(\varepsilon)} [p(x-z) - p(x)]\, dF(z) \leq p(x) + \sigma(\beta) + 0(\varepsilon).$$

Therefore the inclusion below holds:

$$\{R \setminus U\} \cap \left\{ \frac{\int p(x-z)\, dF(z)}{p(x-y)} > e^{\delta} \right\} \cap \{|y|,\, |z| < \beta(\varepsilon)\} \subset$$

$$\subset \{R \setminus U\} \cap \left\{ p(x) < \frac{2\sigma(\beta)}{\delta} \right\} \cap \{|y|,\, |z| < \beta(\varepsilon)\}$$

thus readily leading to

$$Z_1 \leq \iiint\limits_{\left\{p(x) < \frac{2\sigma(\beta)}{\delta}\right\} \cup \{R \setminus U\} \cup \{|y|,\, |z| < \beta(\varepsilon)\}} \left| \int\limits_y^z p'(x-t)\, dt \right| dF(z)\, dF(y)\, dx + 0(\varepsilon) \leq$$

$$\leq \iint (z-y)\, dF(y)\, dF(z) \cdot \int\limits_{\left\{p(x) < \frac{2\sigma(\beta)}{\delta}\right\}} |p'(x)|\, dx + 0(\varepsilon) = 0(\varepsilon) \qquad \text{(a. 5)}$$

if $\dfrac{\sigma(\beta)}{\delta} \to 0$ for $\varepsilon \to 0$.

It is clear from (a.4) and (a.5) that δ may be represented, for example, by the following formula:

$$\delta = \frac{1}{\ln \frac{1}{\varepsilon}} + \sqrt{\sigma(\beta(\varepsilon))}\,.$$

Thus Lemma 3.1 is proved.

References

1. *Прелов В. В.:* Асимптотика пропускной способности непрерывного канала с большим шумом. Проблемы передачи информации, 1970, **6**, 2.
2. *Ибрагимов И. А., Хасьминский Р. З.:* Передача слабого сигнала по каналу без памяти. Проблемы передачи информации 1972, **8**, 4.
3. *Бурнашев М. В.:* Блочный метод передачи слабого сигнала по каналу без памяти. Проблемы передачи информации 1973, **9**, 1.
4. *Хасьминский Р. З.:* Каналы с малой пропускной способностью и посимвольная передача. Проблемы передачи информации, **6**, 2, 58—67.
5. *Пинскер М. С.:* Информация и информационная устойчивость случайных величин и процессов. Проблемы передачи информации, вып. 7, Изд. АН СССР, 1960.

ON NON-EXISTENCE OF PERFECT CODES OVER GALOIS FIELDS

V. A. ZINOV'EV, V. K. LEONT'EV

(Moscow)

The existence problem for perfect codes over Galois fields is solved. The well known Hamming and Golay codes are shown to be the only non-trivial perfect codes over finite Galois fields.

Let E^n be an n-dimensional vector space over the Galois field $GF(q)$, where $q = p^m$ and p is a prime number. A code $A \subseteq E^n$ with minimal Hamming distance $d = 2t + 1$ is referred to as "perfect" if all space E^n may be covered by spheres of the radius t with the centers in the code points.

One code vector of length $n = t$, and, for $q = 2$, two vectors of length $n = 2t + 1$ and distance $d = n$ are trivial examples of perfect codes with any $t \geq 1$. Known Hamming codes [1] with parameters $t = 1$, $n = (q^m - 1)/(q - 1)$, $(m \geq 2)$, and Golay ones [2] with parameters $q = 1$, $t = 3$, $n = 23$ and $q = 3$, $t = 2$, $n = 11$ are non-trivial perfect codes.

This paper is aimed at the proof of the following.

Theorem 1. The above-mentioned Hamming and Golay codes are the only non-trivial perfect codes over the Galois fields $GF(q)$ for all q's.

Various special cases of the theorem have been proved in [3—10]. Proving the theorem we shall essentially use the results of the preceding work [10].

Two necessary conditions of existing of perfect codes are the basis of the proof of the theorem. The first one is due to Hamming and consists of the following. Let $m(A)$ be the number of all code points of a perfect code with distance $d = 2t + 1$. By the definition of perfect code

$$m(A) = q^n \Big/ \sum_{i=0}^{t} (q-1)^i C_n^i.$$

This equality easily implies [5] that

$$\sum_{i=0}^{t} (q-1)^i C_n^i = q^{n-k} \tag{1}$$

where $q^k = m(A)$. Integer solvability of Eq. (1) is a necessary existence condition for perfect codes with the parameters t, q, n.

The second necessary perfect code existence condition is due to Lloyd [3], and is as follows.

*The Lloyd theorem.** If there is a perfect code of length n over a field $GF(q)$ correcting t errors, then the polynomial $P_t(n, \xi)$ of degree t in variable ξ defined by the equality

$$P_t(n, \xi) = \sum_{i=0}^{t} (-1)^i C_{n-\xi}^{t-i} C_{\xi-1}^{i} (q-1)^{t-i},$$

where $C_a^i = a(a-1)\ldots(a-i+1)/i!$ for any a has t different integer roots among the number $1, 2, \ldots, n-1$.

The main idea of deriving Theorem 1 from these necessary conditions lies in the following.

Let a perfect code of length n exist over a field $GF(q)$. Then the polynomial $P_t(n, \xi)$ has t different integer roots $\xi_1, \xi_2, \ldots, \xi_t$ ordered as follows: $\xi_1 < \xi_2 < \ldots < \xi_t$. It may be easily shown [8] that

$$\xi_1 \xi_2 < \ldots \xi_t = \frac{t!}{q^t} \sum_{i=0}^{t} (q-1)^i C_n^i = t!\, q^{n-k-t} \tag{2}$$

$$\xi_1 + \xi_2 + \ldots + \xi_t = t \cdot \frac{q-1}{q} \left(n + 1 - \frac{(t+1)(q-2)}{2(q-1)} \right). \tag{3}$$

We shall use further the following notation:

$$\xi_1 + \xi_2 + \ldots + \xi_t = a(q, t).$$

Eq. (2) holds for the Hamming necessary existence condition for a perfect code, and is therefore an alloy of the Hamming and Lloyd conditions. Comparing the sum and product of the roots $\xi_1, \ldots, \xi_t$ through Eqs. (2) and (3), one can easily verify that (for n large enough) their arithmetic mean is approximately equal to their geometric mean, i.e. that these numbers are approximately identical. But by condition (2) requiring the roots $\xi_1, \ldots, \xi_t$ to be integers, they should be essentially different. By these considerations one can obtain the upper bound for the length n of a possible perfect code [10]. Using the relation between perfect code and tactical configurations the lower bound for the length of possible perfect code may be obtained [10]. Comparison of the bounds yields Theorem 1. Since Theorem 1 has already been proved, for $t \leq 7$

* This wording of the Lloyd theorem is due to van Lint [5].

[5—7] and $t < p$ [8], we shall confine ourselves with $t \geq 8$ and $t \geq P$ although the same method enables one to prove the theorem as a whole.*

Proof. Let A be a perfect code with parameters n, t, and q. Then referring to [10], length n satisfies the inequality

$$n \geq \max\left\{\frac{(t+1)(t+2)}{q-1} + t, qt+1\right\}. \tag{4}$$

Now our aim is to find an upper bound for n which would be less than (4). By the Lloyd theorem the polinomial $P_t(h, \xi)$ has t different roots $\xi_1, \ldots, \xi_t$, therefore any two roots ξ_j and $\xi_{j'}$ ($\xi_j < \xi_{j'}$) are related by $\xi_{j'} = \xi_j \mu_j$ ($\mu_j > 1$). Take $h \leq t/2$ of such intersecting pair of roots $\xi_{j'}, \xi_j$. Thent through (3) one can come as in Ref. [10] to the inequality:

$$\xi_1 \xi_2 \ldots \xi_t \leq \left(\frac{a(q,t)}{l}\right)^t \prod_{j=1}^{h} \frac{4\mu_j}{(1+\mu_j)^2}. \tag{5}$$

Let us find the lower bound of the sum $\sum_{i=0}^{t} (q-1)^i C_n^i$. The inequality $n \geq 4t$ is used to this end which is a trivial consequence of (4). Since for $n \geq 4t$:

$$n(n-t+1) > (n-\alpha t)^2$$

where $\alpha = 2(2-\sqrt{3}) < 0.536$ is a root of the equation $\alpha^2 - 8\alpha + 4 = 0$, we obtain

$$n(n-1)\ldots(n-t+1) > (n-\alpha t)^t > (n-0.536t)^t$$

and consequently

$$\frac{t!}{q^t} \sum_{i=0}^{t} (q-1)^i C_n^i > \left(\frac{q-1}{q}\right)^t t!\, C_n^t > \left(\frac{q-1}{q}\right)^t (n-0.536t)^t. \tag{6}$$

Denoting $\Psi(\mu_1, \mu_2, \ldots, \mu_h) = \prod_{j=1}^{h} \frac{4\mu_j}{(1+\mu_j)^2}$ we obtain from (2), (5) and (6)

$$\left(1 - \frac{t(1+0.072(q-1))+q}{2(n+1)(q-1)-(t+1)(q-2)}\right)^t < \Psi(\mu_1, \ldots, \mu_h). \tag{7}$$

(The rather exact numerical values of the constants in this paper are required to avoid complex computations in the proof of non-existence of perfect codes for initial values of parameters n, t, and q.)

* In effect Theorem 1 is proved also for $q = 2$ [9, 10] and $q = 3, 5, 7$ [10]. We have changed the proof here a little (some more precise upper bound was obtained) and therefore we are repeating it for the aforementioned cases too.

The upper bound for n, which may easily be obtained from (7), depends on the numbers $\mu_1, \ldots, \mu_h$ and decreases monotonically as they increase. Therefore obtaining good lower bounds for these numbers would be of importance. (2) implies that each root ξ_i may be represented as $\xi_j = \alpha_j p^{r_j}$ $((\alpha_j, p)=1)$, and therefore

$$\alpha_1 \alpha_2 \ldots \alpha_t = t!/p^{[t/p]+[t/p^2]+\ldots}$$

where $[x]$ is the integer part of x.

We classify all roots $\xi_1, \ldots, \xi_t$ by placing those with identical α_i into the same class. Let these classes be $L_1, \ldots, L_g$, and x_i be the number of classes of cardinality i. Then

$$\sum_{i \geq 1} x_i = g, \quad \sum_{i \geq 1} i x_i = t. \tag{8}$$

Note that in a class of cardinality $i (i \geq 2)$ the ratio between maximal and minimal roots is at least p^{i-1}. Therefore from the roots $\xi_1, \ldots, \xi^t$ one can separate at least

$$h = \sum_{i \geq 2} x_i$$

non-intersecting root-pairs with ratio $\mu_j \geq p$. Then by estimate (5), taking into consideration that for $i \geq 2$

$$\frac{4p^{i-1}}{(1+p^{i-1})^2} \leq \left(\frac{4p}{(1+p)^2}\right)^{i-1}$$

it follows that

$$\xi_1 \xi_2 \ldots \xi_t \leq \left(\frac{a(q,t)}{t}\right)^t \prod_{i \geq 2} \left(\frac{4p^{i-1}}{(1+p)^2}\right)^{x_i} \leq$$

$$\leq \left(\frac{a(q,t)}{t}\right)^t \left(\frac{4p}{(1+p)^2}\right)^{\sum_{i \geq 2} (i-1) x_i} = \left(\frac{a(q,t)}{l}\right)^t \left(\frac{4p}{(1+p)^2}\right)^{t-g}. \tag{9}$$

Now only the lower bound of the difference $t - g$ is to be found. This estimate was done in essence in [10], therefore the final result is presented here only. Although the another definition of classes was given there, the derivation of lower bound on $t - g$ remains correct, if the number $q = p^m$ would be replaced by p in Lemma 1 (of [10]). Let:

$$\beta(p) = \begin{cases} 5 \text{ for } p = 2, 3 \\ 4 \text{ for } p = 5 \\ \text{maximal integer solution of the} \\ \text{inequality } (\beta - 1)! \leq \beta p \quad \text{for } p \geq 7. \end{cases}$$

Then

$$t - g \geq t \cdot \mu(p \cdot t) \tag{10}$$

$$\mu(p,t) = \begin{cases} \left[\frac{t}{p}\right] \Big/ t & \text{if} \quad p \leq t < \beta(p) \cdot p, \qquad (11) \\ \dfrac{p \log p - (p-1)\log(p-1) - \dfrac{p-1}{t} r(p,t)}{(p-1)\log\left(\dfrac{p}{p-1} \cdot t\right)} & \text{if} \quad l \geq \beta(p) \cdot p \qquad (12) \end{cases}$$

where

$$r(p,t) = \left(\frac{3}{2} + \frac{p-2}{p}\right)\log t - \frac{1}{2}\log\frac{p}{4.2} + \frac{0{,}36}{p}\left(1 + \frac{4(p-2)}{p-1}\right).$$

(Here and further all logarithms are to the base two.)

Eqs. (9) and (10) introduced into (7) and the inequality $(1-\gamma)^a \geq 1 - a\gamma$ ($a \geq 1$ and $0 < \gamma < 1$) lead to the following upper bound for the length n:

$$n + 1 < \left(\frac{p+1}{p-1}\right)^2 \frac{t(1 + 0.072(q-1)) + q}{2(q-1)\,\mu(p,t)} + \frac{(t+1)(q-2)}{2(q-1)}. \tag{13}$$

Since the inequality

$$t\big(1 + 0.072(q-1)\big) + q \leq \left(\frac{1.125}{q-1} + 0.197\right) t(q-1) \tag{14}$$

holds for all $t \geq 8$ and $q \geq 2$, Eq. (13) may be rewritten as follows:

$$n + 1 < \left(\frac{p+1}{p-1}\right)^2 \left(\frac{0.5625}{q-1} + 0.0985\right) \frac{t}{\mu(p,t)} + \frac{(t+1)(q-2)}{2(q-1)}. \tag{15}$$

We compare now the obtained upper bound (15) with the lower bound (4).

Case of $\max\{8\,p\} = t < p \cdot \beta(p)$. The upper bound for this case has the following form:

$$n + 1 < \left(\frac{p+1}{p-1}\right)^2 \left(\frac{0.5625}{q-1} + 0.0985\right) \frac{t^2}{[t/p]} + \frac{(l+1)(q-2)}{2(q-1)}. \tag{16}$$

1. $p = 2$. Since $\beta(2) = 5$, $t = 8.9$ is possible for $p = 2$, therefore $[t/2] = 4$ and the bound (16) may be rewritten as

$$n + 1 < \left(\frac{1.266}{q-1} + 0.224\right) t^2 + \frac{(t+1)(q-2)}{2(q-1)}. \tag{17}$$

For $q = 2$, $n + 1 < 1{,}49 \cdot t^2$ which is less than the lower bound $n + 1 \geq \geq (t + 1) \cdot (t + 3)$ for $t = 8.9$.

For $q = 4$, the possible lengths of perfect codes lie in between

$$\begin{aligned} &at\ t = 8 && 38 \leq n \leq 43 \\ &at\ t = 9 && 46 \leq n \leq 52 \end{aligned} \tag{18}$$

For $q = 2^s \geq 8$

$$n + 1 < 0.405\, t^2 + (t + 1)/2.$$

which is less than the lower bound $n + 1 \geq qt + 2$ *at* $t = 8, 9$.

2. $p = 3$. In this case t is between $8 \leq t < 15$. The bound (16) may be rewritten in this case as follows:

$$n + 1 < \left(\frac{2.25}{q - 1} + 0.394\right) \frac{t^2}{[t/3]} + \frac{(l + 1)(q - 2)}{2(q - 1)}. \tag{19}$$

In the case of $q = 3$, the following inequality holds:

$$n + 1 \geq \frac{(t + 1)(t + 2)}{2} + t + 1 > \frac{1.52\, t^2}{[t/3]} + \frac{t + 1}{4}$$

and in the case of $q = 3^s \geq 9$ holds the inequality:

$$n + 1 \geq q^t + 2 > \frac{0.676\, t^2}{[t/3]} + \frac{t + 1}{2}.$$

Taking into account that $q > t/2$ and $[t/3] \geq 2$ always hold, the latter inequality is evident.

3. $p = 5$. This case also need an individual consideration. For all $8 \leq t < 20$ and $q \geq 5$ the upper bound

$$n + 1 < \left(\frac{1.266}{q - 1} + 0.224\right) \frac{t^2}{[t/5]} + \frac{(t + 1)(q - 2)}{2(q - 1)}$$

is less than the lower bound $n + 1 \geq qt + 2$.

4. $p \geq 7$. In this case, the bound (16) may be rewritten as follows:

$$n + 1 < 0.343 \frac{t^2}{[t/p]} + \frac{l + 1}{2}. \tag{20}$$

The following sequence of evident inequalities

$$n + 1 \geq qt + 2 < pt > \frac{0.5\, t^2}{[t/p]} > 0.343 \frac{t^2}{[t/p]} + \frac{t + 1}{2}$$

points to the fact that the upper bound (20) is always less than the lower bound. Thus for the considered range $\max\{8, p\} \leq t < p\beta(p)$, the only possible lengths of perfect codes for $q = 4$ and $t = 8, 9$ lie within (18). But for these lengths the Hamming condition does not hold [8].

Case of $t \geq p\beta(p)$. For $p = 2, 3, 5, 7$ and with due regard to $t \geq p \cdot \beta(p)$, the function $\mu(p, t)$ may be evaluated as follows:

$$\mu(2, t) > 1.43/\log 2t, \tag{21}$$

$$\mu(3, t) > 0.86/\log \frac{3}{2}t, \tag{22}$$

$$\mu(5, t) > 0.42/\log \frac{5}{4}t, \tag{23}$$

$$\mu(7, t) > 0.37/\log \frac{7}{6}t \tag{24}$$

1. $p = 2$. The upper bound (15) has the following form

$$n + 1 < \left(\frac{3.4}{q-1} + 0.62\right) t \log 2t + \frac{(t+1)(q-2)}{2(q-1)}. \tag{25}$$

For $q = 2$ this bound is less than the lower bound $n + 1 \geq (t + 1)(t + 3)$ for all $t \geq 18$. For $t \leq 18$ the length $n < 380$.

For the bound (25) is less than the lower bound $n + 1 \geq \frac{(t+1)(t+2)}{2(q-1)}$ for all $t \geq 24$.

For $t \leq 24$ the length $n < 240$.

For $q \geq 8$ the upper bound

$$n + 1 < 1.12\, t \log 2t + \frac{t+1}{2}$$

is less than the lower bound (4) for all $t \geq 2 \cdot \beta(2) = 10$.

2. $p = 3$. The upper bound (15) has the following form

$$n + 1 < \left(\frac{2.62}{q-1} + 0.46\right) t \log \frac{3}{2}t + \frac{(t+1)(q-2)}{2(q-1)}. \tag{26}$$

For all $t \geq 3 \cdot \beta(3) = 15$ it is less than the lower bound

$$n \geq \frac{(t+1)(t+2)}{2} + t$$

for $q = 3$, and than the lower bound (4) for $q \geq 9$.

3. $p = 5$. The bound (15) has the following form

$$n + 1 < \left(\frac{3.15}{q-1} + 0.54\right) l \log \frac{5}{4} t + \frac{(t+1)(q-1)}{2(q-1)}. \tag{27}$$

For $q = 5$

$$n + 1 < 1.34\, t \log \frac{5}{4} t + \frac{3}{8}(t+1),$$

being thus less than the lower bound

$$n \geq \frac{(t+1)(l+2)}{4} + t$$

for all $t \geq 5 \cdot \beta(5) = 20$.

For $q \geq 25$

$$n + 1 < 0.67\, t \log \frac{5}{4} t + \frac{t+1}{2}$$

being thus less than (4) for all $t \geq 20$.

4. $p = 7$. In this case

$$n + 1 > 0.93\, l \log \frac{7}{6} t + \frac{t+1}{2} \tag{28}$$

which is less than the bound (4) for all $t \geq 7 \cdot \beta(7) = 35$.

5. $p \geq 11$. For $p \geq 11$ $\mu(p, t)$ may be estimated

$$\mu(p, t) \geq \frac{0.6 \log p}{p \log t}. \tag{29}$$

This gives an upper bound for all $p \geq 11$ and $t \geq 11$ and $t \geq p \cdot \beta(p$:

$$n < 0.4\, tp \log t / \log p \tag{30}$$

We show now that the bound (30) is less than the lower bound (4) for all $p \geq 11$, $q \geq p$, and all $t \geq p \cdot \beta(p)$. In the range $t < q^2$

$$n > qt > 0.4\, tp \log t / \log p$$

because

$$0.4\, tp \frac{2 \log g}{\log p} > 0.4\, tp \frac{\log t}{\log p}$$

and

$$q / \log q > 0,\ 8p / \log p \tag{31}$$

for all $q \geq p$. Now let $t \geq q^2$. Then

$$n \geq \frac{(t+1)(t+2)}{q-1} + t > \frac{t^2}{q} > 0.4\, tp \frac{\log t}{\log p}.$$

The inequality

$$t/\log t > 0.4\, qp/\log p$$

holds because for all $t \geq q^2$ both inequalities

$$t/\log t \geq q^2/2 \log q$$

and (31) hold.

Thus in the case of $t \geq p \cdot \beta(p)$, discussed above, the possible lengths of perfect codes ($q = 1$, $10 \leq t \leq 18$, $n < 380$, and $q = 4$, $10 \leq t \leq 24$, $n < 240$) lie within the range if the Hamming condition is not satisfied [8].

The upper bound (13) is shown to lie always below the lower bound (3) aside from the specified cases, thus indicating that for these values of t there are no non-trivial perfect codes. Q.E.D.

References

1. *Hamming, R. W.:* Error Detecting and Error Correcting Codes. Bell Syst. Techn. J., 1950, 29, 2, 147–160.
2. *Golay, M. J. E.:* Notes on Digital Coding. Proc. IRE, 1949, 37, 6, 657.
3. *Lloyd, S. P.:* Binary Block Coding. Bell Syst. Techn. J., 1957, 36, 2, 517–535.
4. *Shapiro, H. S.:* Slotnick D. L. On the Mathematical Theory of Error Correcting Codes. IBM J. Res. Develop., 1959, 3, 1, 25–34.
5. *Van Lint, J. H.:* In the Nonexistence of Perfect 2- and 3-Hamming-Error-Correcting Codes. Inform. and Control, 1970, 16, 4, 396–401.
6. *Tietäväinen, A.:* On the Nonexistence of Perfect 4-Error-Correcting Codes. "Suomal. tiedeak. toimit.", 1970, AI, 485.
7. *Van Lint, J. H.:* On the Nonexistence of Perfect 5–6 and 7-Hamming-Error-Correcting Codes over CG (q). Report 70-WSK-06, Technological University Eindhoven, 1970.
8. *Van Lint, J. H.:* Error Correcting Codes. Lecture Notes Math., 201, 1971.
9. *Tietäväinen, A.—Perko, A.:* There are no Unknown Perfect Binary Codes. Annales Universitatis Turkuensis, 1971, 148, ser. A.
10. *Зиновьев В. А.—Леонтьев В. К.:* О совершенных кодах. Проблемы передачи информации, 1972, 8, 1, 26–35.

GENERALIZATION OF LLOYD'S THEOREM TO ARBITRARY ALPHABET

L. A. BASSALYGO

(Moscow)

Let $A(n, 2t+1, q)$ be a code of length n defined over an alphabet of q symbols and having the minimal Hamming distance $2t+1$. This code is referred to as "perfect" if its power $|A(n, 2t+1, q)|$ satisfies the following condition:

$$|A(n, 2t+1, q)| = \frac{q^n}{\sum_{i=0}^{t} (q-1)^i C_n^i}. \tag{1}$$

Many papers concerned with the perfect code existence problem over Galois fields $GF(p^m)$ have recently come off the press. The final result here is due to V. A. Zinov'ev and V. K. Leont'ev [1] who have shown that, aside from the well known Hamming and Golay codes, there are no non-trivial perfect codes over Galois field $GF(p^m)$. Both this fact and several preliminary results are based upon Lloyd's theorem [2] proved by Lloyd for $q = 2$ and extented later to $q = p^m$ (p being a prime number) by other authors. For the case where q is not the power of a prime number and $\frac{q^n}{\sum_{i=0}^{i} (q-1)^i C_n^i}$ is an integer, only the following result is known: there is no perfect code for $q = 6$, $n = 7$, and $t = 1$. Extension of Lloyd's theorem to an arbitrary alphabet seems therefore natural, although this aim may be reached only through complication of Lloyd's calculations [2].

Lloyd's theorem. Let $\varphi_s(n, \xi)$ be a polynomial of degree s in ξ given by

$$\varphi_s(n, \xi) = \sum_{r=0}^{s} (-1)^r (q-1)^{s-r} C_\xi^{s-r} C_{n-\xi}^r \tag{2}$$

where $C_u^i = \frac{u(u-1)\dots(u-i+1)}{i!}$. If over an alphabet of q symbols a perfect code of length n exists correcting t errors, $\varphi_t(n-1, \xi)$ has t different integer roots among the numbers $0, 1, 2, \dots, n-1$.

Proof. Since this proof differs from that of Lloyd [2] only in the more complex calculations, only the differences will be indicated without presenting the proof to the full extent. Our notation coincides with that of Lloyd: $V_{j,s} (j = 0, 1, \ldots, n)$ is the number of sequences of weight s which are at a distance j from some code word, $G_j(x) = \sum_{s=0}^{n} V_{j,s} x^s$ is the generating function of the numbers $V_{j,s}$ where x is a complex variable*. The number of sequences at a distance j from a fixed code word of the weight s is

$$\sum_{m=0}^{\infty} C_{n-s}^{j-m} (q-1)^{j-m} C_s^m \sum_{m}^{i} C_m^i (q-2)^{m-i} \quad **$$

hence, $G_j(x) = \sum_{s=0}^{\infty} v_s g_s(j, x)$ where

$$g_s(j, x) = \sum_{m=0}^{\infty} C_{n-s}^{j-m} (q-1)^{j-m} C_s^m \sum_{i=0}^{\infty} C_m^i (q-2)^{m-i} x^{s+j-m-i}.$$

Since

$$\sum_{j=0}^{\infty} y^j G_j(x) = \sum_{j=0}^{\infty} y^j \sum_{s=0}^{\infty} v_s g_s(j, x) =$$

$$= \sum_{s=0}^{\infty} v_s \sum_{m=0}^{\infty} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} C_s^m C_m^i (q-2)^{m-i} x^{s-i} y^m C_{n-s}^{j-m} (q-1)^{j-m} x^{f-m} y^{j-m} =$$

$$= \sum_{s=0}^{\infty} v_s \sum_{m=0}^{\infty} \sum_{i=0}^{\infty} C_s^m x^{s-m} y^m [1 + (q-1)xy]^{s-s} C_m^i (q-2)^{m-i} x^{m-i} =$$

$$= \sum_{s=0}^{\infty} v_s [1 + (q-1)xy]^{n-s} \sum_{m=0}^{\infty} C_s^m x^{s-m} y^m [1 + (q-2)x]^m =$$

$$= \sum_{s=0}^{\infty} v_s [1 + (q-1)xy]^{n-s} [x + y + (q-2)xy]^s =$$

$$= [1 + (q-1)xy]^n \sum_{s=0}^{\infty} v_s \left[\frac{x + y + (q-2)xy}{1 + (q-1)xy} \right]^s =$$

$$= [1 + (q-1)xy]^n G \left(\frac{x + y + (q-2)xy}{1 + (q-1)xy} \right).$$

* Each sequence is taken as many times as there are code words at distance j from it.

** Assume that $C_k^l = 0$ for $l > k$ and $l < 0$.

(To avoid dealing with peculiarities, let us consider x and y small enough, e.g. $|x| < \frac{1}{q-1}$, $|y| < 1$.), then

$$G_j(x) = \frac{1}{2\pi i} \int_{C_0} \frac{[1+(q-1)xy]^n G\left(\frac{x+y+(q-2)xy}{1+(q-1)xy}\right)}{y^{j+1}} dy = Z_j G(x) \quad (3)$$

where the contour C_0 is a small circle centered in the origin and went around in the positive direction. Taking $Z = \frac{x+y+(q-2)xy}{1+(q-1)xy}$, we obtain

$$G_j(x)\, Z_j G(x) = \frac{1}{2\pi i} \int_{cx} \frac{[1+(q-2)x-(q-1)x^2]^{n+1} G(z)}{(z-x)^{j+1}[1+(q-2)x-(q-1)xz]^{n-j+1}} dz =$$

$$= \frac{1}{j!}[1+(q-2)x-(q-1)x^2]^{n+1} \lim_{z\to x} \frac{d^j \frac{Q(z)}{[1+(q-2)x-(q-1)xz]^{n-j+1}}}{dz^j} =$$

$$= \frac{1}{j!}[1+(q-2)x-(q-1)x^2]^{n+1} \sum_{k=0}^{j} C_j^k (j-k)!\, C_{n-k}^{j-k} \times \quad (4)$$

$$\times \frac{[(q-1)x]^{j-k}}{[1+(q-2)x-(q-1)x^2]^{n+1-k}} \frac{d^k G(x)}{dx^k} =$$

$$= \sum_{k=0}^{j} C_{n-k}^{j-k} \frac{[(q-1)x]^{j-k}[1+(q-2)x-(q-1)x^2]^k}{k!} \cdot \frac{d^k G(x)}{dx^k}.$$

Here C_x is small circle enveloping x rather than $\frac{1+(q-2)x}{(q-1)x}$.

Since

$$\sum_{j=0}^{i} G_j(x) = [1+(q-1)x]^n \quad (5)$$

we obtain that $G(x)$ satisfies the non-uniform linear differential equation of order t:

$$MG(x) = \sum_{t=0}^{t} \frac{[1+(q-2)x-(q-1)x^2]^k}{k!} \left(\sum_{r=0}^{t-k} C_{n-k}^r [q-1)x]^r\right) \frac{d^k G(x)}{dx^k} =$$

$$= [1+(q-1)x]^n. \quad (6)$$

Let us assume that

$$f_{n,\xi}(x) = [1+(q-1)x]^{\xi}[1-x]^{n-\xi} = \sum_{s=0}^{\infty} \varphi_s(n,\xi)\, x^s,$$

where

$$\varphi_s(n, \xi) = \sum_{r=0}^{s} (-1)^r (q-1)^{s-r} C_{\xi}^{s-r} C_{n-\xi}^{r}.$$

The equality

$$(1-x)^{-1} [1 + (q-1)x]^{\xi} [1-x]^{n-\xi} = [1 + (q-1)x]^{\xi} [1-x]^{n-1-\xi}$$

implies that $\sum_{k=0}^{s} \varphi_k(n, \xi) = \varphi_s(n-1, \xi)$. Since

$$[1 + (q-1)xy]^n f_{ni\xi}\left(\frac{x+y+(q-2)xy}{1+(q-1)xy}\right) =$$

$$= [1 + (q-1)xy + (q-1)x + (q-1)y + (q-1)(q-2)xy]^{\xi} \cdot$$

$$\cdot [1 + (q-1)xy - x - y - (q-2)xy]^{n-\xi} =$$

$$= [1 + (q-1)x]^{\xi} [1-x]^{n-\xi} [1 + (q-1)y]^{\xi} [1-y]^{n-\xi} = f_{n,\xi}(x) f_{n,\xi}(y),$$

$$L_j f_{n,\xi}(x) = \varphi_j(n\ \xi) f_{n,\xi}(x)$$

hence

$$Mf_{n,\xi}(x) = \sum_{j=0}^{t} \varphi_j(n, \xi) f_{n,\xi}(x) = \varphi_t(n-1, \xi)\ f_{n,\xi}(x).$$

Further argumentation is omitted because it is practically a literal replica of that of Lloyd [2].

Corollary 1. Substituting $n - x$ for ξ Lloyd's theorem may be restated as follows: if over an alphabet of q symbols there is a perfect code of length n correcting t errors, then the polynomial $P_t(n, q, x)$ in x of degree t defined by the equality

$$P_t(n, q, x) = \sum_{r=0}^{t} (-1)^r (q-1)^{t-r} C_{n-x}^{t-r} C_{x-1}^{r} \tag{7}$$

has t different roots among the numbers $1, 2, \ldots, n$.

References

1. *Зиновьев В. А.—Леонтьев В. К.:* Несуществование совершенных кодов над полями Галуа. 1973, **2**, 2, 24—31.
2. *Lloyd, S. P.:* Binary Black Coding. Bell Syst. Techn. J., 1957, **36**, 2, 517—535.

ON PROPERTIES OF ELEMENTS SUITABLE FOR SYNTHESIS OF d-RELIABLE CIRCUITS

M. A. GAVRILOV, V. M. OSTIANU, L. G. BIVOL
(Moscow)

Reliability-oriented design is generally a problem of probability theory because certain specific reliability parameters such as probability of failure-free operation, mean time between failures, etc. should be guaranteed.

For digital control circuits, however, the overwhelming majority of redundancy techniques are based on the so-called (d_0, d_1)-reliability.*

Since d_0 and d_1 are related to the probability of failure-free operation (the form of relation being dependent upon specific redundancy technique), we are going to speak below about design of (d_0, d_1)-reliable circuits.

Implementation complexity of a (d_0, d_1)-reliable circuit depends both on the redundancy technique and hardware elements used.

To the best knowledge of the authors, the literature almost completely lacks works discussing requirements to the inner structure of elements applihable to synthesis of d-reliable circuits with the least possible number of elements.

The present paper confines itself only to combinational (d_0, d_1)-reliable circuits (including combinational parts of sequential circuits). The general statement of the problem is discussed, and certain principal requirements to the inner structure of logic elements are formulated.

Let us describe the operation of a circuit by the system $\Phi\{F_i\}$ of the Boolean functions $F_i (i = 1, 2, \ldots, m)$.

Let the functions F_i be described by state tables with F_i taking value from: 0, 1 and 1/2 (false, true, and don't care).

A set of input variables $x_1, x_2, \ldots, x_t$ for which output j gives 1, 0 or 1/2 will respectively be called operational, non-operational and don't-care state of the output under consideration.

* A digital circuit is referred to as "(d_0, d_1)-reliable" if it operates correctly at "$1 \to 0$" failure in any $t_1 \leq d_1$ elements and "$0 \to 1$" ones in any $t_0 \leq d_0$ elements, and if there is at least one combination of either $d_1 + 1$ "$0 \to 1$" failures at $t_0 \leq d_0$ "$1 \to 0$" ones, or of $d_0 + 1$ "$1 \to 0$" failures at $t_1 \leq d_1$ "$0 \to 1$" ones where operational algorithm cannot be executed. If $d_0 = d_1$, the circuit will be called "d-reliable".

A function $F \in \{F_i\}$ is now considered and M_1 and M_0 denote the sets of its operational and non-operational states respectively. The function may be realized either through operational (M_1), or non-operational (M_0) states. The part of state table (M_1 or M_0) which is used for realization of F is referred to as "realized" and denoted by M_ε, the rest of the state table is called "non-realized" and denoted by $M_{\bar{\varepsilon}}$.

The designer should strive for a possibly best combination of elements implementing circuit outputs, taking into consideration the elements at his disposal.*

Definition 1: The function F will be said to be implemented by element $\varphi(d_0, d_1)$-reliably:

a) if F is implemented by φ which receives either input variables, or their functions, or both, i.e. if there is a circuit A with an output element φ which gives 1 for any state from M_1 and 0 for any state from M_0, and

b) if these conditions are not broken at appearance of at most d_0 of "$1 \rightarrow 0$" and d_1 of "$0 \rightarrow 1$" failures.**

For a circuit implementing Φ to be (d_0, d_1)-reliable, it is evidently sufficient that each $F \in \{F_i\}$ is (d_0, d_1)-reliable.

The output element φ is assumed to have n inputs $(q_1, q_2, \ldots, q_n)$, and its operation is described by the state table μ, with sets of operational and non-operational states being denoted by μ_1 and μ_0, respectively. It will implement function F if a set of the input variables $x_1, x_2, \ldots, x_t$ or their functions is chosen so that, at feeding them into its inputs, condition a) of Definition 1 is satisfied.

The state table of φ is referred to as "reduced" if the inputs $q_1, q_2, \ldots, q_n$ are substituted for by the variables $x_1, x_2, \ldots, x_t$ or their functions fed into φ and assigned to the states of the table $M_1 \cup M_0$ of F.

The method of Ref. [2] has been extended in Ref. [3] to (d_0, d_1)-reliable implementation of F by φ and the even to the case of arbitrary elements, in the following manner:

Let the disjunctive normal form (DNF) of the function implemented by φ and its negation be known. Let g_i deude the i-th conjunctive term of DNF and h_j the j-th conjunctive term of its negation.

We construct tables λ_ε with rows corresponding to the rows of M_ε and columns to the variables g_i, and $\lambda_{\bar{\varepsilon}}$. with rows corresponding to the rows of M_ε and columns to the variables g_i, and $\lambda_{\bar{\varepsilon}}$ with rows corresponding to the

* Method of redundant digital circuit synthesis based on this approach is discussed in Ref. [2] where a procedure of directed search of near-to-optimal circuit is described.

** φ is assumed to be absolutely reliable.

rnw of $M_{\bar{\varepsilon}}$ and columns to the variables h_j. Implementation of F by φ is non-contradictory if each row of λ_ε and $\lambda_{\bar{\varepsilon}}$ contains at least one 1. Implementation of F is d-reliable if each row of these tables contains at least $d + 1$ 1's. The set of conditions, which should be met by the initial variables or extended functions for implementation of F by φ to be non-contradictory and d-reliable, may be derived from λ_ε and $\lambda_{\bar{\varepsilon}}$.

We consider now several requirements to the inner structure of the element, i.e. to the state table μ characterizing its logical properties.

A particular case of these requirements is discussed in Ref. [7]. This problem is considered now in the general form.

In compliance with Ref. [4], the states $s \in \mu_\varepsilon$ and $s'' \in \mu_{\bar{\varepsilon}}$ will be said to have an obligatory variable $\tilde{x}_\varepsilon$ if they differ only in this variable.

Definition. Any state s^+ will be called (d_0, d_1)-neighbouring with the state ϱ if it may be obtained from ϱ through replacement of $t_0 (0 < t_0 < d_0)$ 1's by 0's and $t_1 (0 < t_1 < d_1)$ 0's by 1's, at least one of the numbers t_0 or t_1 being greater than zero.

The set of all (d_0, d_1)-neighbouring states will be denoted as $\{s(d_0, d_1)\}$.

Theorem: For the element φ to implement (d_0, d_1)-reliably function F, it is necessary and sufficient to ment the following conditions.

1) Each set μ_ε and $\mu_{\bar{\varepsilon}}$ of the truth table of φ should have at least one state $s' \in \mu_\varepsilon$ and $s'' \in \mu_{\bar{\varepsilon}}$ with no obligatory variable.

2) Besides s' μ_ε should contain the complete set $\{s'(d_0, d_1)\}$, and besides s'' $\mu_{\bar{\varepsilon}}$ should contain the complete set $\{s''(d_0, d_1)\}$.

Proof:

Sufficiency. Let the output element state table have two states without obligatory variables $s' \in \mu_\varepsilon$ and $s'' \in \mu_{\bar{\varepsilon}}$ and let both μ_ε and $\mu_{\bar{\varepsilon}}$ contain besides s' and s'' all their (d_0, d_1)-neighbouring states. Then for implementation of a given F, only the states s' and s'' will be assigned to the states of its table M. As a result of failures, states $s' \in \mu_\varepsilon$ and $s'' \in \mu_{\bar{\varepsilon}}$ may turn into some states s'^* and s''^*, if the number of "0 → 1" failures does not exceed d_1 and the number of "1 → 0" failures does not exceed d_0, $s'^* \in \{s'(d_0, d_1)\} \in \mu_\varepsilon$ and $s''^* \in \in s''^* \{s''(d_0, d_1)\} \in \mu_{\bar{\varepsilon}}$. Hence for the states s'^* and for s''^* the output will realize the value ε and the value $\bar{\varepsilon}$ respectively. Thus the output will be correct, and the proof of sufficiency is completed.

Necessity. Let F be (d_0, d_1)-reliably realized at the output of φ. This means that to the states $S \in M_\varepsilon$ the states $s' \in \mu_\varepsilon$ are assigned such that by changing in them no more that d_0 1's into 0's and d_1 0's into 1's, they turn into the states $s'^* \in \mu_\varepsilon$. Thus, $\{s'(d_0, d_1) \in \mu_\varepsilon\}$, and there is no obligatory variable in s'. A similar situation exists for $s'' \in \mu_{\bar{\varepsilon}}$ · Q.E.D.

If the table μ has more than two states satisfying the theorem, reliability-oriented synthesis will result in a less complex circuit.

Note that one may deal with (d_0, d_1)-reliable implementation of a subset of values of F (equal to ε or to $\bar{\varepsilon}$) rather than with the function as a whole. In such an event the theorem should be stated with regard to one of the sets μ_ε or $\mu_{\bar{\varepsilon}}$ respectively.

We show now that an arbitrary element φ, implementing (d_0, d_1)-reliably an arbitrary (but given) function F, realizes either the corrector output function, or a sum of two functions, one of them being the corrector output and the other (denoted by ψ) being unfit for (d_0, d_1)-reliable implementation of F.

If a d-reliable circuit is implemented by means of the corrector technique, any corrector output realizes ogeneralized symmetric function [5]:

$$\varphi_{t,t-1,\ldots,t-d}(x_1, x_2, \ldots, x_t)$$

having operating numbers $t, t-1, \ldots, t-d$ and depending on the variables $x_1, x_2, \ldots, x_\tau$. By the virtue of the theorem above, the table of a t-input element φ implementing d-reliably the given function F should contain the complete set $\{s'(d_0, d_1)\}$ in μ_ε besides each $s' \in \mu_\varepsilon$ and the complete set $\{s''(d_0, d_1)\}$ in μc-besides each $s^3 \in \mu_{\bar{\varepsilon}}$. Thus, the given several generating states* $s' \in \mu_\varepsilon$ and $s'' \in \mu_{\bar{\varepsilon}}$, φ implement the function (1).

But since the truth table of φ may contain states other than s', s'', $\{s',(d_0, d_1)\}$, and $\{s''(d_0, d_1)\}$, φ may besides (1) implement also a certain function $(x_1, x_2, \ldots, x_t)$. Thus

$$\varphi(\varkappa_1, \varkappa_2, \ldots, \varkappa_t) = S_{t,t-1,\ldots,t-d}(\varkappa_1, \ldots, \varkappa_x) + \psi(\varkappa_1, \ldots, \varkappa_t).$$

On the other hand, since φ implements F d-reliably, to the states of M only those states of μ are assigned which have all their d-neighbouring states $\{s(d)\}$ in the same table (μ_ε or $\mu_{\bar{\varepsilon}}$) to which s belongs. Thus only those conjunctive terms of $\varphi(\varkappa_1, \varkappa_2, \ldots, \varkappa_t)$ are used which constitute (1).

Let now a $(d_0, d_1,)$-reliable circuit be implemented by the corrector technique. The form of the functions realized at corrector outputs of a (d_0, d_1)-reliable digital circuit built of functional elements, is defined in Ref. [6]. The design of these devices uses the Kim–Freiman code [8] featured by the distance $D(\alpha, \beta)$ between any pair of its words α and β satisfying

$$D(\alpha, \beta) \geq 2(d_0 + d_1 + 1) - (|\alpha| - |\beta|)$$

where $|x|$ is the norm of the code word x.

* For the definition of the generating state we refer to [5].

Let a word γ from the Kim—Freiman code be fed into the corrector inputs. Under $t_0 < d_0$ "$1 \to 0$" failures the word preserves all 0's unchanged, and any t_0 1's will change into 0's. Thus disjunction of corrector input states resulting from t_0 "$1 \to 0$" failures may be represented as conjunction of unchanged letters and an elementary generalized symmetric function having the operating number t_0 and generated by the remaining word symbols. Similar situation exists for $t_1 \leq d_0$ "$0 \to 1$" failures.

As "$1 \to 0$" and "$0 \to 1$" failures may happen simultaneously (their number not exceeding, respectively, d_0 and d_1), the set of corrector input states is defined by the product of two elementary generalized symmetric functions

$$\sigma_{r_0}(\varkappa_{i_1}, \varkappa_{i_2}, \ldots, \varkappa_{i_t})\, \sigma_{t_1}(\varkappa_{i_{k+1}}, \ldots \varkappa_{i_j}), \tag{2}$$

where k is the number of 1's in the code word under consideration. Hence any corrector output realizes a partially symmetric function representing the Boolean sum over all k products of elementary generalized symmetric functions (2):

$$\alpha(\varkappa_1, \ldots, \varkappa_t) = \sum_{k=0}^{t} \sum_{\substack{0 \leq t_0 \leq d_0 \\ 0 \leq t_1 \leq d_1}} \sigma_{j_0}(\alpha_k)\, \sigma_{t_1}(\alpha_{t-k}) = \sum_{k=0}^{t} \sigma_{0,\ldots,d_1}(\alpha_k)\, \sigma_{0,\ldots,d_1}(\alpha_{-k}) \tag{3}$$

where α and α_{t-k} are sets of variables on which σ_{t0} and σ_{t1} of (2) depend respectively. The number and indices of variables on which the functions σ_0, d_0 and σ_0, d_1 depend, are defined by the values of the word γ symbols.

Consider now the function of the element φ implementing (d_0, d_1)-reliably a given function F. By the theorem above, the table μ of φ contains the complete set $\{s'(d_0, d_1)\}$ in μ_ε besides each $s' \in \mu_\varepsilon$, and $s''(d_0, d_1)\}$ in the $\mu_{\bar{\varepsilon}}$ besides each $s^5 \in \mu_{\bar{\varepsilon}}$. Disjunction of all these state sets is nothing but the partially symmetric function (3). Thus the function of φ is as follows

$$\alpha(\varkappa_1, \ldots, \varkappa_t) + \psi(\varkappa_1, \ldots, \varkappa_t).$$

On the other hand, if φ implements $F(d_0, d_1)$-reliably, only those states in μ are assigned to the states of M which have their corresponding (d_0, d_1)-neighbours in the same part of the table (μ_ε or $\mu_{\bar{\varepsilon}}$). Thus among the conjunctive terms of the function φ, only those are used which constitute (3).

Thus, the logic function of φ implementing (d_0, d_1)-reliably a function F (and hence having the afore mentioned properties) is either identical to (3) realized at any output of the (d_0, d_1)-reliable-circuit corrector, or to the sum of (3) and some other function $\psi(\varkappa_1, \varkappa_2, \ldots, \varkappa_t)$ which is not used in the reliability-oriented synthesis.

The element realizing a partially symmetric function is therefore the best suited as an output unit for the purpose of the reliability-oriented synthesis.

References

1. *Гаврилов М. А.—Остиану В. М.—Потехин А. И.:* Надежность дискретных систем. Итоги науки. Теория вероятностей. Математическая статистика. Теоретическая кибернетика, 1969, Москва, 1970, 104 pp.
2. *Гаврилов М. А.—Копыленко В. М.:* Синтез структуры бесконтактных релейных устройств. Bull. Math, de la Soc. Math. de la RSR 1966, 10 (58), 3.
3. *Гаврилов М. А.—Копыленко В. М.:* Инженерные методы синтеза комбинаторных структур на произвольных элементах. Изв. на метаматические институт БАН, 1971, Vol. 12, 211—230.
4. *Гаврилов М. А.:* Структурная теория релейных устройств. Лекция по курсу «Основы теории автоматических и телемеханических систем». Part II. ВЗЭИ. Москва, 1960.
5. *Гаврилов М. А.:* Сигнализация и прогнозирование отказов в дискретных управляющих устройствах со структурной избыточностью. Доклад на II Международномконгрессе по автоматическому управлению, Базель, 1963, М., «Наука», 1965, 211—220.
6. *Гаврилов М. А.—Кришталь В. З.—Остиану В. М.:* Синтез надежных релейных устройств, устойчивых (d_0, d_1)-безотказных относительно внутренних элементов и входов. «Международный конгресс ИФАК по автоматическому управлению, III, Лондон. Рефераты». Лондон, ИФАК, 1966, Vol. 3. Изд. Ин-та инженеров-механиков, 1968.
7. *Бивол Л. Г.:* О надежности схем, построенных из многофункциональных элементов. Материалы докладов VII научно-технической конференции, Кишиневский политехнический институт, Кишинев, 1971.
8. *Kim, W. N., Freiman, G. W.:* Multi-error correcting code for a binary asymmetric channel. IEEE Trans., v. **CT—5**, Special supplement number, 1959.

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

Budapest XI., Goldman György tér 3, Hungary

The authors are requested to retain a third copy of the ubmitted typescript to be able to check the proofs against it

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4–5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction, (outlining the problem), Methods and results, Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которых должен содержать рисунки, таблицы и литературу) направляются

Е. Д. Теряеву – Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский Проспект, 14, Москва В-71, СССР

или

Я. Кочишу – Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета, Будапешт XI, площадь
Голдманн Д., 3, Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре.

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4–5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру гранок.

Статье должна предшествовать аннотация объемом до 50–100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме – на русском языке, если статья написана на английском, и наоборот – должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем разюме не должен превышать 10–15 % объема статьи.

Авторам высылаются гранки статьи, которые они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Meerov, M. V.—Litvak, B. L.: The optimization of multivariable systems and Markovian programming (*Меerов М. В.—Литвак Б. Л.:* Оптимизация многосвязных систем и марковское программирование) 95

Бурнашев М. В.—Хасьминский Р. З.: Асимптотические выражения для информации в канале с разрывной плотностью шума (*Burnashev, M. V.—Has'minsky, R. Z.:* Asymptotical expressions for the information in the channel with discontinuous noise density) 107

Зиновьев В. А.—Леонтьев В. К.: Несуществование совершенных кодов над полями Галуа (*Zinovjev, V. A.—Leontjev, V. K.:* The nonexistence of perfect codes over Galois fields) 123

Бассалыго Л. А.: Обобщение теоремы Ллойда на случай произвольного алфавита (*Bassalygo, L. A.:* The generalizations of Lloyd's theorem on an arbitrary alphabet) 133

Гаврилов М. А.—Остиану В. М.—Бивол Л. Г.: О свойствах элементов, удобных для синтеза d-безотказных структур (*Gavrilov, M. A.—Ostianu, V. M.—Bivol, L. G.:* On properties of elements suitable for the synthesis of d-faultless structures) 139

Gács, P.—Körner, J.: Common information is far less than mutual information (*Гаач П.—Корнер Й.:* Общая информация существенно больше, чем взаимная информация) 149

Przybylo, S.: Relations entre les morphismes et l'equivalence des automates indéterminés (*Пршибило С.:* О связях между морфизмом и эквивалентностью недетерминированных автоматов) 163

Siforov, V. I.: Third International Symposium on Information Theory 173

316.920

VOL. 2 * NUMBER 3–4
TOM ГОМЕР

ACADEMY OF SCIENCES OF THE USSR
HUNGARIAN ACADEMY OF SCIENCES

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

АКАДЕМИЯ НАУК СССР
АКАДЕМИЯ НАУК ВЕНГРИИ

1973

AKADÉMIAI KIADÓ, BUDAPEST

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

— general theory of control systems and system theory

— theory of automata

— information theory

— operation research; theory of complex systems

— theory of economic control; system modelling

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

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

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

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

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

Distributor:

KULTURA

Hungarian Trading Co. for Books and Newspapers

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

The quarterly is published by

AKADÉMIAI KIADÓ

Publishing House of the Hungarian Academy of Sciences

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

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

Журнал создан для обеспечения быстрейшей публикации материалов о новейших результатах научных исследований, разработок и кратких сообщений о достижениях в следующих областях науки:

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

Журнал издается четыре раза в год общим объёмом около 320 печатных страниц.

Журнал в первую очередь будет публиковать научные достижения обеих стран, а также статьи из других стран, представляющие международный интерес.

Распространитель:

KULTURA

Венгерское Общество по распространению книг и журналов

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

Издатель журнала:

AKADÉMIAI KIADÓ

Издательство Академии Наук Венгрии

H-1363 Будапешт, п. о. 24, Венгрия

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITORS

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

DEPUTY EDITORS

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

CO-ORDINATING EDITORS

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

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

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

L. VARGA (Budapest)

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

РЕДАКТОРЫ ЖУРНАЛА

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

ЗАМЕСТИТЕЛИ РЕДАКТОРОВ

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

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

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

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

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

Л. ВАРГА (Будапешт)

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

AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Printed in Hungary
AKADÉMIAI NYOMDA, BUDAPEST

Problems of Control and Information Theory, Vol. 2 (3—4), pp. 183—197 (1973)

COMPROMISE SOLUTIONS, DOMINATION STRUCTURES AND SALUKVADZE'S SOLUTION

P. L. YU

(Rochester)

G. LEITMANN

(Berkeley)

(Received May 10, 1973)

We outline the concepts of compromise solutions and domination structures in such a way that the underlying assumptions and their implications concerning the solution concept suggested by Salukvadze may be clearer. An example is solved to illustrate our discussion.

1. Introduction

In daily decision making we deal quite often with problems involving not only a single criterion. Rather, we may have multiple objective decision problems, or decision problems involving more than one decision maker. For these types of decision problems there is not as yet a universally accepted solution concept, even though there exist quite a few (see Refs. [1—37]). Although we may use a recently derived concept of domination structures [25, 27] to study the assumptions which underly each solution concept, there is no apparent reason to believe that a certain solution concept is superior to others. Although we, as researchers, consultants or decision makers, can use the concept of domination structures to suggest a limitation of decisions to the set of nondominated solutions (which may contain more than one alternative), the final decision must depend on the judgement and/or the relative strength in negotiation or bargaining of the decision makers.

In the search for solution concepts, one often starts with static decision problems (one stage decision problems) rather than with the much more complicated cases of dynamic decision problems. As a consequence, there is a natural gap among different researchers. Those dealing with static problems may be unaware of the implications of solution concepts in dynamic cases. On the other hand, those interested in dynamic problems may be unaware of current developments in solution concepts.

Recently there appeared two interesting papers [32—33] by M. E. Salukvadze, in which he considers dynamic decision problems with multiple objective functionals. He proposes a solution concept and works out the computational details. However, the discussion of the solution concept is not com-

plete. The underlying assumptions as well as their implications are not entirely clear. The solution concept proposed by Salukvadze has been discussed independently in [24] and [34]; in fact, it is a special case of the compromise solutions in [24].

In this note we outline some underlying assumptions and their implications concerning the solution concept suggested by Salukvadze in order that its applicability may be clearer. In particular, in Section 2 we define the decision space and criteria space for both static and dynamic decision problems. We then focus on the concept of compromise solutions and list their known properties. Since Salukvadze's solution is a special case of compromise solutions, all properties described are applicable to his solution. In Section 3, we introduce the concept of domination structure and show how strong the assumption is that underlies compromise solutions (and hence Salukvadze's solution). Finally, in Section 4 we treat the example of [32] to illustrate the points made in Sections 2 and 3 so that the merits as well as the weaknesses of Salukvadze's solution may be more apparent.

2. Compromise solutions

Suppose that we have to make a choice from a set of alternatives $X \subset R^n$, and that we can associate each alternative $x \in X$ with a set of criteria $(f_1(x), \ldots, \ldots, f_l(x))$. Let $f(x) = (f_1(x), \ldots, f_l(x))$. We shall call X the *decision space*, while $Y = \{f(x) \mid x \in X\}$ is the *criteria space*. An element $y \in Y$ is often called the outcome of a decision.

Remark 2.1. Observe that the above definitions of decision and criteria spaces can be extended to dynamic cases and thus the solution concepts we describe here in fact are applicable to such cases. For instance we can denote the set of all admissible strategies or controls μ by X (see Refs [2, 4, 35] for admissible strategies or controls). Then, if for each μ we have $(f_1(\mu), \ldots, f_l(\mu))$ as the performance indices, the criteria space can be specified by $Y = \{(f_1(\mu), \ldots, f_l(\mu)) \mid \text{all admissible strategies (or controls) } \mu\}$. On the other hand, since we can convert integral pay-offs intoter minal pay-offs by adding extra variables (see [2, 35]), we may regard X as the set of all attainable terminal states in the enlarged space. Then Y can be defined accordingly on X. Although conceptually we have no difficulty in defining X and Y in dynamic cases, to actually visualize them is a very difficult task. We shall describe the solution concepts in static cases and leave it to the reader to extend them to dynamic cases. However, in Section 4, we shall supply an example of a dynamic case.

Let $y_j^* = \sup \{f_j(x) \mid x \in X\}$. Then $y^* = (y_1^*, \ldots, y_l^*)$ is called the *utopia point* or *ideal point* of our problem with the interpretation that whenever y^* is feasible, $f^{-1}(y^*)$ simultaneously maximizes each criterion. For simplicity, we shall assume that Y is compact. With some slight modification, the compactness assumption can be relaxed.

Given $y \in Y$, we define a class of *regret functions* by

$$R_p(y) = \|y^* - y\|_p = \Big[\sum_j (y_j^* - y_j)^p\Big]^{\frac{1}{p}}, \quad p \geq 1. \tag{1}$$

Definition 2.1. y^p and $f^{-1}(y^p)$ is the compromise solution with parameter $p \geq 1$ if and only if y^p minimizes $R_p(y)$ over Y.

For convenience, we shall call $R_p(y)$ the group regret of y with respect to p, while $y_j^* - y_j$ is the j^{th} individual regret.

Remark 2.2. The solution concept proposed by Salukvadze is a compromise solution with $p = 2$. The corresponding group regret is associated with the Euclidean norm. The solution resulting from usual goal programming [10] or simple majority rule can be regarded as a compromise solution with $p = 1$ (Ref. [28]). Compromise solutions with $p = \infty$ correspond to a minimax criterion, because $y^{p=\infty}$ solves $\min_y \max_j \{y_j^* - y_j \mid j = 1, \ldots, l\}$.

Remark 2.3. In group decision problems, we may use $y_j = f_j(x)$ to denote the j^{th} decision maker's utility. It is reasonable to assume that compromise solutions cannot be acceptable to each decision maker unless each one has nonnegative utility. Thus in Ref. [24], instead of X and Y, the decision and utility space for compromise solutions are defined respectively by $X_0 = \{x \in X \mid f(x) \geq 0\}$ and $Y_0 = \{f(x) \mid x \in X_0\}$. The utopia point y^* is then the point which has its j^{th} component y_j^* maximizing $f_j(x)$ over X_0. With this modification, compromise solutions enjoy the property of *individual rationality* (i.e., nobody in the group has negative utility). In this note we do not make such a modification because the individual rationality is an unnecessary nicety for one decision maker with multiple criteria. Observe that, except for individual rationality, all properties of compromise solutions (listed below) hold no matter whether or not we introduce the modification. Since compromise solutions can be applied to group decision problems or multicriteria decision problems with one decision maker, $y_j = f_j(x)$ in this note is called the j^{th} objective, criterion or utility, interchangeably. Also since minimizing $f_j(x)$ over X is equivalent to maximizing $-f_j(x)$ over X, if $f_j(x)$ is the level of disutility we may interpret $-f_j(x)$ as the level of utility, and vice versa. Thus without loss of generality, we can focus on the properties of compromise solutions defined in Definition 2.1. The extension to other cases is obvious.

Definition 2.2. Let $\Lambda^{\leqq} = \{d \in R^l \mid d \leqq 0\}$. We say that Y is $\Lambda^{\leqq}$-convex if and only if $Y + \Lambda^{\leqq}$ is convex (For more discussion on cone convexity see [25]).

Compromise solutions enjoy the following properties (see [24]):

(i) *Feasibility.* For each $p \geqq 1$, under the assumption that Y is compact, there is always a compromise solution. Observe that some solution concepts such as *stable set* or *core* may have no feasible solution.

(ii) *Least Group Regret.* Since y^p is the closest point over Y to the utopia point, the group regret is minimized in the sense of distance.

(iii) *No Dictatorship.* That is, the group decision is not completely determined by any one criterion $f_j(x)$. In contrast to lexicographical maximization* in which some f_j may not be considered in the final decision, each f_j is considered in the compromise solution.

(iv) *Pareto Optimality.* For $1 \leqq p < \infty$, each compromise solution is Pareto optimal. That is, there is no other $y \in Y$ such that $y \geqq y^p$ and $y \neq y^p$. This property comes directly from Definition 2.1.

(v) *Uniqueness.* Suppose that Y is $\Lambda^{\leqq}$-convex. Then each y^p, $1 < p < \infty$, is unique.

(vi) *Symmetry or Principle of Equity.* If Y is convex and closed with respect to cyclical rotation, then for each p, $1 < p \leqq \infty$, the y_j^p, $j = 1, \ldots, l$, are identical. For $p = 1$ there is at least one compromise solution y^1 such that the y_j^1, $j = 1, \ldots, l$, are identical. Thus the principle of equity is implicit in the concept of compromise solutions.

(vii) *Independence of Irrelevant Alternatives.* Suppose $X \subset X'$ and $\max_X f_j(x) = \max_{X'} f_j(x)$, $j = 1, \ldots, l$. If a compromise solution with respect to X' happens to be x^0 in X, then x^0 is also a compromise solution with respect to X. Thus the irrelevant alternatives may be discarded from the consideration for compromise solution. For instance for $1 \leqq p < \infty$ we may discard those y of Y which are not Pareto optimal without affecting the final compromise solution.

If we treat p as a parameter of y^p, then $\{y^p \mid p \geqq 1\}$ enjoys the following properties:

* Let $X^0 = X$. For $k = 1, \ldots, l$ define $X^k = \{x^0 \in X^{k-1} \mid f_k(x^0) \geqq f_k(x),\ x \in X^{k-1}\}$. In lexicographical maximization the final decision is on the set X^*. Observe, if X^k contains only one point, the final decision is uniquely determined; $f_{k+1}, \ldots, f_l$ need not to be considered.

(viii) *Continuity.* Suppose that Y is $\Lambda^{\leqq}$-convex. Then, as a function of p, y^p is continuous over $1 < p < \infty$. If y^1 (or y^∞) is unique, then y^p is continuous at $p = 1$ (or $p = \infty$) (see Ref. [36]).

(ix) *Monotonicity and Bounds.* If $l = 2$ and Y is $\Lambda^{\leqq}$-convex, under a very mild condition it can be shown that $\{x_j^p\}$, $j = 1, 2$, is bounded by $[y_j^1, y_j^\infty]$; furthermore it is a monotone function of p (see Ref. [24]; for some generalization of this result for $l > 2$, see Ref. [36]).

(x) *Monotonicity of the Group Utilities and the Individual Regrets.* Under the same assumptions as in (ix) it can be shown that both $\sum_j y_j^p$ and $\max_j \{y_j^* - y_j^p\}$ are decreasing functions of p. Observe that if $y_j = f_j(x)$ is the utility function for the j^{th} decision maker, then $\sum_j y_j^p$ is the sum of the utilities and $\max \{y_j^* - y_j^p\}$ is the maximum individual regret. We may consider $\sum_j y_j^p$ and $\max_j \{y_j^* - y_j^p\}$ as group utility and individual regret, respectively, resulting from the compromise solution with parameter p. Our result says that as p increases, the group utility decreases, however, the individual regret reduces, too. In this sense, the parameter p has a meaning for balancing group utility and individual regret. As a consequence, we see that simple majority rule (see Remark 2.3) is the rule which maximizes the group utility and most neglects the individual regret in the entire domain of compromise solutions (see [28]).

The concept of compromise solutions can be generalized in several directions. For instance, we may replace $R_p(y)$ by $R_p(\alpha, y) = [\sum_j \alpha_j^p (y_j^* - y_j)^p]^{\frac{1}{p}}$. Then most of the above properties with a suitable modification remain the same [24].

Although compromise solutions have merits, they are by no means perfect. Some associated assumptions are discussed in the next section. In addition, compromise solutions implicitly impose an intercomparison among the criteria or utilities through the group regret function (negative side of social welfare function). This imposition is not acceptable on some occasions. One should also observe that compromise solutions are not independent of a positive linear transformation of the $f_j(x)$. In fact, changing the scale of $f_j(x)$ has the same effect as changing the weight α in $R_p(\alpha, y)$. In applying this solution concept, one should be aware of this defect.

3. Domination structures

In this section we shall introduce briefly the concept of domination structures so that we can understand how strong an assumption has been imposed in compromise solutions. (For more detailed discussion we refer to [25]; a simple summary can be found in [16] and [27]. An attempt to apply the concept to solve some practical problems can be found in [37].)

Given two outcomes, y^1 and y^2, in the criteria space Y, we can write $y^2 = y^1 + d$ with $d = y^2 - y^1$. If y^1 is preferred to y^2, written $y^1 \succ y^2$, we can think of this preference as occurring because of d. Now suppose that the nonzero vector d has the additional property that if $y = y^1 + \lambda d$ and $\lambda > 0$, then $y^1 \succ y$. Then d will be called a *domination factor* for y^1. Note that by definition, given a domination factor for y^1, any positive multiple of it is also a domination factor. It follows that, given $y^1 \succ y^2$, it is not necessarily true that $d = y^2 - y^1$ is a domination factor for y^1.

Let $D(y)$ be the set of all domination factors for y together with the zero vector of R^l. The family $\{D(y) \mid y \in Y\}$ is the *domination structure* of our decision problem. For simplicity, the structure will be denoted by $D(\cdot)$.

One important class of domination structures is $D(y) = \Lambda$, Λ a convex cone, for all $y \in Y$. In this case, we shall call Λ the *domination cone*.

Definition 3.1. Given Y, $D(\cdot)$ and two points y^1, y^2 of Y, by y^1 *is dominated by* y^2 we mean $y^1 \in y^2 + D(y^2) = \{y^2 + d \mid d \in D(y^2)\}$. A point $y^0 \in Y$ is a *nondominated solution* if and only if there is no $y^1 \in Y$, $y^1 \neq y^0$, such that $y^0 \in y^1 + D(y^1)$. That is, y^0 is nondominated if and only if it is not dominated by any other outcomes. Likewise, in the decision space X, a point $x^0 \in X$ is a nondominated solution if and only if there is no x^1 in X such that $f(x^0) \neq f(x^1)$ and $f(x^0) \in f(x^1) + D(f(x^1))$. The set of all nondominated solutions in the decision and criteria space will be denoted by $N_X(D(\cdot))$ and $N_Y(D(\cdot))$, respectively. Because of its geometric significance, a nondominated solution with respect to a domination cone Λ is also called a Λ-extreme point. The set of all Λ-extreme points is denoted by $\mathrm{Ext}\,[Y \mid \Lambda]$.

Example 3.1. Let $\Lambda^{\leqq} = \{d \in R^l \mid d \leqq 0\}$. We see that y is Pareto optimal if and only if y is a $\Lambda^{\leqq}$-extreme point. That is, in the concept of Pareto optimality one uses a constant domination cone $\Lambda^{\leqq}$. Observe that $\Lambda^{\leqq}$ is only $1/2^l$ of the entire space. When $l = 6$, for instance, $\Lambda^{\leqq}$ is only $1/64$ of R^6.

Example 3.2. In the additive weight method, one first finds a suitable weight $\lambda = (\lambda_1, \ldots, \lambda_l)$ and then maximizes $\lambda \cdot f(x) = \sum_j \lambda_j f_j(x)$ over X or $\lambda \cdot y = \Sigma \lambda_j y_j$ over Y. Given λ, the solution concept implicitly uses the domi-

nation cone $\{d \in R^l \mid \lambda \cdot d < 0\}$. The concept of the additive weight method is closely related to that of "trade-off" in economic analysis. In fact the latter can be a way of obtaining the weight λ. In order to illustrate this, let us limit ourselves to $l = 2$. The trade-off ratio of $f_2(x)$ with respect to $f_1(x)$ is defined by how many units of $f_2(x)$ we want to sacrifice in order to increase $f_1(x)$ by one unit. Thus the ratio gives the value of $f_2(x)$ in terms of $f_1(x)$. Although the ratio may be nonlinear in reality, in practice one often interprets it as a constant ratio. For the time being, assume that the ratio is constant and given by λ_2/λ_1 with $\lambda_1 > 0$. Clearly our decision problem becomes one of maximizing $f_1(x) + (\lambda_2/\lambda_1) f_2(x)$ over X, which is equivalent to maximizing $\lambda_1 y_1 + \lambda_2 y_2$ over Y. The latter is essentially an additive weight method. To correctly predetermine the ratio λ_2/λ_1 is a very difficult task. In practice, one may first use his experience or judgement to set up its bounds. Once the bounds are set, we have partial information on preference. For instance, suppose that $1 < \lambda_2/\lambda_1 < 3$ is given. Then we obtain valuable information, because it implies that, by using

$$\Lambda = \{(d_1, d_2) \mid d_1 + d_2 \leqq 0,\ d_1 + 3d_3 \leqq 0\}$$

as our domination cone, the optimal solution is contained in Ext $[Y \mid \Lambda]$. (See Fig. 1 and Ref. [27].)

Example 3.3. For compromise solutions with $1 \leqq p < \infty$ we can define the related domination structures by

$$D(y) = \{d \in R^l \mid \nabla R_p(y) \cdot d > 0\} \text{ for } y \in Y,$$

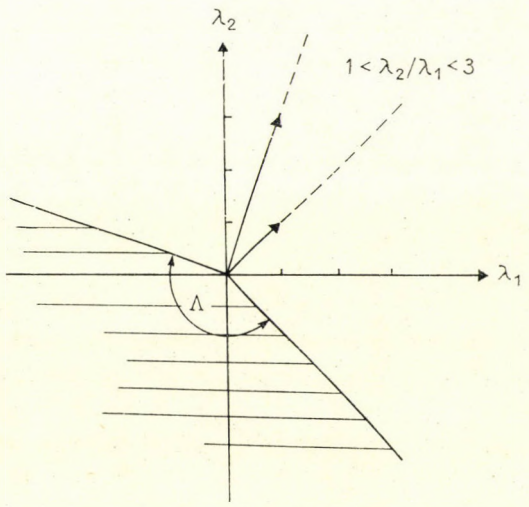


Fig. 1

where $\nabla R_p(y)$ is the gradient of $R_p(y)$ which is given by (1). This is because in compromise solutions the smaller $R_p(y)$ is, the more it is preferred. Since $R_p(y)$ is a differentiable convex function, $D(y)$ forms a domination cone for y. When $p = \infty$, we can again specify its domination structure; we shall not do so here in order not to distract from the main ideas. Observe that each $D(y)$ contains a half space, no matter what the parameter p is.

Remark 3.1. The domination cone induced by Pareto optimality (that is, $\Lambda^{\leqq}$; Example 3.1) is sometimes not large enough to encompass all information possessed by the decision makers. For instance in Example 3.2, Λ is larger than $\Lambda^{\leqq}$; $\Lambda - \Lambda^{\leqq}$ is the valuable information missed by Pareto optimality. On the other hand, the domination structure induced by compromise solutions (Example 3.3) is such that each $D(y)$ contains a half space. To make this possible, too strong an assumption may have been imposed. For instance in Example 3.2, although Λ is much larger than $\Lambda^{\leqq}$, it is much less than a half space. The difference between Λ and the half space of $D(y)$ induced by compromise solutions (when they are comparable) depends on how strong the assumptions are or how much information we have on the preference of the outcomes in Y. From the computational point of view one has no difficulty in computing a compromise solution because the problem reduces to one of mathematical programming or to a control problem. However, because of the strong assumptions imposed or the information required, the compromise solutions may not be acceptable. A researcher may be able to find the compromise solutions, but these solutions may not be actually desirable for the decision maker (see the next section for further discussion). In daily decision problems we are usually faced with problems of partial information. Both Pareto and compromise solutions cannot suitably explain the decision situation. The former does not adequately employ the partial information, while the latter may impose too strong an assumption which is not conformable with the partial information. We believe that one should try to make suitable shoes (mathematical models) for the feet (the decision problems), rather than to cut off the feet in order to wear a given pair of shoes. (One can use domination structures to attack some partial information decision problems, see Refs [16, 27, 37].)

4. Example

In order to illustrate the discussion of the earlier section, consider a problem treated by Salukvadze [32]:

(i) Playing Space: $\{(z_1, z_2) \mid |z_1| \leqq 3\}$

(ii) Dynamic System:
$$\left.\begin{aligned} \dot{z}_1 &= z_2 \\ \dot{z}_2 &= u \end{aligned}\right\} \qquad t \in [0, T]$$

(iii) Control Set: $u \in [-1, 1]$

(iv) Initial Condition: $z_1(0) = 1,\ z_2(0) = 0$

(v) Terminal Condition: $z_1(T) = 0,\ T$ not specified

(vi) Criteria: (A) minimize $f_1(u(\cdot)) = T$

(B) maximize $f_2(u(\cdot)) = z_2(T)$.

As mentioned in Remark 2.1, the decision space X and criteria space Y may be very difficult to visualize. However, in this example, letting X be the set of all admissible controls, the set of all Pareto optimal solutions can be found. Note that (A) requires minimization; to permit use of the results in the earlier sections, we can consider maximizing $-f_1(u(\cdot))$. Rather than doing so, we shall consider the problem as is; the simple modifications required in the results are obvious.

It is readily established that the set of terminal points reachable from $(1, 0)$ along a solution path that remains in the playing space is given by $\{(z_1, z_2) \mid z_1 = 0,\ -\sqrt{6} \leq z_2 \leq \sqrt{6}\}$ (see Fig. 2).

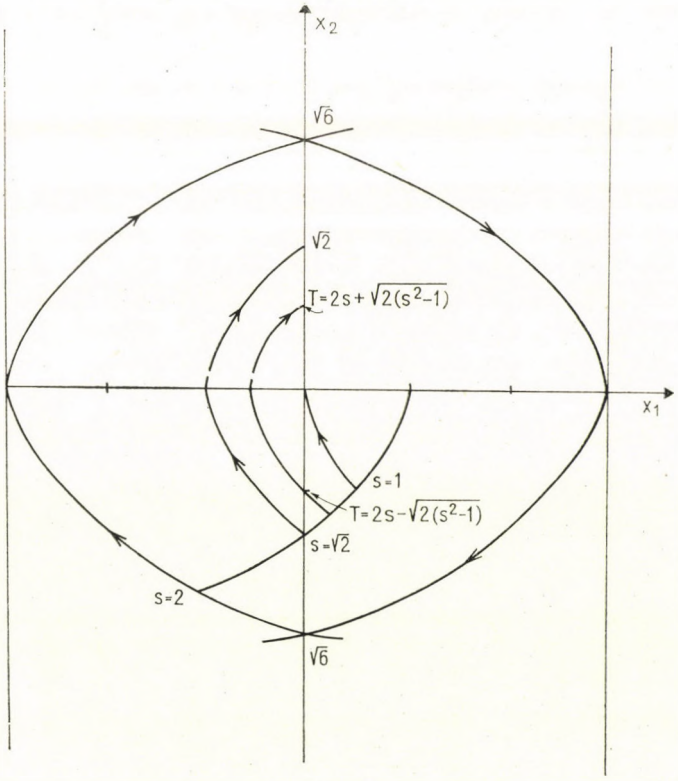


Fig. 2

Let us first consider terminal points on the subset $\{(z_1, z_2) \mid z_1 = 0, -\sqrt{2} \leq z_2 \leq \sqrt{6}\}$ of the set of reachable terminal points. The control

$$u^*(t) = \begin{cases} -1 \text{ for } t \in [0, s] \\ 1 \text{ for } t \in (s, T] \end{cases}$$

renders a minimum of the transfer time T from $(1, 0)$ to a given terminal point $(0, z_2)$, $-\sqrt{2} \leq z_2 \leq \sqrt{6}$, where the switching time s is a function of z_2. Conversely, for a given switching time s, control $u^*(\cdot)$ results in a transfer time T and a corresponding terminal point $(0, z_2(T))$.

The solution path generated by $u^*(\cdot)$ is given by

$$\left.\begin{aligned} z_1(t) &= \frac{1}{2}t^2 + 1 \\ z_2(t) &= -t \end{aligned}\right\} \quad t \in [0, s]$$

and

$$\left.\begin{aligned} z_1(\tau) &= \frac{1}{2}\tau^2 - s\tau + 1 - s^2 \\ z_2(\tau) &= \tau - s \end{aligned}\right\} \quad \tau \geq 0 \tag{2}$$

where $\tau = t - s$. At termination $z_1 = 0$ so that $\tau = s \pm \sqrt{2(s^2 - 1)}$, whence

$$T = 2s \pm \sqrt{2(s^2 - 1)}\,. \tag{3}$$

In order that T be defined, i.e. termination takes place, $s \geq 1$.

Since, by (2), $\dfrac{dz_1}{d\tau} = 0$ for $\tau = s$, it follows that the minimum value of $z_1(\tau)$ is given by $1 - s^2$. But $z_1(\tau) \geq -3$, so that $s \leq 2$. Thus, $s \in [1, 2]$.

Since $\tau \geq 0$, $\tau = s - \sqrt{2(s^2 - 1)} \geq 0$, which is met if and only if $s \leq \sqrt{2}$. Thus, for $s \in [1, \sqrt{2}]$, the portion of the solution path defined by (2) crosses the z_2-axis twice, the first time at $\tau = 2s - \sqrt{2(s^2 - 1)}$ and the second time at $\tau = 2s + \sqrt{2(s^2 - 1)}$. Thus,

$$f_1 = 2s - \sqrt{2(s^2 - 1)},\ f_2 = -\sqrt{2(s^2 - 1)} \tag{4}$$

$$f_1 = 2s + \sqrt{2(s^2 - 1)},\ f_2 = \sqrt{2(s^2 - 1)}\,. \tag{5}$$

For $s \in [\sqrt{2}, 2]$, the portion of the solution path given by (2) intercepts the z_2-axis once. Thus,

$$f_1 = 2s + \sqrt{2(s^2 - 1)},\ f_2 = \sqrt{2(s^2 - 1)}\,. \tag{6}$$

These results are summarized in Table 1.

Table 1

	f_1	f_2	s
Curve 1	$2s - [2(s^2-1)]^{\frac{1}{2}}$	$-[2(s^2-1)]^{\frac{1}{2}}$	$1 \leqq s \leqq \sqrt{2}$
Curve 2	$2s + [2(s^2-1)]^{\frac{1}{2}}$	$[2(s^2-1)]^{\frac{1}{2}}$	$1 \leqq s \leqq 2$

We shall now consider Curves 1 and 2; see Fig. 3. Upon use of

$$\frac{df_2}{df_1} = \frac{df_2/ds}{df_1/ds} \quad \text{and} \quad \frac{d^2f_2}{df_1^2} = \frac{d}{ds}\frac{(df_2)}{df_1}\frac{ds}{df_1} \tag{7}$$

it is readily established that

$$\frac{df_2}{df_1} > 0, \quad \frac{d^2f_2}{df_1^2} < 0 \tag{8}$$

for all points on Curves 1 and 2. Note also that Curve 1 joins Curve 2 smoothly at $(f_1, f_2) = (2, 0)$ corresponding to $s = 1$.

The domination cone for our problem is $\Lambda_1 = \{(d_1, d_2) \mid d_1 \geqq 0, d_2 \leqq 0\}$. It follows directly from Corollary 4.3 of Ref. [25] that a point $(f_1^0, f_2^0) \in \in \text{Ext}\,[Y \mid \Lambda_1]$ (i.e., is noninferior) if and only if f_1^0 is the unique minimum of f_1 for all $(f_1, f_2) \in Y$ and $f_2 \geqq f_2^0$.

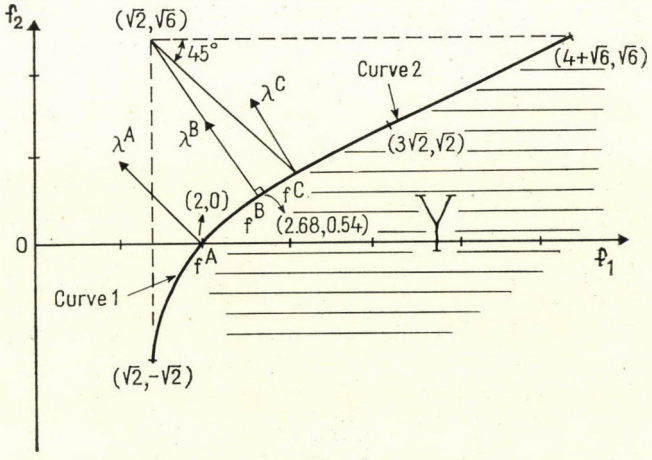


Fig. 3

Since $u^*(\cdot)$ renders the unique minimum of $f_1 = T$ for a given $f_2 = z_2(T) \in [-\sqrt{2}, \sqrt{6}]$, and $df_1/ds > 0$ on Curves 1 and 2, the points on these curves, and only these points, are noninferior among points (f_1, f_2) with $f_2 \in [-\sqrt{2}, \sqrt{6}]$.

On the other hand consider $z_2(T) \in [-\sqrt{6}, -\sqrt{2})$. Among all (f_1, f_2) with $f_2 \geqq z_2(T)$, the unique minimum of f_1 is rendered by control $u^*(t) = -1$. Thus, point $(\sqrt{2}, -\sqrt{2})$ dominates all $(f_1, f_2) \in Y$ with $f_2 < -\sqrt{2}$. Consequently the points on Curves 1 and 2 are the only noninferior points for all $(f_1, f_2) \in Y$.

Finally we observe that, since $\dfrac{d^2 f_2}{df_1^2} < 0$ along Curves 1 and 2, f_2 is a strictly concave function of f_1. Since these curves represent $\mathrm{Ext}[Y \mid \Lambda_1]$, Y is Λ_1-convex. In order to see this point, we shall show that $Y + \Lambda_1$ is a convex set.

Let $y^1, y^2 \in Y + \Lambda_1$. Given α, $0 < \alpha < 1$, we show that $\alpha y^1 + (1-\alpha)y^2 \in \in Y + \Lambda_1$. By hypothesis we can write $y^1 = f^1 + h_1$, $y^2 = f^2 + h_2$ where $f^1, f^2 \in \mathrm{Ext}[Y \mid \Lambda_1]$ and $h_1, h_2 \in \Lambda_1$. Then

$$\alpha y^1 + (1 - \alpha)y^2 = \alpha f^1 + (1 - \alpha)f^2 + \alpha h_1 + (1 - \alpha)\, h_2.$$

Since $\alpha h_1 + (1 - \alpha)\, h_2 \in \Lambda_1$, it suffices to show that $\alpha f^1 + (1 - \alpha)\, f^2 \in Y + \Lambda_1$. Let $f^1 = (f_1^1, f_2(f_1^1))$ and $f^2 = (f_1^2, f_2(f_1^2))$ where $f_2(f_1)$ is specified as in Table 1. Then

$$\begin{aligned}
\alpha f^1 + (1 - \alpha)f^2 &= \big(\alpha f_1^1 + (1 - \alpha)\, f_1^2,\ \alpha f_2(f_1^1) + (1 - \alpha)f_2(f_1^2)\big) \\
&= \big(\alpha f_1^1 + (1 - \alpha)f_1^2,\ f_2(\alpha f_1^1 + (1 - \alpha)f_1^2) - \beta\big) \text{ (where } \beta \geqq 0) \\
&= \big(\alpha f_1^1 + (1 - \alpha)f_1^2,\ f_2(\alpha f_1^1 + (1 - \alpha)f_1^2)\big) + (0, -\beta) \in Y + \Lambda_1
\end{aligned}$$

where the second equality is due to the fact that f_2 is a concave function of f_1.

Q.E.D.

Note that the Λ_1-convexity of Y is very desirable because all the properties listed in Section 2 are applicable. It also allows us to visualize the shape of Y as depicted in Fig. 3.

Referring to Fig. 3, let λ^A, λ^B and λ^C be the normal vectors to ∂Y at the point f^A, f^B and f^C. Observe that at point f^A, the slope of the curve which represents $\mathrm{Ext}\,[Y \mid \Lambda_1]$ is one. Thus, f^A is a compromise solution with $p = 1$. At f^B, the line from the utopia point to f^B is orthogonal to $\mathrm{Ext}[Y \mid \Lambda_1]$. Thus, f^B is a compromise solution with $p = 2$. At f^C, both criteria suffer an equal regret, because $f_1^C - \sqrt{2} = \sqrt{6} - f_2^C$. In fact f^C is the compromise solution with $p = \infty$.

The following may be worth noting:

(i) The solution suggested by Salukvadze is the point f^B which is the compromise solution with parameter $p = 2$ (see Remark 2.2).

(ii) All compromise solutions as functions of p vary continuously and monotonically from f^A to f^C. There is no special reason to pick f^B (Properties (viii) and (ix) of Section 2).

(iii) The compromise solutions have the property that, when p is increased, the sum of the "resulting utility" (i.e., $(-f_1) + f_2$) decreases and the maximum of the individual regrets (i.e., $f_1 - \sqrt{2}$ and $\sqrt{6} - f_2$) decreases, too. How to select a p so that the group utility (the sum of the individual utility) and the individual regret are best balanced remains to be answered (Property (x)).

(iv) Point f^B also corresponds to the unique maximum of $\lambda^B \cdot f = \lambda_1^B f_1 + \lambda_2^B f_2$. For any weight vector $\lambda \neq \lambda^B$, f^B cannot correspond to the maximum of $\lambda \cdot f$ because f_2 is a strictly concave function of f_1 on Ext $[Y \mid \Lambda_1]$. By selecting $p = 2$, we determine implicitly the weight vector λ^B. Why should one not use another λ? Is the resulting solution desirable?

(v) Let $\Lambda_2 = \{d \in R^2 \mid \lambda^A \cdot d \leqq 0,\ \lambda^C \cdot d \leqq 0\}$ (see Fig. 4). One can show that Ext $[Y \mid \Lambda_2] =$ Curve $[f^A, f^C]$. Thus by restricting ourselves to compromise solutions, we assume implicitly that each domination cone $D(y)$, $y \in Y$, contains Λ_2. Is this assumption too strong?

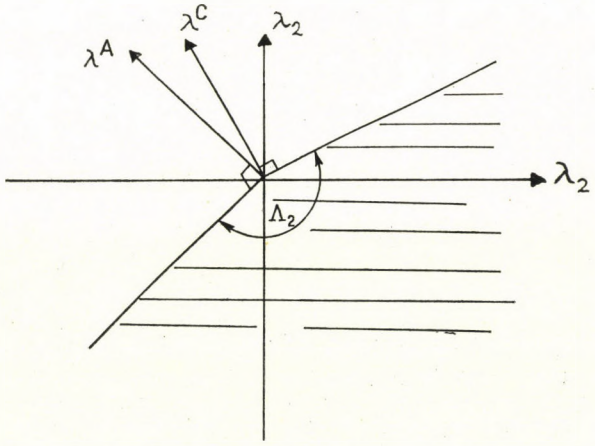


Fig. 4

References

1. *Von Neumann, J., Morgenstern, O.*, Theory of Games and Economic Behavior, Princeton University Press, 1947.
2. *Isaacs, R.*, Differential Games, John Wiley, 1965.
3. *Isaacs, R.*, "Differential Games: Their Scope, Nature and Future," Journal of Optimization Theory and Applications, **3**, 5 (May, 1969), 283–295.
4. *Blaquiere, A., Gerard, F., Leitmann, G.*, Quantitative and Qualitative Games, Academic Press, 1969.
5. *Ho, Y. C.*, Final Report of the First International Conference on the Theory and Applications of Differential Games, Amherst, Massachusetts, January, 1970.
6. *Owen, G.*, Game Theory, Saunders Company, 1968.
7. *Rapoport, A.*, N-Person Game Theory — Concepts and Applications, The University of Michigan Press, 1970.
8. *Arrow, K. J.*, Social Choice and Individual Values, Cowles Commission Monograph 12, 1951.
9. *Luce, R. D., Raiffa, H.*, Games and Decisions, John Wiley and Sons, Seventh Printing, 1967.
10. *Charnes, A., Cooper, W. W.*, Management Models and Industrial Application of Linear Programming, John Wiley and Sons, Vols. I and II, 1961.
11. *Friedman, A.*, Differential Games, Wiley-Interscience, 1971.
12. *Shapley, L. S.*, "A Value for N-Person Games", Contributions to the Theory of Games, II, Annals of Math. Studies 28, Princeton Press, 1953, pp. 307–317. Edited by H. W. Kuhn and A. W. Tucker.
13. *Blackwell, D., Girschick, M. A.*, Theory of Games and Statistical Decisions, John Wiley and Sons, 1954.
14. *Ferguson, T. S.*, Mathematical Statistics, A Decision Theoretic Approach, Academic Press, 1967.
15. *Fishburn, P. C.*, Utility Theory for Decision Making, John Wiley and Sons, 1970.
16. *Freimer, M., Yu, P. L.*, An Approach Toward Decison Problems with Multiobjectives, Center for System Science 72–03, University of Rochester, Rochester, New York, 1972.
17. *Geoffrion, A. M.*, "Proper Efficiency and the Theory of Vector Maximization," Journal of Math. Analysis and Applications, **22** (1968), 618–630.
18. *MacGrimmon, K. R.*, Decision Making Among Multiple-Attribute Alternatives: A Survey and Consolidated Approach, Memorandum RM — 4823 — ARPA, December, 1963, The Rand Corporation, Santa Monica, California.
19. *Raiffa, H.*, Decision Analysis, Addison–Wesley, 1968.
20. *Raiffa, H.*, Preferences for Multi-Attributed Alternatives, Memorandum RM — 5868 — POT/RC, April 1969, The Rand Corporation, Santa Monica, California.
21. *Vincent, T. L., Leitmann, G.*, "Control-Space Properties of Cooperative Games," Journal of Optimization Theory and Application, **6**, 2 (1970), 91–113.
22. *Leitmann, G., Rocklin, S., Vincent, T. L.*, "A Note on Control Space Properties of Cooperative Games," Journal of Optimization Theory and Application, **9**, 6 (1972) 379–390.
23. *Stalford, H. L.*, "Criteria for Pareto Optimality in Cooperative Differential Games", Journal of Optimization Theory and Application, **9**, 6 (1972) 391–398.
24. *Yu, P. L.*, "A Class of Solutions for Group Decision Problems", Management Science, **19**, 8 (1973), 936–946.
25. *Yu, P. L.*, Cone Convexity, Cone Extreme Points and Nondominated Solutions in Decision Problems with Multiobjectives, Center for System Science, 72–02, University of Rochester, Rochester, New York, 1972 (To appear in Journal of Optimization Theory and Application).
26. *Zeleny, M.*, Linear Multiobjective Programming, Ph. D. dissertation, The Graduate School of Management, The University of Rochester, Rochester, New York, 1972.
27. *Yu, P. L.*, Introduction to Domination Structures in Multicriteria Decision Problems, System Analysis Program, F 7219, Graduate School of Management, University of Rochester, 1972 (To appear in the Selected Proceedings of the Seminar on Multicriteria Decision Making, edited by J. L. Cochrane and M. Zeleny).

28. *Freimer, M., Yu, P. L.*, "The Application of Compromise Solutions to Reporting Games," System Analysis Program Series No. F 7220, Graduate School of Management, University of Rochester, Rochester, New York (To appear in a special issue of Theory and Decision, an international journal for philosophy and methodology of the social sciences).
29. *Yu, P. L., Zeleny*, M., The Set of All Nondominated Solutions in the Linear Cases and A Multicriteria Simplex Method, Center for System Science, 73—03, University of Rochester, Rochester, New York.
30. *Zadeh, L. A.*, Optimality and Non-Scalar-Valued Performance Criteria, IEEE Transactions on Automatic Control, pp. 59—60, January 1963.
31. *Da Cuncha, N. O., Polak, E.*, "Constrained Minimization Under Vector-Valued Criteria in Finite Dimensional Space", Journal of Math. Analysis and Applications, **19** (1967),103—124.
32. *Salukvadze, M. E.*, "Optimization of Vector Functionals. I. Programming of Optimal Trajectories", Avtomatika i Telemekhanika, **8** (August 1971) 5—15 (translation pp. 1169—1178).
33. *Salukvadze, M. E.*, "Optimization of Vector Functionals. II. The Analytic Construction of Optimal Control," Avtomatika i Telemekhanika, **9** (September 1971) 5—15 (translation pp. 1347—1357).
34. *Huang, S. C.*, "Note on the Mean-Square Strategy of Vector Values Objective Function," Journal of Optimization Theory and Applications, **9**, 5 (1972).
35. *Leitmann, G.*, An Introduction to Optimal Control, McGraw-Hill Book Company, New York, 1966.
36. *Freimer, M., Yu, P. L.*, "Some New Results on Compromise Solutions", (forthcoming).
37. *Yu, P. L.*, Nondominated Investment Policies in Stock Markets (Including an Empirical Study), Systems Analysis Program F 7222, Graduate School of Management. University of Rochester, 1973.
38. *Shubik, M.*, Readings in Game Theory and Political Behavior, Doubleday, 1954,

Компромиссные решения, доминирующие структуры и решение Салуквадзе

П. Л. ЙУ – Г. ЛЕЙТМАНН
(США)

Рассматривается обобщение основных понятий по теории решений, предложенных Салуквадзе, применительно к статическим и динамическим задачам. Особое внимание уделяется компромиссным решениям и связанным с ними доминирующим структурам. Приводится пример.

P. L. Yu
Graduate School of Management
University of Rochester
Rochester, New York, USA

and

G. Leitmann
University of California at Berkeley
Berkeley, California 94720, USA

Problems of Control and Information Theory, Vol. 2 (3–4), pp. 199–217 (1973)

MEAN SQUARE STABILITY CRITERIA FOR LINEAR WHITE NOISE STOCHASTIC SYSTEMS

J. L. WILLEMS

(Gent)

(Received December 10, 1972)

In this paper Lyapunov theory is used to analyse the mean square stability properties of linear systems with white noise stochastic coefficients in the state equations. Discrete time as well as continuous time systems with a finite dimensional state vector are considered. First the single noise case is discussed, then the case of multiple but uncorrelated noise elements is analysed, and finally the general case of multiple correlated noise coefficients is dealt with.

Introduction

An exact model of many dynamical systems includes stochastic elements, not only in an additive way (noisy inputs), but also in a multiplicative way (random coefficients). Whereas the analysis of deterministic systems with random input perturbations is rather straightforward, and is not much different from deterministic theory, the same is not true for systems with stochastic coefficients; the random coefficients are for example due to unknown or unpredictable variations in actuator gains, to the presence of a human operator in a control loop, or to the quantization and round-off error in digital control systems. Stochastic load variations in structural systems also lead to differential equations with stochastic coefficients [1]; the same is true for the equations describing the motion of a satellite if random fluctuations of the geomagnetic field [2] or the atmospheric density [3] are included in the mathematical model.

Consider a finite dimensional linear time-invariant discrete time system whose coefficients depend (linearly) on a white noise stochastic gain; the system equation is

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + f(k)\ \mathbf{D}\mathbf{x}(k) \tag{1}$$

where $\mathbf{x}(k)$ is the state of the system and $f(k)$ is a scalar stationary white noise process whose mean can be assumed zero without loss of generality:

$$E(f(k)) = 0,\ E(f(k)f(m)) = \sigma^2 \delta_{km} \tag{2}$$

where E denotes the expectation value and δ_{km} is the Kronecker delta. Let n be the number of components of $\mathbf{x}(k)$. The corresponding continuous time sys-

tem with a single white noise gain is described by the stochastic Ito equation:

$$d\mathbf{x} = \mathbf{A}\mathbf{x}\,dt + \mathbf{B}\mathbf{x}\,dw \tag{3}$$

where $\mathbf{x}(t)$ is the system state and w is a Wiener process with independent increments:

$$E(dw) = 0,\ E(dw^2) = \sigma^2 dt\,. \tag{4}$$

If the Ito interpretation of white noise is used, then the matrix $\mathbf{A}$ does not depend on the noise variance; the same model (3) can be used for the Stratonovitch interpretation of white noise, but in that case the matrix $\mathbf{A}$ depends on σ^2.

This paper deals with stability analysis of the stochastic systems (1) and (3); the number of possible stability definitions is very large for stochastic systems [4]. This paper is only concerned with mean square stability properties.

Definition 1. The linear stochastic system (1) or (3) is mean square stable if $\mathbf{Q}(k) = E(\mathbf{x}(k)\,\mathbf{x}(k)')$ or $\mathbf{Q}(t) = E(\mathbf{x}(t)\,\mathbf{x}(t)')$ is bounded for positive k or t whatever the initial state be. It is asymptotically mean square stable if moreover the second mean matrix $\mathbf{Q}(k)$ or $\mathbf{Q}(t)$ tends to zero as k or t tends to infinity.

The mean square asymptotic stability implies that all motions converge to zero with probability one, but it is a stronger form of asymptotic stability. Moreover, since

$$\mathbf{Q}(k) = E(\mathbf{m}(k)\mathbf{m}(k)') + E\big((\mathbf{x}(k) - \mathbf{m}(k))\ (\mathbf{x}(k) - \mathbf{m}(k))'\big)$$

where $\mathbf{m}(k) = E(\mathbf{x}(k))$, mean square (asymptotic) stability implies the corresponding property for the first mean vector $\mathbf{m}(k)$ as well as for the covariance matrix. There are several ways of analysing the stability properties in the mean square of a stochastic system:

(i) In some cases, such as the linear systems considered here, the difference or differential equation governing the evolution of the second mean matrix can be derived; the stability properties can then be studied by means of methods available for deterministic equations [5], in particular by means of Lyapunov's direct method.

(ii) The direct method of Lyapunov has been extended to be applied directly to stochastic system equations [6]. Its application is in particular very interesting for white noise elements.

Both techniques are used in this paper. For system (1) the first mean vector and the second mean matrix evolutions are governed by the difference equations:

$$\mathbf{m}(k+1) = \mathbf{A}\mathbf{m}(k) \tag{5}$$

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\,\mathbf{A}' + \sigma^2\,\mathbf{D}\mathbf{Q}(k)\,\mathbf{D}' \tag{6}$$

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

since $\mathbf{x}(k)$ is uncorrelated with $f(k-1), f(k-2), \ldots$. The stability properties of the mean $\mathbf{m}(k)$ are hence equivalent to those of the "average" system, or the system without noise; they are necessary conditions for the corresponding stability properties of the second mean matrix $\mathbf{Q}(k)$. Since the matrix $\mathbf{Q}(k)$ is symmetric, it can be represented by an $n(n+1)/2$-dimensional vector $\mathbf{q}(k)$, and (6) is equivalent to a vector difference equation

$$\mathbf{q}(k+1) = \mathbf{M}\mathbf{q}(k) \tag{7}$$

and the stability properties follow from criteria for deterministic discrete time systems: a necessary and sufficient condition for asymptotic stability of (6) is that all eigenvalues of $\mathbf{M}$ are less than 1 in modulus; a necessary and sufficient condition for stability of (6) is that $\mathbf{M}$ has no eigenvalues larger than 1 in modulus and that the eigenvalues of $\mathbf{M}$ with modulus 1 are simple zeros of the minimal polynomial of $\mathbf{M}$.

This criterion requires the application of a stability criterion for a linear system of order $n(n+1)/2$; this criterion can, however, be simplified considerably if the properties presented in the Appendix are used. The mapping $\mathbf{Q} \rightarrow \mathbf{L}(\mathbf{Q})$ defined by

$$\mathbf{L}(\mathbf{Q}) = \mathbf{A}\mathbf{Q}\mathbf{A}' + \sigma^2 \mathbf{D}\mathbf{Q}\mathbf{D}'$$

always has a positive real eigenvalue which is at least as large in modulus as all the other eigenvalues. Moreover, this real eigenvalue increases with σ^2. Hence [7]:

Criterion 1. The system (1) is mean square asymptotically stable if and only if all eigenvalues of $\mathbf{A}$ are smaller than 1 in modulus, and if $0 \leq \sigma^2 < \sigma_0^2$, where σ_0^2 is the smallest value of σ^2 such that the mapping $\mathbf{L}$ has an eigenvalue equal to one, hence such that $\det(\mathbf{M} - \mathbf{I})$ vanishes, where $\mathbf{I}$ denotes the identity matrix. It is mean square unstable if $\sigma_0^2 < \sigma^2$, or if $\mathbf{A}$ has an eigenvalue with modulus larger than 1. If σ^2 is equal to σ_0^2, then the system is either mean square stable but not asymptotically stable, or mean square unstable.

Although this criterion already simplifies considerably the stability analysis, it still requires the evaluation of a determinant of a matrix of order $n(n+1)/2$; this matrix is of much higher order than the system matrices $\mathbf{A}$ and $\mathbf{D}$. It is shown in this paper that the criterion can still be simplified in many interesting cases. First, however, it is pointed out how the same result can be obtained by means of Lyapunov theory. Therefore we either use $\mathrm{tr}(\mathbf{Q}(k)\mathbf{P})$, with $\mathbf{P}$ positive definite, as a Lyapunov function for the deterministic difference equation (6), or the quadratic form $\mathbf{x}'\mathbf{P}\mathbf{x}$ as a stochastic Lyapunov function for the stochastic system (1). Note that the function $\mathrm{tr}(\mathbf{Q}(k)\mathbf{P})$ is positive definite, although it is linear, because $\mathbf{Q}(k)$ is at least positive semidefinite.

In the latter case the sign of the stochastic difference

$$W(\mathbf{x}(k)) = E\big(V(\mathbf{x}(k+1)/\mathbf{x}(k)) - V(\mathbf{x}(k))\big)$$

should be considered, where $V(\mathbf{x}(k))$ denotes the Lyapunov function. In the former case the usual deterministic difference along the solutions of (6) is to be considered. In both cases the negative (semi-) definiteness of the difference is equivalent to the positive (semi-) definiteness of the matrix $\mathbf{P} - \mathbf{A}'\mathbf{P}\mathbf{A} - \sigma^2\mathbf{D}'\mathbf{P}\mathbf{D}$. Considering the equation

$$\mathbf{P} - \mathbf{A}'\mathbf{P}\mathbf{A} - \sigma^2\,\mathbf{D}'\mathbf{P}\mathbf{D} = \mathbf{R} \tag{8}$$

for a given positive definite $\mathbf{R}$ yields the asymptotic stability part of Criterion 1. For $\sigma^2 = \sigma_0^2$, there is a solution $\mathbf{P}$ of

$$\mathbf{P} - \mathbf{A}'\mathbf{P}\mathbf{A} - \sigma^2\,\mathbf{D}'\mathbf{P}\mathbf{D} = 0 \tag{9}$$

such that $\mathrm{tr}\,(\mathbf{P}\mathbf{Q}(k))$ is constant along any solution of (6), which proves either stability but no asymptotic stability, or instability as in Criterion 1. The instability for $\sigma^2 > \sigma_0^2$ is also easily derived.

Similarly, stability results are obtained for multiple noise systems, described by

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + f_1(k)\,\mathbf{D}_1\mathbf{x}(k) + \ldots + f_m(k)\,\mathbf{D}_m\mathbf{x}(k) \tag{10}$$

with uncorrelated white noise coefficients

$$E(f_i(k)) = 0,\ E(f_i(k)f_j(m)) = \sigma_i^2\,\delta_{km}\,\delta_{ij} \tag{11}$$

or with correlated noise elements

$$E(f_i(k)) = 0,\ E(f_i(k)\,f_j(m)) = r_{ij}\,\delta_{km} \tag{12}$$

where the matrix $\mathbf{R}$ of the elements r_{ij} is positive definite and symmetric. The stability boundary in the r_{ij}-space is obtained by expressing that the linear mapping $\mathbf{Q} \to \mathbf{L}(\mathbf{Q})$:

$$\mathbf{L}(\mathbf{Q}) = \mathbf{A}\mathbf{Q}\mathbf{A}' + \sum_{i,j} r_{ij}\mathbf{D}_i\mathbf{Q}\mathbf{D}_j'$$

has an eigenvalue equal to 1. Note that the difference equation of the second mean matrix is

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\mathbf{A}' + \sum_{i,j} r_{ij}\mathbf{D}_i\mathbf{Q}(k)\mathbf{D}_j' \tag{13}$$

and that the correlated noise case can always be reduced to the uncorrelated noise case by means of the transformation

$$\mathbf{F} = \sum s_{ij}\,\mathbf{D}_i \tag{15}$$

where the matrix $\mathbf{S}$ of the elements s_{ij} is a symmetric positive semidefinite matrix such that $\mathbf{S}^2 = \mathbf{R}$.

Similar results are obtained for continuous time systems (1) and the system with multiple noise elements

$$\mathrm{d}\mathbf{x} = \mathbf{A}\mathbf{x}\,\mathrm{d}t + \sum_i \mathbf{B}_i\,\mathbf{x}\mathrm{d}w_i \tag{15}$$

where the processes w_i are zero mean Wiener processes:

$$E(\mathrm{d}w_i) = 0,\ E(\mathrm{d}w_i(t)\,\mathrm{d}w_j(t)) = r_{ij}\,\mathrm{d}t\,. \tag{16}$$

By means of the properties given in the Appendix, and the differential equation of the second mean matrix

$$\mathrm{d}\mathbf{Q}(t)/\mathrm{d}t = \mathbf{A}\mathbf{Q}(t) + \mathbf{Q}(t)\mathbf{A}' + \sum_{i,j} r_{ij}\mathbf{B}_i\,\mathbf{Q}(t)\,\mathbf{B}'_j \tag{17}$$

we obtain the boundary of the asymptotic stability region in the r_{ij}-space, of the eigenvalues of $\mathbf{A}$ have negative real parts, by expressing that the linear mapping $\mathbf{Q} \to \mathbf{L}(\mathbf{Q})$ defined as

$$\mathbf{L}(\mathbf{Q}) = \mathbf{A}\mathbf{Q} + \mathbf{Q}\mathbf{A}' + \sum_{i,j} r_{ij}\mathbf{B}_i\,\mathbf{Q}\,\mathbf{B}'_j$$

has a zero eigenvalue. The direct application of deterministic theory to (17) would require to compute when the linear mapping $\mathbf{L}$ has an eigenvalue with zero real part.

2. Single-variable noisy feedback

In this section the particular case of (1) is considered, where the matrix $\mathbf{D}$ has rank 1; it can hence be written as a dyad $\mathbf{D} = \mathbf{b}\mathbf{c}$ with $\mathbf{b}$ a column vector, and $\mathbf{c}$ a row vector. The difference equation

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + f(k)\,\mathbf{b}\mathbf{c}\mathbf{x}(k) \tag{18}$$

can be seen as the description of a feedback system with a linear time-invariant element with transfer function $\mathbf{c}(\mathrm{I}z - \mathbf{A})^{-1}\mathbf{b} = \mathbf{G}(z)$ in the forward path and a feedback with white noise stochastic gain $f(k)$. Let us suppose first that the triple $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ is a minimal realization (that is, observable and controllable) of the transfer function $\mathbf{G}(z)$. The difference equation for the second mean matrix is

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\,\mathbf{A}' + \sigma^2\mathbf{b}\mathbf{c}\,\mathbf{Q}(k)\,\mathbf{c}'\mathbf{b}'. \tag{19}$$

Using the Lyapunov function $V(\mathbf{Q}) = \mathrm{tr}\ (\mathbf{P}\mathbf{Q})$, we obtain the difference

$$V(\mathbf{Q}(k+1)) - V(\mathbf{Q}(k)) = \mathrm{tr}\,(\mathbf{Q}(k)\,\mathbf{M}) \tag{20}$$

with

$$\mathbf{M} = \mathbf{A}'\mathbf{P}\mathbf{A} - \mathbf{P} + \sigma^2(\mathbf{b}'\mathbf{P}\mathbf{b})\mathbf{c}'\mathbf{c}. \tag{21}$$

Therefore we use the particular $\mathbf{P}$ which satisfies

$$\mathbf{A}'\mathbf{P}\mathbf{A} - \mathbf{P} = -\mathbf{c}'\mathbf{c}. \tag{22}$$

This equation has a positive semidefinite solution $\mathbf{P}$ if the eigenvalues are smaller than 1 in modulus; $\mathbf{P}$ is positive definite, since it can explicitly be expressed as

$$\mathbf{P} = \sum_{n=0}^{\infty} (\mathbf{A}')^k \mathbf{c}'\mathbf{c}\mathbf{A}^k.$$

This is positive definite if the system $\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k)$, $\mathrm{y} = \mathbf{c}\mathbf{x}$, is observable. Hence

Criterion 2. The system (18) is mean square asymptotically stable iff the eigenvalues of $\mathbf{A}$ are smaller than one in modulus and $\sigma^2\mathbf{b}'\mathbf{P}\mathbf{b} < 1$, where $\mathbf{P}$ is the solution of (22); it is mean square stable but not asymptotically stable iff the eigenvalues of $\mathbf{A}$ are smaller than 1 in modulus, and $\sigma^2\mathbf{b}'\mathbf{P}\mathbf{b}$ equals 1. It is mean square unstable iff either the eigenvalues of $\mathbf{A}$ are smaller than 1 in modulus and $\sigma^2\mathbf{b}'\mathbf{P}\mathbf{b} > 1$, or the matrix $\mathbf{A}$ has at least one eigenvalue with modulus equal to or larger than one.

The instability part of the criterion is proved by observing that the conditions of instability imply the existence of a positive r larger than one such that the difference equation governing $\mathbf{x}(k)r^{-k}$ is mean square stable but not asymptotically stable.

Remark 1. The condition $\sigma^2\mathbf{b}'\mathbf{P}\mathbf{b} < 1$ can be expressed in several equivalent ways; the following equalities can be proved [8]:

$$\mathbf{b}'\mathbf{P}\mathbf{b} = \sum_{k=0}^{\infty} w(k)^2 = \oint_C |\mathbf{G}(z)|^2 \, \mathrm{d}z/2\pi jz = h_n$$

where $w(k)$ is the inverse Z-transform of the function $\mathbf{G}(z)$, or the impulse response of the forward path element, C denotes the unit circle and the integral is taken in the anticlockwise direction, and h_n is the coefficient of z^n in the polynomial $h(z)$ satisfying

$$h(z)\,p(z^{-1}) + h(z^{-1})\,p(z) = 2q(z)\,q(z^{-1})$$

with $\mathbf{G}(z) = q(z)/p(z)$. The computation of h_n requires the evaluation of the determinant of two square matrices of order n. The criterion is hence much simpler than would follow from a straightforward application of Criterion 1.

Remark 2. If system (18) is non-observable [9], then the statement on asymptotic stability of criterion 2 remains unaltered; this can be shown by using the Lyapunov function $V(\mathbf{Q}) = \mathrm{tr}\,(\mathbf{PQ} + u\mathbf{MQ})$, where $\mathbf{P}$ is the solution of (22), $\mathbf{M}$ is an arbitrary positive definite matrix, and u is a sufficiently small positive constant. The controllability does not have to be assumed. If the system is not observable, then the system can either be unstable or weakly stable if some unobservable modes have an eigenvalue with modulus one; the same happens when some uncontrollable modes have an eigenvalue with modulus one, and the other eigenvalues are smaller than 1. It is unstable if one or more eigenvalues are larger than one in modulus, and also if $\sum_{k=0}^{\infty} w(k)^2$ either diverges or is larger than $1/\sigma^2$. It is possible to make the distinction between the weak stability and the instability behavior in the critical cases pointed out above by means of techniques similar to the criteria used in deterministic theory, or by checking if the solution $\mathbf{P}$ of (22) is positive definite or only positive semi-definite. Since the uncontrollability and unobservability of the realization does not add essential difficulties, minimality is assumed in the sequel.

Remark 3. Suppose the realization is minimal. Then in the case of asymptotic mean square stability, the second mean matrix decreases at least as fast as r^{2k}, where r is a positive number, smaller than 1 but larger than the modulus of any eigenvalue of $\mathbf{A}$ and is such that the matrix $\mathbf{P}$ satisfying $\mathbf{A}'\mathbf{PA} - r^2\mathbf{P} = -\mathbf{c}'\mathbf{c}$ satisfies $\sigma^2\mathbf{b}'\mathbf{Pb} = 1$; an equivalent requirement is that r is such that

$$\sigma^2 \sum_{n=0}^{\infty} \left(w(k)/r^k\right)^2 = 1\,. \tag{23}$$

In the case of mean square stability the second mean matrix does not increase faster than r^{2k}, where r is a positive number larger than 1 and larger than the modulus of any eigenvalue of $\mathbf{A}$ such that (23) holds, or such that the solution $\mathbf{P}$ of $\mathbf{A}'\mathbf{PA} - r^2\mathbf{P} = -\mathbf{c}'\mathbf{c}$ satisfies $\sigma^2\mathbf{b}'\mathbf{Pb} = 1$.

Remark 4. The stability result can also be obtained as follows. The difference equation (19) can be seen as the description of the feedback system [10]

$$\mathbf{Q}(k+1) = \mathbf{AQ}(k)\mathbf{A}' + \mathbf{bub}',\ y(k) = \mathbf{cQ}(k)\mathbf{c}',\ u(k) = \sigma^2 y(k) \tag{24}$$

which is a feedback system with a linear system with impulse response

$$\mathbf{cA}^{k-1}\mathbf{bb}'(\mathbf{A}')^{k-1}\mathbf{c}' \quad (\text{for } k \geq 1)$$

and a positive feedback where the feedback element is an amplifier with gain σ^2. Since the impulse response is always positive, its Z-transform assumes its largest value on the unit circle for $Z = 1$; therefore the Nyquist criterion for

discrete time systems [5] proves asymptotic stability if that Z-transform is smaller than $1/\sigma^2$ at $Z = 1$, and if the open loop is stable; this corresponds to Criterion 2. The instability part of the theorem also follows from the Nyquist criterion.

Similar results can be obtained for the analogous continuous time system

$$\mathrm{d}\mathbf{x} = \mathbf{A}\mathbf{x}\,\mathrm{d}t + \mathbf{b}\mathbf{c}\mathbf{x}\,\mathrm{d}w. \tag{25}$$

For discussion of the criteria the reader is referred to the literature, where it is shown that the mean square asymptotic stability can be checked by applying the Routh Hurwitz criterion to the average system and by means of an additional inequality involving the largest Hurwitz determinant and a similar determinant [11—14].

Remark 5. Criterion 2 could also be obtained by means of the circle criterion for nonlinear nonstationary feedback systems. Considering the feedback system introduced in Remark 4, the circle criterion [5] proves the global asymptotic stability in the mean square for all $\sigma^2 < \sigma_0^2$ such that the Z-transform $\mathbf{G}^*(z)$ of the impulse response of the forward path element satisfies the inequality

$$\mathrm{Re}\,(\mathbf{G}^*(z) - 1/\sigma_0^2) \geq 0$$

on the unit circle; since $\mathbf{G}^*(z)$ is the Z-transform of a nonnegative function of k, the real part and the modulus are largest for $z = 1$. The circle criterion hence requires

$$\sigma^2 \sum_{k=0}^{\infty} w(k)^2 < 1$$

as in criterion 2. However the circle criterion is also valid for nonstationary and nonlinear feedback; this proves that the stochastic system [8]

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + g(\mathbf{x}(k), k)\,\mathbf{b}\mathbf{c}\mathbf{x}\,(k)\,f(k)$$

with $f(k)$ as above, is mean square asymptotically stable in the large if the average system is asymptotically stable, and if

$$\sigma^2(\sup g(\mathbf{x}(k), k))^2 \sum_{k=0}^{\infty} (w(k))^2 < 1.$$

3. Multiple noise elements and D-matrices of rank one

In this section the mean square stability properties of the stochastic systems described by the difference equation

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \sum_{=1}^{m} \mathbf{b}_i\mathbf{c}_i\,\mathbf{x}(k)\,f_i(k) \tag{26}$$

is considered, which corresponds to (10) with matrices $\mathbf{D}_i$ of rank 1; it is assumed that the noise elements are uncorrelated and zero mean processes:

$$E(f_i(k)\, f_j(h)) = \sigma_i^2\, \delta_{ij}\, \delta_{hk} \tag{27}$$

and that and that we have a minimal realization of the matrix of the transfer functions

$$G_{ij}(z) = \mathbf{c}_j(\mathbf{I}\, z - \mathbf{A})^{-1}\, \mathbf{b}_i\,.$$

Inspired by the results of the previous section, the Lyapunov function $V(\mathbf{Q}(k)) = \mathrm{tr}\,(\mathbf{Q}(k)\,\mathbf{P})$ is used, where $\mathbf{P} = \sum_i a_i \mathbf{P}_i$, with $\mathbf{P}_i$ being the solution of

$$\mathbf{A}'\,\mathbf{P}_i\,\mathbf{A} - \mathbf{P}_i = -\mathbf{c}_i'\,\mathbf{c}_i$$

which is positive definite if the eigenvalues of $\mathbf{A}$ are smaller than 1 in modulus, and the numbers a_i are nonnegative constants, of which at least one is different from zero. The difference of the Lyapunov function along the solutions of the equation of the second mean matrix

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\,\mathbf{A}' + \sum_i \sigma_i^2\, \mathbf{b}_i\mathbf{c}_i\, \mathbf{Q}(k)\, \mathbf{c}_i'\mathbf{b}_i' \tag{28}$$

is

$$V(\mathbf{Q}(k+1)) - V(\mathbf{Q}(k)) = -\mathrm{tr}\,(\mathbf{Q}(k)\mathbf{M})$$

with

$$\mathbf{M} = \sum_i a_i \mathbf{c}_i'\mathbf{c}_i - \sum_i \sum_j \mathbf{a}_j \mathbf{b}_i'\, \mathbf{P}_j \mathbf{b}_i \mathbf{c}_i'\mathbf{c}_i\, \sigma_i^2\,.$$

We attempt to select the constants a_i in such a way that

$$\mathbf{M} = \sum_i d_i \mathbf{c}_i'\mathbf{c}_i$$

for nonnegative constants d_i. This is true if the constants d_i satisfy

$$a_i - \sum_j \mathbf{b}_i'\, \mathbf{P}_j \mathbf{b}_i a_j = d_i\,. \tag{29}$$

If $\mathbf{a}$ and $\mathbf{d}$ denote the vectors with elements a_i and d_i respectively, then (29) is equivalent to

$$(\mathbf{I} - \mathbf{M})\,\mathbf{d} = \mathbf{d} \tag{30}$$

where I is the identity matrix of order m, and $\mathbf{M}$ the matrix with elements

$$m_{ij} = \sum_i \sigma_i^2 \mathbf{b}_i'\, \mathbf{P}_j \mathbf{b}_i = \sum_{k=0}^{\infty} \sigma_i^2\, w_{ij}(k)^2$$

where $w_{ij}(k)$ is the inverse Z-transform of the transfer function $G_{ij}(z)$. Since $\mathbf{M}$ is a matrix with positive entries, the following stability result is obtained by means of the properties of the Appendix:

Criterion 3. System (26) is mean square asymptotically stable if the eigenvalues of **A** are smaller than 1 in modulus, and if the largest eigenvalue of **M**, which is necessarily a positive real eigenvalue, is less than one. It is weakly stable, but not asymptotically stable, if the eigenvalues of **A** are smaller than 1 in modulus, and the largest eigenvalue of **A** equals 1. It is mean square unstable if either **A** has an eigenvalue which is larger than 1, or the eigenvalues of **A** are smaller than 1 but the largest eigenvalue of **M** is larger than one.

Criterion 3 hence requires a stability check of the average system and the solution of the equation

$$\det(\mathbf{M} - \mathbf{I}) = 0$$

which is much smaller than the determinental equation which would follow from a straightforward application of Criterion 1, since **M** is a square matrix of order m (the number of independent noise elements) which is usually much smaller than $n(n+1)/2$. For continuous time systems this criterion has been obtained by Levit and Yacubovich [15].

Remark 6. If $\sigma_{10}^2, \ldots, \sigma_{m0}^2$ is a point on the stability boundary, then system (26) is mean square asymptotically stable if $\sigma_i^2 \leq \sigma_{i0}^2$, and the inequality is not an equality for at least one value of the index. This follows from the property given in the Appendix concerning the eigenvalues of matrices with positive elements.

Remark 7. The result of Criterion 3 could also be obtained by considering the feedback system equivalent of Eq. (19):

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\mathbf{A}' + \sum_i \mathbf{b}_i u_i \mathbf{b}_i'$$

$$y_i = \mathbf{c}_i \mathbf{Q}(k) \mathbf{c}_i'$$

$$u_i = \sigma_i^2 y_i .$$

The impulse response matrix of the forward path element is

$$\mathbf{W} = [w_{ij}(k)^2]$$

and the stability condition $\inf_{|z|\leq 1} |\det(\mathbf{H}(z) - \mathbf{I})| > 0$, where $\mathbf{H}(z)$ is the Z-transform of $\mathbf{W}(k)$, is equivalent to the stability condition of Criterion 3 since the largest eigenvalue of $\mathbf{H}(z)$ is not larger than the largest eigenvalue of $\mathbf{H}(1)$, if $|z| \leq 1$.

Remark 8. Note that the assumption that the white noise stochastic elements $f_i(k)$ are uncorrelated is very essential. Although it is possible to transform correlated white noise elements into uncorrelated elements, this

does not mean that Criterion 3 is also valid for correlated elements. Indeed consider

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \sum_i \mathbf{b}_i \mathbf{c}_i f_i(k)$$

with $E(f_i(k)) = 0$, $E(f_i(k) f_j(q)) = r_{ij}(k)\,\delta_{kq}$. Let $\mathbf{R}$ be the matrix with elements r_{ij}; since $\mathbf{R}$ is positive semidefinite, it is possible to define a symmetric matrix $\mathbf{S}$ such that $\mathbf{R} = \mathbf{S}^2$. Let

$$\mathbf{D}_i = \sum s_{ij} \mathbf{b}_i \mathbf{c}_i.$$

Then the second mean matrix difference equation

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\,\mathbf{A}' + \sum_{i,j} r_{ij} \mathbf{b}_i \mathbf{c}_i\,\mathbf{Q}(k)\,\mathbf{c}_j' \mathbf{b}_j'$$

is equivalent to

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\,\mathbf{A}' + \sum_i \mathbf{D}_i\,\mathbf{Q}(k)\,\mathbf{D}_i' \tag{31}$$

which does not belong in general to the class of difference equations (28), since $\mathbf{D}_i$ is in general not a dyad. The stability analysis of this equation requires the techniques of the next section. Only in a few interesting particular cases does (31) belong to class (28): (i) if all b_i-vectors are the same; (ii) if all c_i-vectors are equal. This is obviously also true if $\mathbf{R}$ is diagonal.

Remark 9. The elements of the matrix $\mathbf{M}$ can be computed by means of the methods pointed out in Remark 1 of Section 2. The elements can be expressed as the ratio of two determinants; the denominator is the same for all elements; only one column of the numerator determinant depends on the element, one actually computes.

Remark 10. In a number of interesting cases Criterion 3 can be considerably simplified, since the square matrix $\mathbf{M}$ of order m does not have to be evaluated. The idea is that instead of $\mathbf{M}$ a lower order matrix can be analysed if one a priori knows that the rank of $\mathbf{M}$ is less than m, such that some zero eigenvalues are known. Suppose that all vectors b_i are equal; then $\mathbf{M}$ is of rank one, and it is sufficient to check whether the only non-zero eigenvalue is smaller, larger, or equal to 1, for proving mean square asymptotic stability, instability or weak stability. We have

$$\mathrm{tr}(\mathbf{M}) \sum_{j=1}^{m} \sum_{k=0}^{\infty} \sigma_j^2\, w_j(k)^2$$

where

$$\mathbf{c}_j(\mathbf{I}z - \mathbf{A})^{-1}\,\mathbf{b} = Z(w_j(k))$$

Let $q_j(z)$ denote the numerator of $\mathbf{G}_j(z)$, $\mathbf{q}(z)$ the column vector with components $q_j(z)$, and $p(z)$ the denominator of all $\mathbf{G}_j(z)$. Let $\mathbf{R}$ be the diagonal matrix with elements σ_i^2. Then $\mathrm{tr}(\mathbf{M})$ is the coefficient of $h(z)$ in the polynomial satisfying

$$p(z)\,q(z^{-1}) + p(z^{-1})\,q(z) = 2\mathbf{q}(z)'\,\mathbf{R}\mathbf{q}(z^{-1}).$$

This is also valid for correlated noise elements, where $\mathbf{R}$ is now the second mean matrix of the white noise elements [8]. Hence in this case only one inequality is to be checked in addition to the stability of the average system.

The same occurs when all vectors $c\mathbf{l}$ are equal to c; then

$$\mathrm{tr}(\mathbf{M}) = \sum_{i=1}^{n} \sum_{k=1}^{\infty} \sigma_i^2\, w_i(k)^2$$

where

$$Z(w_i(k)) = \mathbf{c}(\mathbf{I}z - \mathbf{A})^{-1}\mathbf{b}_i.$$

This can be calculated as in the case of equal $\mathbf{b}$-vectors. This can also be obtained directly by using the tentative Lyapunov function $\mathrm{tr}(\mathbf{PQ}(k))$, with $\mathbf{P}$ the solution of $\mathbf{A}'\mathbf{P}\mathbf{A} - \mathbf{P} = -\mathbf{c}'\mathbf{c}$, since all $\mathbf{P}_i$-matrices used in the proof of Criterion 3 are equal in this case.

If the result is applied to a higher order difference equation with white noise coefficients

$$x(k+n) + (a_{n-1} + f_{n-1}(k))\,x(k+n-1) \ldots + (a_0 + f_0(k))\,x(k) = 0$$

with uncorrelated noises, and $E(f_i(k) f_i(1)) = \sigma_i^2\,\delta_{k1}$, we obtain the necessary and sufficient conditions for mean square asymptotic stability: (i) $p(z) = z^n + s_{n-1}z^{n-1} + \ldots + a_0$ should have all its zeros within the unit circle; (ii) if C denotes the unit circle:

$$\left(\sum_i \sigma_i^2\right) \frac{1}{2\pi j} \oint_C \frac{\mathrm{d}z}{x\,|p(z)|^2} < 1 \tag{32}$$

which only depends on the sum of the variances! The stability in the mean square of this system is hence equivalent to the stability of the system where only the lowest order coefficient is noisy, but with a variance which is equal to the sum of the variances.

Remark 11. There are also other interesting cases where the rank of $\mathbf{M}$ is lower than m, and where the conditions of Criterion 3 (and hence also of the Levit and Yacubovich result for the continuous time case) can be considerably simplified. In general the rank of $\mathbf{M}$ is not higher than the number of independent vectors $\mathbf{b}_i$, and also of the number of independent vectors $\mathbf{c}_i$. Consider

for example the case where there are only two different vectors c_i; then there are only two different matrices $\mathbf{P}_i$ in the Lyapunov function and hence only two coefficients a_i. Using the tentative Lyapunov function $\mathrm{tr}(a_1\mathbf{P}_1\mathbf{Q}(k) + a_2\mathbf{P}_2\mathbf{Q}(k))$, where $\mathbf{P}_1$ and $\mathbf{P}_2$ are defined as above, we obtain as necessary and sufficient conditions for asymptotic stability in the mean square: (i) the eigenvalues of $\mathbf{A}$ must be smaller than 1 in modulus; (ii) the largest eigenvalue of

$$\mathbf{N} = \begin{bmatrix} \sum_{i=1}^{r} \sum_{k=0}^{\infty} \sigma_i^2 w_{i1}(k)^2 & \sum_{i=1}^{r} \sum_{k=0}^{\infty} \sigma_i^2 w_{i2}(k)^2 \\ \sum_{i=r+1}^{m} \sum_{k=0}^{\infty} \sigma_i^2 w_{i1}(k)^2 & \sum_{i=r+1}^{m} \sum_{k=0}^{\infty} \sigma_i^2 w_{i2}(k)^2 \end{bmatrix}$$

which is positive and real, should be smaller than 1. In this expression the following symbols are used: the system equation is

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{r} \mathbf{b}_i \mathbf{c}_1 \mathbf{x}(k) f_i(k) + \sum_{i=r+1}^{m} \mathbf{b}_i \mathbf{c}_2 \mathbf{x}(k) f_i(k)$$

with $E(f_i(k) f_j(q)) = \sigma_i^2 \delta_{kq} \delta_{ij}$, and $w_{ij}(k)$ is the inverse Z-transform of $c_j(\mathbf{I}z - \mathbf{A})^{-1} b_i$ $(i = 1, \ldots, m;\ j = 1, 2)$. Condition (ii) can also be obtained by proving that a similarity transformation exists which transforms $\mathbf{M}$ in the matrix

$$\begin{bmatrix} \mathbf{0}_{m-2\,;\,m-2} & \mathbf{X} \\ \mathbf{0}_{2,\,m-2} & \mathbf{N} \end{bmatrix}$$

where the indices denote the number of rows and columns, and where $\mathbf{X}$ is a matrix which does not have to be computed for evaluating the eigenvalues. Condition (ii) is equivalent to

$$\mathrm{tr}\,(\mathbf{N}) - \det(\mathbf{N}) < 1,\ \mathrm{tr}\,(\mathbf{N}) < 1.$$

where $\mathrm{tr}(\mathbf{N})$ and $\det(\mathbf{N})$ denote the trace and the determinant of a matrix respectively. This shows that here a considerable simplification of the conditions of Criterion 3 is obtained. The same result is obtained if there are only two different vectors $\mathbf{b}_i$; a similar development can be made for three or more different $\mathbf{b}$ or $\mathbf{c}$ vectors.

Consider indeed the case of d different vectors $\mathbf{b}_i$ which correspond respectively to $r_1, \ldots, r_d$ terms in Eq. (26). Then the system equation has the form

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{d} \mathbf{b}_i \sum_{j=s_i+1}^{s_{i+1}} \mathbf{c}_j \mathbf{x}(k) f_j(k)$$

where $s_i = r_1 + r_2 + \ldots + r_{i-1}$, $s_1 = 0$, $s_{d+1} = m$, $E(f_j(k)) = 0$, $E(f_j(k) f_i(p)) = \sigma_j^2 \delta_{ij} \delta_{pk}$. It is readily proved that correlation between noise elements can be taken in consideration without changing the form of the equation if only noises corresponding to the same vectors b_i are correlated. Let $p(z) = \det(z\mathbf{I} - \mathbf{A})$, $G_{ij}(z) = \mathbf{c}_j(\mathbf{I}z - \mathbf{A})^{-1}\, \mathbf{b}_i = q_{ij}(z)/p(z)$, and let the impulse response $w_{ij}(k)$ be the inverse Z-transform of $G_{ij}(z)$. The matrix $\mathbf{M}$ of Criterion 3 has $(m - d)$ zero eigenvalues in addition to the eigenvalues of the matrix $\mathbf{N}$ with positive entries defined by

$$n_{ij} = \sum_{u=s_j+1}^{s_{j+1}} \sum_{k=0}^{\infty} \sigma_u^2 w_{iu}(k)^2 \qquad (i, j = 1, \ldots, d).$$

The matrix $\mathbf{M}$ of Criterion 3 can hence be replaced by the lower dimensional matrix $\mathbf{N}$ for stability analysis. The elements of $\mathbf{N}$ can be computed as sums of elements of $\mathbf{M}$ as shown above; n_{ij} also equals the coefficient of z^n in the polynomial $h_{ij}(z)$ satisfying

$$\frac{1}{2}\big(p(z)\, h_{ij}(z^{-1}) + p(z^{-1})\, h_{ij}(z)\big) = \sum_{u=s_j+1}^{s_{j+1}} \sigma_u^2 q_{iu}(z)\, q_{iu}(z^{-1}).$$

The same result can also be used in the case where the noise elements corresponding to the same vector $\mathbf{b}_i$ are correlated, but uncorrelated with noise elements corresponding to another vector $\mathbf{b}_i$. Then the right hand side of the equation determining $h_{ij}(z)$ should be replaced by $\mathbf{v}_{ij}(z)' \mathbf{R}_j \mathbf{v}_{ij}(z^{-1})$, where $\mathbf{R}_j$ is the covariance matrix of the noise elements in the terms involving $\mathbf{b}_j$, and where $v_{ij}(z)$ is the vector of the elements $q_{iu}(z)$, for $\mathbf{u} = s_j + 1, \ldots, s_{j+1}$.

A similar result can be derived for equal vectors c_i; however the easiest way to derive the stability result in this case is to observe that the mean square stability properties of (26) are equivalent to the corresponding properties of

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{lm} \mathbf{c}_i' \mathbf{b}_i' \mathbf{x}(k)\, f_i(k)\,.$$

Remark 12. One may wonder if Criterion 3 can also be obtained by means of the multivariable circle criterion with respect to the feedback system considered in Remark 7. The answer is positive only if $w_{ij}(k) = w_{ji}(k)$ for all $i, j; k$. Only then is it possible to generalize the result of Criterion 3 to nonstationary and nonlinear variances, as in Remark 5.

4. General cases

Consider now the case of a single white noise stochastic element and the system equation

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + f(k)\, \mathbf{D}\mathbf{x}(k),\ E(f(k)) = 0,\ E(f(k) f(q)) = \sigma^2 \delta_{kq}$$

where is not of rank one, as in Section 2. Let r be the rank of $\mathbf{D}(r \leq n)$. Then $\mathbf{D}$ can be written as $\mathbf{D} = \mathbf{BC}$, where $\mathbf{B}$ and $\mathbf{C}$ are $(n \times r)$ and $(r \times n)$ matrices respectively, such that the system equation is

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + f(k)\,\mathbf{BC}\mathbf{x}(k). \tag{33}$$

The difference equation for the second mean matrix is

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\mathbf{A}' + \sigma^2\,\mathbf{BCQ}(k)\,\mathbf{C}'\mathbf{B}'. \tag{34}$$

We assume that the triple $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ is a minimal realization of $\mathbf{G}(z) = \mathbf{C}(\mathbf{I}z - \mathbf{A})^{-1}\mathbf{B}$, and use the tentative Lyapunov function $\mathrm{tr}(\mathbf{PQ}(k))$, where $\mathbf{P}$ is the solution of

$$\mathbf{A}'\mathbf{PA} - \mathbf{P} = -\mathbf{C}'\mathbf{UC}$$

where $\mathbf{U}$ is some positive definite matrix. The difference of the Lyapunov function along the solutions of (34) is $\mathrm{tr}(\mathbf{MQ}(k))$, where

$$\mathbf{M} = -\mathbf{C}'\mathbf{UC} + \mathbf{C}'\mathbf{B}'\mathbf{PBC}\,\sigma^2.$$

Trying to choose $\mathbf{U}$ such that $\mathbf{M} = -\mathbf{C}'\mathbf{VC}$, where $\mathbf{V}$ is positive semidefinite, the following criterion is obtained by means of the Appendix:

Criterion 4. Let $\mathbf{L}$ denote the linear mapping $\mathbf{Q} \to \mathbf{L}(\mathbf{Q})$ defined by

$$\mathbf{L}(\mathbf{Q}) = \sum_{k=0}^{\infty} \sigma^2 \mathbf{W}(k)'\,\mathbf{Q}\mathbf{W}(k)$$

where $\mathbf{W}(k)$ is the impulse response matrix of a system with transfer matrix $\mathbf{G}(z)$, i.e. $\mathbf{C}(\mathbf{I}z - \mathbf{A})^{-1}\mathbf{B} = \mathbf{Z}(\mathbf{W}(k))$. Then a necessary and sufficient condition for mean square asymptotic stability of (33) is that the largest eigenvalue of $\mathbf{L}$ which is real and positive, is smaller than 1. It is unstable in the mean square if the largest eigenvalue is larger than 1 in modulus; it is mean square stable but not asymptotically stable, or unstable, if the largest eigenvalue of $\mathbf{L}$ equals 1.

Remark 13. The elements of the matrix corresponding to the linear mapping $\mathbf{L}$ are explicitly of the form $\sum_{k=0}^{\infty} (w_{ij}(k) - w_{pq}(k))$. If $\mathbf{G}(z) = \dfrac{1}{p(z)}\mathbf{Q}(z)$, where $\mathbf{Q}(z)$ is a matrix of polynomials, and $p(z) = \det(z\mathbf{I} - \mathbf{A})$, then the above coefficient is the coefficient of z^n of the polynomial $h(z)$ satisfying

$$p(z)\,h(z^{-1}) + p(z^{-1})\,h(z) = q_{ij}(z)\,q_{pq}(z^{-1}) + q_{ij}(z^{-1})\,q_{pq}(z)$$

where q_{ij} and w_{ii} are the (i, j) entries of the matrices $\mathbf{Q}$ and $\mathbf{W}$.

Finally we consider the most general case of the stochastic difference equation

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{m} f_i(k)\,\mathbf{D}_i\,\mathbf{x}(k)$$

where the noise elements are white but may be correlated, and where the matrices $\mathbf{D}_i$ are of rank possibly larger than 1. It has been shown before how this case can be reduced to that of the uncorrelated noise (in this transformation the matrices $\mathbf{D}_i$ are changed). Let this be done, and let the resulting $\mathbf{D}_i$ matrices be factored $\mathbf{D}_i = \mathbf{B}_i\mathbf{C}_i$, where $\mathbf{B}_i$ and $\mathbf{C}_i$ are $(n \times r_i)$ and $(r_i \times n)$ matrices respectively, r_i is the rank of $\mathbf{D}_i$. Let $\mathbf{G}_{ij}(z) = \mathbf{C}_j(\mathbf{I}z - \mathbf{A})^{-1}\mathbf{B}_i$ be minimally realized by $(\mathbf{A}, \mathbf{B}_i, \mathbf{C}_j)$, and let the impulse response matrix $\mathbf{W}_{ij}(k)$ be the inverse Z-transform. Let $\mathbf{L}$ be the linear mapping from $\mathbf{R}^s \to \mathbf{R}^s$ (with $s = \sum_{i=1}^{m} r_i(r_i+1)/2$), mapping the symmetric matrices $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m$, respectively of order $r_1, r_2, \ldots, r_m$, into the matrices $\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_m$, which are also symmetric matrices or orders $r_1, r_2, \ldots, r_m$ respectively, defined by

$$\mathbf{V}_i = \sigma_i^2 \sum_{k=0}^{\infty} \sum_{j=0}^{m} \mathbf{W}_{ij}(k)'\,\mathbf{S}_j\,\mathbf{W}_{ij}(k)\,. \tag{35}$$

The difference equation of the stochastic system is

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \sum_{i=1}^{m} \mathbf{B}_i\mathbf{C}_i\mathbf{x}(k)\,f_i(k) \tag{36}$$

where without loss of generality it is assumed that $E(f_i(k)) = 0$, $E(f_i(k)f_i(q)) = \sigma_i^2\,\delta_{ij}\,\delta_{kq}$. The difference equation of the second mean is

$$\mathbf{Q}(k+1) = \mathbf{A}\mathbf{Q}(k)\,\mathbf{A}' + \sum_{i=1}^{m} \mathbf{B}_i\,\mathbf{C}_i\,\mathbf{Q}(k)\,\mathbf{C}_i'\,\mathbf{B}_i'\,. \tag{37}$$

Along the same lines as above, we obtain:

Criterion 5. The stochastic system (36) is mean square asymptotically stable if and only if the largest eigenvalue (which is real and positive) of the linear transformation L defined by (35) is less than 1. It is either stable but not asymptotically stable, or unstable, if that eigenvalue equals 1. It is mean square unstable if thet eigenvalue exceeds 1.

Remark 14. It is clear that Criterion 4 is only interesting if r is small with respect to n. Otherwise Criterion 1 may be useful. Similarly Criterion 5 is only interesting if m and r are small with respect to n.

Remark 15. Criterion 15 can be simplified if some of the elements $\mathbf{B}_i$ or some of the elements $\mathbf{C}_i$ are equal as in Section 3.

Remark 16. Criteria 4 and 5 can also be proved by means of the multivariable Nyquist criterion by modelling the second mean equations as multivariable feedback systems, and by observing that only $z = 1$ should be considered, because of the properties of the eigenvalues of the linear mappings involved; this is why the criteria obtained here are simpler than the corresponding results of Willems and Blankenship [11].

Remark 17. In Criteria 4 and 5 it is not possible in the critical cases to decide upon weak stability or instability, unlike in Criteria 2 and 3. This is indeed not possible as is shown by the following example

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{A}\mathbf{x}(k)\, f(k)$$

and hence

$$\mathbf{Q}(k+1) = (1+\sigma^2)\, \mathbf{A}\mathbf{Q}(k)\, \mathbf{A}'.$$

In the critical case (that is if the largest eigenvalue of $\mathbf{A}$ equals $1/\sqrt{(1+\sigma^2)}$) the distinction depends on the multiplicity of the largest eigenvalue of $\mathbf{A}$ as a zero of the minimal polynomial. In a critical case instability is such that the second mean matrix increases as a power of k as $k \to \infty$.

APPENDIX

In this appendix some interesting properties of a number of linear mappings used in the present paper are summarized. Let Q and P denote square symmetric matrices of order n. They belong to a vector space of dimension $s = n(n+1)/2$. Consider a linear mapping $L : R^s \to R^s : P = L(Q)$ such that a symmetric Q is transformed into a symmetric P. Suppose this mapping maps the cone of positive semidefinite matrices into the cone of positive semidefinite matrices. Then:

(i) The mapping has an eigenvalue λ^* which is real and nonnegative, and which is not exceeded by any other eigenvalue in modulus.

(ii) The eigenfunction Q^*, corresponding to λ^*, i.e. $L(Q^*) = \lambda^* Q^*$, is positive semidefinite. These properties can easily be proved by considering a repeated application of the linear transformation, and by using the Jordan canonical form of the matrix of the linear mapping.

Let A, D_i, and $W(k)$ for $k = 0, 1, 2, \ldots$ be square matrices of order n; then

$$L_1(Q) = A'QA,\ L_2(Q) = A'QA + \sigma^2 D'QD,$$

$$L_3(Q) = A'QA + \sigma_1^2 D_1'QD_1 + \sigma_2^2 D_2'QD_2 + \ldots + \sigma_m^2 D_m'QD_m,$$

$$L_4(Q) = \sum_{h=0}^{\infty} W(k)'\, QW(k)$$

are linear mappings of the class considered above, and hence have properties (i) and (ii). Moreover,

(–) The eigenvalues of L_1 are the products of eigenvalues of A.

(–) The maximum eigenvalue of L_2 increases with σ^2; the maximum eigenvalue of L_3 increases with $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$.

(—) If the maximum eigenvalue of L (and hence also L_1, L_2, L_3, L_4) is less than 1, then the inverse mapping of $P = Q - L(Q)$ can be expressed as follows:

$$Q = \sum_{i=0}^{\infty} L^i(P)$$

with $L^i(P) = L(L(\ldots L(P)\ldots))$, $L^0(P) = P$. This inverse mapping maps strictly positive definite matrices into strictly positive definite matrices. Similarly, the mapping $L_5 : R^s \to R^s : L_5(Q) = AQ + QA' + \sum_{i=1}^{m} \sigma_i^2 D_i Q D_i'$ can be proved to have a real eigenvalue which is at least as large as the real part of any other eigenvalue. That eigenvalue increases with $\sigma_1^2, \ldots, \sigma_m^2$.

Let M be a square matrix of order n whose entries are all positive, and x a vector with n components. The linear mapping $R^n \to R^n$:

$$L(x) = Mx$$

has a *simple* real positive eigenvalue r which is larger in modulus than any other eigenvalue. There is a corresponding eigenvector with only positive entries. The eigenvalue r increases strictly with the entries of M. If S is a complex square matrix of order n such that

$$|s_{ij}| \leq m_{ij}$$

then no eigenvalue of S is larger in modulus than the eigenvalue r of the matrix M. These properties of matrices with positive elements are proved in Gantmacher's book [16].

References

1. *Wedig, W.*, Regions of instability for a linear system with random parametric excita tion. Stability of Stochastic Systems, pp. 160—172, Lecture Notes in Mathematics, vol. 294, Springer Verlag, New York, 1972.
2. *Sagirow, P.*, The stability of a satellite with parametric excitation by the fluctuations of the geomagnetic field. Stability of Stochastic Systems, pp. 311—316, Lecture Notes in Mathematics, vol. 294, 1972.
3. *Ruhrmann, I.*, Stabilität der stochastischen Roll-Gierschwingungen eines Satelliten. Zeitschrift für Angewandte Mathematik und Mechanik, **52** (1972) T68—Z70.
4. *Kozin, F.*, A survey of stability of stochastic systems. Automatica, **5** (1969) 95—112.
5. *Willems, J. L.*, Stability Theory of Dynamical Systems. Nelson, London, 1972.
6. *Kushner, H. J.* Stochastic Stability and Control. Academic Press, New York, 1967.
7. *Heess, G.*, A mean square stability criterion for linear stochastic systems. Proceedings of the J. A. C. C., pp. 67—71, 1971.
8. *Willems, J. L.*, Mean square stability criteria for stochastic feedback systems. Intern. Journal of System Science, **4** (1973). 545—564,
9. *Brockett, R. W.*, Finite Dimensional Linear Systems. Wiley, New York, 1970.
10. *Willems, J. C., Brockett, R. W.*, Average value criteria for stochastic stability. Stability of Stochastic Systems, pp. 252—272, Lecture Notes in Mathematics, vol. 294, 1972.
11. *Willems, J. C., Blankenship, G. L.*, Frequency domain stability criteria for stochastic systems. IEEE Trans. Automatic Control, **AC—16** (1971) 292—299.
12. *Willems, J. L.*, Lyapunov functions and global frequency domain stability criteria for a class of stochastic feedback systems. Stability of Stochastic Systems, pp.131—146, Lecture Notes in Mathematics, vol. 294, 1972.
13. *Nakamizo, T., Sawaragi*, X., Analytical study of n-th order linear system with stochastic coefficients. Stability of Stochastic Systems, pp. 173—185, Lecture Notes in Mathematics, vol. 294, 1972.
14. *Nevel'son, B., Khas'minskii, Z.*, Stability of a linear system with random disturbances of its parameters, Journal of Applied Math. and Mech., **30** (1966) 487—490.
15. *Levit, M. V., Yacubovich, V. A.*, Algebraic criterion for stochastic stability of linear systems with parametric excitation of the white noise type. *P. M. M.*, **36** (1972) 142—148.
16. *Gantmacher, F. R.*, Theory of Matrices. vol II, Chapter XIII, Chelsea Publ. Co, 1959.

Квадратичные критерии устойчивости для линейных стохастических систем с белыми шумами

Й. Л. ВИЛЛЕМС

(Гент)

Применяется теория функций Ляпунова для анализа устойчивости в среднем квадратическом линейных систем со случайными коэффициентами в виде белых шумов.

Рассматриваются как непрерывные, так и дискретные линейные системы конечной размерности.

В первой части работы изучен случай одного случайного коэффициента, во второй – случай нескольких некоррелированных случайных коэффициентов, в последней части – общий случай.

J. L. Willems
Engineering School
University of Gent
Plateaustraat 22
B-9000 Gent
Belgium

Problems of Control and Information Theory, Vol. 2 (3–4), pp. 219–234 (1973)

ASYMPTOTIC PROPERTIES OF MARKOV CHAIN PARAMETER ESTIMATES

D. S. APOKORIN

(Moscow)

(Received November 16, 1971)

Asymptotic normality and asymptotic effectiveness of Bayes and generalized Bayes estimates of an unknown Markov chain parameter are established for a certain smoothness of the transition density and a sufficiently quick intermixing. The result obtained might be useful in the development of optimal reception techniques for communication channels with dependent noise.

Many technical problems (e.g. identification, information transmission through communication channel, etc.) require observation-based estimation of an unknown parameter distribution.

Study into properties of various estimates corresponds, in the information theory, to the study of various demodulators. Moreover, the estimation problems may be directly applied to the field of identification.

The literature contains a number of papers on asymptotic properties of Markov chain parameter estimates. Roussas [1, 2] has studied the asymptotic properties of the likelihood relation and maximal likelihood estimates for the case where observations are connected in an ergodic Markov chain. It has been proved [2], among other things, that under certain assumptions the maximal likelihood estimate $\hat{\theta}_n$ of a one-dimensional parameter is asymptotically normal and asymptotically effective. The present paper establishes a similar result for the generalized Bayes estimate. In doing so, some ideas of [3] are used, and some of its results are generalized.

1. Let $\xi_1, \ldots, \xi_n$ be a union of unknown values whose joint distribution density $P_n(x_1, \ldots, x_n; \theta)$ exists and depends on one-dimensional parameter $\theta \in \Theta$ (Θ being the open set of a straight line R^1) whose unknown "true" value will be denoted by θ_0. The asymptotic behaviour of the generalized Bayes estimates of θ which are known to have the following form

$$\hat{\theta}_n = \frac{\int u P_n(\xi_1, \ldots, \xi_n; u)\, \pi(u)\, du}{\int P_n(\xi_1, \ldots, \xi_n; u)\, \pi(u)\, du} \tag{1}$$

will be studied under the assumption that $\pi(u)$ is a bounded continuous function with $\pi(\theta_0) > 0$.

It will be assumed that for each $\theta \in \Theta$ the values $\xi_1, \ldots, \xi_n$ constitute a uniform-in-time stationary Markov chain with values from R^1. Denote by $P(x, y, \theta)$ its transition density, and by $\widetilde{P}(x, \theta)$ its stationary distribution density with respect to the Lebesque measure on the straight line.

The symbols $m_\theta \xi$ and $\Pr_\theta(A)$ indicate that the mathematical expectation of ξ and the probability of A (measurable with respect to $\xi_1, \ldots, \xi_n$) are computed for the parameter value equal to θ. Denote, for sake of briefness, $m_{\theta_0} \xi = m\xi$ and $\Pr_{\theta_0}(A) = \Pr(A)$.

Below we expect to meet the following conditions:

1. Functions $\widetilde{P}(x; \theta)$ and $P(x, y; \theta)$ are twice differentiable with respect to θ, they and their derivatives with respect to θ as well, are limited, and their second derivatives satisfy the Lipschitz condition uniformly over all x and y.

2. For a certain $\alpha > 0$ the moment $m_{\theta_0} \mid \xi_1 \mid^{2+\alpha}$ is finite.

3. Let ξ be a random value measurable with respect to $\{\xi_k : k \leq t\}$, η be a random value measurable with respect to $\{\xi_k : k \geq t + \tau\}$, with $m \mid \xi \mid^p < \infty$, $m \mid \xi \mid^q < \infty$ where $p, q > 0, \dfrac{1}{p} + \dfrac{1}{q} = 1$. Then for certain $\beta > 0$ and $c > 0$ the following holds for any τ:

$$| m\xi\eta - m\xi m\eta | \leq Ce^{-\beta\tau}(m \mid \xi \mid^p)^{\frac{1}{p}} (m \mid \eta \mid^q)^{\frac{1}{q}}. \tag{2}$$

4. There are $c > 0$, $\gamma > 0$ such that for all x, y

$$P(x, y; \theta_0) P(x, y; \theta_0 + z) < c \mid z \mid^{-\gamma}.$$

5.1. The sets

$$A\{x, \widetilde{P}(x, \theta) = 0\}, \quad B\{(x, y) : P(x, y; \theta) = 0\}$$

are θ-independent.

5.2. For any $\theta' \neq 0$ and $x \notin A$

$$\int | P(x, y; \theta) - P(x, y; \theta') | dy > 0.$$

5.3. For all x the equality $\int P(x, y; \theta) dy = 1$ is twice differentiable with respect to θ under the integration sign, i.e.

$$\int \frac{\partial^i P(x, y; \theta)}{\partial \theta_i} dy = 0 \qquad i = 1, 2.$$

5.4. For all $\theta \in \Theta$

$$\int \left| \frac{\partial}{\partial \theta} P(x; \theta) \right| d\alpha < c < \infty.$$

6. The function

$$g(z) = m\left[\frac{P(\xi_1, \xi_2; \theta_0 + z)}{P(\xi_1, \xi_2; \theta_0)}\right]^{1/2}$$

is continuous and twice continuously differentiable at zero, with differentiation being possible under the sign of mathematical expectation.

7. Introduce the functions

$$I(x, y; \theta) = \frac{\dfrac{\partial P(x, y; \theta)}{\partial \theta}}{P(x, y; \theta)}; \quad J(x, y; \theta) = \frac{\dfrac{\partial^2}{\partial \theta^2} P(x, y; \theta)}{P(x, y; \theta)}.$$

Then there are constants $c > 0$ and $\beta > 0$ such that for all θ

$$m_\theta |I(\xi_1, \xi_2; \theta)|^{2+\beta} \leq c; m_\theta |J(\xi, \xi_2; \theta)|^{1+\beta} \leq c. \tag{3}$$

A propos of Condition 3, note that it is satisfied when the Markov chain meets either the well-known Deblin condition [4], or the condition of uniformly intense intermixing [5]. The present paper aims at the proof of the following:

Theorem. Let Conditions 1 – 7 above be satisfied. Then the sequence of random variables $\varepsilon_n = (\hat{\theta}_n - \theta_0)\sqrt{n}$ is asymptotically normal $(0, \sigma^2)$ where $\sigma^2 = I^{-1}$, $I = mJ(\xi_1, \xi_2; \theta_0)$.

Together with [2] this theorem leads to the conclusion that the generalized Bayes estimate is asymptotically effective.

2. The proof is based on a number of auxiliary results representing some estimates related to $\hat{\theta}_n$. For sake of convenience we are going to state them prior to their proof.

First of all, let us write (1), in accordance with [3], in the following form:

$$\hat{\theta}_n = \frac{\displaystyle\int_{-\infty}^{+\infty} \frac{P_n(\xi_1, \ldots, \xi_n; u)}{P_n(\xi_s, \ldots, \xi_n; \theta_0)} \pi(u)\, du}{\displaystyle\int_{-\infty}^{+\infty} \frac{P_n(\xi_1, \ldots, \xi_n; u)}{P_n(\xi_1, \ldots, \xi_n; \theta_0)} \pi(u)\, du} = \theta_0 +$$

$$+ \frac{1}{\sqrt{n}} \cdot \frac{\displaystyle\int_{-\infty}^{+\infty} z \frac{P_n(\xi_1, \ldots, \xi_n; \theta_0 + z/\sqrt{n})}{P_n(\xi_1, \ldots, \xi_n; \theta_0)} \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz}{\displaystyle\int_{-\infty}^{+\infty} \frac{P_n(\xi_1, \ldots, \xi_n; \theta_0 + z/\sqrt{n})}{P_n(\xi_1, \ldots, \xi_n; \theta_0)} \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz} = \tag{4}$$

$$= \theta_0 + \frac{1}{\sqrt{n}} \cdot \frac{\int_{-\infty}^{+\infty} z Z_n(z)\, \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz}{\int_{-\infty}^{+\infty} Z_n(z)\, \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz}$$

where

$$Z_n(z) = \frac{\widetilde{P}\left(\xi_1 ; \theta_0 + \frac{z}{\sqrt{n}}\right)}{\widetilde{P}(\xi_1 : \theta)} \prod_{i=2}^{n} \frac{P\left(\xi_{i-1}, \xi_i ; \theta_0 + \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} = \exp\left(Y_n(z)\right)$$

and

$$Y_n(z) = \ln Z_n(z).$$

By virtue of Condition 5.1, $Y_n(z)$ is a random variable. Represent it in the following form:

$$Y_n(z) = z\eta_n - \frac{z^2}{2} d_n + w_n \tag{5}$$

where

$$\eta_n = \frac{1}{\sqrt{n}} \sum_{i=2}^{n} J(\xi_{i-1}, \xi_i ; \theta_0) ; d_n = \frac{1}{n} \sum_{i=2}^{n} [J(\xi_{i-1}, \xi_i ; \theta_0) - J(\xi_{i-1}, \xi_i ; \theta_0)];$$

$$w_n = \ln \frac{\widetilde{P}\left(\xi_1 ; \theta_0 + \frac{z}{\sqrt{n}}\right)}{\widetilde{P}(\xi_1 ; \theta_0)} + v_n ;$$

$$v_n = \sum_{i=2}^{n} \ln \frac{P\left(\xi_{i-1}, \xi_i ; \theta_0 + \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} - z\eta_n + \frac{z^2}{2} d_n .$$

The statements below are of special importance in the proof.

Lemma 1. For certain $c > 0$, $\Delta > 0$

$$g(z) = m\left[\frac{P(\xi_1, \xi_2 ; \theta_0 + z)}{P(\xi_1, \xi_2 ; \theta_0)}\right]^{1/2} \leq \frac{1}{(1 + cz^2)\,\Delta}$$

at all z.

Lemma 2. For all sufficiently small $\varepsilon > 0$ and an arbitrary $K > 0$ the inequality $mZ_n^{1/3}(z) \leq |z|^{-k}$ holds in the domain $n^{\varepsilon} \leq z \leq c^{n\varepsilon}$ beginning from a certain $n = n(k, \varepsilon)$.

Lemma 3. There is a constant $c > 0$ such that for any a and b

$$m \mid Z_n(b) - Z_n(a) \mid < c \mid b - a \mid, \quad n = 1, 2, \ldots$$

Lemma 4. For any $\varepsilon > 0$ the integral

$$\int\limits_{|z|<n^{\varepsilon}} |z| \, Z_n(z) \, \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz$$

converges to zero with respect to probability under $n \to \infty$.

Lemma 5. For any sufficiently small ε

$$\Pr\{|w_n(z)| > n^{-\varepsilon}\} = O(n^{-\varepsilon})$$

is uniform in the domain $|z| < n^{\varepsilon}$.

The theorem stems directly from the following result whose proof is based essentially on Lemmata 1 – 5.

Lemma 6. Under $n \to \infty$,

$$\int\limits_{-\infty}^{+\infty} z \, Z_n(z) \, \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz - \pi(\theta_0) \int\limits_{-\infty}^{+\infty} z \exp\left(z\eta_n - \frac{z^2}{2} dn\right) dz$$

and

$$\int\limits_{-\infty}^{+\infty} Z_n(z) \, \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz - \pi(\theta_0) \int\limits_{-\infty}^{+\infty} \exp\left(z\eta_n - \frac{z^2}{2} dn\right) dz$$

converge to zero with respect to probability.

Lemma 6 implies by virtue of (4) that

$$\varepsilon_n - \frac{\eta_n}{d_n} = \frac{\int z \, Z_n(z) \, \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz}{\int Z_n(z) \, \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz} - \frac{\int z \exp\left(z\eta_n - \frac{z^2}{2} d_n\right) dz}{\int \exp\left(z\eta_n - \frac{z^2}{2} d_n\right) dz} \to 0$$

with respect to probability. Furthermore, it follows from our assumptions [1] that η_n is asymptotically normal with the parameters (0, 1), and d_n converges to 1 with the probability equal to one. Thus, the required asymptotic normality results from the stated estimates.

Let us prove these statements.

3. Proof of the lemmata

Proof of Lemma 1. It is obvious that

$$g(z) = g_1(z) + g_2(z) \quad .$$

where

$$g_1(z) = \iint\limits_K \tilde{P}(x;\, \theta_0)\, P^{1/2}(x, y;\, \theta_0)\, P^{1/2}(x, y;\, \theta_0 + z)\, dx\, dy,$$

$$g_2(z) = \iint\limits_{\overline{K}} \tilde{P}(x;\, \theta_0)\, P^{1/2}(x, y;\, \theta_0)\, P^{1/2}(x, y;\, \theta_0 + z)\, dx\, dy,$$

domain K is a square $\{|x| \leq |z|^{\delta}, |y| \leq |z|^{\delta}\}$, $\overline{K}$ is the complement of K. Estimate separately $g_1(z)$ and $g_2(z)$ under large z.

In accordance with Cauchy—Buniakowski inequality

$$g_1(z) \leq [\iint\limits_K \tilde{P}^2(x, \theta_0)\, dx\, dy]^{1/2}\ [\iint\limits_K P(x, y;\, \theta_0)\, P(x, y;\, \theta_0 + z)\, dx\, dy]^{1/2}.$$

Take $0 < \delta < \frac{\sigma}{4}$. Then by virtue of (4) and the boundedness of $\tilde{P}(x, \theta_0)$, positive c', Δ' will be found such that

$$g_1(z) \leq c' |z|^{-\Delta'}. \tag{6}$$

Applying once again the Cauchy—Buniakowski inequality, obtain

$$g_2(z) \leq [\iint\limits_{\overline{K}} \tilde{P}(x;\, \theta_0)\, p(x, y;\, \theta_0 + z)\, dx\, dy]^{1/2}\, [\iint\limits_{\overline{K}} \tilde{P}(x;\, \theta_0)\, P(x, y;\, \theta_0)\, dx\, dy]^{1/2} \leq$$

$$\leq \iint\limits_{\overline{K}} \tilde{P}(x;\, \theta_0)\, P(x, y;\, \theta_0)\, dx\, dy]^{1/2}.$$

The latter integral is the probability that either $|\xi_1| > |z|^{\delta}$ or $|\xi_2| > |z|^{\delta}$, and it follows from the Chebyshev inequality by virtue of Condition 2 that for certain $c'' > 0$, $\Delta'' > 0$

$$g_2(z) \leq c'' |z|^{-\Delta''}. \tag{7}$$

Study now the behaviour of $g(z)$ under small z. It is easy to verify by Conditions 6 and 5.3 that

$$g'(0) = 0, \quad g''(0) = -\frac{1}{4} I\,.$$

Hence, under all z small enough

$$g(z) \leq \frac{1}{1 + c'''\, z^2} \tag{8}$$

for a certain $c''' > 0$.

To estimate $g(z)$ for the remaining values of z note that, in accordance with the Cauchy—Buniakowski inequality and with due regard to Condition 5.2 for $z \neq 0$, the following strict inequality holds

$$g(z) < [\int\int \overline{P}(x;\theta_0)\, P(x,y;\theta_0)\, dx\, dy]^{1/2}\, [\int\int \overline{P}(x;\theta_0)\, P(x,y;\theta_0+z)\, dx\, dy]^{1/2} = 1.$$

Consequently, owing to the continuity of $g(z)$ and inequalities (6) and (7)

$$\sup_{0>z_0<|z|} g(z) = q < 1\,. \tag{9}$$

Thus, by virtue of (6)—(9) the inequality

$$g(z) < \frac{1}{(1+cz^2)^{\Delta}}$$

holds for certain c and Δ under all z.

Proof of Lemma 2. Transform the estimated expression using Condition 3. Let $\delta > 0$ be sufficiently small. Define X by the following inequality:

$$Z_n(z) = \frac{\widetilde{P}\left(\xi_i\,;\theta_0+\frac{z}{\sqrt{n}}\right)}{\widetilde{P}(\xi_i\,;\theta_0)} \prod_{i=2}^{n} \frac{P\left(\xi_{i-1},\xi_i\,;\theta_0+\frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1},\xi_i\,;\theta_0)} = \Big(\prod_{i=1}^{[n^{1-\delta}]} \eta_i^2\Big)\, X\,,$$

where

$$\eta_i = \left(\frac{P\left(\xi_i\,[n^\delta],\xi_i\,[n^\delta]+1\,;\theta_0+\frac{z}{\sqrt{n}}\right)}{P(\xi_i\,[n^\delta],\xi_i\,[n^\delta]+1\,;\theta_0)}\right)^{1/2},$$

and $[m]$ denotes the integer part of m. Through the Hölder inequality obtain the following:

$$m Z_n^{1/3}(z) \leq \Big(m \prod_{i=2}^{[n^{1-\delta}]} \eta_i\Big)^{2/3} (mx)^{1/3} = \Phi_n(z)\,.$$

Note that $mx = 1$ because for any integers $k_1, \ldots, k_l$

$$\int \ldots \int \widetilde{P}(x_1\,;\theta)\, p_{k1}(x_1,x_2\,;\theta)\, p_{k_2}(x_2,x_3\,;\theta) \ldots p_{k_l}(x_l,x_{l+1},\theta)\, dx_1 \ldots dx_{l+1} = 1$$

where $p_n(x,y;\theta)$ is the transition density after k steps. Similarly

$$m(\eta_{i_1} \ldots \eta_{i_j})^2 = 1$$

for any unequal $i_1, \ldots, i_j$. To estimate $\Phi_n(z)$ use Condition 3 and take $p = q = 2$. Let $\tau = [n^\delta] - 1$, $l = [n^{1-\delta}]$.

Obtain then by (10)

$$m_{\eta_1 \dots \eta_l} \leq m_{\eta_1} m_{\eta_2 \dots \eta_e} + Cl^{-\beta r} \leq m_{\eta_1} m_{\eta_2} m_{\eta_3} \dots \eta_n +$$

$$+ Cl^{-\beta r}(1 + m_{\eta_1}) \leq \dots \leq (m_{\eta_1})^l + Cl^{-\beta\tau} \sum_{i=0}^{l-1} (m_{\eta_i})^i$$

or, taking (9) into consideration,

$$m_{\eta_1 \dots \eta_l} \leq (m_{\eta_1})^l + Cl^{-\beta\tau}$$

for $|z| > z_0$ and a certain $\tilde{C} > 0$. Thus

$$\Phi_n(z) \leq \left[M \left\{ \frac{P(\xi_1, \xi_2; \theta_0 + \frac{z}{\sqrt{n}})}{P(\xi_1, \xi_2; \theta_0)} \right\}^{1/2} \right]^{2/3 [n^{1-\delta}]} + \tilde{\tilde{C}} l^{-\beta [n^\delta]}.$$

Applying now Lemma 1 and taking, for sake of definiteness, $\delta = \frac{3}{2}\varepsilon$, one can verify that for any sufficiently small ε

$$\Phi_n(z) < |z|^{-k}$$

in the domain $n^\varepsilon < |z| < l^{n^\varepsilon}$ beginning from a certain $n(k, \varepsilon)$. The lemma is proved.

Proof of Lemma 3. Let, for sake of definiteness, $a < b$. Obtain:

$$m |Z_n(b) - Z_n(a)| = m \left| \int_a^b Z_n'(z)\, dz \right| < \int_a^b m |Z_n'(z)|\, dz$$

The following identity is readily verified:

$$Z_n(z) = \frac{1}{\sqrt{n}} Z_n(z) \left[\tilde{J}(\xi_1; \theta_0 + \frac{z}{\sqrt{n}}) + \sum_{i=2}^{n} J\left(\xi_{i-1}, \xi_i; \theta_0 + \frac{z}{\sqrt{n}}\right) \right],$$

where $\tilde{J}(\xi_1; \theta) = \dfrac{\frac{\partial}{\partial\theta} \tilde{P}(x, \theta)}{\tilde{P}(x, \theta)}$.

By virtue of the definition of $Z_n(z)$, it follows evidently that:

$$m_{\theta_0} |Z_n'(z)| = m_{\theta_0} + \frac{z}{\sqrt{n}} \left(\frac{1}{\sqrt{n}} \left| \tilde{J}\left(\xi_1; \theta_0 + \frac{z}{\sqrt{n}}\right) + \sum_{i=2}^{n} J\left(\xi_{i-1}, \xi_i; \theta_0 + \frac{z}{\sqrt{n}}\right) \right| \right).$$

It is easy to see by Condition 5.3 that the values $J(\xi_{i-1}, \xi_i; \theta)$ are mutually non-correlated with respect to m_0. Indeed, if $k < l$ then

$$m_\theta J(\xi_{k-1}, \xi_k; \theta) J(\xi_{l-1}, \xi_l; \theta) = \int \dots \int J(x_{k-1}, x_k; \theta) P(x_{k-1}; \theta),$$

$$P(x_{l-2}, x_{l-1}; \theta) \dots P(x_{l-2}, x_{l-1}; \theta)\, dx_{k-1} \dots dx_{l-1} \int \frac{\partial P(x_{l-1}, x_l; \theta)}{\partial \theta}\, dx_l = 0 .$$

Hence

$$m_\theta |Z_n'(z)| \leq \frac{1}{\sqrt{n}} m_\theta + \frac{z}{\sqrt{n}} \tilde{J}\left(\xi_1 ; \theta_0 + \frac{z}{\sqrt{n}}\right) +$$

$$+ \frac{1}{\sqrt{n}} \sqrt{\sum_{i=2}^{n} m_{\theta_0} + \frac{z}{\sqrt{n}} \left| J\left(\xi_{i-1}, \xi_i ; \theta_0 + \frac{z}{\sqrt{n}}\right)^2 \right|}$$

and 5.4 and 7 imply that $m_{\theta_0} | Z_n'(z) |$ is z-bounded. Thus the lemma is proved.

Proof of Lemma 4. Since the function $\pi(\theta)$ is bounded, it is sufficient to estimate

$$\int_{n^\varepsilon}^{\infty} z Z_n(z)\, dz .$$

The integral over the domain $(-\infty, -n^\varepsilon)$ is estimated similarly. Let ε satisfy Lemma 2.

Estimate first

$$\int_{n^\varepsilon}^{e^{n\varepsilon}} z Z_n(z)\, dz .$$

Denote

$$\Pr_N = \Pr\left\{ \int_N^{N+1} z Z_n(z)\, dz > \frac{1}{N^2} \right\} (N \geq 1) .$$

Break up the segment $[N, N+1]$ into N^m-equal intervals Δ_k, and let $Z_k \in \Delta_k$. Then

$$\Pr_n \leq \Pr\left\{ \frac{1}{N^m} \sum_{k=1}^{N^m} Z_n(z_k) + \sum_{k=1}^{N^m} \int_{\Delta_k} |Z_n(z) - Z_n(z_k)|\, dz > \frac{1}{2N^3} \right\} \leq$$

$$\Pr\left\{ \frac{1}{N^m} \int_{k=1}^{N^m} Z_n(z_k) > \frac{1}{4N^3} \right\} + \Pr\left\{ \sum_{k=1}^{N^m} \int_{\Delta_k} |Z_n(z) - Z_n(z_k)|\, dz > \frac{1}{4N^3} \right\} .$$

Estimate each term of the latter expression. By virtue of the Chebyshev inequality and Lemma 3,

$$\Pr\left\{ \sum_{k=1}^{N^m} \int_{\Delta_k} |Z_n(z) - Z_n(z_k)|\, dz > \frac{1}{4N^3} \right\} \leq c_1 N^{-m+3} \tag{11}$$

for a certain $c_1 > 0$. Further,

$$\Pr\left\{\frac{1}{4N^3}\sum_{k=1}^{N^m} Z_n(z_k) > \frac{1}{4N^3}\right\} \leq \sum_{k=1}^{N^m} \Pr\left\{Z_n(z_k) > \frac{1}{4N^3}\right\} \leq (4N^3)^{1/3} \sum_{k=1}^{N^m} m\, Z^{1/3}(z_k).$$

By virtue of Lemma 2, the latter expression does not exceed

$$C_2 \frac{N^{m+1}}{N^k}$$

under all N in the interval $(n^\varepsilon, e^{n^\varepsilon} - 1)$ for $n > n^{(k,\varepsilon)}$. Assuming that $m = 5$, $k = 8$, and taking (11) into consideration find:

$$\Pr_N \leq \frac{C_3}{N^2} \tag{12}$$

under sufficiently large n.

Break up the segment $(n^\varepsilon, e^{n^\varepsilon} - 1)$ into smaller segments Δ^k by means of the points $n^\varepsilon + k$; $k = 1, 2, \ldots, n^\varepsilon + k < l^{n^\varepsilon}$. Then for sufficiently large n one obtains, by virtue of (12):

$$\Pr\left\{\int_{n^\varepsilon}^{e^{n^\varepsilon}} z\, Z_n(z)\, dz > \delta\right\} \leq \sum_k \Pr\left\{\int_{\Delta_k} z Z_n(z)\, dz > \frac{1}{(n^\varepsilon + k)^2}\right\} \leq C_3 \sum_{k=0}^{\infty} \frac{1}{n^\varepsilon + k} = 0(1)$$

for $n \to \infty$. Show now that $U_n = \int_{e^{n^\varepsilon}}^{\infty} z Z_n(z) dz$ also converges to zero with respect to probability. Write U_n in the following form:

$$U_n = \frac{\int_{e^{n^\varepsilon}}^{\infty} z\widehat{P}\left(\xi_1 ; \theta_0 + \frac{z}{\sqrt{n}}\right) \prod_{i=2}^{\infty} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \frac{z}{\sqrt{n}}\right) \prod_{i=2}^{\infty} \widetilde{P}(\xi_i ; \theta_0)}{\left[\prod_{i=1}^{n-1} \widehat{P}(\xi_i ; \theta_0)\, P(\xi_i, \xi_{i+1} ; \theta_0)\right]^2} \times$$

$$\times \frac{\prod_{i=1}^{n-1} \widetilde{P}(\xi_i ; \theta_0)\, P(\xi_{i-1}, \xi_i ; \theta_0)\, dz}{\left[\prod_{i=1}^{n-1} \widehat{P}(\xi_i ; \theta_0)\, P(\xi_i, \xi_{i+1} ; \theta_0)\right]^2} = \frac{a_n}{b_n}.$$

By inequality (15) obtained below, it is readily seen that

$$m \mid \ln [\widetilde{P}(\xi_1; \theta)\, P(\xi_1, \xi_2; \theta_0)] \mid < \infty.$$

Ergodicity of the Markov chain $\xi_1, \ldots, \xi_n$ leads to the conclusion that

$$\frac{1}{n} \sum_{i=1}^{n} \ln [P(\xi_1 ; \theta) P(\xi_i, \xi_{i+1} ; \theta_0]$$

converges with respect to probability, and, hence, there is a constant $c_0 > 0$ for which the following relation

$$\Pr \left\{ \left[\prod_{i=1}^{n} \widetilde{P}(\xi_i ; \theta_0) P(\xi_{i-1}, \xi_i ; \theta_0) \right]^2 > \frac{1}{c_0^n} \right\} \to 1$$

holds under $n \to \infty$. Thus, the convergence of $c_0^n a_n$ to zero with respect to probability is sufficient to prove that U_n converges to zero with respect to probability. Let us evaluate its mathematical expectation. By Condition 4 and the boundedness of $\widetilde{P}(x; \theta)$ one can obtain that

$$m c_0^n V_n = c_0 \sqrt{n} \int\limits_{1/\sqrt{n} \cdot e^{n^\varepsilon}}^{\infty} dz \int \ldots \int P(x_1 ; \theta_0 + z) \prod_{i=2}^{n-1} P(x_{i-1}, x_i ; \theta_0 + z) \times$$

$$\times \left[\prod_{i=1}^{n-1} P(x_i ; \theta_0) P(x_i ; x_{i+1} ; \theta_0) \right]^2 dx_1 \ldots dx_n \leq c \sqrt{n} \int\limits_{1/\sqrt{n} \cdot e^{n^\varepsilon}}^{n} z^{-(n-1)\gamma} dz \int \ldots$$

$$\ldots \int P(x_1 ; \theta_0) \ldots P(x_{n-1}, x_n ; \theta_0) dx_1 \ldots dx_n = o(1) \quad \text{for} \quad n \to \infty .$$

Taking the above remarks into account, the proof of the lemma is completed. Note that $\int\limits_{|z| > n^\varepsilon} Z_n(z) A\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz$ also converges to zero with respect to probability by virtue of this lemma. The following estimates may be shown to follow from the moment ξ_1 of order α-finiteness: there is $\delta > 0$ such that for certain c_1, c_2, c_3 under all $0 < h < 1$

$$\Pr \{\widetilde{P}(\xi_1; \theta_0) < h\} < c_1 h^\delta \tag{13}$$

$$\Pr \{P(\xi_1, \xi_2; \theta_0) < h\} < c_2 h^\delta \tag{14}$$

$$\Pr \{\widetilde{P}(\xi_1; \theta_0) P(\xi_1, \xi_2; \theta_0) < h\} < c_3 h^\delta . \tag{15}$$

Indeed, denoting $A_h = \{x : \widetilde{P}(x; \theta_0) < h\}$ we obtain for any $\gamma > 0$

$$\Pr\{\widetilde{P}(\xi_1 ; \theta_0) < h\} = \int\limits_{A_h} \widetilde{P}(x ; \theta_0) dx \leq \int\limits_{A_h \cap \{|x| < h^{-\gamma}\}} \widetilde{P}(x ; \theta_0) dx +$$

$$+ \int\limits_{\{|x| > h^{-\gamma}\}} P(x ; \theta_0) dx \leq 2h^{1-\gamma} m |\xi_1|^\alpha h^{\gamma\alpha} \leq (2 + m |\xi_1|^\alpha) h^{\min(1-\gamma, \gamma\alpha)} .$$

In particular, by taking $\gamma = \dfrac{1}{1+\alpha}$ we verify that (13) holds for $\delta = \dfrac{\alpha}{1+\alpha}$. Similarly, denoting $B_h = \{(x, y) : P(x, y; \theta) < h\}$ obtain for $\gamma = \dfrac{1}{1+\alpha}$ that

$$\Pr\{P(\xi_1, \xi_2; \theta_0) < h\} = \iint\limits_{B_h} \widetilde{P}(x; \theta_0)\, P(x, y; \theta_0)\, dx\, dy \leq$$

$$\leq \iint\limits_{B_h \cap \{|x| < h^{-\gamma}, |y| < h^{-\gamma}\}} \widetilde{P}(x; \theta_0)\, P(x, y; \theta_0)\, dx\, dy +$$

$$+ 2P\{|\xi_1| > h^{-\gamma}\} \leq (2 + 2m\,|\xi_1|^{\alpha})\, h^{\frac{\alpha}{1+\alpha}}.$$

(15) may similarly be verified. Note that if Condition 2 is met, (13), (14) and (15) are correct for $\delta < 2/3$.

Proceed now to the proof of the following statement.

Proof of Lemma 5. It is sufficient to check it separately for

$$\ln \frac{\widetilde{P}\left(\xi_1; \theta + \dfrac{z}{\sqrt{n}}\right)}{\widetilde{P}(\xi_1; \theta_0)} \quad \text{and} \quad v_n(z).$$

Make use of the boundedness of $\dfrac{\partial P(x; \theta)}{\partial \theta}$. Then for $|z| < n^{\varepsilon}$

$$\Pr\left\{\left|\ln \frac{\widetilde{P}\left(\xi_1; \theta_0 + \dfrac{z}{\sqrt{n}}\right)}{\widetilde{P}(\xi_1; \theta_0)}\right| > n^{-\varepsilon}\right\} \leq$$

$$\leq \Pr\left\{\frac{\left|\widetilde{P}\left(\xi_1; \theta_0 + \dfrac{z}{\sqrt{n}}\right) - P(\xi_1; \theta_0)\right|}{\widetilde{P}(\xi_1; \theta_0)} > \frac{1}{2}\, n^{-\varepsilon}\right\} \leq \Pr\{\widetilde{P}(\xi_1; \theta_0) < 2cn^{-\frac{1}{2}+\varepsilon}\}$$

for sufficiently large n, and by virtue of (13), for $\varepsilon < \dfrac{1}{2(1+2\beta)}$

$$\Pr\left\{\left|\ln \frac{\widetilde{P}\left(\xi_1; \theta_0 + \dfrac{z}{\sqrt{n}}\right)}{\widetilde{P}(\xi_1; \theta_0)}\right| > \frac{1}{n^{\varepsilon}}\right\} = 0(n^{-\varepsilon}) \tag{16}$$

uniformly over $|z| < n^{\varepsilon}$.

To estimate $v_n(z)$, transform it as follows:

$$v_n(z) = \sum_{i=2}^{n} \left[\ln\left(1 + \frac{z}{\sqrt{n}} J(\xi_{i-1}, \xi_i ; \theta_0) + \frac{z^2}{2n} \cdot \frac{\frac{\partial^2}{\partial\theta^2} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)}\right) - \right.$$

$$\left. - \frac{z}{\sqrt{n}} J(\xi_{i-1} ; \xi_i ; \theta_0) + \frac{z^2}{2n} J^2(\xi_{i-1}, \xi_i ; \theta_0) - \frac{z^2}{2n} J(\xi_{i-1}, \xi_i ; \theta_0) \right], \quad |\varkappa| < 1. \quad (17)$$

Make use of the expansion $\ln(1 + x) = x - \frac{x^2}{2} + o(x^3)$ taking place uniformly over $\{|x| \leq q < 1\}$. Show to this and that under sufficiently small ε

$$\Pr\left\{ \bigcup_{i=2}^{n} \left\{ \left| \frac{z}{\sqrt{n}} J(\xi_{i-1}, \xi_i ; \theta_0) + \frac{z^2}{2n} \frac{\frac{\partial^2}{\partial\theta^2} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} \right| > q \right\} \right\} =$$

$$= 0(n^{-\varepsilon}) \quad (18)$$

uniformly in the interval $|z| < n^{\varepsilon}$.

Asymptotic equality (18) stems from the following relations:

$$\Pr\left\{ \bigcup_{i=2}^{n} \left\{ \left| \frac{z}{\sqrt{n}} J(\xi_{i-1}, \xi_i ; \theta_0) \right| > \frac{q}{2} \right\} \right\} \leq \sum_{i=2}^{n} \Pr\left\{ \left| \frac{z}{\sqrt{n}} J(\xi_{i-1}, \xi_i ; \theta_0) \right| > \frac{q}{2} \right\} \leq$$

$$\leq \left(\frac{2}{q}\right)^{2+\beta} \sum_{i=2}^{n} \left(\frac{n^2}{\sqrt{n}}\right)^{2+\beta} m\,|J(\xi_{i-1}, \xi_i ; \theta_0)^{2+\beta} = 0(n^{-\varepsilon}) \quad \text{under} \quad \varepsilon < \frac{\beta}{2(3+\beta)}$$

$$\Pr\left\{ \bigcup_{i=2}^{n} \left\{ \left| \frac{z^2}{n} \frac{\frac{\partial^2}{\partial\theta^2} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} \right| > q \right\} \right\} \leq$$

$$\sum_{i=2}^{n} \Pr\left\{ \frac{z^2}{n} \left| \frac{\frac{\partial^2}{\partial\theta} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} - J(\xi_{i-1}, \xi_i ; \theta_0) \right| > \frac{q}{2} \right\} +$$

$$+ \sum_{i=2}^{n} \Pr\left\{ \frac{z^2}{n} |J(\xi_{i-1}, \xi_i ; \theta_0)| > \frac{q}{2} \right\} \leq \sum_{i=2}^{n} \Pr\left\{ \frac{cz^3}{n^{3/2}} \cdot \frac{1}{P(\xi_{i-1}, \xi_i ; \theta_0)} > \frac{q}{2} \right\} +$$

$$+ \left(\frac{2}{q}\right)^{1+\beta} \sum_{i=2}^{n} \left(\frac{z^2}{n}\right)^{1+\beta} m\,|J(\xi_{i-1}, \xi_i ; \theta_0)|^{1+\beta} = 0(n^{-\varepsilon})$$

obtained through Conditions 1, 7, and inequality (14) under an appropriate choice of θ_0. Since in (14) $\delta > 2/3$ if Condition 2 is met.

Now $v_n(z)$ may be estimated as follows:

$$v_n(z) = \sum_{i=2}^{n} \frac{z^2}{2n} \left[\frac{\frac{\partial^2}{\partial\theta^2} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} - \sum_{i=2}^{n} J(\xi_{i-1}, \xi_i ; \theta_0) \right] -$$

$$- \sum_{i=2}^{n} \frac{z^3}{n^{3/2}} J(\xi_{i-1}, \xi_i ; \theta_0) \frac{\frac{\partial^2}{\partial\theta^2} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} - \tag{18a}$$

$$- \frac{1}{2} \sum_{i=2}^{n} \frac{z^4}{n^2} \left[\frac{\frac{\partial^2}{\partial\theta^2} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} \right]^4 + \bar{v}_n(z)$$

where

$$\bar{v}_n(z) \leq C \sum_{i=2}^{n} \left[\frac{z}{\sqrt{n}} J(\xi_{i-1}, \xi_i ; \theta_0) + \frac{z^2}{2n} \frac{\frac{\partial^2}{\partial\theta^2} P\left(\xi_{i-1}, \xi_i ; \theta_0 + \varkappa_i \frac{z}{\sqrt{n}}\right)}{P(\xi_{i-1}, \xi_i ; \theta_0)} \right]^3.$$

The first term in (18a) has been estimated above. To estimate the remaining terms of (18a), Conditions 1, 2 and 7 should be used. We are not going to present the detailed calculations since they are quite similar to those given just now.

Proof of Lemma 6. Use Lemma 4 to estimate the following difference

$$\int_{-\infty}^{\infty} z Z_n(z) \pi\left(\theta_0 + \frac{z}{\sqrt{n}}\right) dz - \pi(\theta_0) \int_{-\infty}^{\infty} z l^{z\eta_n - \frac{z^2}{2} dn} dz .$$

Moreover, we have:

$$\int_{n^\varepsilon}^{\infty} z l^{z\eta_n - \frac{z^2}{2} dn} dz = l^{\eta_n^2/2dn} \int_{n^\varepsilon}^{\infty} z l^{z - \frac{d_n}{2}\left(z - \frac{\eta_n}{dn}\right)} dz ,$$

and for sufficiently large n

$$\Pr\left\{ \int_{n^\varepsilon}^{\infty} z l^{-\frac{dn}{2}\left(z - \frac{\eta_n}{dn}\right)^2} dz > \delta \right\} \leq \Pr\left\{ \frac{\eta_n}{dn} > \frac{n^\varepsilon}{2} \right\} + \Pr\left\{ dn < \frac{I}{2} \right\}.$$

Since η_n is asymptotically normal, and $dn \to I$ nearly for sure, the integral $\int_{n^\varepsilon}^{\infty} ze^{z\eta_n - \frac{z^2}{2}dn}\,dz$ converges to zero with respect to probability. This holds also for $\int_{-\infty}^{-n^\varepsilon} ze^{z\eta_n - \frac{z^2}{2}dn}\,dz$. Thus, owing to continuity of $\pi(\theta)$, it is sufficient to show that

$$\int_{|z|<n^\varepsilon} z\,Z_n(z)\,dz - \int_{|z|<n^\varepsilon} ze^{z\eta_n - \frac{z^2}{2}dn}\,dz \to 0 \tag{19}$$

with respect to probability for some $\varepsilon > 0$.

Since the function $Z_n(z)$ satisfies on the average the Lipschitz condition for z uniformly over n, its integral may be replaced by a sum. Namely, break up the segment $[-n^\varepsilon, n^\varepsilon]$ into intervals Δl of length not exceeding $n^{k\varepsilon}$. Then, by Lemma 3,

$$\int_{-n^\varepsilon}^{n^\varepsilon} z\,Z_n(z)\,dz = \frac{1}{n^{k\varepsilon}} \sum_e z_e\,Z_n(z_e) + \mu_n \tag{20}$$

where μ_n converges to zero with respect to probability, and z_l are fixed non-random points in the interval Δ_l. Similarly, one may readily verify the equality for the function $e^{z\eta_n - \frac{z^2}{2}dn}$. Note that, by Lemma 5, an ε will be found such that

$$\Pr\left\{\left|Z_n(z) - \exp\left(z\eta_n - \frac{z^2}{2}dn\right)\right| > \frac{1}{n^{m\varepsilon}} = 0\left(\frac{1}{n^{m\varepsilon}}\right)\right\}$$

uniformly for $|z| < n^\varepsilon$. Hence, (19) follows by virtue of (20). The second part of the lemma is proved similarly.

References

1. *Roussas, G. G.*, Asymptotic inference in Markov processes, Annals of Math. Stat., **36,** *3* (1965), 978–992.
2. *Roussas, G. G.*, Some applications of the asymptotic distribution of likelihood functions to the asymptotic efficiency of estimates. Z. Wahrscheinlichkeitstheorie und Verw., Geb., **3,** *10* (1968), 252–260.
3. *Ибрагимов И. А., Хасьминский Р. З.* Асимптотическое поведение некоторых статистических оценок. «Теория вероятностей и ее применения», *3,* 1972.
4. *Doob, J. L.*, Stochastic Processes. John Wiley & Sons, New York, Chapman & Hall, London, 1953.
5. *Ибрагимов И. А., Линник Ю. В.* Независимые и стационарные связанные величины. «Наука», М., 1965.

Оценка параметра марковской цепи

Д. С. АПОКОРИН

(Москва)

Пусть $\xi_1, \ldots \xi_n, \ldots$ — случайные величины, образующие марковскую цепь. Функция $P(x, y, \theta)$ — стационарная переходная плотность и $\tilde{P}(x, \theta)$ — плотность стационарного распределения марковской цепи, где $\theta \in \Theta$ — одномерный параметр, Θ — открытое множество на прямой, θ_0 — неизвестное «истинное» значение θ.

В настоящей работе исследуется асимптотическое поведение обобщенной бейесовской оценки θ_0. Такая оценка имеет вид:

$$\hat{\theta}_n = \frac{\int u P_u(\xi_1, \ldots \xi_n, u)\, \pi(u)\, du}{\int P(\xi_1 \ldots \xi_n, u)\, \pi(u)\, du},$$

где $P_n(x_1, \ldots, x_n, \theta) = \tilde{P}(x_1, \theta) \cdot \prod_{i=1}^{n} P(x_{i-1}, x_i, \theta)$ — плотность совместного распределения $\xi_1, \ldots \xi_n$. Цель работы — доказательство следующего результата.

Теорема. Пусть выполнены условия 1—7, тогда случайная величина $E_n = \sqrt{n}\,(\theta_n - \theta_0)$ имеет асимптотически гауссовское распределение с параметрами $(0, \sigma^2)$, где

$$\sigma^2 = I^{-1},$$

$$I = MJ(\xi_1, \xi_2, \theta_0),$$

$$J(x, y, \theta) = \frac{\dfrac{\partial}{\partial \theta} P(x, y, \theta)}{P(x, y, \theta)}.$$

Из результатов теоремы и работы [2] следует, что бейесовская оценка асимптотически эффективна. Этот результат может быть использован при построении оптимальных методов передачи информации по каналам с зависимыми шумами.

Problems of Control and Information Theory, Vol. 2 (3–4), pp. 235–241 (1973)

SYNTHESIS OF ALGORITHMS FOR DATA COMPRESSION

V. A. VITTIKH

(Kuibyshev)

(Received February 7, 1972)

The problem of data compression is stated in the terms of measuring systems invariant to a given class of signals. Such systems are shown to be effectively synthesized through measurement signal structural models. Synthesis of linear algorithms for adaptive signal sampling is given as an example.

1. Introduction

A priori data on processes or phenomena being studied enable one to compress the measurement information. If the processes follow a certain model, the data are not recorded and transmitted. Deviation of a parameter is recorded only when it surpasses the allowable margin.

Since on the basis of a mathematical model the measurement signals may be classified in a way, the compression problem may be formulated as that of signal membership estimation. If a signal belongs to a given class (with an accuracy of ε), there should be no information at the system output. Putting it in other way, the measurement system should be invariant to a given class of signals.

Thus, solution of the data compression problem is related to implementation of invariant measuring systems. It should be noted that the invariance theory is already used extensively for the implementation of automatic control systems [1]. Papers have appeared recently discussing its application to the field of the communication theory [2].

Implementation techniques for measuring devices invariant to arbitrary external disturbances are suggested in [3]. Invariance is shown there to be an important factor in synthesis of algorithms and systems for measurement information compression. This problem is solved here through structural signal models which furnish a convenient vehicle for synthesis [4].

2. Problem statement and basic definitions

Figure 1 shows the generalized block-diagram of a measurement information compressor [5]. Block 1 estimates the deviation of the measurement signal $s(t)$ from the adopted model $q(t) \in Q$, i.e. it checks whether the signal belongs

to the given class Q. Function $\varrho(t)$ characterizing the degree of mismatch is compared with the allowable error ε in Block 2. Current values of the signal parameters $\{a_n(t)\}_{n=0}^N$ are computed in Block 3. If $\varrho(t) < \varepsilon$ they do not reach the output of the gating circuit (Block 4). If at time t_k the mismatch becomes equal to ε, values of parameters $\{a_n(t_k)\}_{n=0}^N$ are recorded in Block 5 upon the command k. Parameters a_n may be either the coefficients of the ex-

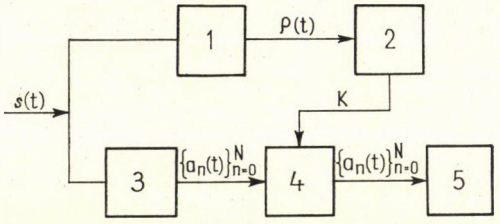


Fig. 1

pansion of the signal $s(t)$ into a series upon a chosen system of basic functions, readings $s(t_k)$ of the signal itself, or parameters of probability distributions, correlation functions, etc. Eventually, it is the aim of study that defines the information to be registered.

The main task is to choose a Block 1 operation algorithm, i.e. synthesize a signal membership estimation system.

Measurement signal $s(t)$ may be represented in the interval $[t_{k-1}, t_k]$ $(k = 1, 2, \ldots)$ as a combination of the signal $q(t) \in Q$ (model) and the discrepancy function $r(t) \in R$. There may be various kinds of combinations of model and discrepancy $\bar{s} = s(q, r)$, but addition — $s = q + r$ — is the most frequent one.

Let the output signal ϱ of a system (Block 1) be related to its input s through an operator A:

$$\varrho = As = A(q, r) .$$

The system will be called invariant with respect to the signal set Q if the following equality:

$$\varrho = A(q, r) = A(0, r) = A(r), t_{k-1} \leq t \leq t_k \tag{2.1}$$

holds for all $q(t) \in Q$.

Discrepancy $r(t)$ characterizes signal deviation $s(t)$ from the model $q(t)$. In the implementation of a compression system the norm of this deviation $\| r \|$ is bounded by a value of Δ which is to be checked during measurements. We can estimate directly, however, only the norm $\| \varrho \| = \varepsilon$. This gives rise to the problem of estimation of Δ through ε.

An invariant system where the following condition holds:

$$|| r || = || \varrho || \tag{2.2}$$

is the best in the terms of accuracy of estimation of $|| r ||$. In practical systems, however, this estimate is not always feasible. Therefore, in the general case, the estimate may be represented by the following inequalities:

$$b_{\min} || \varrho || \leq || r || \leq b_{\max} || \varrho ||. \tag{2.3}$$

Then, in a given class of invariant systems, a system is to be found for which

$$\frac{b_{\max}}{b_{\min}} = \min. \tag{2.4}$$

The basic definitions introduced, the compression problem may be stated as that of synthesis of measurement systems invariant to the given class of signals Q. Discrepancy norm $|| r ||$ characterizing measurement signal deviation from the model q is to be estimated through the output signal norm $|| \varrho ||$. If at a time t_k the value of mismatch $|| \varrho ||$ becomes equal to ε, signal parameters $\{a_n(t_k)\}_{n=0}^{N}$ are recorded; if $|| \varrho || < \varepsilon$, data are neither recorded, nor transmitted. It is this process which is called "information compression".

3. Synthesis of invariant measurement systems through structural signal models

Structural model A of signal $s(t)$ is a combination (by certain rule Φ) of results of signal $s(t)$ transformations by operators A_m [4]:

$$AS = \Phi[A_0(s),\ A_1(s),\ \ldots,\ A_M(s)].$$

Transformations $\{A\}_{m=0}^{M}$ and Φ may be both linear and nonlinear, the former being of greater practical interest.

If the class of operators A_m is linear, norm $|| r ||$ may be estimated through $|| \varrho ||$ as follows. Since $Ar = 0$ only for $r = 0$, it means that operator A should be non-degenerate, i.e. it should have only a single inverse operator A^{-1} equal to $r = A^{-1}\varrho$. Estimate norms of operators A and A'

$$|| A || = \sup_{r} \frac{|| Ar ||}{|| r ||} \geq \frac{|| Ar ||}{|| r ||}, \quad (r \neq 0) \tag{3.1}$$

$$|| A^{-1} || = \sup_{\varrho} \frac{|| A_{\varrho}^{-1} ||}{|| \varrho ||} \geq \frac{|| A_{\varrho}^{-1} ||}{|| \varrho ||}, \quad (\varrho \neq 0). \tag{3.2}$$

Taking into account that $|| r || = \Delta$ and $|| \varrho || = \varepsilon$, (3.1) and (3.2) imply that

$$\frac{\varepsilon}{|| A ||} \leq \Delta \leq \varepsilon || A^{-1} || . \tag{3.3}$$

By (2.4) operators A and A^{-1} should satisfy the following condition:

$$\frac{b_{\max}}{b_{\min}} = || A || \cdot || A^{-1} || = \min \tag{3.4}$$

i.e. the product of the direct and inverse operator norms should be minimal in the chosen class of transformations.

The choice of operators A_m depends on the class of signals ϱ, random noise, implementation complexity, etc. Therefore, additional restrictions should be introduced for the discussion below. As an example, take synthesis of adaptive signal sampling algorithms based upon signal structural models.

Adaptive signal sampling is a powerful means of measurement information compression. In doing so, the class of algebraic polynomials defined over a finite interval

$$q(\tau) = \sum_{n=0}^{N} a_n \tau^n \quad 0 \leq \tau \leq t$$

is used as model. For sake of simplicity of the physical realization, the degree of the approximating polynomial is chosen to be 1, i.e. $q(\tau) = a_0 + a_1 \tau$. Assuming that the approximating straight line goes through the origin $q(0)$, without loss of generality the synthesis may be considered for the following signal model:

$$q(\tau) = a_1 \tau. \tag{3.5}$$

The problem lies in the design of a system invariant to the class of linear signals of type (3.5).

To make the choice of operators A_m, possible additional restrictions should be introduced. Let $\xi(\tau)$ be a stationary white noise; $x = s + \xi$, and $s(\tau)$ be representable as a power series

$$s(\tau) = \sum_{n} a_n \tau^n \quad (0 \leq \tau \leq t) .$$

Stipulate that operators A_m guarantee the minimum r.m.s. error at the invariant system output for zero dynamical error of reproduction of the m-th term τ^m of the initial signal $s(\tau)$, i.e.

$$\sigma^2 = \min \quad \text{for} \quad t^m - A_m(\tau^m) = 0 . \tag{3.6}$$

It is shown in [6] that operators A_m satisfying (3.6) have the following form

$$A_m(s) = \frac{2m+1}{t^{m+1}} \int_0^t \tau^m(s)\, d\tau . \tag{3.7}$$

We are, thus, going to look for an invariant system in the class

$$\varrho = A(\varrho) = \sum_{m=0}^{M} C_m A_m(s) \tag{3.8}$$

where $A_m(s)$ are operators of the type (3.7). (The unit operator is assumed to be included in the class.)

The following relations

$$\begin{aligned} &\varrho = Aq = a_1 A(\tau) = 0 \\ &\sum_{m=0}^{M} C_m A_m(s = \tau) = 0 \end{aligned} \tag{3.9}$$

serve as necessary conditions for implementation of invariant system. But to find all coefficients C_m ($m = 0, 1, \ldots, M$), $M + 1$ equations should be written, i.e. M more restrictions should be imposed. These may be of various kinds: smoothing of certain components of signal t^m, filtering of noise of a certain class, etc. The value of M should be limited for the sake of simplicity of the invariant system implementation

$$A(s) = s(t) + C_m \frac{2m+1}{t^{m+1}} \int_0^t \tau^m \varrho(\tau)\, d\tau . \tag{3.10}$$

For $s(0) = 0$ and $q(\tau) = \tau$ obtain

$$A(q) = t + C_m \frac{2m+1}{t^{m+1}} \int_0^t \tau^{m+1}\, d\tau$$

thus

$$C_m = -\frac{m+2}{2m+1} .$$

Substituting the expression for C_m in (3.10) obtain

$$\varrho = s(t) - \frac{m+2}{2m+1} \int_0^t \tau^m s(\tau)\, d\tau \tag{3.11}$$

and taking into account that $s = q + r = t + r$ rewrite it as follows:

$$\varrho = A(r, q) = \tau(t) - \frac{m+2}{t^{m+1}} \int_0^t r(\tau)\, \tau^m\, d\tau. \tag{3.12}$$

Now estimates of the type (3.3) are to be obtained. Through computation of norms of the direct and inverse operators (A and A^{-1}) find the following inequality

$$\frac{m+1}{m+2} \| \varrho \| \leq \| r \| \leq \| \varrho \| (m+3).$$

Let us determine m satisfying (3.4):

$$\frac{b_{\max}}{b_{\min}} = \frac{(m+3)(m+2)}{(m+1)} = \min.$$

By means of simple calculations one can find that $m = 0.41$. By taking $m = 0$, obtain through this round-off the following invariant system operation algorithm:

$$\varrho(t) = s(t) - \frac{2}{t} \int_0^t s(\tau)\, d\tau. \tag{3.13}$$

Norm $\| \varrho \| = \varepsilon$ should be chosen on the basis of the following condition

$$\varepsilon = \frac{\| r \|}{m+3} = \frac{\Delta}{3}. \tag{3.14}$$

Error norm $\| r \|$ will not, in this case, exceed the allowable value Δ. Estimate (3.14) being approximate, the compression ratio becomes lower. But algorithm (3.13) is readily implemented and has smoothing action.

Conclusions

Design of data compression systems may be reduced to synthesis of measuring systems invariant to a given class of signals. This problem is solved effectively by structural signal models. The technique discussed enables one to synthesize algorithms on the basis of given restrictions.

To solve the synthesis problem, one has to choose a class of signal procession operators. Technically and mathematically speaking, linear structural models lead to the simplest solutions. In this case requirements of invariance are stated as conditions for structural model coefficients resulting in zero or sufficiently small system output for a given signal model.

References

1. Современные методы проектирования систем автоматического управления (под редакцией *Б. Н. Петрова*). Машиностроение, М., 1967.
2. *Окунев Ю. Б.* Системы связи, инвариантные к помехам. «Радиотехника», *8*, 1971.
3. *Уланов Г. М.* Статистические и информационные вопросы управления по возмущению. «Энергия», М., 1970.
4. *Виттих В. А., Заездный А. М.* Сжатие измерительной информации на основе использования ее структурных свойств. Известия ВУЗов, раздел «Приборостроение», **XIV**, *8*, 1971.
5. *Виттих В. А., Гинзбург А. Н.* Некоторые общие вопросы теории сокращенного представления измерительных сигналов. Автометрия, *3*, 1968.
6. *Солодовников В. В., Семенов В. В.* Синтез и оценка точности фильтров, самонастраивающихся по входному сигналу. Сборник «Автоматическое управление и вычислительная техника», *9*, «Машиностроение», М., 1968.

Синтез алгоритмов сжатия данных

В. А. ВИТТИХ

(Куйбышев)

Построение систем сжатия данных сводится к синтезу измерительных систем, инвариантных к заданному классу сигналов. Эта задача решается на основе использования структурных моделей сигналов. Предлагаемая методика позволяет производить синтез алгоритмов сжатия, исходя из заданных условий. Линейные структурные модели обеспечивают наибольшую простоту решений в техническом и математическом смысле.

Рассматривается пример синтеза линейного алгоритма адаптивной дискретизации сигналов.

Problems of Control and Information Theory, Vol. 2 (3—4), pp. 243—256 (1973)

ЭНТРОПИЯ, ИНФОРМАЦИЯ И РАБОТА В КВАНТОВОЙ СТАТИСТИКЕ

В. В. МИТЮГОВ

(Поступила в редакцию 29 октября 1971 г.)

(Горький)

Обсуждается роль необратимых процессов при разветвлении сигнала. Показано, что размножение (дупликация) информации связано с возрастанием энтропии и сопровождается потерями термодинамической работы.

1. Введение

Многочисленные исследования взаимосвязи между физическим и информационным содержаниями понятия энтропии [1—7], берущие начало от работ Бриллюэна и Габора, привели к пониманию того факта, что энтропия, определенная как функционал $H_{кл} = -\sum_i p_i \ln p_i$ распределения p_i вероятности состояний физической системы, есть одновременно и мера неопределенности микросостояния этой системы, и характеристика степени ее неравновесности, и мера количества информации, которое может быть передано при заданных макро-ограничениях, если рассматриваемая система используется в качестве материального носителя сигналов. Варьируя этот функционал при заданной средней энергии, можно показать, что он достигает максимума на равновесном распределении, а этот максимум имеет смысл термодинамической энтропии.

Таким образом, единство различных аспектов понятия энтропии в настоящее время представляется очевидным для широкого круга задач, в которых состояние физической системы может быть описано с помощью распределения вероятности p_i, т. е. для классической статистики и тех квантовых задач, в рамках которых нет необходимости учитывать некоммутативность различных переменных, а квантовая природа материи проявляется лишь в дискретности энергетического спектра.

Однако ситуация не столь ясна при последовательно квантовом описании, когда распределение p_i заменяется оператором плотности, а упомянутый функционал $H_{кл}$ — его естественным обобщением $H = -Sp\,\hat{\varrho} \ln \hat{\varrho}$, инвариантным по отношению к унитарным преобразованиям. Различные свойства энтропии, определенной через оператор плотности, исследовались еще Фон-Нейманом [8] и Эльзасером [9]. В сравнительно недавних работах [5—7]

выяснена роль этого функционала в задачах физической теории информации. Кроме того, величина H так же, как и $H_{кл}$ достигает максимума при заданной средней энергии $E = Sp\,\hat{\varrho}\,\hat{E}$ на распределении Гиббса, что позволяет связать формализм матрицы плотности с равновесной термодинамикой [10]. Тем не менее, при обсуждении свойств неравновесных систем часто отказываются отождествлять H с физической энтропией, оставляя за этой величиной лишь ее информационный смысл.

Основной причиной подобного подхода является тот факт, что замкнутость физической системы не изменяется со временем, то есть не подчиняется H — теореме физической кинетики. С другой стороны, если систему можно разбить на большое число N подсистем, то величина $-\sum_{\lambda=1}^{N} Sp\,\hat{\varrho}^{(\lambda)} \ln \hat{\varrho}^{(\lambda)}$ может со временем возрастать [4], что делает отождествление именно этой величины с физической энтропией более резонным, хотя это отождествление и не приводит к ясной и единой концепции, поскольку при этом остается неясным, что же называть «истинной» энтропией сложной физической системы.

Тем не менее, выяснение упомянутых вопросов необходимо для правильного понимания природы необратимых процессов, для рассмотрения потерь информации и работы, обусловленных возрастанием физической энтропии. При решении вопроса об универсальности того или иного из упомянутых определений энтропии, наряду с его информационным содержанием, критерием следует считать вопрос о максимальной работе, которую можно получить за счет неравновесности системы, поскольку именно этот вопрос и исторически, и с точки зрения логической структуры теории является краеугольным камнем термодинамики.

2. Обмен работой и информацией при взаимодействиях

Представляется очевидным, что при взаимодействии физических систем между ними происходит обмен энергией и информацией. Постараемся проанализировать эти соображения более строго, используя определения средней энергии, работы и информации, принятые в статистической механике и квантовой теории передачи сигналов. При этом будем рассматривать лишь взаимодействия типа столкновений, в результате которых суммарные значения всех аддитивных интегралов движения систем не изменяются, т. к., во-первых, процессы рассеяния света, релаксации при столкновениях частиц и т. д., которые нас интересуют, носят именно такой характер, и во-вторых, что более важно с принципиальной точки зрения, и те задачи, в которых имеет место «включение» и «выключение» взаимодействия, способное изменить значение суммарной энергии, также можно свести к рассмотрению столкно-

вений, если провести последовательно квантовое описание физического фактора, обуславливающего это «включение».

Рассмотрим процесс переноса информации и работы при взаимодействии двух систем, каждая из которых может содержать любое число степеней свободы и быть, в свою очередь, разбитой на подсистемы. Пусть до столкновения их состояния были статистически независимы и описывались операторами плотности $\hat{\varrho}^{(1)}$ и $\hat{\varrho}^{(2)}$ с энергиями и энтропиями

$$E_\lambda = Sp\,\hat{\varrho}^{(\lambda)}\widehat{E}_\lambda\,;\; H_\lambda = -\,Sp\,\hat{\varrho}^{(\lambda)}\ln\hat{\varrho}^{(\lambda)}\,;\quad \lambda = 1,2\,, \tag{2.1}$$

где $\widehat{E}_\lambda$ — гамильтонианы систем без взаимодействия. С течением времени оператор плотности сложной системы $\hat{\varrho}_{(t)}$ изменяется по закону

$$\hat{\varrho}_{(t)} = \widehat{U}_{(t)}^{-1}\,\hat{\varrho}^{(1)}\,\hat{\varrho}^{(2)}\,\widehat{U}_{(t)}\,, \tag{2.2}$$

где $\widehat{U}^{(t)}$ — эволюционный унитарный оператор. Например, в представлении Шредингера при наличии взаимодействия $\widehat{V}_{(t)}$ он выражается через оператор полной энергии $\widehat{E} = \widehat{E}_1 + \widehat{E}_2 + \widehat{V}_{(t)}$ соотношением $\widehat{U}_{(t)} = \exp\,(-i\hbar^{-1}\widehat{E}t)$. Энергии и энтропии систем, вообще говоря, тоже изменяются

$$E_{\lambda(t)} = Sp\,\hat{\varrho}_{(t)}^{(\lambda)}\,\widehat{E}_\lambda\,;\; H_{\lambda(t)} = -\,Sp\,\hat{\varrho}_{(t)}^{(\lambda)}\ln\hat{\varrho}_{(t)}^{(\lambda)}\,, \tag{2.3}$$

где $\hat{\varrho}_{(t)}^{(1)} = Sp_y\,\hat{\varrho}_{(t)}$, $\hat{\varrho}_{(t)}^{(2)} = Sp_x\,\hat{\varrho}_{(t)}$, а под x и y понимаются полные наборы переменных первой и второй систем соответственно. По окончании взаимодействия справедливы соотношения

$$E_{1(t)} + E_{2(t)} = E = E_1 + E_2\,;\; H_{1(t)} + H_{2(t)} \geq H = H_1 + H_2\,. \tag{2.4}$$

Доказательство последнего неравенства, представляющего собой квантовое обобщение хорошо известного соотношения шенноновской теории [11], легко получается с использованием леммы Клейна [12].

Для определенности будем считать, что система 1 служит источником информации и работы по отношению ко второй системе, получающей их при взаимодействии.

Сначала разберем вопрос об обмене работой между подсистемами 1 и 2, которые в дальнейшем будем называть соответственно источником и адоптером.[1] Для этого предположим, что при $t = 0$ адоптер находится в состоянии с энергией E_2 и нулевой энтропией, а под работой R, получаемой им в процессе взаимодействия с источником, будем понимать увеличение энергии $E_{2(t)} - E_2$, не сопровождающееся возрастанием энтропии. Если $H_{\max}^{(1)}(E_1)$ — макси-

[1] От англ. to adopt — принимать. Мы воздерживаемся от употребления термина приемник (экв. англ. receiver), поскольку под последним в квантовой теории связи понимается измерительный прибор в смысле Фон-Неймана.

мальная энтропия источника при заданной средней энергии, то с помощью (2.4) нетрудно показать, что

$$R \leq E_1 - E^{(1)}_{\min}(H_1)\,, \tag{2.5}$$

где E_1 и H_1 — средняя энергия и энтропия состояния источника при $t = 0$, а $E^{(1)}_{\min}(H)$ — функция,[2] обратная $H^{(1)}_{\max}(E)$. Если этому состоянию соответствует равновесный оператор плотности вида

$$\varrho_{EE'} = Z^{-1} \exp(-E/T)\,\delta_{EE'}. \quad \text{то} \quad R = 0\,.$$

В случае, когда источник представляет собой классическое однородное тело, а неравновесность его состояния связана с отклонением температуры в каком-то малом объеме v от среднего значения T, с помощью (2.5) приходим к известной формуле

$$R \leq {}_V C_V\, T \ln(1 + \Delta T/T)\,. \tag{2.6}$$

Продолжая рассмотрение подобных примеров, нетрудно показать, что операторное определение энтропии с использованием неравенства (2.4) естественным образом приводит к известной в термодинамике теореме о максимальной работе. Заметим что вопрос о достижимости равенства в соотношении (2.5) является весьма тонким и, по-видимому, не может быть решен без конкретизации энергетических спектров систем и взаимодействия между ними. Однако, отвлекаясь от этого обстоятельства, для удобства дальнейших рассуждений введем величину $R_{\max}$, соответствующую знаку равенства в (2.5).

Выведем одно полезное следствие соотношения (2.5). Будем рассматривать сложную систему, состоящую из источника и адоптера, в качестве источника работы, которая может быть извлечена некоторым внешним устройством. Если системы 1 и 2 в прошлом не взаимодействовали и полная энтропия $H = H_1 + H_2$, то максимальное значение этой работы $R_{\max} = E - E_{\min}(H)$, где $E_{\min}(H)$ — равновесная энергия сложной системы, отвечающая значению энтропии $H_1 + H_2$. Кроме того, можно говорить о количествах работы R_λ, содержащихся по отношению к внешнему устройству в каждой из систем в отдельности. Тогда с помощью (2.5) нетрудно показать, что

$$R_{1\max}(t) + R_{2\max}(t) = R_{\max} + E_{\min}(H) - E^{(1)}_{\min}(H_1) - E^{(2)}_{\min}(H_2)\,. \tag{2.7}$$

Ниже используем это соотношение при обсуждении потерь работы и переноса информации в необратимых процессах.

Теперь перейдем к вопросу об обмене информацией при взаимодействии систем. Пусть сигналами служат состояния источника, описываемые опера-

[2] В системах с конечным числом уровней эта функция может оказаться двузначной; тогда под $E_{\min}(H)$ следует понимать наименьшее значение, отвечающее тепловому равновесию.

торами плотности $\hat{\varrho}_i^{(1)}$ и посылаемые с вероятностями p_i. Тогда шенноновское количество информации, которое может быть получено в результате измерений полного набора переменных источника при $t = 0$

$$I_x = -\sum_{i,x} p(x|i)\, p_i \ln \sum_i p(x|i)\, p_i + \sum_i p_i \sum_x p(x|i) \ln p(x|i), \qquad (2.8)$$

где $p(x|i)$ — диагональные элементы оператора $\hat{\varrho}_i^{(1)}$ в x-представлении. Количеством информации I_1, заключенным в состояниях источника, можно называть $\mathrm{Sup}_x I_x$ по всевозможным полным наборам x. Естественно, измерение в момент $t = 0$ дополнительно к x переменных адоптера y не дает дополнительной информации о сигнале.

В результате взаимодействия между системами ситуация может измениться. Действительно, в момент $t > 0$ и оператор плотности каждого из сигналов для сложной системы $\hat{\varrho}_{i(t)} = \hat{U}_{(t)}^{-1}\, \hat{\varrho}_i^{(1)}\, \hat{U}_{(t)}$, и оператор их ансамбля $\hat{\varrho}_{(t)} = \sum_i p_i \hat{\varrho}_{i(t)}$ не распадаются, вообще говоря, на произведение $\hat{\varrho}_{(t)}^{(1)}\, \hat{\varrho}_{(t)}^{(2)}$. Поэтому любые динамические переменные x и y, характеризующие состояния источника и адоптера, будет статистически связаны, а их совместное вероятностное распределение может быть найдено путем вычисления диагональных элементов $\varrho_{xy,xy}(t)$ оператора плотности всей системы в соответствующем представлении. Следует отметить, что энтропия связи между переменными x и y

$$H_{\text{св}}(x, y) = -\sum_x \varrho_{xx}^{(1)} \ln \varrho_{xx}^{(1)} - \sum_y \varrho_{yy}^{(2)} \ln \varrho_{yy}^{(2)} + \sum_{x,y} \varrho_{xy,xy} \ln \varrho_{xy,xy} \qquad (2.9)$$

характеризует степень статистической необратимости процесса взаимодействия между источником и адоптером, но отнюдь не является мерой количества переданной информации. Действительно, даже при $H_1 = 0$, когда источник находится в определенном квантовом состоянии и никаких сообщений не посылается, в результате взаимодействия $H_{\text{св}}(x, y)$ может оказаться отличной от нуля.

Для вычисления количества информации $I_y(t)$, которое может быть извлечено путем измерения переменных y адоптера, заметим, что при посылке i-го сигнала и последующего взаимодействия с источником адоптер оказывается в состоянии с оператором плотности $\hat{\varrho}_i^{(2)}(t) = Sp_x\, \hat{\varrho}_i(t)$, поэтому

$$I_y(t) = H(y, t) - \overline{H}_i(y, t) =$$

$$= -\sum_y \varrho_{yy}^{(2)}(t) \ln \varrho_{yy}^{(2)}(t) + \sum_i p_i \sum_y \varrho_{iyy}^{(2)}(t) \ln \varrho_{iyy}^{(2)}(t)\,. \qquad (2.10)$$

Точно таким же образом количество информации $I_x(t)$, получаемое при измерении переменных источника в момент t, выражается через элементы матриц $\varrho_{ixx}^{(1)}$.

Посмотрим, каким закономерностям подчиняется перераспределение информации между источником и адоптером при взаимодействии. Для этого сложим величины $I_x(t)$ и $I_y(t)$, а затем воспользуемся очевидным неравенством $H_i(x,t) + H_i(y,t) \geq H_i(x,y,t)$, откуда

$$I_x(t) + I_y(t) \leq H_1(x,t) + H_2(y,t) - \overline{H}_i(x,y,t)\,. \tag{2.11}$$

Для того, чтобы оценить $\overline{H}_i(x, y, t)$ заметим, что набор x, y в общем случае не является тем набором переменных, на котором достигается $\mathrm{Sup}_{x,y} I_{x,y} = I$,[3] поэтому $I_{xy}(t) = H(x,y,t) - H_i(x,y,t) \leq I$, что дает

$$I_x(t) + I_y(t) \leq I + H(x,t) + H(y,t) - H(x,y,t)\,. \tag{2.12}$$

Неравенство (2.12) уже не содержит в явном виде каких-либо характеристик состояний источника, используемых в качестве сигналов, но предполагает заданными наборы переменных x и y, с измерением которых связана процедура извлечения информации. Для того, чтобы избавиться от этого ограничения, что было бы удобно при обсуждении общих закономерностей переноса информации при взаимодействиях, воспользуемся сформулированным ниже утверждением.

Теорема

Энтропия связи любых переменных x и y различных систем $H_{\text{св}}(x, y) = H(x) + H(y) - H(x, y)$ не превышает величины $H_1 + H_2 - H$.

Доказательство приведено в Приложении.

Используя это неравенство совместно с (2.12), найдем

$$I_1(t) + I_2(t) \leq I + H_1(t) + H_2(t) - H\,. \tag{2.13}$$

Воспользуемся соотношениями (2.13) и (2.7) для анализа процессов переноса энергии, работы и информации при взаимодействиях. Рассмотрим случай $H_2 = 0$, который физически означает, что устройство, предназначенное для аккумуляции полезной работы или информации, тем или иным способом избавлено от влияния посторонних шумов и флюктуаций.

Прежде всего отметим, что получение адоптером работы и информации от источника в некотором смысле взаимно исключают друг друга. Действительно, при получении максимальной работы, отвечающей знаку равенства в формуле (2.5), адоптер по окончании взаимодействия оказывается в состоя-

[3] Легко показать, что $I(x,y,t) \leq H_1$, однако вопрос о существовании на первый взгляд очевидного неравенства $I_{xy}(t) \leq \mathrm{Sup}_x I_x(0)$ до сих пор является дискуссионным, хотя ег орешение имело бы важное принципиальное значение в теории косвенных измерений.

нии с $H_2(t) = 0$, что с помощью (2.10) дает $I_y(t) = 0$, поскольку при этом операторы $\hat{\varrho}_i^{(2)}(t)$ не зависят от того, какой из сигналов был послан.

Теперь рассмотрим промежуточный случай $H_1(t)$, $H_2(t) \neq 0$, для чего обратимся к формулам (2.7) и (2.13). Во-первых, если взаимодействие удовлетворяет условию статистической обратимости $H_1(t) + H_2(t) = H$, то из (2.13) следует $I_1(t) + I_2(t) \leq I$, то есть сумма количеств информации, извлекаемых из частей, не может превысить полной информации, содержащейся в системе. Для анализа соотношения (2.7) удобно ввести эффективные температуры $T_1^{эфф}(t)$, $T_2^{эфф}(t)$, $T^{эфф}$, однозначно связанные с энтропиями $H_1(t)$, $H_2(t)$ и H посредством соотношений

$$H_\lambda(t) = \int_0^{T_\lambda^{эфф}} \frac{C_\lambda(T')}{T'}\, dT' \,;\; H = \int_0^{T^{эфф}} \frac{(C_1(T') + C_2(T'))}{T'}\, dT', \qquad (2.14)$$

где C_1 и C_2 — соответственно теплоемкости источника и адоптера. Если $T_1^{эфф} \neq T_2^{эфф}$, то $E_{\min}(H) < E_{\min}^{(1)}(H_1) + E_{\min}^{(2)}(E_2)$ потому, что установление равновесия в полной системе связано с выравниванием температур $T_1^{эфф}$ и $T_2^{эфф}$. Если же $T_1^{эфф} = T_2^{эфф}$, то при соблюдении условия обратимости вместо (2.7) получим

$$R_{1\max} + R_{2\max} = R_{\max}.$$

Теперь проанализируем поведение количеств работы и информации в необратимых процессах. Как видно из неравенства (2.13), возможность размножения сведений (репродукции), когда $I_1(t) + I_2(t) > I$, связана именно с возрастанием энтропии и существует лишь при необратимом характере взаимодействия источника с адоптером. Напротив, суммарное количество работы $R_{1\max} + R_{2\max}$, содержащейся в каждой из частей в отдельности, при этом падает, поскольку даже в случае $T_1^{эфф} = T_2^{эфф}$ справедливо $E_{\min}(H_1 + H_2) < E_{\min}^{(1)}(H_1) + E_{\min}^{(2)}(H_2)$, если $H_1 + H_2 > H$. Найдем связь между величинами потерь суммарной работы и информации. Вводя обозначения $\Delta I = I_1 + I_2 - I$, $\Delta R = R_{1\max} + R_{2\max} - R_{\max}$, дифференцируя неравенства (2.7), (2.13) и замечая, что $(dE_{\min}/dH) = T^{эфф}$, найдем

$$-d(\Delta R) = T_1^{эфф}\, dH_1 + T_2^{эфф}\, dH_2 \qquad (2.15)$$

$$d(\Delta I) \leq dH_1 + dH_2.$$

Если $T_1^{эфф} = T_2^{эфф}$ из неравенств (2.15) следует

$$-d(\Delta R) \geq T^{эфф}\, d(\Delta I), \qquad (2.16)$$

откуда видно, что репродукция информации связана с возрастанием энтропии и неизбежно оплачивается потерей суммы количеств полезной работы, содержащихся по отдельности в каждой из взаимодействующих систем.

3. Репродуцирование информации и потеря работы в многоканальных системах

Важной проблемой, неоднократно привлекавшей внимание не только физиков и математиков, но и философов, считается следующая: является ли потеря термодинамической работы при релаксации за счет столкновений частиц принципиально неустранимой, или же она обусловлена просто нашей неспособностью вычислить фазовые траектории всех частиц при заданных начальных условиях. Не претендуя на окончательное решение этого вопроса, обсудим его в связи с проблемой статистической необратимости, сопровождающей процесс разветвления сигналов.

Рассмотрим сложную физическую систему, состоящую из N подсистем, состояния которых при $t = 0$ статистически независимы и характеризуются значениями средней энергии и энтропии E_λ, H_λ $(1 \leq \lambda \leq N)$. В результате столкновений между подсистемами значения полной энергии $E = \sum_\lambda E_\lambda$ и энтропии $H = \sum_\lambda H_\lambda$ всей системы не изменяются, однако сумма $\sum_\lambda H_\lambda(t) \geq H$, вообще говоря, возрастает и в пределе при достаточно длительном перемешивании и $N \gg 1$ может достичь значения $H_{\max}(E)$, соответствующего статистическому равновесию.

Попытаемся выяснить, происходит ли в процессе релаксации, протекающей за счет внутренних столкновений в системе, потеря полезной работы и какая из величин H или $\sum_\lambda H_\lambda(t)$ при этом играет решающую роль. Соотношения (2.4) на любой стадии эволюции состояния системы не препятствуют получению работы $R_{\max} = E - E_{\min}(H)$; по крайней мере для каждого конечного N можно придумать достаточно сложный способ воздействия на систему, по эффективности приближающийся к этому пределу.

Однако не следует забывать, что все выводы статистической термодинамики по вопросу о максимальной работе развиты применительно к устройствам, осуществляющим макроскопическое воздействие на систему и поэтому основаны на рассмотрении «одночастичных» функций распределения или матриц плотности, игнорирующем корреляции между подсистемами. К устройствам упомянутого типа принадлежат все системы и аппараты, применяемые в настоящее время для извлечения полезной работы из неравновесных сред: например, поршень машины Карно, осуществляющий давление на рабочий объем газа; лазер, использующий для генерации монохроматического света

спектральные свойства излучения подкачки, но игнорирующий корреляции в спектре этого излучения и т. д. При таком подходе становится несущественным, произошла релаксация за счет внутренних столкновений в системе или за счет ее взаимодействия с подходящим термостатом. Поэтому работа $R^{(T)}$, извлекаемая устройствами указанного класса, связана лишь с неравновесностью, заключенной в одночастичных матрицах плотности и лимитируется неравенством

$$R_{(t)}^{(T)} \leq E - E_{\min}\left(\sum_{\lambda} H_{\lambda}(t)\right). \tag{3.1}$$

В том случае, когда состояния одной или нескольких подсистем при $t = 0$ содержат информацию о каком-либо сигнале, внутренние взаимодействия в системе за достаточно длительное время приводят к распределению этой информации по всем частям, т. е. к разветвлению сигнала. При этом количества информации $I_{\lambda}(t)$, содержащиеся в каждой из подсистем, связаны неравенством, непосредственно обобщающим (2.13):

$$\sum_{\lambda} I_{\lambda}(t) \leq I - H + \sum_{\lambda} H_{\lambda}(t), \tag{3.2}$$

где I — по-прежнему Sup количества информации по полным наборам переменных всей системы.

Разумеется, под рассматриваемой совокупностью N подсистем не следует непременно понимать систему частиц; этот пример лишь наиболее традиционен при обсуждении процессов релаксации. В теории передачи информации более естественно отождествлять эти подсистемы с ветвями многоканальной сети, взаимодействие между которыми обуславливает перенос информации и энергии при разветвлении сигнала.

Рассмотрим случай, когда при $t = 0$ носителем информации является лишь одна из подсистем (припишем ей номер 1), соответствующая каналу, по которому сигнал поступает в сеть. Посмотрим, какое количество информации может содержаться в каждой из ветвей системы в произвольный момент $t > 0$ при различных предположениях.

Во-первых, пусть исходный ансамбль сигналов обладает статистическими свойствами равновесного, чему соответствует $E_1 = E_{\min}^{(1)}(H_1)$ а остальные подсистемы в начальный момент находятся в состояниях с нулевыми энергиями и энтропиями, то есть не содержат собственных шумов. Учитывая, что $I \leq H_1$ (см. [18]), для этого случая с помощью (3.2) нетрудно получить

$$\sum_{\lambda} I_{\lambda}(t) \leq H_{\max}(E_1), \tag{3.3}$$

где $H_{\max}(E_1)$ — равновесная энтропия всей системы, соответствующая энергии E_1. Таким образом, при отсутствии в многоканальной системе дополни-

тельных помех неравенство (3.2) по сути не добавляет новых ограничений к очевидному $I \leq H$. Например, если все N ветвей представляют собой линейные осцилляторы одинаковой частоты ω, то в квазиклассическом пределе $E_1 \gg \hbar\omega N$ имеем

$$H^{(1)}_{\max}(E) = \ln eE - \ln \hbar\omega, \quad H_{\max}(E) = N(\ln eE - \ln \hbar\omega N),$$

и вместо (3.3) получим

$$\sum_\lambda I_\lambda(t)/N \leq \ln \frac{eE}{\hbar\omega} - \ln N, \tag{3.4}$$

т. е. при любом конечном N возможна асимптотически полная репродукция, когда в каждой ветви в принципе может содержаться практически полная информация о сигнале.

Напротив, в квантовом случае $E_1 < \hbar\omega$ имеем $H^{(\lambda)}_{\max}(E_\lambda) \approx (E_\lambda/\hbar\omega) \ln \times$ $\times \hbar\omega\, e/E_\lambda$, поэтому $\sum_\lambda I_\lambda \leq \frac{E_1}{\hbar\omega} \ln \frac{\hbar\omega\, Ne}{E_1}$ и количество информации в среднем на каждой из ветвей не превышает $N^{-1}\left(\frac{E_1}{\hbar\omega} \ln \frac{\hbar\omega\, e}{E_1} + \ln N\right)$, что указывает на неизбежность потерь информации при разветвлении сигнала даже в отсутствие дополнительных помех. Заметим, что если исходному состоянию сообщены те же энтропия и информация, но большая энергия, то (3.3) не препятствует уменьшению этих потерь, откуда можно сделать вывод о целесообразности предварительного усиления квантового сигнала или придания ему нетеплового статистического характера.

Если же в каждой из ветвей многоканальной системы присутствуют тепловые шумы, статистически подобные сигналу, то все эффективные температуры $T_\lambda^{\text{эфф}}$ равны между собой, процесс взаимодействия термодинамически обратим и репродукция информации вообще невозможна, поскольку $\sum_\lambda I_\lambda(t) \leq I$.

Проиллюстрируем обсуждавшиеся выше закономерности на примере физической системы связи, использующей когерентные сигналы. Пусть передатчик создает состояния $< \alpha_\lambda|$ осцилляторов поля внутри полосы частот $(\omega_0 - \Delta/2,\ \omega_0 + \Delta/2)$, посылая их с вероятностями $p(A_\lambda,\ \Phi_\lambda) = D^{-1} A_\lambda \exp \times$ $\times(-A_\lambda^2/2D)$, где $D = 2\pi W/\hbar\omega_0\Delta$, $A_\lambda = |\alpha_\lambda|, \Phi_\lambda = -i \ln(\alpha_\lambda/A_\lambda)$; а W — полная мощность сигнала. Тогда скорость передачи информации r, извлекаемой при одновременном измерении канонических координат и импульсов сигнала (например, методом расщепления пучка [14]), равна $\frac{\Delta}{2\pi} \ln(1 + 2\pi W/\hbar\omega_0\Delta)$. Если поток излучения от передатчика разветвляется по двум одинаковым

нешумящим каналам, то в каждом из них снова оказываются когерентные сигналы [15] с амплитудами $A_{\lambda\nu} = A_\lambda/\sqrt{2}$ ($\nu = 1,2$) и фазами $\Phi_{\lambda\nu} = \Phi_\lambda$, распределенными по закону $p(A_{\lambda\nu}\,\Phi_{\lambda\nu}) = (D/2)^{-1}\,\mathrm{A}_{\lambda\nu}\exp(-A^2_{\lambda\nu}/D)$.

Поэтому поток информации на выходе каждого их этих каналов получается из той же формулы и равен $\dfrac{\Delta}{2\pi}\ln(1+\pi W/\hbar\omega_0\Delta)$. Если же в канале, используемом для отвода половины мощности сигнала содержится тепловой шум, статистически подобный сигналу, то потоки информации r, r_ν соответственно в исходном канале и в выходных ветвях определяются соотношениями (см. [14])

$$r = \frac{\Delta}{2\pi}\ln(1+2\pi W/\hbar\omega_0\Delta)\,;\, r_\nu = \frac{\Delta}{2\pi}\ln\left(\frac{\hbar\omega_0\Delta + 2\pi W}{\hbar\omega_0\Delta + \pi W}\right), \qquad (3.5)$$

из которых видно, что $r_1 + r_2 < r$. Нетрудно заметить, что приведенный пример позволяет проследить все характерные особенности репродукции информации при разветвлении сигнала, следующие из неравенства (3.2) и обсуждавшиеся выше.

В заключение вкратце остановимся на вопросе о взаимосвязи между потерями термодинамической работы и репродукцией информации при взаимодействиях подсистем. Дифференцируя (3.1) и (3.2), вводя, как и прежде, эффективную температуру $T^{\text{эфф}}_{(i)} = (dH_{\max}/dE)_{\sum_\lambda H_{\lambda(i)}}$, найдем

$$-dR^{(T)}_{\max} \geq T^{\text{эфф}}\,d\left(\sum_\lambda I_\lambda - I\right), \qquad (3.6)$$

что при $N = 2$ сводится к (2.16), поскольку при $T^{\text{эфф}}_1 = T^{\text{эфф}}_2$, $R^{(T)}_{\max}$ и $R_{1\max} + R_{2\max}$ из (2.7) совпадают. Таким образом, процесс репродуцирования данных при взаимодействии подсистем неизбежно связан с термодинамически необратимым процессом и сопровождается потерей полезной работы $R^{(T)}$.

Приложение

Докажем сформулированную теорему. Обозначим $\varrho_{nm,n'm'}$ элементы матрицы плотности всей системы в представлении, диагонализирующем $\hat{\varrho}^{(1)}$ и $\hat{\varrho}^{(2)}$, ϱ_i — ее собственные значения, а $\varrho^{(1)}_n$ и $\varrho^{(2)}_n$ — соответственно собственные значения матриц плотности подсистем. Если $V^{(1)}_{nx}$ и $V^{(2)}_{my}$ — унитарные матрицы, связывающие «n,m — и x,y — представления», то энтропия связи любых переменных x и y запишется в виде

$$H_{\text{св}}(x,y) = H(x) + H(y) - H(x,y) = -\sum_x p(x)\ln p(x) -$$

$$-\sum_y p(y)\ln p(y) + \sum_{x,y} p(x,y)\ln p(x,y) \qquad (\text{п. 1})$$

где

$$p(x) = \sum_n \varrho_n^{(1)} |V_{nx}^{(1)}|^2 ; p(y) = \sum_m \varrho_m^{(2)} |V_{my}^{(2)}|^2 ; p(x,y) = \\ = \sum_{n,m} V_{xn}^{(1)} V_{ym}^{(2)} \varrho_{nm,n'm'} V_{xn}^{(1)*} V_{ym}^{(2)*}.$$

Кроме того, введем величину $\tilde{H}_{\text{св}}(x,y)$, отличающуюся от (п. 1) заменой $H(x,y)$ на $\tilde{H}(x,y) = -\sum_{x,y,n,m} \varrho_{nm,nm} |V_{nx}^{(1)} V_{my}^{(2)}|^2 \ln \sum_{m,n} \varrho_{nm,nm} |V_{nx}^{(1)} V_{my}^{(2)}|^2$.

Покажем, что $\tilde{H}_{\text{св}}(x,y) \leq H_{\text{св}}(n,m) = -\sum_n \varrho_n^{(1)} \ln \varrho_n^{(1)} - \sum_m \varrho_m^{(2)} \ln \varrho_m^{(2)} + \\ + \sum_{n,m} \varrho_{nm,nm} \ln \varrho_{nm,nm}$. Записывая совместное распределение вероятностей для четырех случайных величин $\tilde{p}(x, n, m, y) = |V_{nx}^{(1)} V_{my}^{(2)}|^2 \varrho_{nm,nm}$, нетрудно убедиться, что они образуют марковскую последовательность, в которой x и y являются крайними элементами, откуда и следует $\tilde{H}_{\text{св}}(x,y) \leq H_{\text{св}}(n,m)$.

Прежде, чем вернуться к обсуждению свойств связанных систем, рассмотрим следующую абстрактную схему. Пусть ϱ_i — собственные значения матрицы плотности некоторой физической системы, а U_{ik} и V_{kz} — унитарные матрицы, произведение которых $\hat{U}\hat{V} = \hat{S}$ фиксировано. Введем энтропии

$$H(k) = -\sum_k p(k) \ln p(k) ; \quad p(k) = \sum_i \varrho_i |U_{ik}|^2 , \\ H(z) = -\sum_z p(z) \ln p(z) ; \quad p(z) = \sum_i \varrho_i |S_{iz}|^2 , \qquad \text{(п. 2)} \\ \tilde{H}(z) = -\sum_z \tilde{p}(z) \ln \tilde{p}(z) ; \quad \tilde{p}(z) = \sum_{i,k} \delta_i |U_{ik} V_{kz}|^2 , \\ H = -\sum_i \varrho_i \ln \varrho_i$$

и сравним разности $H(z) - H$, $\tilde{H}(z) - H(k)$ в предположении, что $H \leq H(k) \leq H(z)$, а унитарные операторы $\hat{U}$ и $\hat{V}$ принадлежат непрерывной последовательности $\hat{U}_{(\beta)}$, $\hat{V}_{(\beta)}$ $(0 \leq \beta \leq 1$, $\hat{U}_{(0)} = \hat{1}$, $\hat{V}_{(1)} = \hat{1})$, для которой $\partial H(k,\beta)/\partial\beta > 0$. Нетрудно заметить, что $H(z) - H$ не зависит от β, в то время как $\tilde{H}(z) - H(k)$ с изменением β от 0 до 1 убывает от $H(z) - H$ до нуля. Путем непосредственного варьирования можно показать, что при наших предположениях это убывание монотонно, поэтому справедливо неравенство

$$\tilde{H}(z) - H(k) \leq H(z) - H . \qquad \text{(п. 3)}$$

Теперь вернемся к задаче об энтропии связи двух систем. Из (п. 3) следует, что

$$\tilde{H}(x,y) - H(n,m) \leq H(x,y) - H . \qquad \text{(п. 4)}$$

Действительно, унитарная матрица $U_{inm,jn'm'}$, приводящая $\varrho_{nm,n'm'}$ к диагональному виду, определена с точностью до перестановки собственных значений ϱ_i, и всегда может быть выбрана в форме, допускающей ее непрерывную деформацию к единичной с монотонным изменением $H(n, m)$. Аналогичные рассуждения справедливы и для матриц $V^{(1)}_{nx}$, $V^{(2)}_{my}$, диагонализирующих $\varrho^{(1)}_{xx'}$ и $\varrho^{(2)}_{yy'}$, а, следовательно, и для их прямого произведения, осуществляющего переход от x, y-представления к n, m-представлению, что позволяет удовлетворить требованиям, обеспечивающим выполнение (п. 3). Комбинируя (п. 4) с неравенством $\hat{H}_{\text{св}}(x, y) \leq H_{\text{св}}(n, m)$ и учитывая, что $\varrho^{(1)}_{nn'} = \varrho^{(1)}_{n}\,\delta_{nn'}$, $\varrho^{(2)}_{mm'}\,\varrho^{(2)}_{m}\,\delta_{mm'}$, получим

$$H_{\text{св}}(x, y) \leq H_1 + H_2 - \sum_i \varrho_i \ln \varrho_i \,. \qquad \text{(п. 5)}$$

Теорема доказана.

Литература

1. *Gabor, D.*, Comunication theory and physics. Phil. Mag., **41,** 7, (1950) 1161.
2. *Бриллюэн Л.* Наука и теория информации. Физматгиз, М., 1960.
3. *Szilard, L.*, Über die Entropieverminderung in einem termodynamischen System bei Eingriff intelligenter Wesen. Z. Physik, **53,** 5 (1929) 840—856.
4. *Файн В. М., Ханин Я. И.* Квантовая радиофизика. «Сов. радио», М., 1965.
5. *Лебедев Д. С., Левитин Л. Б.* Перенос информации электромагнитным полем. Сб. «Теория передачи информации», «Наука», М., 5—20, 1964.
6. *Боровицкий С. И., Митюгов В. В.* О свойствах неравновесного излучения и термодинамически обратимых преобразователях. Изв. вузов. «Радиофизика», **7,** *5* (1964) 854—864.
7. *Gordon, J. P.*, Quantum effects in communication system. Proc. IRE, **50,** 9 (1962) 1898—1908.
8. *Фон-Нейман И.* Математические основы квантовой механики. «Наука», М., 1964.
9. *Elsasser, W. M.*, Quantum measurement and statistical method. Phys. Rev., **52,** 9 (1937) 987—1001.
10. *Ingarden, R. S.*, Higher-Order Temperatures and Coherence of Sight. Bull. Acad Polon. Sci., Ser. Math. Astr. Phys. **16** (1968) 57—59.
11. *Шеннон К.* Математическая теория связи. Сб. «Работы по теории информации и кибернетике», изд-во Иностр. лит., М., 1963, 433—460.
12. *Стратонович Р. Л.* Количество информации, передаваемое квантовым каналом связи. I, II. Изв. вузов. «Радиофизика», **8** *1* (1965) 115—141.
13. *Левитин Л. Б.* О квантовой мере количества информации. Труды IV конференции по теории кодирования и передачи информации, секц. II, Москва—Ташкент, 1969.
14. *Левитин Л. Б., Митюгов В. В.* Прием когерентных сигналов методом расщепления луча. Тезисы докладов 1 Всесоюзной конференции по проблемам передачи информации лазерным излучением. Киев, 1969.
15. *Глаубер Р.* Сб. Квантовая оптика, квантовая радиофизика, «Мир», М., 1966.

Entropy, information and work in quantum statistics

V. V. MITYUGOV

(Gor'ky)

On the basis of the definitions approved in statistic mechanics and the quantum theory of signal transmission, general laws of energy, information and work transfer at physical interaction are analyzed.

The states of a certain physical system interpreted as a material carrier of information are considered as the signals used for data transmission, while the reception is reduced to measuring the values of a certain dynamic variable x. The diagonal elements of the density matrix of the i-th state (signal) in the x-representation are simply transient probabilities $p(x \mid i) = \varrho_{xx}^{(i)}$ of the channel, which allows to apply Shannon's measure of information and to define the informativity I_x (1.8) of the variable x measurement. The quantity of information contained in the ensemble of the states is defined as $I = \mathrm{Sup}_x I_x$. The evolution of states of two interacting systems is analyzed assuming that originally they had been independent and their energies had been E_λ, ($\lambda = 1,2$) and the ensemble of messages containing the information is connected with the states of the first system only. The inequality (2.13) limiting the sum of the quantities of information in the systems after the interaction is derived. In particular, under the conditions of convertibility, $I_1(t) + I_2(t) \leq I$ is true, hence only redistribution but not reproduction of the information is possible.

Incidentally, the theorem (see the Appendix) stating that the correlation entropy (2.9) of any interacting systems does not exceed the increment of the entropy $H_1(t) + H_2(t) - H$ is proved.

Nonconvertibility of the process of the data reproduction allows to connect it with the losses of the total useful work by means of a differential inequality (2.16). These results are generalized for the case of an arbitrary number N of interacting systems which may be considered as gas molecules and/or branches of a multi-channel net as well. For the case when the net branches do not contain their own noise, an example with a system of oscillators shows that in quasiclasics reproduction may, practically, not be accompanied by losses (3.4), while in the quantum case they are inevitable. If every branch comprises noise which is statistically similar to a signal, as for $N = 2$, only redistribution of the original information is possible, since $\sum_{\lambda=1}^{N} I_\lambda(t) \leq I$. In this case an effective reproduction under interaction takes place, $\sum_\lambda I_\lambda(t) - I > 0$, which must inevitably effect in a loss of the useful work $R^{(T)}$ (3.1), which could be obtained on the account of an unequilibrium of the system state by the proper macroscopic influence.

В. В. Митюгов
Горьковский Государственный университет
СССР, Горький,
просп. Гагарина, 23.

Problems of Control and Information Theory, Vol. 2 (3—4), pp. 257 – 272 (1973)

DISPERSIONAL IDENTIFICATION EQUATION OF NONLINEAR PLANTS

A. L. BUNICH, N. S. RAIBMAN

(Moscow)

(Received March 3, 1972)

1. Introduction

The main result of this paper is dispersional identification of nonlinear plants in a class of models with finite memory T. Thus the result is an extension of the identification equation to the class of noninertial models of [1]. The result allows to obtain the cross regressional function estimation of the optimal model response in relation to the input disturbance.

This estimation error does not exceed the optimal (in terms of mean square deflection) error of linear estimation for a plant, whose input signal has linear autoregression, the dispersional and optimal linear estimation coincide. In a plant whose input and output are stationary in a wide range and are stationarily related, the dispersional weighting function is shown to coincide with the pulse transitional function of the optimal linear model.

In planned experiment for any deterministic plant it is possible to find a test signal such that the mean squared estimation of cross regression function will be as small as desired.

The concept of d-stationary and d-stationarily related random processes is introduced which is an analogue of stationarity in a wide sense for processes with nonlinear regression.

In conclusion examples of nonlinear plants are considered for which the dispersional identification method leads to an accurate result.

The main characteristic of the stochastic relation between two random processes $x(t)$ and $y(t)$ is a cross multiple dispersional function which may be defined as follows. Let T be a finite set of points on the positive semi-axis and $x_{s-\tau}$ a section of function $x(t)$ in the point $s - \tau$, $\tau \in T$, then the conditional mathematical expectation of a section $y(t)$ with respect to sections $x_{t-\tau}$, $\tau \in T$ is a function of sections $x_{s-\tau}$, $\tau \in T$ and of t and s. Since the values of the sections $x_{s-\tau}$, $\tau \in T$ are random variables, the conditional mathematical expectation $M(y_t \mid x_{s-\tau},\ \tau \in T)$ is a random function. The dispersion of this random function is termed multiple cross dispersional function of the random process

$y(t)$ in relation to the random process $x(s)$:

$$d^T_{y\bar{x}}(t, s) = M\{M(y_t/x_{s-\tau}, \tau\in T) - M_{y_t}\}^2 = DM(y_t/x_{s-\tau}, \tau\in T), \qquad (1.1)$$

where M is a symbol of mathematical expectation, and D — symbol of dispersion.

In a similar way we may find the autodispersional function of the process $x(t)$:

$$d^T_{x\bar{x}}(t, s) = M\{M(x_t/x_{s-\tau}, \tau\in T) - M_{x_t}\}^2 = DM(x_t/x_{s-\tau}, \tau\in T). \qquad (1.2)$$

We shall also use the terms generalized cross dispersional function and generalized autodispersional function. A generalized cross dispersional function

$$\Theta^T_{y\bar{x}}(t, \sigma) = \operatorname{cov}\{M(y_t/x_{t-\tau}, \tau\in T) - M_{y_t}; M(x_\sigma/x_{t-\tau}, \tau\in T) - Mx_\sigma\}, \qquad (1.3)$$

where cov is the symbol of covariation. The generalized autodispersional function is

$$\Theta^T_{x\bar{x}}(s,\sigma; t) = \operatorname{cov}\{M(x_s/x_{t-\tau}, \tau\in T) - Mx_s; M(x_\sigma/x_{t-\tau}, \tau\in T) - Mx_\sigma\}. \qquad (1.4)$$

In the case when the set T consists of one point all the above definitions coincide with those in [1]. In all the above definitions we use moments of up to the second order.

The section y_t is termed undispersional with the sections

$$x_{s-\tau}, \tau\in T \quad \text{if} \quad M(y_t/x_{s-\tau}, \tau\in T) - My_t = 0.$$

This may be written in an equivalent form: $d^T_{y\bar{x}}(t, s) = 0$.

Independent sections y_t and $x_{s-\tau}$, $\tau \in T$ are undispersional. The reverse cannot be generally fulfilled, that is the properties of stochastic independence are stronger than undispersionality. On the other hand, undispersionality is stronger than lack of correlation, that is why the use of dispersional characteristics of random processes allows to reveal nonlinear forms of dependence, in fact undetectable by correlation analysis. Let the value of y_t be undispersional with $x_{s-\tau}$, $\tau \in T$ and $\psi(x_{s-\tau}, \tau \in T)$ a certain non-random function of the section $x_{s-\tau}$, $\tau \in T$ then

$$\operatorname{cov}\{y_t, \psi(x_{s-t}, \tau\in T)\} = M\{[\psi(x_{s-\tau}, \tau\in T) - M\psi(x_{s-\tau}, \tau\in T)]\,[M(y_t/x_{s-\tau}, \tau\in T) - My_t]\} = 0. \qquad (1.5)$$

It should be observed that the random function $M(y_t(x_{s-\tau}, \tau \in T) - My_t$ will not vary if we add a non-random function of the argument t to the section y_t. Therefore the process y_t will be assumed in the further discussion centered.

Section y_t can be presented as a sum of a conditional mathematical expectation and a random variable ζ_t, undispersed with $x_{s-\tau}$, $\tau \in T$

$$y_t = M(y_t/x_{s-\tau}, \tau \in T) + \xi_t . \tag{1.6}$$

If $x(t)$ is a random process at the input of the plant and $y(t)$ is its response then the first item in this decomposition will be considered as the useful component of the signal at the output of the plant and ζ_t will be considered as noise. According to Eq. (1.5) the useful components of the signal and noise are uncorrelated.

With regard to Eq. (1.5) and our assumption that the process $y(t)$ is centered, we obtain the following formula for multiple generalized cross dispersional function

$$\Theta^{T}_{y\bar{x}}(t, \sigma) = \operatorname{cov}\{y_t, M(x_\sigma/x_{t-\tau}, \tau \in T)\} = \operatorname{cov}\{x_\sigma, M(y_t/x_{t-\tau}, \tau \in T)\}. \tag{1.7}$$

We shall also use the formula for dispersion decomposition [2]

$$D\zeta_t = MD(\zeta_t/x_{s-\tau}, \tau \in T) + DM(\zeta_t/x_{s-\tau}, \tau \in T), \tag{1.8}$$

which can be obtained by replacing the conditional expectation mean by its analytical expression through conditional distribution densities. The first member in the right-hand side of this decomposition can be interpreted as a measure of the stochasticity of the plant with the input x_t and output ζ_t [2]. The measure of the plant stochasticity vanishes if and only if the output ζ_t is a uniform function of the input signal.

The random process x_t is called d-stationary under the following conditions [4]:

1) The joint distribution function of the sections, does not depend on the time t, that is

$$F_{\bar{x}_t}(u_\tau \tau \in T) = F_{\bar{x}\bar{s}}(u_\tau \tau \in T)$$

where

$$\bar{x}_t = (x_{t-\tau}, \tau \in T), \bar{x}_s = (x_{s-\tau}, \tau \in T). \tag{1.9}$$

2) The regression function depends exclusively on the difference of arguments t and s

$$M(x_t/x_{s-\tau}, \tau \in T) = \varphi_{t-s}(x_{s-\tau}, \tau \in T); M(x_s/x_{t-\tau}, \tau \in T = \varphi_{s-t}(x_{t-\tau}, \tau \in T). \tag{1.10}$$

In the case when the set T consists of one point we shall obtain the definition of a random process, stationary in terms of dispersion (see part 1). We may prove that as in the case where a stationary random process is in terms of dispersion, the mathematical expectation of the process is constant and its autocorrelation function depends only on the difference of arguments. Similar-

ly we may introduce the concept of d-stationarily related random processes and prove that also the cross correlation function of such processes depends only on the difference of variables. The random processes $x(t)$, $y(t)$ are called d-stationarily related if they satisfy conditions 1) and 2) and besides

$$M(y_t/x_{s-\tau}, \tau \in T) = \varphi_{t-s}(x_{s-\tau}, \tau \in T). \tag{1.11}$$

The generalized autodispersional function of d-stationary random processes can be obtained in the following way

$$\Theta^T_{x\bar{x}}(t-s, t-\sigma; t) = \Theta^T_{x\bar{x}}(s, \sigma). \tag{1.12}$$

The independence of the right-hand member of the definition of t follows from the definition of d-stationary relation. Thus the dispersional function of stationarily related random processes is a function of two variables.

As an example for d-stationary random process one can mention any random process, stationarily in the narrow sense. For stationarily related random processes the generalized cross dispersional function can be naturally found by analogy with Eq. (1.12) as

$$\Theta^T_{y\bar{x}}(t, \sigma) = \Theta^T_{y\bar{x}}(0, t-\sigma; t). \tag{1.13}$$

2. A dispersional equation of identification for a model with finite memory

The conditional mathematical expectation may be considered as smoothing characteristic of a random process. Naturally it is more convenient to deal not with random processes themselves but with their regression functions. The role of the plant correlation characteristics in construction of a linear model is played by the dispersional functions of random processes. Thus, instead of the concept of a linear plant it is convenient to introduce the concept of a plant linear on the average. The equation of this plant will be:

$$y_t = \int_{-\infty}^{t} g(t, s)\, x_s\, ds + \zeta_t, \tag{2.1}$$

where noise is undispersional with the input, i.e.

$$M(\zeta_t/x_{t-\tau}, \tau \in T) = 0.$$

The equation may be written in terms of conditional mathematical expectations

$$M(y_t/x_{t-\tau}, \tau \in T) = \int_{-\infty}^{t} g(t, s)\, M(x_s/x_{t-\tau}, \tau \in T)\, ds. \tag{2.2}$$

One can see from Eq. (1.2) that the useful component of the process at the output of a plant linear on the average is found through the weighting function and autoregression of the input signal. We shall call the plants of this kind dispersional models.

Multiply both parts of Eq. (1.2) by $M(x_\sigma \mid x_{t-\tau},\ \tau \in T)$ and use the mathematical expectation in the expression obtained. Then using the definitions of dispersional functions we shall obtain the following dispersional equation

$$\int_{-\infty}^{t} g(t, s)\, \Theta^{T}_{\bar{x}\bar{x}}(s, \sigma; t)\, ds = \Theta^{T}_{\bar{y}\bar{x}}(t, \sigma). \tag{2.3}$$

If the value of the argument t is fixed the equation is an integral Fredholm equation of the first kind with respect to the weighting function $g(t, s)$. This equation is incorrect and numerically solvable by regularization algorithms or by additional physical studies of the nature of a solution that leads to plotting, for the weighting function, a region of correctness and solving this dispersional equation within this region.

The useful component of the random process at the output of the plant may be decomposed into two components

$$M(y_t/x_{t-\tau}, \tau \in T) = (\Phi x)(t) + \zeta_t \tag{2.4}$$

where the nonlinear operator φ is determined by Eq. (2.11) and ζ_t is noise, noncorrelated with x_s, $s < t$. The decomposition (2.4) is equivalent to the dispersional Eq. (2.3). When we multiply both members of (2.4) by the section x_σ and introduce the mathematical expectation, we obtain (2.3). On the other hand Eq. (2.3) may be written as

$$\begin{aligned} &\operatorname{cov}\{M\{y_t - \int_{-\infty}^{t} g(t, s)\, x_s\, ds / x_{t-\tau}, \tau \in T\}, x_\sigma\} \\ &\quad = \operatorname{cov}\{M/y_t(x_{t-\tau}, \tau \in T) - (\Phi x)(t); x_\sigma\} = 0; \end{aligned} \tag{2.5}$$

this is equivalent to the decomposition (2.4).

Let us take time τ, so that $\tau \in T$ then Eq. (2.5) may be written in the following form

$$K_{yx}(t, t-\tau) = \int_{-\infty}^{t} g(t, s)\, K_{xx}(s, t-\tau)\, ds, \quad \tau \in T. \tag{2.6}$$

Suppose that the random processes $x(t)$, $y(t)$ are stationary in the wide sense and are stationarily related and the weighting function $y(t, s)$ depends only on the difference of arguments $t - s$. After substitution of the variable in Eq. (2.6) we obtain

$$K_{y,x}(\tau) = \int_{0}^{\infty} g(z)\, k_x(\tau - z)\, dz, \ \tau \in T. \tag{2.7}$$

If the input signal is deterministic $x_\sigma = M(x_\sigma \mid x_{t-\tau},\ \tau \in T)$ or if $T = (0, \infty)$, then Eq. (2.6) is fair for arbitrary τ and in this case Eqs (2.3) and (2.6) coincide. If the output and input signals are stationarily related and stationary in wide sense, then (2.7) and (2.3) are equivalent and the dispersional weighting function coincides with the weighting function of the linearized model.

If the input signal is stochastic, then the integral equation (2.6) is valid only for the finite set of values $\tau \in T$ and the dispersional weighting function and the weighting function of the plant do not coincide.

Let us see how a dispersional identification equation is obtained using optimality criterion.

In solving the identification problem one should confine oneself to a certain class of input actions. For this purpose we shall take a set of random functions with a specified autoregression function

$$M(x_s/x_{t-\tau},\ \tau \in T) = f(s, t, x_{t-\tau},\ \tau \in T). \tag{2.8}$$

Such a constraint on the nature of the input signal is quite natural. On the other hand, the choice of the signal remains arbitrary to a great extent because the conditional mathematical expectation does not determine unambiguously the sequence of n-dimensional probability density functions ($n = 1, 2, \ldots$) of a random process which is known to be its exhaustive characteristic.

The section linear approximation y_t of the process at the plant output which corresponds to t may be expressed by the integral

$$w_t = \int_{-\infty}^{t} g(t, s)\, x_s\, ds, \tag{2.9}$$

where a factor $g(t, s)$ is the weight with which the section x_s is included into the integral. The upper limit of integration is the observation time t which follows from the physical feasibility. However, we are interested in the estimation of the section y_t in terms of the sections $x_{r-\tau}$, $\tau \in T$. Since the set T is generally the eigensubset of the positive semi-axis and since the best mean squared estimate of the random variable w_t is its conditional mathematical expectation, this estimate is:

$$\hat{y}_t = M(w_t/x_{t-\tau},\ \tau \in T). \tag{2.10}$$

The right-hand member of this equation is a nonlinear operator which influences the input random process x_t

$$(\Phi_x)\,(t) = \int_{-\infty}^{t} g(t, s)\, M(x_s/x_{t-\tau},\ \tau \in T)\, ds. \tag{2.11}$$

Nonlinear operator Φ depends first of all on the autoregression of the random process at the input which, as shown above, is assumed to be a known function, and on the parameter $g(t, s)$. The latter must be chosen so that the mean squared identification error

$$\Delta(g) = D(y_t - \hat{y}_t) \tag{2.12}$$

be minimal $\Delta(g) = \min$.

Before deriving the integral identification equation (2.3) from the optimum criterion, rewrite mean squared functional in a more convenient form. For this purpose let us represent the section y_t in the form of Eq. (1.6).

Noise ζ_t is easily seen to be nondispersed with the section $x_{t-\tau}$, $\tau \in T$ and thus noncorrelated with the useful component of the signal at the output y_t. This is valid for any function of the sections $x_{t-\tau}$, $\tau \in T$. Consequently the dispersion of the sum

$$\Delta(g) = D\{\zeta_t + [M(y_t/x_{t-\tau}, \tau \in T) - \int_{-\infty}^{t} g(t, s)\, M(x_s/x_{t-\tau}, \tau \in T)\, ds]\},$$

is equal to the sum of dispersions of the members

$$\Delta(g) = D\,\zeta_t + DM\{[y_t - \int_{-\infty}^{t} g(t, s)\, x_s\, ds]/x_{t-\tau}, \tau \in T\}. \tag{2.13}$$

Let us introduce a new notation

$$y_t - \int_{-\infty}^{t} g(t, s)\, x_s\, ds = \zeta_t\,, \tag{2.14}$$

then the mean squared identification error is the sum of two members, where the first one does not depend on the weighting parameter $g(t, s)$ and the second one is a cross dispersional function $d^T_{\zeta\bar{x}}(t)$. Therefore the minimization of $\Delta(g)$ leads to the minimization of functional $I(g) = d^T_{\zeta\bar{x}}(t)$ squared over g.

The typical problem of mean squared functional minimization is quite similar to the derivation to the identification integral from the optimality criterion in the correlation theory (see Eq. (2.3)) has been touched in part 1.

The final result formulated in terms of dispersion function leads to the multiple dispersion equation of identification (2.3) whereby to find the weighting function $g(t, s)$.

Thus we obtained the dispersion equation identification (2.2) from the optimality criterion.

3. Dispersion identification error and its comparison with the linearization error

Let $g'(t, s)$ be the weighting function. Then

$$D\{M(y_t/x_{t-\tau}, \tau\in T) - \int_{-\infty}^{t} g'(t, s)\, x_s\, ds\} = D\{M(y_t/x_{t-\tau}, \tau\in T) - \\ - \int_{-\infty}^{t} g'(t, s)\, M(x_s/x_{t-\tau}, \tau\in T) - \zeta_t\} \tag{3.1}$$

where $\zeta_t = \int_{-\infty}^{t} g'(l, s)\, [x_s - M(x_s/x_{t-\tau}, \tau \in T)]\, ds$ is nondispersed with $x_{t-\tau}, \tau\in T$. Consequently the right-hand part of (3.1) may be rewritten

$$D\{M(y_t/x_{t-\tau}, \tau\in T) - \int_{-\infty}^{t} g'(t, s)\, M(x_s/x_{t-\tau})\, \tau\in T) - \zeta_t\} = \\ = DM\{[y_t - \int_{-\infty}^{t} g'(t, s)\, x_s\, ds]/x_{t-\tau}, \tau\in T\} + D\zeta_t .$$

Let $g(t, s)$ be a dispersional weighting function of the plant, then on the strength of the optimality criterion

$$DM\{[y_t - \int_{-\infty}^{t} g(t, s)\, x_s\, ds]/x_{t-\tau}, \tau\in T\} \leq DM\{[y_t - \int_{-\infty}^{t} g'(t, s)\, x_s\, ds]/x_{t-\tau},\ \tau\in T\},$$

because

$$D\{M(y_t/x_{t-\tau}, \tau\in T) - \int_{-\infty}^{t} g'(t, s)\, x_s\, ds\} = DM\{[y_t - \int_{-\infty}^{t} g'(t,s) x_s ds]/x_{t-\tau}, \tau\in T\} + D\zeta_t .$$

The second member in the right-hand part of this equality is nonnegative and the following inequality holds

$$I(g) = DM\{[y_t - \int_{-\infty}^{t} g(t, s)\, x_s\, ds]/x_{t-\tau}, \tau\in T\} \leq \\ \leq D\{M(y_t/x_{t-\tau}, \tau\in T) - \int_{-\infty}^{t} g'(t, s)\, x_s\, ds\}. \tag{3.2}$$

In case of a deterministic plant described by the equation

$$y_t = \varphi(t, x_{t-\tau}, \tau\in T) \tag{3.3}$$

the right-hand part is a mean squared error of optimal linearization in terms of minimal mean squared error and the left-hand part is the dispersional identification error. Consequently for deterministic plants this method of estimation of the plant operator is more accurate than the method of linearization. A similar property is also characteristic of a wide class of stochastic plants described by the equation

$$y_t = \psi(t, x_{t-\tau}, \tau\in T) + \eta_t , \tag{3.4}$$

where noise η_t is nondispersed with x_s, $s \leq t$. In this case

$$A(g) = MD(y_t x_{t-\tau}, \tau \in T) + I(g) \leq MD(y_t/x_{t-\tau}, \tau \in T) + D\{M(y_t/x_{t-\tau}, \tau \in T - \\ - \int_{-\infty}^{t} g'(t, s)\, x_s\, ds\} = D\{y_t - \int_{-\infty}^{t} g'(t, s)\, x_s\, ds\}. \tag{3.5}$$

In a general case one always can choose input random process with such autoregression that the left-hand part of this inequality (3.2) will be as small as desired. Then the relation between the input and the output of a plant is described by the equation $y_t = \psi(x_{t-\tau},\ \tau \in T) + \zeta_t$, where ζ_t is a disturbance. Let us take a certain positive number a such that $a \in T$ and the weighting function

$$g(\tau) = \begin{cases} 1/\varepsilon & a < \tau < a + \varepsilon \\ 0 & \tau \in (a, a + \varepsilon)\,. \end{cases}$$

Feed to the plant input d-stationary random process whose autoregression function satisfies the following condition

$$M(x_{t-a}/x_{t-\tau}, \tau \in T) = \psi(x_{t-\tau}, \tau \in T)\,.$$

Then

$$(\Phi x)(t) = \frac{1}{\varepsilon} \int_{t-a-\varepsilon}^{t-a} M(x_{t-s}/x_{t-\tau}, \tau \in T)\, ds = M(x_{t-a}/x_{t-\tau}, \tau \in T) + o(\varepsilon),$$

where the second item $o(\varepsilon)$ in the right-hand part is infinitely small in relation to ε. Thus $(\Phi x)(t) = \psi(x_{t-\tau},\ \tau \in T) + o(\varepsilon)$. So the cross regression function estimation may be as accurate as desired.

4. A dispersional equation in case of signals at the input with linear autoregression

It is natural to expect that in the case when the random process at the input of the plant has linear regression, the solutions of the dispersional equation and identification equation obtained by minimization of the mean squared error coincide. Let

$$M(x_s/x_{t-\tau}, \tau \in T) = \sum_{\tau \in T} a(t, s, \tau)\, x_{t-\tau}, \tag{4.1}$$

where $a(t, s, \tau)$ are expansion autoregression coefficients which depend on the arguments t and s. For generalized cross dispersional and generalized auto-dispersional function we shall obtain the following equations

$$\Theta^{T}_{y\bar{x}}(t, s) = \operatorname{cov}\{y_t, M(x_s/x_{t-\tau}, \tau \in T)\} = \sum_{\tau \in T} a(t, s, \tau)\, K_{yx}(t, t - \tau)\,, \tag{4.2}$$

$$\Theta^{T}_{x\bar{x}}(\sigma, s, t) = \operatorname{cov}\{x_\sigma,\ M(x_s/x_{t-\tau}, \tau \in T)\} = \sum_{\tau \in T} a(t, s, \tau)\, K_x(\sigma, t - \tau). \tag{4.3}$$

Let $g(t, \sigma)$ be the dispersional model weighting function, then

$$\Theta^{T}_{y\bar{x}}(t, s) - \int_{-\infty}^{t} g(t, \sigma)\, \Theta^{T}_{x\bar{x}}(\sigma, s; t)\, d\sigma =$$

$$= \sum_{\tau \in T} a(t, s, \tau) \left[K_{yx}(t, t-\tau) - \int_{-\infty}^{t} g(t, \sigma)\, K_x(\sigma, t-\tau)\, d\sigma \right]. \qquad (4.4)$$

Because of the dispersional equation (2.3) the left-hand member of (4.4) becomes identically zero at all values of the arguments t and s then,

$$K_{yx}(t, t-\tau) - \int_{-\infty}^{t} g(t, \sigma)\, K_{xx}(\sigma, t-\tau)\, d\sigma = 0. \qquad (4.5)$$

Consequently, if the weighting function $g(t, \sigma)$ satisfies the integral equation (4.5) then according to Eq. (4.4) it satisfies the dispersional equation as well. Thus the random process y_t^* at the output of the dispersional model is $y_t^* = \sum_{\tau \in T} b(t, \tau)\, x_{t-\tau}$, where $b(t, \tau) = \int_{-\infty} g(t, s) a(t, s, \tau)\, ds$. Consequently, the dispersional model responds to random disturbances with linear autoregression as a linear plant and its weighting function coincides with that of the plant.

5. Constructing a model of an actual plant

The random process $y(t)$ at the output of the plant may be introduced as a sum of a useful component and the noise according to (1.6). If the one-dimensional probability density $\frac{1}{\sigma_\zeta} \varphi\left(\frac{\zeta}{\sigma_\zeta}\right)$ of the noise is assumed and $M(y_t/x_{t-\tau_1}, \tau \in T)$, $D(x_t/x_{t-\tau_1}, \tau \in T)$ are sufficient statistics relating to $x_{t-\tau}$, $\tau \in T$ in the sense of Fisher—Neumann, then it is not difficult to obtain the following analytical equation for the distribution conditional density:

$$f(y_t/x_{t-\tau_1 \tau \in T}) = \frac{1}{\sqrt{D(y_t/x_{t-\tau_1 \tau \in T})}}\, \varphi\left(\frac{y_t - M(y_t/x_{t-\tau_1, \tau \in T})}{\sqrt{D(y_t/x_{t-\tau_1}, \tau \in T)}} \right). \qquad (5.1)$$

Indeed, the right-hand side of this equation is the density function for distribution of some random variables, and its mathematical expectation and dispersion are $M(y_t/x_{t-\tau}, \tau \in T)$ and $D(y_t/x_{t-\tau}, \tau \in T)$, respectively.

Since $M\xi_t = 0$ and $D\zeta_t = M[y_t - M(y_t/(x_{t-\tau}, \tau \in T)]^2$, the noise ζ is a random process with zero mean and dispersion $MD(y_t/x_{t-\tau}, \tau \in T)$. The last variable for a certain class of input disturbances may become zero and for this class of input signals the plant will behave like a deterministic system. Thus the classification of plants into deterministic and stochastic ones is relative and depends on the structure of the input signal.

If the noise dispersion is known, then the plant identification is reduced to find the useful component of a signal at its output, i.e. $M(y_t/x_{t-\tau},\ \tau \in T)$.

In its turn the function $M(y_t/x_{t-\tau},\ \tau \in T)$ is estimated through autoregression of the input signal and the dispersional weighting function of the plant according to (2.7). A general diagram of a dispersional model is shown in Fig. 2.

Unit G is a linear filter, whose weighting function satisfies the dispersional equation of identification (2.3).

The useful component of the signal at the output of the block G is calculated in the unit M. Unit D analyses the random process curves at the input and output of the plant to calculate conditional dispersion of the plant response to a disturbance at the input.

Unit Φ generates a random sequence with the density $\varphi(\zeta)$ and the parameters $(M\zeta_t = 0,\ D\zeta_t = D(y_t/x_{t-\tau},\ \tau \in T))$.

The random process y_t^* at the output of a model has the same conditional density of distribution as the random process y_t at the output of the plant.

Thus the plant dispersional model is constructed in three stages:

1. The determination of the input signal autoregression

$$M(x_s/x_{t-\tau},\ \tau \in T).$$

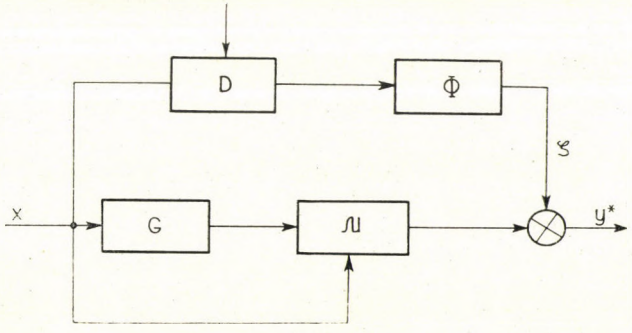


Fig. 1

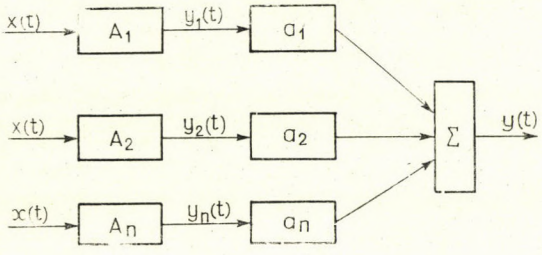


Fig. 2

2. The determination of the generalized autodispersional and cross dispersional functions. In order to find in this case the generalized cross dispersional function, the calculation of regression at the output of the plant in relation to the input is unnecessary, because $\theta^T_{y\bar{x}}(t, s) = \operatorname{cov}\{y_t, M(x_{t-s}/x_{t-\tau}, \tau \in T)\}$.

3. The solution of dispersional equation (2.3) and the construction of the useful component of the process at the output of the plant according to (2.11).

Example 1

Let us consider a three-level telegraph signal which takes the values -1.0 and 1. The equation of this random process is $x^3(t) = x(t)$. The random process transition from one state to another is represented in a third-order transition matrix. The random process $x(t)$ has a parabolic autoregression

$$M(x_t/x_s) = \alpha(t, s) + \beta(t, s)\, x_s + \gamma(t, s) x_s^2$$

where the function $\alpha(t, s)$, $\beta(t, s)$, $\gamma(t, s)$ satisfies the following linear equations

$$\alpha = M(x_t/x_s = 0),$$

$$\alpha + \beta + \gamma = M(x_t/x_s) = 1,$$

$$\alpha - \beta + \gamma = M(x_t/x_s) = -1.$$

Let T consist of one point: $T = \{0\}$.

Let us consider a noninertial plant whose equation is in the form

$$y_t = A(t)\, x_t^2 + B(t)\, x_t + C(t),$$

where the coefficients $A(t)$, $B(t)$, $C(t)$ are determined in the following way

$$C(t) = \int_{-\infty}^{t} g(t, s)\, \alpha(t, s)\, ds,$$

$$B(t) = \int_{-\infty}^{t} g(t, s)\, \beta(t, s)\, ds,$$

$$A(t) = \int_{-\infty}^{t} g(t, s)\, \gamma(t, s)\, ds.$$

The method of dispersional identification in this case leads to a model which gives a full representation of the plant since

$$y_t = \int_{-\infty}^{t} g(t, s)\, M(x_s/x_{t-\tau}, \tau \in T)\, ds.$$

On the other hand the linearization method inevitably leads to error, since the plant is nonlinear. The coefficients may be selected such that the error is very large.

Example 2

A stochastic plant is described by a two-dimensional simultaneous distribution density

$$f(y_t, x_s) = \left[1 + \frac{\sin y_t \cos x_s}{1 + (t-s)^2}\right] \varphi(y_t)\, \psi(x_s)\,,$$

where φ, ψ are nonnegative even functions satisfying the norming condition. The conditional mathematical expectation of the plant response relative to the input action is $M(y_t/x_{t-\tau}) = \dfrac{A}{1+\tau^2} \cos x_{t-\tau}$, where $A = \int\limits_{-\infty}^{\infty} x \sin x\varphi(x)\, dx$.

On the other hand, the mathematical expectation of the input action relative to the plant response is zero

$$M(x_s/y_t) = 0\,.$$

Thence in particular, the cross correlation function of the response and the input action also vanishes

$$K_{yx}(t, s) = M\{y_t M(x_s/y_t)\} = 0\,.$$

The pulse transient function of such a plant is consequently zero, but the dispersional pulsed transient function is not, since the right-hand part of the dispersional equation is

$$\Theta^T_{y\bar{x}}(t, s) = \frac{A}{1+\tau^2} \int\limits_{-\infty}^{\infty} x_s \cos x_{t-\tau}\, p(x_s, x_{t-\tau})\, dx_s\, dx_{t-\tau}\,,$$

where $p(x_s, x_{t-\tau})$ is the simultaneous function of the cross sections x_s and $x_{t-\tau}$:

Example 3

One example of a d-stationary random process with essentially nonlinear self-regression is a harmonic signal with a random phase

$$x_t = \sin(t - \varphi).$$

This signal is a deterministic process and is unambiguously reconstructed on the knowledge of any of its cross sections

$$x_t = \sin(t-s)\sqrt{1 - x_s^2} + \cos(t-s)\, x_s.$$

Consequently, the self-regression function of such a process is

$$M(x_t/x_s) = \sin(t-s)\sqrt{1 - x_s^2} + \cos(t-s)\, x_s.$$

Let y_t be the plant response to a harmonic signal x_t. If the plant is stationary and its pulse transient function is $g(s)$, then for the operator Φ we have the following expression

$$(\Phi x)(t) = \int_0^\infty g(s) \left[\cos(s-\tau)\, x_{t-\tau} - \sin(s-\tau) \sqrt{1 - x_{t-\tau}^2}\right] ds\,, \qquad \text{(E1)}$$

where $A = \int_0^\infty g(s) \cos(s-\tau)\, ds$, $B = \int_0^\infty g(s) \sin(s-\tau)\, ds$. Note that the expression for the operator Φ would look quite differently if we solved the identification problem in a class of linear dynamic models. In this case

$$(\Phi x)(t) = A x_{t-\tau} - B x_{t-\tau} \qquad \text{(E2)}$$

in other words Φ is a linear differential operator with constant factors. Purely formally there is no difference between expressions (E1) and (E2) since, for a harmonic signal, their right-hand parts coincide. Consequently, the interpretation of the model operator should be confined to a certain class of plants and a certain class of inputs.

6. The superpositional principle for dispersional models

If the linear plant is connected in parallel and $g_i(t, s)$ is a weighting function of i plant then the total weighting function of the system is

$$g(t, s) = \sum_{i=1}^{n} g_i(t, s)\,.$$

Suppose that after every plant with the ordinal number the signal is transformed by the amplifier with the constant gain a_i. Then the weighting function is the same as linear weighting function combinations of individual elements with the gains

$$g(t, s) = \sum_{i=1}^{n} a_i\, g_i(t, s)\,. \qquad (6.1)$$

The analogous equations are valid for the nonstationary linear plants.

For nonlinear plants the principle of superposition is not valid, i.e. the weighting function of optimal linear model is not equal to the linear combination of the weighting functions of optimal linear models. Nevertheless, the generalized superpositional function is valid for the dispersional models. In that Eq. (6.1) is valid on the average.

Let the dispersional model A be obtained by parallel connection of dispersional models $A_1, A_2, \ldots, A_n$; the random signal, which has been transmitted through the model, goes to the amplifier with the gain a_i.

The weighting function of the dispersional model A_i is denoted by $g_i(t, s)$. Then the useful component of the random signal that goes to the adder from the amplifier is equal to

$$M(y_t^i/x_{t-\tau}, \tau \in T) = \int_{-\infty}^{t} a_i g_i(t, s) M(x_s/x_{t-\tau}, \tau \in T) ds, i = 1, 2 \ldots n. \quad (6.2)$$

Consequently, the random signal at the output of the adder is

$$M(y_t/x_{t-\tau}, \tau \in T) = \int_{-\infty}^{t} g(t, s) M(x_s/x_{t-\tau}, \tau \in T) ds,$$

where $g(t, s) = \sum_{i=1}^{n} a_i g_i(t, s)$.

Thus for the dispersional models the principle of superposition is not valid for the random processes because of the noise at the output, which is nondispersed with the input. But it is valid on the average and may be formulated in terms of conditional mathematical expectations.

Now we can formulate the principle of superposition in a form similar to Eq. (6.1).

Let $\varphi_1, \varphi_2, \ldots, \varphi_r$ be the dispersional model operators. Then the model whose block-diagram is shown in Fig. 2 is also a dispersional model, and the operator corresponding to it is

$$(\Phi x)(t) = \sum_{i=1}^{n} a_i (\Phi_i x)(t). \quad (6.3)$$

It should be emphasized that, unlike linear theory, the generalized superpositional principle takes place only for a quite certain random process at the input of the plant. What is more if we feed a random process with an unknown internal structure to the input of the dispersional model, then this plant may be not linear on the average and in this case (6.3) may be untrue. For example, any linear deterministic plant with a Gaussian random process at the input is not linear on the average, because for this plant the operator Φ is linear, and the cross regression function is nonlinear.

As have been proved above, it is always possible to select a random input with such internal structure that the same plant with the chosen input disturbance will be a plant linear on the average.

Thus the dispersional equation for estimate of mutual regression function was obtained and proved that the error of this estimation does not exceed the error of linear optimal estimate.

The authors are indebted to V. S. Pugachev for discussion and remarks which helped to improve the manuscript.

References

1. *Bunich A. L., Raibman N. S.* Dispersional identification equation of nonlinear plants, *1*, 1972. Theory of information and control.
2. *Райбман Н. С.* Что такое идентификация. «Наука», 1970.
3. *Пугачев В. С.* Теория случайных функций и ее применение к задачам автоматического управления. Физматгиз, 1962.
4. *Райбман Н. С., Ханш О. Ф.* Дисперсионные методы идентификации многомерных нелинейных объектов. Автоматика и телемеханика, *5*, 1967.

Дисперсионное уравнение идентификации нелинейных объектов

А. Л. БУНИЧ, Н. С. РАЙБМАН

(Москва)

Настоящая статья представляет собой обобщение результатов, полученных в [1] на нелинейные динамические объекты с конечной памятью. Так как оператор условного математического ожидания дает наилучшее в смысле критерия минимума среднего квадрата ошибки приближение реакции объекта в классе функций входного сигнала, то задача идентификации сводится к оценке взаимной регрессионной функции случайного процесса на выходе системы относительно наблюдаемых сечений случайного процесса на ее входе и к определению точности такой оценки. По принятому критерию получено дисперсионное уравнение идентификации (2.3), а средняя квадратическая ошибка оценивается уравнением (2.14).

При идентификации динамических объектов с существенными нелинейностями класс линейных моделей оказывается слишком узким, поэтому оценку взаимной регрессионной функции необходимо искать в классе нелинейных операторов. С этой целью вводится понятие объекта, линейного в среднем, которое включает в себя как частный случай понятие линейного объекта. Идентифицируемому объекту сопоставляется некоторый объект, линейный в среднем, причем выбор модели производится таким образом, чтобы среднеквадратическая ошибка уклонения выхода модели от выхода реального объекта была минимальна. Полученный таким образом критерий оптимальности приводит к интегральному уравнению для определения весовой функции модели объекта. Доказано (2,7), что для объектов с детерминированным входным сигналом дисперсионная весовая функция совпадает с весовой функцией оптимальной линейной модели. Введено определение α-стационарных и α-стационарно связанных случайных процессов. Так как класс объектов, линейных в среднем шире, чем класс линейных объектов, то точность оценки взаимной регрессионной функции выше, чем точность линейной оценки, причем дисперсионная идентификация наиболее эффективна для входных сигналов с нелинейной внутренней структурой. В том случае, когда случайный процесс на входе объекта обладает линейной авторегрессией (см. раздел 4), полученные результаты совпадают с линейными оценками по методу наименьших квадратов. Таким образом методы наилучших линейных оценок, основанные на применении корреляционной теории случайных функций, можно рассматривать как частный случай дисперсионной идентификации, соответствующей входным сигналам с линейной авторегрессионной функцией. При построении модели объекта мы ограничились условными моментами первого и второго порядка. По этой причине естественно предположить, что условное математическое ожидание и условная дисперсия реакции объекта относительно сечений входного сигнала являются достаточными статистиками. Если одномерная плотность вероятности шума на выходе объекта известна, то, применив теорему о факторизации Фишера—Неймана, можно получить аналитическое выражение (5.1) для условной плотности распределения. Идентификация нелинейных объектов предложенным методом в классе безынерционных моделей приводит к результатам, полученным в [1].

Problems of Control and Information Theory, Vol. 2 (3–4), pp. 273–281 (1973)

РЕАЛИЗАЦИЯ ЧАСТИЧНО-СИММЕТРИЧЕСКИХ ФУНКЦИЙ СХЕМАМИ ИЗ ФУНКЦИОНАЛЬНЫХ ЭЛЕМЕНТОВ

А. В. МАРКОВСКИЙ, Л. Б. ШИПИЛИНА

(Москва)

(Поступила в редакцию 30 декабря 1971 г.)

В работе рассматривается реализация частично-симметрических функций схемами из функциональных элементов. Получено асимптотическое выражение функции Шеннона. Рассматриваются два частных вида частично-симметрических функций, допускающих более простую, чем в общем случае, реализацию. Приводятся асимптотические выражения функции Шеннона для этих случаев.

Данная работа является еще одним примером применения принципа локального кодирования, изложенного в фундаментальной статье О. Б. Лупанова [1, 2], и существенно использует понятия и обозначения, введенные в ней.

Булева функция $f(\mathbf{x})$ называется частично-симметрической (ЧСФ) [3, 4], если существует разбиение множества ее аргументов $\mathbf{x} = (x_1, \ldots, x_n)$ на непустые подмножества $\mathbf{x}_i = (x_{i_1}, \ldots, x_{in_i})$; $i = 1, 2, \ldots, k$, $n = \sum_{i=1}^{k} n_i$ (называемые группами симметрии), такие, что $f(\mathbf{x})$ не меняются при произвольной перестановке аргументов, внутри любой из групп $\mathbf{x}_i$. Без потери общности будем считать, что

$$\mathbf{x}_1 = (x_1, \ldots, x_{n_1}), \mathbf{x}_2 = (x_{n_1+1}, \ldots, x_{n_1+n_2}), \ldots, \mathbf{x}_k = \\ = (x_{n_1+\ldots+n_{k-1}+1}, \ldots, x_{n_1+\ldots+n_k}).$$

Числа $n_1, \ldots, n_k$ будем называть параметрами ЧСФ $f(x)$, а самое ее обозначать через $f_{n_1, \ldots, n_k}(\mathbf{x})$. Очевидно, что любую булеву функцию от n аргументов можно считать частично-симметрической, полагая $n_i = 1$, $i = 1, 2, \ldots, k$, $k = n$. Другим частным видом ЧСФ являются симметрические функции [5].

Будем рассматривать реализацию систем ЧСФ схемами из функциональных элементов над конечным базисом [6]. Под сложностью $L(S)$ схемы S над заданным базисом понимается сумма весов (цен) входящих в нее элементов. Под сложностью $L(F^m)$ системы $F^m(\mathbf{x})$, состоящей из m ЧСФ $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ будем понимать минимум сложностей схем над данным базисом, реализующих $F^m_{(x)}$.

Пусть $\mathfrak{F}^m_{n_1,\ldots,n_k}$ — класс всех систем $F^m_{n_1,\ldots,n_k}(\mathbf{x})_m$ ЧСФ с параметрами $n_1, \ldots, n_k$. Функцию $L(n_1,\ldots,n_k,m) = \max L(F^m_{n_1,\ldots,n_k}(\mathbf{x}))$ по всем $F^m_{n_1,\ldots,n_k}(\mathbf{x}) \in \mathfrak{F}^m_{n_1,\ldots,n_k}$ назовем функцией Шеннона для класса $\mathfrak{F}^m_{n_1,\ldots,n_k}$.

Пусть $H(n_1, \ldots, n_k) = \prod_{i=1}^{k} (n_i + 1)$, тогда имеет место следующая теорема о поведении функции $L(n_1, \ldots, n_k, m)$.

Теорема 1. Для всякой последовательности

$$\mathfrak{F}^{m_1}_{n_{11},\ldots,n_{k1}}, \mathfrak{F}^{m_2}_{n_{12},\ldots,n_{k2}}, \ldots, \mathfrak{F}^{mt}_{n_{1t},\ldots,n_{kt}}, \ldots$$

классов ЧСФ, такой, что $n = \sum_{i=1}^{k} n_i \to \infty$,

$$\frac{\log m}{H} \to 0, \text{[1]} \tag{1}$$

и

$$\frac{n \log mH}{mH} \to 0 \tag{2}$$

имеет место

$$L(n_1, \ldots, n_k, m) \sim^{\text{[2]}} \varrho \frac{mH(n_1, \ldots, n_k)}{\log mH(n_1 \ldots, n_k)}, \tag{3}$$

где ϱ – минимум приведенных весов элементов базиса. Приведенный вес определяется только для элементов ε, имеющих не менее двух входов и равняется отношению $\frac{P(\varepsilon)}{i_\varepsilon - 1}$, где $P(\varepsilon)$ — вес, а i_ε — число входов элемента ε [6].

Доказательство. Пусть $F^m_{n_1,\ldots,n_k}(\mathbf{x})$ — система m ЧСФ $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ с параметрами $n_1, \ldots n_k$. Рассмотрим совокупность всевозможных наборов $\mathbf{b} = (b_1, \ldots, b_k)$ целых чисел b_i, где $0 \leq b_i \leq n_i$; $i = 1, 2, \ldots, k$. Число таких наборов равно $H(n_1, \ldots, n_k)$.

Упорядочим множество наборов $\mathbf{b}$ лексикографически справа налево и занумеруем их подряд натуральными числами от 0 до $H-1$. Обозначим через $B(\mathbf{b})$ номер набора $\mathbf{b}$:

$$\begin{aligned} B(\mathbf{b}) = {} & (n_1 + 1)(n_2 + 1)\cdot \ldots \cdot (n_{k-1} + 1) b_k + \ldots \\ & + (n_1 + 1)(n_2 + 1)\cdot \ldots \cdot (n_{i-1} + 1) b_i + \ldots \\ & + (n_1 + 1)(n_2 + 1) b_3 + (n_1 + 1) b_2 + b_1. \end{aligned} \tag{4}$$

[1] Индексы опускаем. Логарифм берется по основанию 2.

[2] $a \sim b$ означает, что при $a, b \to \infty \frac{a}{b} \to -1$.

Пусть $0 \leq B \leq H-1$. Обозначим через $\mathbf{b}(B) = \{b_1(B), \ldots, b_i(B), \ldots, b_k(B)\}$ набор $\mathbf{b}$, имеющий номер B. Положим $H_i = \prod_{j=2}^{i} n_{j-1} + 1)$ для $i = 2, 3, \ldots, k_n$, $H_1 = 1$, тогда $B(\mathbf{b}) = \sum_{i=1}^{k} H_i b_i$.

Система $F^m_{n_1,\ldots,n_k}(\mathbf{x})$ полностью определяется последовательностью $\Pi = \tilde{\pi}_0, \tilde{\pi}_1, \ldots \tilde{\pi}_{H-1}$ длины mH, где $\pi_B = f_1, (\mathbf{b}(B)), f_2(\mathbf{b}(B)), \ldots, f_m(\mathbf{b}(B))$ — набор значений ЧСФ $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ на (любом) наборе значений аргументов, содержащем в группе $\mathbf{x}_i$ ($i = 1, 2, \ldots, k$) $b_i(B)$ единиц и $n_i - b_i(B)$ нулей. Очевидно, что последовательность Π может быть однозначно построена по системе $\mathfrak{F}^m_{n_1,\ldots,n_k}(\mathbf{x})$, причем разным системам из $\mathfrak{F}^m_{n_1,\ldots,n_k}$ соответствуют разные последовательности Π. Поэтому последовательность Π может служить кодом системы $\mathfrak{F}^m_{n_1,\ldots,n_k}(\mathbf{x})$ в классе $\mathfrak{F}^m_{n_1,\ldots,n_k}$. Следовательно, класс $\mathfrak{F}^m_{n_1,\ldots,n_k}$ содержит 2^{mH} систем $\mathfrak{F}^m_{n_1,\ldots,n_k}(\mathbf{x})$. В силу (1) и (2) получаем (теорема Д 1 [1]) мощностную нижнюю оценку для $L(n_1, \ldots, n_k, m)$

$$L(n_1 \ldots, n_k, m) \gtrsim^{3} \varrho \frac{mH(n_1, \ldots, n_k)}{\log mH(n_1, \ldots, n_k)}. \tag{5}$$

Верхнюю оценку для $L(n_1, \ldots, n_k, m)$ получим пользуясь принципом локального кодирования [1]. Пусть $F^m_{n_1,\ldots,n_k}(\mathbf{x})$ — произвольная система из $\mathfrak{F}^m_{n_1,\ldots,n_k}$ и $\Pi = \tilde{\pi}_0, \pi_1, \ldots, \pi_{H-1}$ — код системы $F^m_{n_1,\ldots,n_k}(\mathbf{x})$. Пусть $\tilde{\sigma}$ — произвольный набор значений переменных $x_1, \ldots, x_n$, а $\tilde{\sigma}_i$, $i = 1, 2, \ldots, k$ поднабор набора $\tilde{\sigma}$, содержащий значение переменных группы $\mathbf{x}_i$ и пусть b_i — число единиц в $\tilde{\sigma}_i$. Набор $F^m_{n_1,\ldots,n_k}(\tilde{\sigma}) = \{f_1(\tilde{\sigma}), \ldots, f_m(\tilde{\sigma})\}$ будем вычислять следующим образом:

1) для каждого $\tilde{\sigma}_i$ вычисляем b_i.

2) для набора $\mathbf{b} = (b_1, \ldots, b_k)$ вычисляем $B(\mathbf{b})$ по формуле (4).

3) из кода Π выделяем кусок $\tilde{\pi}_B$, который совпадает с $F^m_{n_1}, \ldots, n_k(\tilde{\sigma})$.

Обозначим через $\tilde{\beta}_i$ двоичный набор длины $]\log(n_i + 1)[$, изображающий число b_i, через $\tilde{\nu}_i$ — двоичный набор длины $]\sum_{i=1}^{k} \log(n_i + 1)[$, изображающий число H_i, и через $\tilde{B}(\mathbf{b})$ — двоичный набор длины $]\log H[$, изображающий число $B(\mathbf{b})$.

Схема S для $F^m_{n_1,\ldots,n_k}$ строится из следующих блоков (рис. 1).

1) Блоки $Б_i$, $i = 1, 2, \ldots, k$ подсчитывают число единиц b_i в наборах $\tilde{\sigma}_i$. Блок $Б_i$ имеет n_i входов и $]\log(n_i + 1)[$ выходов. Согласно теореме 3.1 из

[3] $a \gtrsim b$ означает, что при $a, b \to \infty$ $\lim \frac{a}{b} \geq 1$.

[1] имеем:

$$\sum_{i=1}^{k} L(\bar{\text{Б}}_i) = O\left(\sum_{i=1}^{k} n_i\right) = O(n)\,.^{4} \tag{6}$$

2) Блок B по набору $\mathbf{b} = (b_1, \ldots, b_k)$ вычисляет $B(\mathbf{b})$. Блок B имеет $\sum_{i=1}^{k}]\log(n_i + 1)[$ входов и $]\log H[=]\sum_{i=1}^{k} \log(n_i + 1)[$ выходов. Блок B

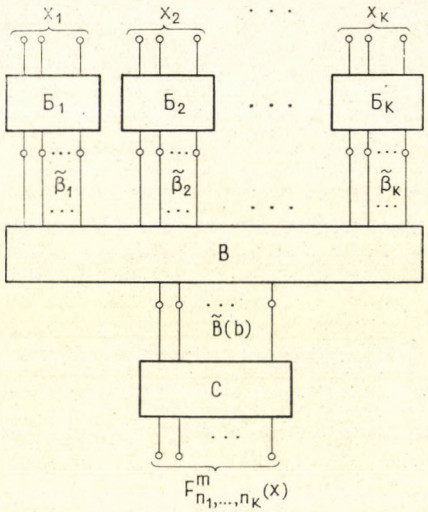


Рис. 1

перемножает двоичные числа $\tilde{\beta}_i$ и $\tilde{\nu}_i$, а затем суммирует их произведения. Для этого требуется k схем умножения и k схем сложения двух не более чем $]\log H[$ разрядных двоичных чисел. Схема умножения может быть выбрана со сложностью не превышающей $C_1 \log^2 H$ (в основу такой схемы можно положить, например, обычный алгоритм умножения двоичных чисел). Схему сложения можно выбрать со сложностью не более чем $C_2 \log H$ [1]; $C_1, C_2 =$ $=$ const. Отсюда следует, что

$$L(B) = O(k \log^2 H)\,. \tag{7}$$

3) Блок C по числу $B(\mathbf{b})$ выдает кусок $\tilde{\pi}_B$ кода Π. Блок C имеет $]\log H[$ входов, по которым поступает число $B(\mathbf{b})$ и m выходов, на которых выдается набор $\tilde{\pi}_B = F^m_{n_1, \ldots, n_k}(\tilde{\sigma})$ значений функций $f_1, \ldots, f_m$ на наборе $\tilde{\sigma}$.

Заметим, что число всевозможных наборов значений $]\log H[$ входов блока C равно $2^{]\log H[}$. В действительности на входах блока C появляются

[4] $a = O(b)$ означает, что при $a, b \to \infty$ $\frac{a}{b} \leq k$, где k — константа.

только первые[5] H наборов, поэтому можно считать, что блок C реализует систему функций $\varphi_1, \ldots, \varphi_m$ от $]\log H[$ аргументов, произвольно заданных на первых H наборах значений аргументов и равных нулю на остальных. Согласно теореме Д8 [1] схему блока C можно выбрать со сложностью

$$L(C) \lesssim \varrho \frac{mH}{\log mH}. \tag{8}$$

Из (6), (7) и (8) имеем

$$L(S) \lesssim \varrho \frac{mH}{\log mH} + O(n) + O(k^2 \log^2 H). \tag{9}$$

Используя очевидное соотношение

$$H(n_1, \ldots, n_k) \geq 2^k,$$

имеем

$$\frac{k \log^3 H}{mH} = \frac{1}{m} \cdot \frac{K}{H^{1/2}} \cdot \frac{\log^3 H}{H^{1/2}} \leq \frac{1}{m} \cdot \frac{K}{\sqrt{2^k}} \cdot \frac{\log^3 H}{\sqrt{H}} \to 0.$$

Отсюда, используя (1), (2) и (9), получаем

$$L(n_1, \ldots, n_k, m) \leq L(F^m_{n_1, \ldots, n_k}) \leq L(S) \lesssim \varrho \frac{mH}{\log mH}, \tag{10}$$

следовательно из (10) и (5) получаем (3). Теорема доказана.

Отметим, что для системы m симметрических функций (в этом случае $H(n) = n + 1 \sim n$) формула (3) теоремы 3.2 [1] переходит в формулу

$$L(n) \sim \varrho \frac{mn}{\log mn}.$$

Для системы произвольных функций (в этом случае $k = n$, $n_1 = n_2 = \ldots = n_k = 1$, $H(n_1, \ldots, n_k) = 2^n$), формула (3) принимает вид (формула Д10 [1])

$$L(n) \sim \varrho \frac{m 2^n}{n + \log m}.$$

Рассмотрим два частных случая, когда ЧСФ реализуются более просто, чем в общем случае.

1. Пусть $F^{(1)m}_{n_1, \ldots, n_k} = \{f_1, \ldots, f_m\}$ — система ЧСФ, каждая из которых представима в виде:

$$f_i = \varphi_i(S_1(\mathbf{x}_1), S_2(\mathbf{x}_2), \ldots, S_k(\mathbf{x}_k)), \qquad i = 1, 2, \ldots, m,$$

где $S_1, \ldots, S_k$ — симметрические функции, φ — некоторая булева функция от k аргументов и множества $\mathbf{x}_i$, $i = 1, 2, \ldots, m$ содержат n_i аргументов.

[5] Считаем, что наборы упорядочены в порядке возрастания чисел, двоичными изображениями которых они являются.

Обозначим через $\mathfrak{F}^{(1)m}_{n_1,\ldots,n_k}$ множество всех систем $F^{(1)m}_{n_1,\ldots,n_k}(\mathbf{x})$, а через $L^{(1)}(n_1,\ldots,n_k,m)$ — функцию Шеннона для класса $\mathfrak{F}^{(1)m}_{n_1,\ldots,n_k}$.

Теорема 2. Если при $n\to\infty$, $k\to\infty$,

$$\frac{\log m}{2^k}\to 0 \tag{11}$$

и

$$\frac{k\cdot n\cdot\log m}{m2^k}\to 0\,, \tag{12}$$

то

$$L^{(1)}(n_1,\ldots,n_k,m)\sim\varrho\,\frac{m2^k}{k+\log m}\,. \tag{13}$$

Доказательство. Обозначим через $N^{(1)}(n_1,\ldots,n_k,m)$ число систем $F^{(1)m}_{n_1,\ldots,n_k}$. Очевидно, что $N^{(1)}(n_1,\ldots,n_k,m)=2^{n_1+1}\cdot\ldots\cdot 2^{n_k+1}\cdot 2^{m2^k}=2^{n+k}\cdot 2^{m2^k}$. Первый сомножитель показывает, сколькими способами можно выбрать функции $S_1,\ldots,S_k$; второй — число способов выбора функций $\varphi_1,\ldots,\varphi_m$. Обозначим через $H^{(1)}(n_1,\ldots,n_k,m)$ двоичный логарифм числа $N^{(1)}(n_1,\ldots,n_k,m)$. Тогда в силу (12) имеем:

$$H^{(1)}(n_1,\ldots,n_k,m)=n+k+m\cdot 2^k\sim m2^k\,. \tag{14}$$

Из (14) и теоремы Д1 [1] в силу (11) и (12) получаем:

$$L^{(1)}(n_1,\ldots,n_k,m)\gtrsim\varrho\,\frac{H^{(1)}(n_1,\ldots,n_k,m)}{\log H^{(1)}(n_1,\ldots,n_k,m)}\sim\partial\,\frac{m2^k}{\log m2^k}\,. \tag{15}$$

Схема S, реализующая произвольную систему $F^{(1)m}_{n_1\ldots,n_k}$, показана на рис. 2. Блоки $S_1,\ldots,S_k$ реализуют симметрические функции $S_1(\mathbf{x}_1),\ldots,S_k(\mathbf{x}_k)$. Из т. 3.1 [1] имеем:

$$L(S_i)=O(n_i)\,. \tag{16}$$

Блок D имеет k входов и m выходов и реализует функции $\varphi_1,\ldots,\varphi_m$. В силу (11) и теоремы Д8 [1] имеем:

$$L(D)\gtrsim\varrho\,\frac{m2^k}{\log m2^k}\,.$$

Отсюда и из (16) следует, что

$$L^{(1)}(n_1,\ldots,n_k,m)\le L(S)\gtrsim\varrho\,\frac{m2^k}{\log m2^k}+O(n_1)+\ldots+O(n_k)=$$

$$=\varrho\,\frac{m2^k}{\log m2^k}+O(n)\,.$$

В силу (12) и (15) получаем (13). Теорема 2 доказана.

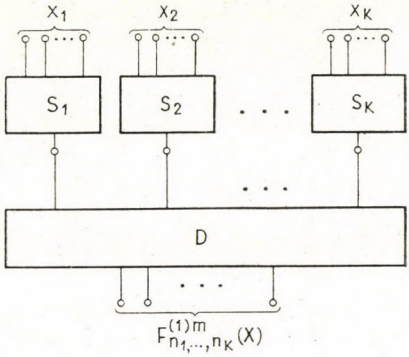


Рис. 2

2. Пусть $F^{(2)m}_{n_1,\ldots,n_k} = \{f_1, \ldots, f_m\}$ — система ЧСФ, каждая из которых представлена в виде

$$f_j = C_j(S_1(\mathbf{x}_1), S_2(\mathbf{x}_2), \ldots, S_k(\mathbf{x}_k)), \qquad j = 1, 2, \ldots, m,$$

где $S_1, \ldots, S_k, C_1, \ldots, C_m$ — симметрические функции и множество $\mathbf{x}_i$ содержит n_i аргументов.

Обозначим через $\mathfrak{F}^{(2)m}_{n_1,\ldots,n_k}$ множество всех систем $F^{(2)m}_{n_1,\ldots,n_k}(\mathbf{x})$, через $N^{(2)}(n_1, \ldots, n_k, m)$ — мощность множества $\mathfrak{F}^{(2)m}_{n_1,\ldots,n_k}$ и через $L^{(2)}(n_1, \ldots, n_k, m)$ — функцию Шеннона для класса $\mathfrak{F}^{(2)m}_{n_1,\ldots,n_k}$. Пусть $H^{(2)}(n_1, \ldots, n_k, m) = \log N^{(2)}(n_1, \ldots, n_k, m)$.

Теорема 3. Если $n \to \infty$, $k \to \infty$ и

$$\frac{\log k}{m} \to 0, \tag{17}$$

$$\frac{\log m}{k} \to 0, \tag{18}$$

$$\frac{n \log (km)}{km} \to 0, \tag{19}$$

то

$$L^{(2)}(n_1, \ldots, n_k, m) \sim \varrho \frac{km}{\log km}. \tag{20}$$

Доказательство. Очевидно, что $N^{(2)}(n_1, \ldots, n_k, m) = 2^{n+k} \cdot 2^{m(k+1)} = 2^{n+k+m+mk}$. Тогда в силу (19) $H^2(n_1, \ldots, n_k, m) \sim mk$. Отсюда, согласно теореме Д1 [1], получается нижняя оценка

$$L^2(n_1, \ldots, n_k, m) \geq \varrho \frac{mk}{\log (mk)}. \tag{21}$$

Схема, реализующая произвольную систему $F^{(2)m}_{n_1 \ldots, n_k}(\mathbf{x})$, показана на рис. 2. Блоки $S_1, \ldots, S_k$ здесь такие же, как и в предыдущем случае. Блок D имеет k входов и m выходов и реализует функции $C_1, \ldots, C_m$. В силу (17) и (18), согласно т. 3.2. [1], имеем

$$L(D) \lesssim \varrho \frac{mk}{\log mk}.$$

Отсюда и из (16) следует, что

$$L^2(n_1, \ldots, n_k, m) \leq L(S) \lesssim \varrho \frac{mk}{\log mk} + O(n).$$

С силу (19) и (21) получаем (20). Теорема доказана.

Литература

1. *Лупанов О. Б.* Об одном подходе к синтезу управляющих систем — принципе локального кодирования. Сб. «Проблемы кибернетики», *14*, (1965), 31—110.
2. *Лупанов О. Б.* О принципе локального кодирования и реализации функций из некоторых классов схемами из функциональных элементов. Доклады АН СССР, **140**, *2*, (1961), 322—325.
3. *Mukhopadhyay, A.*, Detection of Total or Partial Symmetry of a Switching Function with the Use of Decomposition Charts. IEEE Trans. on Electr. Comp., **EC—12**, N. 5 October 1963, pp. 553—561.
4. *Dietmeyer, D. L., Schneider, P. R.*, Identification of Symmetry, Redundancy and Equivalence of Boolean Functions. IEEE Trans. on Electr. Comp., **EC—16**, N. 6 December 1967, pp. 804—817.
5. *Колдуэлл С.* Логический синтез релейных устройств. М., ИЛ., 1962.
6. *Лупанов О. Б.* Об одном классе схем из функциональных элементов. Сб. «Проблемы кибернетики», вып. 7, (1962), 61—114.

Realization of partially symmetric functions by networks of functional elements

A. V. MARKOVSKII, L. B. SHIPILINA

(Moscow)

The Boolean function $f(\mathbf{x})$ is defined to be partially symmetric function (PSF), if there exists a partition of the set $\mathbf{x} = (x_1, \ldots, x_n)$ into such nonempty subsets $X_i = (x_{i_1}, \ldots, x_{i_n})$, $i = 1, 2, \ldots, k$, $n = \sum_{i=1}^{k} n_i$ (to be defined as classes of symmetry), that an arbitrary permutation of function's arguments preserving the intrasubset X_i keeps $f(\mathbf{x})$ unchanged.

The class of all the multiple output PSF with parameters $n_1, \ldots, n_k, m$ (m is the quantity of outputs in multiple output PSF) is denoted as $F^m_{n_1, \ldots, n_k}$.

The realization of the multiple output partially symmetric function by functionally packaged logic with the finite totality of functions realized by packages is considered.

Let Shannon's function $L(n_1, \ldots n_k, m)$ be maximum of the complexity realizing m-output PSF from $\mathfrak{F}^m_{n_1, \ldots, n_k}$. There is the theorem about behaviour $L(n_1, \ldots, n_k, m)$.

Theorem 1. For any sequence

$$\mathfrak{F}^{m_1}_{n_{11},\ldots,n_{k1}},\ \mathfrak{F}^{m_2}_{n_{12},\ldots,n_{k_2}},\ \ldots,\ \mathfrak{F}^{m_t}_{n_{1t},\ldots,n_{kt}}$$

and provided that

$$n \to \sum_{i=1}^{k} n_i \to \infty,\ \frac{\log m}{H(n_1,\ldots,n_k)} \to 0,\ \frac{m\log\left(mH(n_1,\ldots,n_k)\right)}{mH(n_1,\ldots,n_k)} \to 0,$$

the asymptotical equality exists

$$L(n_1,\ \ldots,\ n_k,\ m) \sim \varrho\,\frac{mH(n_1,\ \ldots,\ n_k)}{\log\left(mH(n_1,\ \ldots,\ n_k)\right)}$$

here

$$H(n_1,\ldots,n_k,\ m) = \prod_{i=1}^{k}(n_i+1)$$

and ϱ is minimum of normalized weights of packages, $\varrho = \min\dfrac{P(\varepsilon)}{i_\varepsilon - 1}$, $P(\varepsilon)$ is the weight of package ε, $i_\varepsilon \geq 2$ is the number of package inputs.

It is of interest to consider two special cases when multiple output PSF is realized fairly simply.

1) Let $\mathfrak{F}^{(1)m}_{n_1,\ldots,n_k}$ be the class of PSF

$$f_i = \varphi_i\left(S_1(\mathbf{x}_1),\ S_2(\mathbf{x}_2),\ \ldots,\ S_k(\mathbf{x}_k)\right)$$

whose $S_1, \ldots, S_k$ are the symmetric functions and φ_i is the arbitrary Boolean k-input function.

Then for multiple output PSF belonging to the class $\mathfrak{F}^{(1)m}_{n_1,\ldots,n_k}$ and provided that $n \to \infty$ $k \to \infty$, $\dfrac{\log m}{2^k} \to 0$, $\dfrac{nk\log m}{m2^k} \to 0$ the asymptotical equality exists

$$L^{(1)}(n_1,\ \ldots,\ n_k,\ m) \sim \varrho\,\frac{m\,2^k}{k+\log m}.$$

2) Let $\mathfrak{F}^{2)m}_{n_1,\ldots,n_k}$ be the class of PSF

$$f_i = C_i(S_i(\mathbf{x}_1),\ldots,S_j(\mathbf{x}_j),\ldots,S_k(\mathbf{x}_k))$$

whose C_i, S_j $(i, j = 1, \ldots, k)$ are the symmetric functions.

Then for multiple output PSF, belonging to the class $\mathfrak{F}^{(2)m}_{n_1,\ldots,n_k}$ and provided that $n \to \infty$, $k \to \infty$, $\dfrac{\log m}{k} \to 0$,

$$\frac{m}{\log k} \to \infty,\ \frac{n\log(km)}{km} \to 0$$

the asymptotical equality exists

$$L^{(2)}(n_1,\ldots,n_k,\ m) \sim \varrho\,\frac{km}{\log km}.$$

А. В. Марковский, Л. Б. Шипилина
Институт проблем управления (автоматики и телемеханики)
СССР, Москва В—485
Профсоюзная ул. 81

Problems of Control and Information Theory, Vol. 2 (2—4), pp. 283—304 (1973)

ON THE CAPACITY OF NOISY CHANNELS WITH ARBITRARY SIGNAL COSTS

I. CSISZÁR

(Budapest)

(Received November 30, 1972)

Four types of information capacity of general noisy channels with arbitrary input signal costs are considered and theorems on their asymptotic equality are proved, both for unconstrained and constrained inputs. The considered capacities are relevant for transmission of fixed and variable cost, respectively. In particular, for the transmission length cost assignment, the results amount to a general proof of the intuitive statement that the capacity bound for the length of reliable transmission is valid, asymptotically, also for the average length of variable length transmission. Examples with memoryless channels are given, including channels with feedback.

1. Definitions and motivation

A *noisy channel* is characterized by the sets V and $\overline{V}$ of possible input and output signals, respectively, and by the probability distributions governing the output for each possible input $v \in V$. Typically, V consists either of strings of symbols from a given alphabet Y, or of functions with domain $[0, t]$ and range Y, where t runs from 0 to ∞. These two cases are referred to as *discrete time* and *continuous time* channels, respectively.[1] In both cases, the structure of $\overline{V}$ is, as a rule, the same as that of V; for the purposes of this paper, however, the structure of $\overline{V}$ is irrelevant.

Formally, by a *channel* Π we shall mean a family

$$\Pi = \{\mathcal{S}_v : v \in V\} \tag{1}$$

of PD's (probability distributions) on $\overline{V}$, more exactly, on the measurable space $(\overline{V}, \overline{\mathcal{Y}})$, where $\overline{\mathcal{Y}}$ is a σ-algebra of subsets of $\overline{V}$ considered to be a priori given.

As a rule, there are certain costs associated with the possible input signals $v \in V$. In the typical examples mentioned in the first paragraph, the most common cost assignment is the *length* of the signal, i.e. the number of symbols

[1] This means, that we have in mind channels operating for a finite duration of time, rather than continuously operating channels. A model of the latter type would arise by letting V consist of doubly infinite sequences of functions with domain $(-\infty, \infty)$ but in that case the introduction of the concept of costs would be awkward.

in the string $v \in V$ or the length of the domain of the function $v \in V$. There may be, however, also other natural cost assignments; e.g. in discrete time channels the elementary symbols may be of different duration (as dot, dash and space in telegraphy), in which case the duration of the signal is a cost assignment significantly different from length.

Formally, a *cost function* l is any non-negative valued function defined on V.

Remark. In models of real communication channels, the PD's $\mathcal{S}_v$, see (1), always satisfy certain consistency conditions. Moreover, they have, as a rule, the property that for each $B \in$, $\overline{\mathcal{V}}\ \mathcal{S}_v(B)$ is a measurable function of v, with respect to some a priori given σ-algebra $\mathcal{V}$ of subsets of V. For our purposes, however, no such assumptions will be needed.

If the input signal of a channel Π is selected from a finite subset $\{v_1, \ldots, v_m\}$ of V with probabilities $p_1, \ldots, p_m$ $(\sum_{i=1}^{m} p_i = 1)$, the average amount of transmitted information can be measured by the mutual information $I(\eta; \overline{\eta})$, where the RV's (random variables) η and $\overline{\eta}$ describe the input and output, respectively, i.e. $\mathbf{P}\{\eta = v_i\} = p_i$, $i = 1, \ldots, m$ and

$$\mathbf{P}\{\overline{\eta} \in B \mid \eta = v\} = P_v(B) \text{ for all } v \text{ with } \mathbf{P}\{\eta = v\} > 0. \tag{2}$$

Recall that the *mutual information* $I(\xi; \eta)$ of discrete RV's ξ and η is defined as

$$\begin{aligned} I(\xi\,;\eta) &= \sum_{x,y} \mathbf{P}\{\xi = x, \eta = y\} \log_2 \frac{\mathbf{P}\{\xi = x, \eta = y\}}{\mathbf{P}\{\xi = x\}\,\mathbf{P}\{\eta = y\}} = \\ &= \sum_{x} \mathbf{P}\{\xi = x\} \sum_{y} \mathbf{P}\{\eta = y \mid \xi = x\} \log_2 \frac{\mathbf{P}\{\eta = y \mid \xi = x\}}{\mathbf{P}\{\eta = y\}}\,; \end{aligned} \tag{3}$$

if η is arbitrary (i.e. if η is a RV[2] taking values in a set Y with underlying σ-algebra $\mathcal{Y}$, $I(\xi; \eta)$ is defined, as a straightforward generalization of (3), by

$$I(\xi\,;\eta) = \sup_{\mathfrak{B}} \sum_{x} \mathbf{P}\{\xi = x\} \sum_{j=}^{n} \mathbf{P}\{\eta \in B_j \mid \xi = x\} \log_2 \frac{\mathbf{P}\{\eta \in B_j \mid \xi = x\}}{\mathbf{P}\{\eta \in B_j\}}\,, \tag{4}$$

where $\mathfrak{B} = (B_1, \ldots, B_n)$ ranges over all finite measurable partitions of Y (a partition $\mathfrak{B}$ is measurable if $B_j \in \mathcal{Y}$, $j = 1, \ldots n$). The sum on the right hand side means the average amount of information obtainable for ξ from

[2] A RV in general sense is defined as a measurable mapping of a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ into a measurable space $(Y, \mathcal{Y})$.

observing η "with precision $\mathfrak{B}$". It never decreases if the partition $\mathfrak{B}$ is refined, because if B_j is subpartitioned into sets $B_{jk} \in \mathfrak{Y}$, then[3]

$$\mathbf{P}\{\eta \in B_j \mid \xi = x\} \log_2 \frac{\mathbf{P}\{\eta \in B_j \mid \xi = x\}}{\mathbf{P}\{\eta \in B_j\}} \leqq$$

$$\leqq \sum_k \mathbf{P}\{\eta \in B_{jk} \mid \xi = x\} \log_2 \frac{\mathbf{P}\{\eta \in B_{jk} \mid \xi = x\}}{\mathbf{P}\{\eta \in B_{jk}\}}. \tag{5}$$

In view of (2) and (4), the average amount of information transmitted over the channel Π in the case considered in the paragraph containing (2) equals

$$I(\mathfrak{S}, \Pi) = \sup_{\mathfrak{B}} \sum_{i=1}^{m} p_i \sum_{j=1}^{n} \mathfrak{S}_{v_i}(B_j) \log_2 \frac{\mathfrak{S}_{v_i}(B_j)}{\overline{\mathfrak{S}}(B_j)}, \tag{6}$$

where $\mathfrak{B} = (B_1, \ldots, B_n)$ ranges over the finite measurable partitions of $\overline{V}$ and

$$\overline{\mathfrak{S}} = \sum_{i=1}^{m} p_i \mathfrak{S}_{v_i} \quad \text{i.e.,} \quad \overline{\mathfrak{S}}(B) = \sum_{i=1}^{m} p_i \mathfrak{S}_{v_i}(B), \quad B \in \overline{\mathfrak{V}}. \tag{7}$$

Let us define four types of *t-cost capacity* of a channel Π with cost function l as the maximum average amount of information trsnsmissible with

(i) cost exactly t
(ii) cost not exceeding t
(iii) average cost t
(iv) average cost not exceeding t.

Formally, we set

$$C_t^{(k)}(\Pi) = \sup_{\mathfrak{S}} I(\mathfrak{S}, \Pi) \qquad (k = 1, 2, 3, 4) \tag{8}$$

where $\mathfrak{S}$ ranges over the *PD*'s assigning probabilities $p_i \geqq 0$ to certain $v_i \in V$, $i = 1, \ldots, m$ ($\sum_{i=1}^{m} p_i = 1$) such that

(i) $$l(v_i) = t, \; i = 1, \ldots, m$$

(ii) $$l(v_i) \leqq t, \; i = 1, \ldots, m$$

(iii) $$\sum_{i=1}^{m} p_i l(v_i) = t$$

(iv) $$\sum_{i=1}^{m} p_i l(v_i) \leqq t,$$

[3] Since $t \log_2 t$ is convex, $(\Sigma \alpha_k t_k) \log_2 \Sigma \alpha_k t_k \leqq \Sigma \alpha_k t_k \log_2 t_k$ if $\Sigma \alpha_k = 1$; apply this to $t_k = \frac{\mathbf{P}\{\eta \in B_{jk} \mid \xi = x\}}{\mathbf{P}\{\eta \in B_{jk}\}}$ and $\alpha_k = \frac{\mathbf{P}\{\eta \in B_{jk}\}}{\mathbf{P}\{\eta \in B_j\}}$.

as $k = 1, 2, 3, 4$, respectively. If such $\mathcal{S}$ does not exist, the corresponding capacity is undefined.

For mathematical convenience, in the definition of t-cost capacities we restricted attention to the case that the input signal is governed by a PD with finite support. As the following lemma shows, this constraint is, in fact, irrelevant.

For arbitrary RV's ζ and η (taking values in sets X and Y, with underlying σ-algebras $\mathfrak{X}$ and $\mathfrak{Y}$, respectively) the mutual information $I(\xi; \eta)$ can be defined by the obvious generalization of (4) as[4]

$$I(\xi;\eta) = \sup_{\mathfrak{B}} \int \left(\sum_{j=1}^{n} \mathbf{P}\{\eta \in B_j \mid \xi = x\} \log_2 \frac{\mathbf{R}\{\eta \in B_j \mid \xi = x\}}{\mathbf{P}\{\eta \in B_j\}} \right) P_\xi(dx) \tag{9}$$

where the PD $\mathcal{S}_\xi$ is defined by $\mathcal{S}_\xi(A) = \mathbf{P}\{\xi \in A\}$ $(A \in \mathfrak{X})$ and the conditional probabilities $\mathbf{P}\{\eta \in B \mid \xi = x\}$ are defined as the essentially unique $\mathfrak{X}$-measurable functions of x satisfying $\int_A \mathbf{P}\{\eta \in B \mid \xi = x\} P_\xi(\mathrm{d}x) = \mathbf{P}\{\zeta \in A,\ \eta \in B\}$ for each $A \in \mathfrak{X}$.

Lemma 1. Let η and $\overline{\eta}$ denote arbitrary RV's taking values in V and $\overline{V}$, with underlying σ-algebras $\mathfrak{V}$ and $\overline{\mathfrak{V}}$, respectively, such that for each fixed $B \in \overline{\mathfrak{V}}$,

$$\mathbf{P}\{\overline{\eta} \in B \mid \eta = v\} = \mathcal{S}_v(B) \text{ for } \mathcal{S}_\eta - \text{almost all } v \in V. \tag{10}$$

Then under the constraints (i) $l(\eta) = t$, (ii) $l(\eta) \leqq t$, (iii) $El(\eta) = t$, (iv) $El(\eta) \leqq t$, we have

$$I(\eta;\overline{\eta}) \leqq C_t^{(k)}(\Pi), \tag{11}$$

with $k = 1, 2, 3, 4$, respectively.

Here $\mathfrak{V}$ can be any σ-algebra of subsets of V; the constraints (iii) and (iv) include that $l(\eta)$ is a RV, i.e., a measurable function on the underlying probability space.

Proof. In view of (9) and (10),

$$I(\eta;\overline{\eta}) = \sup_{\mathfrak{B}} \int_{V'} \left(\sum_{j=1}^{n} \mathcal{S}_v(B_j) \log_2 \frac{\mathcal{S}_v(B_j)}{\overline{\mathcal{S}}(B_j)} \right) \mathcal{S}_\eta(dv) \tag{12}$$

with

$$\overline{\mathcal{S}}(B_j) = \int_{V'} \mathcal{S}_v(B_j)\big((\mathcal{S}_\eta(dv)\big), \qquad j = 1, \ldots, n, \tag{13}$$

where $V' \in \mathfrak{V}$ is a set, possibly depending on the partition $\mathfrak{B} = (B_1, \ldots, B_n)$,

[4] It is an easy consequence of standard results (see [5]) that (9) is equivalent to the standard general definition of mutual information.

such that $\mathcal{S}_\eta(V') = 1$ and $\mathbf{P}\{\overline{\eta} \in B_j \mid \eta = v\} = \mathcal{S}_v(B_j)$ holds for all $v \in V'$ and $j = 1, \ldots, n$.

Given a partition $\mathfrak{B} = (B_1, \ldots, B_n)$ of $\overline{V}$ and $\varepsilon > 0$ arbitrarily small, there exists a measurable partition $\mathfrak{A} = (A_1, \ldots, A_m)$ of V' such that for $p_i = \mathcal{S}_\eta(A_i)$ and arbitrary $v_i \in A_i$, the sum in (6) (with $\overline{\mathcal{S}}$ defined as in (7)) is not less than the integral in (12) (with $\overline{\mathcal{S}}$ defined as in (13)) minus ε.

This proves the assertion for the cases (i) and (ii). Moreover, one can select two elements v_{i1} and v_{i2} in each A_i, as well, assigning to them probabilities p_{i1} and p_{i2} with $p_{i1} + p_{i2} = \mathcal{S}(A_i)$, and the above argument is still valid. But by proper choice, $p_{i1} l(v_{i1}) + p_{i2} l(v_{i2}) = \int\limits_{\{\omega:\eta(\omega) \in A_i\}} l(\eta(\omega))\, \mathbf{P}(\mathrm{d}\omega)$ can always be ensured, proving the assertion also for the cases (iii) and (iv).

The motivation for considering the t-cost capacities $C_t^{(k)}(\Pi)$, $k = 1, 2, 3, 4$ is that they provide simple necessary conditions for the possibility of reliable transmission of messages over the channel Π with a certain cost t. The transmission of messages from a given ensemble is described by four RV's ξ, η, $\overline{\eta}$, $\overline{\xi}$, forming a Markov chain, where ξ stands for the message to be transmitted, η denotes the input signal of the channel obtained by encoding ξ, $\overline{\eta}$ is the output signal and $\overline{\xi}$ denotes the result of a decoding operation performed on $\overline{\eta}$ (both encoding and decoding may be randomized, as well).

The minimum average amount of information needed to represent messages from the given ensemble with prescribed fidelity is called the *entropy* of the message ensemble with fidelity criterion. E.g., if the fidelity criterion is given in terms of a loss function and a threshold ε for tolerable average loss then this entropy is known as *ε-entropy* or *rate-distortion function*. In view of Lemma 1, it follows by a standard argument[5] that a necessary condition for the possibility of reliable transmission with

(i) cost t
(ii) cost not exceeding t
(iii) average cost t
(iv) average cost not exceeding t

consists in

$$H \leqq C_t^{(k)}(\Pi) \tag{14}$$

with $k = 1, 2, 3, 4$, respectively, where H denotes the entropy of the given message ensemble with fidelity criterion. This necessary condition will be referred to as the *converse Shannon theorem*.

[5] The so called data processing theorem states that for RV's ξ, η, $\overline{\eta}$, $\overline{\xi}$ forming a Markov chain $I(\xi, \overline{\xi}) \leqq I(\eta; \overline{\eta})$; if ξ, η, $\overline{\eta}$, $\overline{\xi}$ describe transmission of messages of the desired type then $I(\xi; \overline{\xi}) \geqq H$ and $I(\eta; \overline{\eta}) \leqq C_t^{(k)}(\Pi)$; cf. e.g. [3], § 2.10.

In the typical cases mentioned in the first paragraph, with the length of the input signal as cost function, one speaks of *transmission of length t* if the input signal is constrained to $l(\eta) = t$. By the converse Shannon theorem, for a message ensemble with fidelity criterion of entropy H reliable transmission of length t is possible only if $H \leqq C_t^{(1)}(\Pi)$. In this context, it is not usual to consider transmission of variable length thus, as a rule, it is $C_t^{(1)}(\Pi)$ what one means by "capacity". More exactly, the limit

$$C(\Pi) = \lim_{t \to \infty} \frac{1}{t} C_t^{(1)}(\Pi) \tag{15}$$

is commonly called the capacity (information capacity) of the channel Π) in most cases of practical interest, this limit exists. If $C(\Pi)$ is positive, (14 implies for any $\varepsilon > 0$ that reliable transmission of length t is possible for H sufficiently large only if

$$t \geqq \frac{H}{C(\Pi)} (1 - \varepsilon) . \tag{16}$$

The capacity (15) is a particularly significant characteristic of the channel in those cases when the *direct Shannon theorem* is true, i.e., if the necessary condition (16) of t-length transmissibility, with $-\varepsilon$ replaced by $+\varepsilon$, is sufficient, as well.[6] Still, the question seems to not have been considered, whether the lower bound (16) is valid also for transmission of variable length, i.e., for the constraint $l(\eta) = t$ replaced by $l(\eta) \leqq t$, $El(\eta) = t$ or $El(\eta) \leqq t$. This would immediately follow if the capacities $C_t^{(k)}(\Pi)$, $k = 1, 2, 3, 4$ were asymptotically equal as $t \to \infty$.

One aim of this paper is to show for general cost functions — under some regularity conditions — that not even $C_t^{(4)}(\Pi)$ can be essentially larger than $C_t^{(1)}(\Pi)$ (asymptotically, as $t \to \infty$). Let us point out that for the above considered cases this means that transmission of variable length cannot offer essential gain, if the direct Shannon theorem is valid. The situation is different if the direct Shannon theorem does not hold, as e.g. for messages coming from a non-ergodic source. In the latter case the lower bound (16) is not exact under the constraint $El(\eta) > t$ but, typically, it becomes exact under the weaker constraint $El(\eta) = t$. A closer study of these problems, however, is beyond the scope of this paper.

[6] For a rigorous statement and proof of the direct Shannon theorem under very general conditions, see Dobrushin [3].

2. Results

Given a channel Π with cost function l, let R_l denote the range of l and write

$$l_* = \inf R_l = \inf_{v \in V} l(v), \qquad l^* = \sup R_l = \sup_{v \in V} l(v). \tag{17}$$

The domain of the capacity functions $C_t^{(k)}(\Pi)$, $k = 1, 2, 3, 4$, see (8), is as follows:

(i) $C_t^{(1)}(\Pi)$ is defined for $t \in R_l$;

(ii) $C_t^{(2)}(\Pi)$ is defined in the open interval (l_*, ∞) or, if the inf in (17) can be replaced by min, in the semiclosed interval $[l_*, \infty]$;

(iii) $C_t^{(3)}(\Pi)$ is defined in the smallest interval containing R_l, i.e., in (l_*, l^*) or $[l_*, l^*)$ or $(l_*, l^*]$ or $[l_*, l^*]$;

(iv) the domain of $C_t^{(4)}(\Pi)$ is the same as that of $C_t^{(2)}(\Pi)$.

Obviously, $C_t^{(1)}(\Pi)$ is never larger and $C_t^{(4)}(\Pi)$ is never smaller than the other three types of capacity. Intuitively, the capacity functions $C_t^{(k)}(\Pi)$ are expected to be increasing. Actually, this is trivial for $k = 2, 4$ but not necessarily true for $k = 1$ (as trivial though artificial examples show). The corollary of the next lemma shows that if $l^* = \infty$, as in real communication channels, then also $C_t^{(3)}(\Pi)$ is an increasing function.

Lemma 2. $C_t^{(3)}(\Pi)$ is a continuous concave function of t in the open interval (l_*, l^*), provided that it is not identically infinite in this interval.

Corollary. $C_t^{(3)}(\Pi)$ is non-decreasing continuous concave function of $t \in (l^*, \infty)$ and in the case $l^* = \infty$ $C_t^{(3)}(\Pi) = C_t^{(4)}(\Pi)$ for all $t \in (l_*, \infty)$.

Proof. $I(\mathcal{P}, \Pi)$ defined by (6) is well known to be a concave function of $\mathcal{P}$, i.e., for $\mathcal{P}_1$ and $\mathcal{P}_2$ assigning probabilities p_{1i} and p_{2i} to certain v_i's, $i = 1, \ldots, m$ (the latter may be considered to be the same in both cases since 0 probabilities are permitted) and for any $0 < \alpha < 1$, the PD $\mathcal{P}$ assigning probabilities $p_i = \alpha p_{1i} + (1 - \alpha) p_{2i}$ to the v_i's satisfies

$$I(\mathcal{P}, \Pi) \geqq \alpha I(\mathcal{P}_1, \Pi) + (1 - \alpha) I(\mathcal{P}_2, \Pi). \tag{18}$$

(In fact, the analogous inequality is immediately checked for the sums in (6), for any fixed partition $\mathcal{B}$; but any partitions $\mathcal{B}_1$ and $\mathcal{B}_2$ associated with $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively, may be replaced by their common refinement $\mathcal{B}$, without decreasing the corresponding sum, cf. (5), thus (18) follows.) The concavity of $C_t^{(3)}(\Pi)$, i.e., the inequality

$$C^{(3)}_{\alpha t_1 + (1-\alpha) t_2}(\Pi) \geqq \alpha C_{t_1}^{(3)}(\Pi) + (1 - \alpha) C_{t_2}^{(3)}(\Pi) \tag{19}$$

is an immediate consequence of (19) and the definition of $C_t^{(3)}(\Pi)$, because if $\mathcal{S}_1$ and $\mathcal{S}_2$ satisfy the constraint (iii) following (8), with certain t_1 and t_2, the the above $\mathcal{S}$ satisfies the same constraint with $t = \alpha t_1 + (1 - \alpha) t_2$.

Now define

$$D_t = \sup_{El(\eta)=t} I\big(l(\eta) ; \bar{\eta}\big) \tag{20}$$

where the sup refers to RV's η taking values in finite subsets of V and $\bar{\eta}$ is connected with η by (2).

Theorem 1. Let $l^* = \infty$ and $\sup_{t<t_0} C_t^{(1)}(\Pi) < \infty$ for each $t_0 \in (l_*, \infty)$. Then

$$\lim_{t\to\infty} \frac{1}{t} D_t = 0 \tag{21}$$

implies

$$\lim_{t\to\infty} \frac{1}{t} C_t^{(4)}(\Pi) = \overline{\lim_{t\to\infty}} \frac{1}{t} C_t^{(1)}(\Pi) . \tag{22}$$

The existence of the limit in (22) follows from the corollary of Lemma 2. Since $C_t^{(4)}(\Pi) = C_t^{(3)}(\Pi) \geq C_t^{(2)}(\Pi) \geq C_t^{(1)}(\Pi)$, Theorem 1 means that the common value of the two sides of (22) deserves the name *capacity per unit cost of the channel* Π.

Before proving Theorem 1, let us discuss condition (21). It means, intuitively, that the cost of the input signal contributes only an asymptotically negligible amount to the information transmitted by the channel. With the signal length cost assignment, this is always true for discrete time channels (see Lemma 3 below), but for continuous time channels (21) does not hold if one insists that the output signal be of the same length as the input one, i.e., that the PD's $\mathcal{S}_v$ be concentrated on the sets $\{\bar{v} : l(\bar{v}) = l(v)\}$. In the latter case, the pathological situation obtains that by encoding the messages to be transmitted just into the length of the input signal, an arbitrarily large amount of information is transmissible with arbitrarily small average cost. Of course, the mentioned seemingly natural assumption on the $\mathcal{S}_v$'s is unrealistic, because, in practice, the receiver never can exactly tell the length of the input signal from the output signal. This analysis reveals the fact that reasonable models of continuous time channels must incorporate synchronization errors; still, the calculation of capacity per unit transmission length as the limit of the maximal mutual information of an input signal of length t and the t-length segment of the corresponding output signal, divided by t, is often justified, namely if the neglected segment of the output signal can be shown to contribute asymptotically negligible additional amount of information on the input.

Lemma 3. Suppose that the possible values of the cost function l form a sequence $l_* = l_0 < l_1 < l_2 < \dots$ such that $\frac{l_n}{\log n} n \to \infty$ as $n \to \infty$. Then condition (21) is always fulfilled.

Proof. For any discrete RV ξ and arbitrary RV η, we have

$$I(\xi;\ \eta) \leqq H(\xi) \tag{23}$$

where $H(\xi) = -\sum_x \mathbf{P}\{\xi = x\} \log_2 \mathbf{P}\{\xi = x\}$ (this well known fact immediately follows from (4) since $\frac{\mathbf{P}\{\eta \in B_j \mid \xi = x\}}{\mathbf{P}\{\eta \in B_j\}} \leqq \frac{1}{\mathbf{P}\{\xi = x\}}$).

Under the constraint $\sum_{n=0}^{\infty} p_n l_n = t$, the entropy $H = -\sum_{n=0}^{\infty} p_n \log_2 p_n$ is upper bounded by $\alpha t + \log_2 \sum_{n=0}^{\infty} 2^{-\alpha l_n}$ where $\alpha > 0$ is arbitrary; this follows from the inequality $\sum_{n=0}^{\infty} p_n \log_2 \frac{2^{-\alpha l_n}}{p_n} \leqq \log_2 \sum_{n=0}^{\infty} 2^{-\alpha l_n}$, implied by the concavity of the function $\log_2 t$. Thus in view of (20) and (23) we arrive at

$$D_t \leqq \alpha t + \log_2 \sum_{n=0}^{\infty} 2^{-\alpha l_n}. \tag{24}$$

By the assumption $\frac{l_n}{\log n} \to \infty$, the series in (24) converges for arbitrarily small $\alpha > 0$, thus the assertion follows.

Proof of Theorem 1. Let $\mathcal{S}$ be any P be any PD assigning probabilities p_i to certain elements $v_i \in V$, $i = 1, \dots, m$ ($\sum_{i=1}^{m} p_i = 1$), let η and $\bar{\eta}$ denote an input and output RV, respectively, with $\mathbf{P}\{\eta = v_i\} = p_i$, $i = 1, \dots, m$ and $\bar{\eta}$ connected with η by (2). We claim for $\sum_{i=1}^{m} p_i l(v_i) = t$

$$I(\mathcal{S}, \Pi) = I(\eta\,;\, \bar{\eta}) \leqq EC^{(1)}_{l(\eta)}(\Pi) + D_t. \tag{25}$$

To see this, let $l_1, \dots, l_k$ denote the different costs associated with the v_i's; for $h = 1, \dots, k$, set $q_h = \mathbf{P}\{l(\eta) = l_h\} = \sum_{i:l(v_i)=l_h} p_i$, let $\mathcal{S}_h$ denote the PD assigning the probabilities $\frac{p_i}{q_h}$ to the v_i's with $l(v_i) = l_h$ (i.e., the conditional distribution of η given $l(\eta) = l_h$) and write $\bar{\mathcal{S}} = \sum_{i:l(v_i)=l_h} \frac{p_i}{q_h} \mathcal{S}_{v_i}$ (i. e., $\mathcal{S}_h(B) = \mathbf{P}\{\bar{\eta} \in B \mid l(\eta) = l_h\}$, $B \in \bar{\mathfrak{B}}$).

Then, for any measurable partition $B = (B_1, \ldots, B_n)$ of $\overline{V}$

$$\sum_{i=1}^{m} p_i \sum_{j=1}^{n} \mathcal{S}_{v_i}(B_j) \log_2 \frac{\mathcal{S}_{v_i}(B_j)}{\overline{\mathcal{S}}(B_j)} =$$

$$= \sum_{k=1}^{k} \sum_{i:l(v_i)=l_k} p_i \sum_{j=1}^{n} \mathcal{S}_{v_i}(B_j) \left(\log_2 \frac{\mathcal{S}_{v_i}(B_j)}{\overline{\mathcal{S}}_h(B_j)} + \log_2 \frac{\overline{\mathcal{S}}_h(B_j)}{\overline{\mathcal{S}}(B_j)} \right) =$$

$$= \sum_{h=1}^{k} q_h \sum_{i:l(v_i)=l_k} \frac{p_i}{q_h} \sum_{j=1}^{n} \mathcal{S}_{v_i}(B_j) \log_2 \frac{\mathcal{S}_{v_i}(B_j)}{\overline{\mathcal{S}}_h(B_j)} + \tag{26}$$

$$+ \sum_{h=1}^{k} q_h \sum_{j=1}^{n} \overline{\mathcal{S}}_h(B_j) \log_2 \frac{\overline{\mathcal{S}}_h(B_j)}{\overline{\mathcal{S}}(B_j)} .$$

Taking sup for $\mathfrak{B}$, in view of (6) and (4) we obtain

$$I(\mathcal{S}, \Pi) \leq \sum_{h=1}^{k} q_h I(\mathcal{S}_h, \Pi) + I(l/\eta) ; \overline{\eta}) . \tag{27}$$

Since $\mathcal{S}_h$ is supported by v_i's with $l(v_i) = l_h$, we have $I(\mathcal{S}_h, \Pi) \leq C_{l_h}^{(1)}(\Pi)$, and since $q_h = \mathbf{P}\{l(\eta) = l_h\}$, (25) follows from (27) and (20).

Now write $\overline{\lim_{t\to\infty}} \frac{1}{t} C_t^{(1)}(\Pi) = C$ and, given $\varepsilon > 0$, let t_0 be so large that $C_t^{(1)}(\Pi) < t(C + \varepsilon)$ for $t \geq t_0$. Then $C_{l(\eta)}^{(1)}(\Pi) \leq l(\eta)\ (C + \varepsilon) + \sup_{t<t_0} C_t^{(1)}(\Pi)$ thus

$$EC_{l(\eta)}^{(1)}(\Pi) \leq t(C + \varepsilon) + \sup_{t<t_0} C_t^{(1)}(\Pi) \quad \text{if} \quad El(\eta) = t . \tag{28}$$

Since $C_t^{(3)}(\Pi)$ is, by definition, the supremum of $I(\mathcal{S}, \Pi)$ under the constraint $\sum_{i=1}^{m} p_i l(v_i) = El(\eta) = t$, in view of the conditions of Theorem 1, (25) and (28) give rise to

$$\lim_{t\to\infty} \frac{1}{t} C_t^{(3)}(\Pi) \leq C + \varepsilon . \tag{29}$$

But $C_t^{(4)}(\Pi) = C_t^{(3)}(\Pi) \geq C_t^{(1)}(\Pi)$, and in (29) $C = \overline{\lim_{t\to\infty}} \frac{1}{t} C_t^{(1)}(\Pi)$ and $\varepsilon < 0$ is arbitrary, thus the proof of Theorem 1 is complete.

Remark. Using that the sums in (4) do not decrease when refining the partition $\mathfrak{B}$, inequality (27) turns out, in fact, to be an equality. This is the identity $I(\eta; \overline{\eta}) = I(\eta; \overline{\eta} \mid l(\eta)) + I(l(\eta); \overline{\eta})$ which — as a particular case of Kolmogorov's identity — is valid also for arbitrary η. What does not seem possible for arbitrary η — without imposing some regularity condition — is the upper bounding of $I(\eta; \overline{\eta} \mid l(\eta))$ in terms of the capacity function $C_t^{(1)}(\Pi)$.

Thus the approach of defining capacity through discrete input RV's was essentially used in our proof.

Theorem 2. If $l^* = \infty$,

$$\lim_{t\to\infty} \frac{1}{t} C_t^{(4)}(\Pi) = \overline{\lim_{t\to\infty}} \frac{1}{t} C_t^{(2)}(\Pi) \tag{30}$$

holds without any further condition.

Proof. Consider the quantized cost function $\bar{l}$ defined by $\bar{l}(v) = n$ if $n \leq l(v) < n + 1$ $(n = 0, 1, \ldots)$; denote the corresponding t-cost capacities by $\bar{C}_t^{(k)}(\Pi)$, $k = 1, 2, 3, 4$. Then $0 \leq l(v) - \bar{l}(v) < 1$ implies $\bar{C}_n^{(1)}(\Pi) \leq \bar{C}_n^{(2)}(\Pi) \leq C_{n+1}^{(2)}(\Pi)$ and $C_t^{(4)}(\Pi) \geq \bar{C}_{t-1}^{(4)}(\Pi)$. For $\bar{l}$, (20) is fulfilled by Lemma 3, and $\sup_{n<t_0} \bar{C}_n^{(1)}(\Pi) < \infty$ also holds if $C_{t_0}^{(2)}(\Pi) < \infty$ (but (30) is trivial if $C_{t_0}^{(2)}(\Pi) = \infty$ for some t_0).

Thus Theorem 1 applies for $\bar{l}$ and we have

$$\lim_{t\to\infty} \frac{1}{t} C_t^{(4)}(\Pi) \leq \lim_{t\to\infty} \frac{1}{t} \bar{C}_t^{(4)}(\Pi) = \overline{\lim_{n\to\infty}} \frac{1}{n} \bar{C}_n^{(1)}(\Pi) \leq \overline{\lim_{t\to\infty}} \frac{1}{t} C_t^{(2)}(\Pi)\,. \tag{31}$$

Since $C_t^{(2)}(\Pi) \leq C_t^{(4)}(\Pi)$ for all $t \in (l_*, \infty)$, the proof of Theorem 2 is complete.

Our next aim is to generalize the above results for channels with constrained input. Since input constraints are very common in communication problems, such a generalization is desirable, particularly in case of Theorem 1. Recall that — as pointed out in § 1 — it is the identity (22) on the basis of which one can assert, both for discrete and continuous time channels with the signal length cost assignment, that the capacity bound for the length of reliable transmission is valid also for variable length transmission, if $t = El(\eta)$ denotes the average transmission length.

Suppose that beside the cost function $l(v) \geq 0$ there is also a constraining function $m(v) \geq 0$ defined on V. Intuitively, it is most convenient to think of $l(v)$ and $m(v)$ as the length and power of the signal $v \in V$.

Constraints of type $m(\eta) \leq M$ with probability 1 (where M is a constant) are not interesting in the present context, because such a constraint can be eliminated by giving the role of V to its subset $V' = \{v : m(v) \leq M\}$. We shall be interested in constraints of type

$$\frac{El(\eta)\, m(\eta)}{El(\eta)} \leq M\,, \tag{32}$$

if l and m denote signal length and power, the left hand side of (32) is the expected energy of the input divided by the expected length, i.e., (32) is an average power constraint.

Define the constrained t-cost capacities as

$$C^{(k)}_{t,M}(\Pi) = \sup_{\mathcal{S}} I(\mathcal{S}, \Pi) \qquad (k = 1, 2, 3, 4) \tag{33}$$

where $\mathcal{S}$ ranges over the PD's with finite support satisfying the corresponding constraint indicated after (8), as well as the additional constraint

$$\frac{\sum_{i=1}^{m} p_i l(v_i)\, m(v_i)}{\sum_{i=1}^{m} p_i l(v_i)} \leq M\,. \tag{34}$$

Of course, in the case $k = 1$ (34) reduces to

$$\sum_{i=1}^{m} p_i m(v_i) \leq M\,. \tag{35}$$

Assuming $l^* = \infty$, write

$$V_t = \{v : l(v) = t\}, \quad m_t = \inf_{v \in V_t} m(v), \quad \underline{m} = \lim_{t \to \infty} m_t, \quad \overline{m} = \overline{\lim_{t \to \infty}}\, m_t\,; \tag{36}$$

here t ranges over R_l, the range of the cost function l. Let $R_l(M)$ denote the subset of R_l where $m_t < M$ or $m_t = M$ and $m_t = m(v)$ for some $v \in V_t$. Then $C^{(1)}_{t,M}(\Pi)$ is defined for $t \in R_l(M)$, while $C^{(k)}_{t,M}(\Pi)$, $k = 2, 3, 4$, as functions of t, are defined in intervals, the description of which may be omitted.

Lemma 1 remains valid for the constrained case and so do Lemma 2 and its corollary (with the same proof). The only change needed in Lemma 2 is to replace (l_*, l^*) by the open interval where $C^{(3)}_{t,M}(\Pi)$ is defined; in particular, the condition $l^* = \infty$ in the corollary turns to this interval being infinite, and that is certainly true if $M > \underline{m}$.

Let us observe that $C^{(1)}_{t,M}(T)$ and $C^{(3)}_{t,M}(\Pi)$ are concave functions of M, as well (defined in an infinite interval, if at all), for any fixed t. This follows by the same argument[7] as Lemma 2.

In order to generalize Theorem 1 to constrained capacity functions, write

$$C(M) = \overline{\lim_{t \to \infty}} \frac{1}{t} C^{(1)}_{t,M}(\Pi) \qquad (M > m) \tag{37}$$

where t runs to infinity through $R_l(M)$ (the assumption $M > \underline{m}$ ensures $\sup R_l(M) = \infty$, see (36). If in (37) actually the limit exists then $C(M)$ is concave. In general, however, this need not be the case, therefore we shall

[7] The concavity of $C^{(1)}_{t,M}(\Pi)$ as a function of M is also a formal consequence of the corollary of Lemma 2: consider the channel $\Pi_t = \{\mathcal{S}_v : v \in V_t\}$ with cost function $m(v)$.

consider the concave envelope of $C(M)$, i.e., the infimum $\bar{C}(M)$ of all concave functions $\tilde{C}(M) \geq C(M)$, $M \in (\underline{m}, \infty)$. We set $\bar{C}(\underline{m}) = \lim_{M \to \underline{m}+0} \bar{C}(M)$.

Let us make the regularity assumptions that $\overline{m} < \infty$ and that to each t_0 and $\varepsilon > 0$ there exists a finite $K = K(t_0, \varepsilon)$ such that

$$C^{(1)}_{t,M}(\Pi) < t\big(\bar{C}(M) + \varepsilon M\big) + K \quad \text{if} \quad t < t_0, \quad t \in R_t \tag{38}$$

for all $M > \max(\underline{m}, m_t)$.

Theorem 3. Under the above assumptions and condition (21), we have

$$\lim_{t \to \infty} \frac{1}{t} C^{(3)}_{t,M}(\Pi) = \bar{C}(M) \quad \text{for all} \quad M > \underline{m}. \tag{39}$$

Remark. If $m^* = \sup_{v \in V} m(v) < \infty$ then $C^{(1)}_{t,M}$ as a function of M is constant for $M \geq m^*$, thus condition (38) is trivially fulfilled if $\sup_{t<t_0} C^{(1)}_t(\Pi) < \infty$. Moreover, in the case $l_* > 0$, a useful sufficient condition of the validity of assumption (38) is $\lim_{M \to \infty} \frac{1}{M} \sup_{t<t_0} C^{(1)}_{t,M}(\Pi) = 0$ fo all t_0.

Proof of Theorem 3. We proceed as in the proof of Theorem 1. In addition to the notations used there, set $M_h = \sum_{i:l(v_i)=l_h} \frac{p_i}{q_h} m(v_i)$ $h = 1, \ldots k$. Then the PD's $\mathfrak{S}_h$ satisfy the constraint (35) with M_h in the role of M, thus $I(\mathfrak{S}_h, \Pi) \leq C^{(1)}_{l_h, M_h}(\Pi)$. In this way (27) gives rise to

$$I(\mathfrak{S}, \Pi) \leq \sum_{h=1}^{k} q_h C^{(1)}_{l_h, M_h}(\Pi) + D_t. \tag{40}$$

Now suppose that $\mathfrak{S}$ satisfies the constraints in the definition of $C^{(3)}_{t,M}(\Pi)$ i.e. constraint (iii) after (8) and (34). In the present notation, this means

$$\sum_{h=1}^{k} q_h l_h = t\,; \quad \frac{\sum_{h=1}^{k} q_h l_h M_h}{\sum_{h=1}^{k} q_h l_h} \leq M. \tag{41}$$

We have to upper bound the sum in (40) in terms of $\bar{C}(M)$. This is a matter of a routine but somewhat tedious reasoning.

By considering $\underline{m} = a_0 < \ldots < a_r = \overline{m} + 1 = b_0 < \ldots < b_s = \infty$ such that the variation of $\bar{C}(M)$ in each $[a_{i-1}, a_i)$ and the variation of $\frac{\bar{C}(M)}{M - \overline{m}}$

in each $[b_{j-1}, b_j)$ in less than $\frac{\varepsilon}{2}$, from $\overline{\lim_{t\to\infty}} \frac{1}{t} C^{(1)}_{t,M}(\Pi) \leqq \bar{C}(M)$ one readily concludes that the nondecreasing functions $f_t(M) = \frac{1}{t} C^{(1)}_{t,M}(\Pi)$ are below $\bar{C}(M) + \varepsilon$ in $[\underline{m}, \overline{m} + 1]$ and the nonincreasing[8] functions $g_t(M) = \frac{1}{M - \overline{m}} \cdot \frac{1}{t} C^{(1)}_{t,M}(\Pi)$ are below $\frac{\bar{C}(M)}{M - \overline{m}} + \varepsilon$ in $[\overline{m} + 1, \infty)$, if t is sufficiently large. This together with the assumption (38) implies that to any $\varepsilon > 0$ there exists $K = K(\varepsilon)$ such that

$$C^{(1)}_{t,M}(\Pi) < t\big(\bar{C}(M) + \varepsilon(M + 1)\big) + K \tag{42}$$

for all t and $M \in [\underline{m}, \infty)$ where the left hand side is defined.

Let $\delta > 0$ be arbitrary and let t_0 be such that $m_t > \underline{m} - \delta$ for $t \in R_t$, $t > t_0$. Then $M_h \leqq \underline{m} - \delta$ implies $l_h \leqq t_0$ (because, by definition, $m_{l_h} \leqq M_h$). Thus

$$\sum_{h:\, M_h < \underline{m}} q_h l_h (\underline{m} - M_h) \leqq \delta \sum_{h=1}^{k} q_h l_h + \underline{m} t_0 . \tag{43}$$

On account of (42) and the fact that $C^{(1)}_{t,M}(\Pi)$ is a nondecreasing function of M, we have

$$\sum_{h=1}^{k} q_h C^{(1)}_{l_h, M_h}(\Pi) < \sum_{h:\, M_h \geqq \underline{m}} q_h l_h \big(\bar{C}(M_h) + \varepsilon(M_h + 1)\big) + \\ + \sum_{h:\, M_h < \underline{m}} q_h l_h \big(\bar{C}(\underline{m}) + \varepsilon(\underline{m} + 1\big) + K . \tag{44}$$

Since $\bar{C}(M)$ is concave, by Jensen's inequality

$$\sum_{h:\, M_h \geqq m} q_h l_h \bar{C}(M_h) + \sum_{h:\, M_h < \underline{m}} q_h l_h \bar{C}(m) \leqq \\ \leqq \left(\sum_{h=1}^{k} q_h l_h \right) \bar{C} \left(\frac{\sum_{h=1}^{k} q_h l_h M_h + \sum_{h:\, M_h < \underline{m}} q_h l_h (\underline{m} - M_h)}{\sum_{h=1}^{k} q_h l_h} \right). \tag{45}$$

In view of (41), (43) and the continuity of the function $\bar{C}$, the rigth hand side of (45) is $\leqq t\big(\bar{C}(M) + \varepsilon\big)$ if t is large enough.

[8] Here we use that $C^{(1)}_{t,M}(\Pi)$ is a concave function of M and its domain certainly includes $[\overline{m} + 1, \infty)$ if t is large enough.

Hence (44) gives for sufficiently large t

$$\sum_{h=1}^{k} q_h C^{(1)}_{l_h, M_h}(\Pi) < t\big(\bar{C}(M) + \varepsilon\big) + \varepsilon(Mt + t) + \varepsilon(\underline{m} + 1)\, t + K\,. \tag{46}$$

Here $\varepsilon > 0$ has been arbitrary, thus from (40), (46) and (21) follows

$$\lim_{t\to\infty} \frac{1}{t} C^{(3)}_{t,M}(\Pi) \leqq \bar{C}(M) \quad \text{for all} \quad M > \underline{m}. \tag{47}$$

On the other hand, $\lim\limits_{t\to\infty} \frac{1}{t} C^{(3)}_{t,M}(\Pi)$ is a concave function of M which is obviously $\geq \overline{\lim\limits_{t\to\infty}} \frac{1}{t} C^{(1)}_{t,M}(\Pi) = C(M)$. Thus, since $\bar{C}(M)$ denotes the concave envelope of $C(M)$, the converse of inequality (47) is trivial.

The proof of Theorem 3 is complete.

The generalization of Theorem 2 to constrained capacity functions will be omitted.

3. Examples

Example 1. Let Y be a finite or countable alphabet, let Y^* denote the set of all finite strings of elements of Y and let V be a subset of Y^*. Suppose that $\overline{V} = V$ and that for each $v \in V$, $\mathcal{S}_v$ is the unit mass at v. Then $\Pi = \{\mathcal{S}_v : v \in V\}$ is a *noiseless channel* with alphabet Y; if l is a given cost function on V, of which we assume $l^* = \infty$ and $l(u) \leqq l(v)$ for $u \prec v$, we speak of a noiseless channel (V, l) (cf. [1] and [2]). In this case

$$C^{(1)}_t(\Pi) = \log_2 N(t), \quad C^{(2)}_t(\Pi) = \log_2 \overline{N}(t)\,, \tag{48}$$

where $N(t)$ and $\overline{N}(t)$ denote the number of strings $v \in V$ with cost $l(v) = t$ and $l(v) \leqq t$, respectively; furthermore, $C^{(3)}_t(\Pi)$ is the supremum of the Shannon entropy $H(\eta)$ of RV's η taking values in V with $El(\eta) = t$. Let us call

$$C = \overline{\lim_{t\to\infty}} \frac{1}{t} \log_2 \overline{N}(t) \tag{49}$$

(the *capacity* of the noiseless channel (V, l). Theorem 2 implies that $\lim\limits_{t\to\infty} \frac{1}{t} C^{(3)}(\Pi) = C$. Since transmission of average length L of messages represented by a RV ξ is possible only if $H(\xi) \leqq C^{(3)}_L(\Pi)$ (having the fidelity criterion in mind that the received message must coincide with the original one), it is possible to assert, at least in an asymptotic sense, that the average transmission length L cannot be smaller than $\frac{H(\xi)}{C}$; this is the converse Shan-

non theorem for noiseless channels. If the cost function l satisfies the condition of Lemma 2, Theorem 1 implies that

$$C = \overline{\lim_{t\to\infty}} \frac{1}{t} \log_2 N(t) . \tag{50}$$

The condition of Lemma 2 is certainly fulfilled if Y is finite and the cost function is additive, i.e.

$$l(v) = \sum_{i=1}^{n} l(y_i) \quad \text{for} \quad v = y_1 \dots y_n \tag{51}$$

and, more generally, also if there exists a finite set of non-negative numbers such that the cost of each n-length string $v \in V$ is the sum of n terms from this set; the latter condition is fulfilled, e.g., for finite-state channels.

(Different possible definitions of capacity of noiseless channels and the direct and converse Shannon theorems for such channels were discussed in detail in [2].)

Example 2. Let $\overline{Y}$ and Y be arbitrary sets, let $V = \bigcup_{n=1}^{\infty} Y^n$, $\overline{V} = \bigcup_{n=1}^{\infty} \overline{Y}^n$ and let $\overline{\mathfrak{y}}$ be the σ-algebra spanned by $\bigcup_{n=1}^{\infty} \overline{\mathfrak{y}}^n$, where $\overline{\mathfrak{y}}$ is a given σ-algebra of subsets of Y. Let the PD's $\mathfrak{S}_v$ for $v = (y_1, \dots, y_n) \in Y^n$ be concentrated on $\overline{Y}^n$ such that

$$\mathfrak{S}_v(B_1 \times \dots \times x\, B_n) = \prod_{i=1}^{n} \mathfrak{S}_{y_i}(B_i) \quad (B_i \in \overline{\mathfrak{y}},\ i = 1, \dots, n) . \tag{52}$$

In this case $\Pi = \{\mathfrak{S}_v : v \in V\}$ is called a *memoryless channel*, generated by the elementary channel $\Pi_1 = \{\mathfrak{S}_y : y \in Y\}$. Let $l(v)$ be the length of the string $v \in V$. Then $R_l = \{1, 2, \dots\}$ and, as well known, $C_n^{(1)}(\Pi) = nC(\Pi_1)$ for each n; here $C(\Pi_1) = \sup_{\mathfrak{S}} I(\mathfrak{S}, \Pi_1)$, $\mathfrak{S}$ running over all PD's with finite support on Y. Moreover, if m is a constraining function of form

$$m(v) = \frac{1}{n} \sum_{i=1}^{n} m(y_i) \qquad \text{for} \qquad v = (y_1, \dots, y_n) , \tag{53}$$

the constrained capacity $C_{n,M}^{(1)}(\Pi)$ equals $nC_M(\Pi_1)$; here $C_M(\Pi_1) = \sup_{\mathfrak{S}} I(\mathfrak{S}, \Pi_1)$ $\mathfrak{S}$ running over all PD's with finite support $\{y_1, \dots, y_k\} \subset Y$ satisfying the constraint $\sum_{i=1}^{k} p_i m(y_i) \leq M$ (cf. e.g. [4], p. 325).

In this case the conditions of Theorems 1 and 3 are fulfilled, thus

$$\lim_{t\to\infty}\frac{1}{t}C_t^{(3)}(\Pi)=C(\Pi_1)\,;\quad \lim_{t\to\infty}\frac{1}{t}C_{t,M}^{(3)}(\Pi)=C_M(\Pi_1)\,. \tag{54}$$

Observe that in our case (24) gives $\left(\text{with the choice } \alpha=\frac{t}{t-1}\right) D_t \leqq$ $\leqq t\log_2\frac{t}{t-1}+\log_2(t-1)<\log_2(et)$, and therefore (54) can be sharpened to

$$\begin{aligned} tC(\Pi_1)\leqq C_t^{(3)}(\Pi)<tC(\Pi_1)+\log_2(et)\,,\\ tC_M(\Pi_1)\leqq C_{t,M}^{(3)}(\Pi)<tC_M(\Pi_1)+\log_2(et)\,. \end{aligned} \tag{55}$$

In particular, for a message ensemble with fidelity criterion of entropy H, reliable transmission of average length t is possible only if $H<tC(\Pi_1)+$ $+\log_2(et)$, where, in the constrained case, $C(\Pi_1)$ is to be replaced by $C_M(\Pi_1)$.

Example 3. Let Π be a memoryless channel generated by $\Pi_1=$ $=\{\mathcal{S}_y\colon y\in Y\}$ as in Example 2, but now with an arbitrary additive cost function l, see (51). Then, by definition, $C_t^{(1)}(\Pi)=\sup_{\mathcal{S}} I(\mathcal{S},\Pi)$, where $\mathcal{S}$ runs over the PD's with finite support $\{v_1,\ldots,v_k\}\subset V$ satisfying $\sum_{i=1}^{k}p_i l(v_i)\leqq t$. Consider first the supremum under the additional constraint $\sum_{i=1}^{k}p_i l_0(v_i)=\tau$, where $l_0(v)$ denotes the length of v. This is the same as $C_{\tau,M}^{(3)}(\Pi)$ for cost function $l_0(v)$, constraining function $m(v)=\frac{l(v)}{l_0(v)}$ and $M=\frac{t}{\tau}$, thus in view of (55), we have

$$\sup_M\frac{t}{M}C_M(\Pi_1)\leqq C_t^{(4)}(\Pi)\leqq\sup_M\left(\frac{t}{M}C_M(\Pi_1)+\log_2\left(e\frac{t}{M}\right)\right); \tag{56}$$

here $C_M(\Pi_1)=\sup_{\mathcal{S}} I(\mathcal{S},\Pi_1)$, $\mathcal{S}$ running over all PD's with finite support $\{y_1,\ldots,y_k\}\subset Y$ satisfying the constraint $\sum_{i=1}^{k}p_i\,l(y_i)\leqq M$ and M runs over all positive numbers for which such a $\mathcal{S}$ exists.

In the case $l_*>0$ we obtain from (56) and Theorem 2 the intuitively expected relation[9]

$$\lim_{t\to\infty}\frac{1}{t}C_t^{(4)}(\Pi)=\lim_{t\to\infty}\frac{1}{t}C_t^{(2)}(\Pi)=\sup_M\frac{C_M(\Pi_1)}{M}\,, \tag{57}$$

[9] For memoryless channels with additive cost function, $\lim_{t\to\infty}\frac{1}{t}\cdot C_t^{(k)}(\Pi)$ exists also for $k=1,2$; this follows from the easily checked relation $C^{(k)}(\Pi)+C_2^{(k)}(\Pi)\leqq$ $\leqq C_{t_1+t_2}^{(k)}(\Pi)$.

and, by Theorem 1, under condition (21) also $\lim_{t\to\infty}\frac{1}{t}C_t^{(1)}(\Pi)$ has the same value; observe that the sufficient condition of the validity of (21) given in Example 1 is often useful for noisy channels, too. Since $C_M(\Pi)$ is a concave function of M (by the corollary of Lemma 2), typically there is a unique $M = M_0$ maximizing $\frac{C_M(\Pi_1)}{M}$. In order to transmit maximum information per unit cost, it is suggested to work at average cost M_0 per one channel use.

In the case $l_* = 0$ we have the pathological situation $C_t^{(4)}(\Pi) = \infty$ (and also $C_t^{(1)}(\Pi) = \infty$) because one can construct signals of arbitrary length of arbitrarily small cost, and the signal length itself can represent the information to be transmitted. One way to overcome this problem is to consider channels with synchronization errors. Another possibility is to set up a constraint on the ratio of the length of the input signal and its cost; considering in (57) such constrained capacities instead of the unconstrained ones, the limit will differ from $\sup_M \frac{C_M(\Pi_1)}{M} = \lim_{M\to 0}\frac{C_M(\Pi_1)}{M}$ arbitrarily little, if the mentioned constraint is loose enough. Thus it is meaningful to call $\sup_M \frac{C_M(\Pi_1)}{M}$ *capacity per unit cost* in the case $l_* = 0$, as well. For example, for channels with additive Gaussian noise (the $\mathfrak{S}_y$'s being Gaussian, with mean y and variance N, the noise power) and the energy cost function $l(y) = y^2$, one has $C_M(\Pi_1) = \frac{1}{2}\log_2\left(1+\frac{M}{N}\right)$; the capacity per unit energy in the above sense equals $\frac{\log_2 e}{2N}$ and this can be approached by transmitting information with a very small signal power.

Example 4. Let us be given a memoryless channel with the transmission length cost assignment, as in Example 2, but this time with feedback; this is the case when transmission of variable length is most often used in practice (the author is indebted to Prof. T. Kailath for calling his attention to this important example). In the presence of feedback, the channel capacity can not be reasonably defined in terms of the mutual information of the input and output signals $\eta = (\eta_1, \ldots, \eta_\tau)$ and $\bar{\eta} = (\bar{\eta}_1, \ldots, \bar{\eta}_\tau)$; rather, to keep the intuitive meaning of this concept, the mutual information $I(\xi; \bar{\eta})$ is to be maximized, where ξ stands for the message to be transmitted. Formally, a method of transmission for ξ (taking values in a set X with underlyin σ-algebra $\mathfrak{X}$) is defined by prescribing probabilities $p_{x,y_1,\ldots,y_{k-1},\bar{y}_1,\ldots,\bar{y}_{k-1}}$ as well as PD's $\mathfrak{S}_{x,y_1,\ldots,y_{k-1},\bar{y}_1,\ldots,\bar{y}_{k-1}}$ on $(X, \mathfrak{Y})$, where $\mathfrak{Y}$ is an a priori given σ-algebra of subsets of Y; both $p_{x,y_1,\ldots,y_{k-1}}$ and $\mathfrak{S}_{x,y_1,\ldots,y_{k-1},\bar{y}_1,\ldots,\bar{y}_{k-1}}(A)$ (for any fixed $A \in \mathfrak{Y}$) are required

to be jointly measurable functions of $x, y_1, \ldots, y_{k-1}, \bar{y}_1, \ldots, \bar{y}_{k-1}$. Here $p_{x,y_1,\ldots,y_{k-1},\bar{y}_1,\ldots,\bar{y}_{k-1}}$ is the probability that the transmission is to be continued (i.e., $\tau \geqq k$) if $\xi = x$ and the first $(k-1)$ input and output symbols were $\eta_1 = y_1, \ldots, \eta_{k-1} = y_{k-1}$, $\bar{\eta}_1 = \bar{y}_1, \ldots, \bar{\eta}_{k-1} = \bar{y}_{k-1}$, while $\mathcal{S}_{x,y_1,\ldots,y_{k-1},\bar{y}_1\ldots,\bar{y}_{k-1}}$ is the PD governing the selection of the kth input symbol η_k in that case.

For channels with feedback, lack of memory cannot be defined by requiring $\mathbf{P}\{\bar{\eta}_1 \in B_1, \ldots, \bar{\eta}_n \in B_n \mid \eta_1 = y_1, \ldots, \eta_k = y_n\} = \prod_{j=1}^{n} \mathcal{S}_{y_j}(B_j)$. Instead of this, the proper postulate is

$$\mathbf{P}\{\bar{\eta}_k \in B \mid \xi = x, \eta_1 = y_1, \ldots, \eta_k = y_k, \bar{\eta}_1 = \bar{y}_1, \ldots, \bar{\eta}_{k-1} = \bar{y}_{k-1}\} = \mathcal{S}_{y_k}(B) \quad (58)$$

for each $B \in \mathfrak{B}$; in order to avoid measure-theoretical difficulties, now let us make the regularity assumption that $\mathcal{S}_y(B)$ is a measurable function of $y \in Y$ for each $B \in \mathfrak{B}$.

Let us point out that the distribution $\mathcal{S}_\xi$ of ξ, the method of transmission and the elementary channel $\Pi_1 = \{\mathcal{S}_y : y \in Y\}$ uniquely determine the joint distribution of $\xi, \eta = (\eta_1, \ldots, \eta_\tau)$ and $\bar{\eta} = (\bar{\eta}_1, \ldots, \bar{\eta}_\tau)$. Moreover, letting $\varkappa_k = 1$ or 0 according as $\tau \geqq k$ or $\tau < k$, one may write

$$\mathbf{P}\{\eta_k \in A \mid \bar{\eta}_1 = y_1, \ldots, \bar{\eta}_{k-1} = y_{k-1}, \varkappa_k = 1\} = \mathcal{S}_{\bar{y}_1,\ldots,\bar{y}_{k-1}}(A) \quad (A \in \mathfrak{y}) \quad (59)$$

where $\mathcal{S}_{\bar{y}_1,\ldots,\bar{y}_k}$ is a PD for each fixed $\bar{y}_1, \ldots, \bar{y}_{k-1}$. These PD's can be determined in a straightforward though lengthy way from $\mathcal{S}_\xi$

$$p_{x,y_1\ldots,y_{k-1},\bar{y}_1,\ldots,\bar{y}_{k-1}}, \quad , y_1 \ldots, y_{k-1}, \bar{y}_1, \ldots, \bar{y}_{k-1}$$

and the PD's $\mathcal{S}_y$ defining the channel.

For transmission of fixed length $\tau = n$, the capacity $C_n^{(1)}(\Pi)$ is defined as the supremum of the mutual information $I(\xi, \bar{\eta})$ for all ξ (with values in any set) and all methods of transmission. This capacity is well known to equal $nC(\Pi_1)$, just as in the non-feedback case. Let us define the capacity $C_t^{(3)}(\Pi)$ as the supremum of $I(\xi; \bar{\eta})$ for all ξ and all methods of transmission with average transmission length $E\tau = t$; the supremum under the additional constraint $\frac{E(\tau m(\eta))}{E\tau} \leqq M$ — where m is a given constraining function of form (53) — will be denoted by $C_{t,M}^{(3)}(\Pi)$. Since variable length transmission is commonly used for channels with feedback, the following theorem showing that feedback does not increase capacity (of memoryless channels) for transmission of variable length, either, is of considerable interest.

Theorem 4. The inequalities (55) remain valid if $C_t^{(3)}(\Pi)$ and $C_{t,M}^{(3)}(\Pi)$ denote the capacities with feedback as defined above.

Proof. In order to avoid lengthy arguments, in this proof we assume (in contrary to the other parts of the paper) that the reader is familiar with the properties of mutual information and conditional mutual information for general RV's, see e.g. [3], [6]. As a general definition of conditional mutual information $I(\xi, \eta \mid \zeta)$[10] one may use — to the analogy of (9) —

$$I(\xi\,;\eta\,|\,\zeta) = \tag{60}$$

$$= \sup_{\mathfrak{B}} \int \left(\sum_{j=1}^{n} \mathbf{P}\{\eta \in B_j \,|\, \xi = x, \zeta = z\} \log_2 \frac{\mathbf{P}\{\eta \in B_j \,|\, \xi = x, \zeta = z\}}{\mathbf{P}\{\eta \in B_j \,|\, \zeta = z\}} \right) P_{\xi\zeta}\big(d(x,z)\big)\,.$$

The input and output signals η and $\bar{\eta}$ may be formally written as infinite sequences $\eta = (\eta_1, \eta_2, \ldots)$, $\bar{\eta} = (\bar{\eta}_1, \bar{\eta}_2, \ldots)$ setting $\eta_k = \bar{\eta}_k = y_0$, $y_0 \notin Y \cup \bar{Y}$ for $\varkappa_k = 0$ (i.e., if the transmissions has stopped before time k). Then $\varkappa_k$ is a function of $\bar{\eta}_k$, and we have

$$I(\xi\,;\bar{\eta}_1, \ldots, \bar{\eta}_n) = I(\xi\,;\varkappa_1, \bar{\eta}_1, \varkappa_2, \bar{\eta}_2, \ldots, \varkappa_n, \bar{\eta}_n) =$$

$$= I(\xi\,;\varkappa_1) + I(\xi\,;\eta_1\,|\,\varkappa_1) + I(\xi\,;\varkappa_2\,|\,\varkappa_1, \bar{\eta}_1) + \ldots + I(\xi\,;\bar{\eta}_n\,|\,\varkappa_1, \bar{\eta}_1, \ldots \bar{\eta}_{n-1}, \varkappa_n) =$$

$$= \sum_{k=1}^{n} I(\xi\,;\varkappa_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}) + \sum_{k=1}^{n} I(\xi\,;\bar{\eta}_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k)\,. \tag{61}$$

Here $I(\xi;\, \varkappa_k \mid \bar{\eta}_1, \ldots, \bar{\eta}_{k-1}) \leqq H(\varkappa_k \mid \bar{\eta}_1, \ldots, \bar{\eta}_{k-1}) \leqq H(\varkappa_k \mid \varkappa_1, \ldots, \varkappa_{k-1})$ and $\sum_{k=1}^{\infty} H(\varkappa_k \mid \varkappa_1, \ldots, \varkappa_{k-1}) = H(\varkappa_1, \varkappa_2, \ldots) = H(\tau)$ (recall that $\varkappa_k = 1$ if $\tau \geqq k$ and $\varkappa_k = 0$ if $\tau < k$). Thus, on account of the limit relation

$$I(\xi;\, \bar{\eta}) = \lim_{n \to \infty} I(\xi; \bar{\eta}_1, \ldots, \bar{\eta}_n)\,,$$

(61) gives rise to

$$I(\xi\,;\bar{\eta}) \leqq H(\tau) + \sum_{k=1}^{\infty} I(\xi\,;\bar{\eta}_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k)\,. \tag{62}$$

But

$$I(\xi\,;\bar{\eta}_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k) \leqq I(\xi, \eta_k\,;\bar{\eta}_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k) =$$

$$= I(\eta_k\,;\bar{\eta}_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k) + I(\xi\,;\bar{\eta}_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k, \eta_k) =$$

$$= I(\eta_k\,;\bar{\eta}_k\,|\,\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k)\,, \tag{63}$$

[10] In [3] and [6], the notation $MI(\xi, \eta \mid \zeta \mid$ is used.

because the other term vanishes on account of (58). Using (58), (59) and definition (60) of conditional mutual information we obtain

$$I(\eta_k ; \bar{\eta}_k \,|\, \bar{\eta}_1, \ldots, \bar{\eta}_{k-1}, \varkappa_k) = \mathbf{P}\{\varkappa_k = 1\} \cdot \tag{64}$$

$$\cdot \sup_{\mathcal{B}} \int \left[\int \left(\sum_{j=1}^{n} \mathcal{S}_y(B_j) \log_2 \frac{\mathcal{S}_y(B_j)}{\mathcal{S}_{\bar{y}_1, \ldots, \bar{y}_{k-1}}(B_j)} \right) \mathcal{S}_{\bar{y}_1, \ldots, \bar{y}_{k-1}}(dy) \right] \mathcal{S}'\big(d(\bar{y}_1, \ldots, \bar{y}_{k-1})\big),$$

where $\mathcal{S}_{\bar{y}_1, \ldots \bar{y}_{k-1}}(\mathcal{B}_j) = \int \mathcal{S}_y(B_j)\, \mathcal{S}_{\bar{y}_1, \ldots, \bar{y}_{k-1}}(dy)$ and $\mathcal{S}'$ denotes the joint distribution of $\bar{\eta}_1, \ldots, \bar{\eta}_{k-1}$ under the condition $\varkappa_k = 1$. Since

$$\sup_{\mathcal{B}} \int \left(\sum_{j=1}^{n} \mathcal{S}_y(B_j) \log_2 \frac{\mathcal{S}_y(B_j)}{\mathcal{S}_{\bar{y}_1, \ldots, \bar{y}_{k-1}}(B_j)} \right) \mathcal{S}_{\bar{y}_1, \ldots, \bar{y}_{k-1}}(dy)$$

is the mutual information of such RV's connected by the channel $\Pi_1 = \{\mathcal{S}_y : y \in Y\}$ for which the input RV has distribution $\mathcal{S}_{\bar{y}_1, \ldots, \bar{y}_{k-1}}$, in view of Lemma 1 the [ ] term in (64) is $\leqq C(\Pi_1)$. Thus (62), (63) and (64) give rise to

$$I(\xi ; \bar{\eta}) \leqq H(\tau) + C(\Pi_1) \sum_{k=1}^{\infty} \mathbf{P}\{\varkappa_k = 1\} = H(\tau) + C(\Pi_1)\, E\tau. \tag{65}$$

If the average transmission length is $E\tau = t$, we have, as in Example 2, $H(\tau) < \log_2(et)$ and (65) gives $C_l^{(3)}(\Pi) < tC(\Pi_1) + \log_2(et)$. This proves Theorem 4 for the unconstrained case.

In the constrained case it should be noted that the [ ] term in (64) is actually $\leqq C_{M_{\bar{y}_1, \ldots, \bar{y}_{k-1}}}(\Pi_1)$, where $M_{\bar{y}_1, \ldots, \bar{y}_{k-1}} = \int m(y)\, \mathcal{S}_{\bar{y}_1, \ldots, \bar{y}_{k-1}}(dy)$. In view of the concavity of the function $C_M(\Pi_1)$, it follows that the right hand side of (64) is $\leqq \mathbf{P}\{\varkappa_k = 1\}\, C_{M_k}(\Pi_1)$, where

$$M_k = \int M_{\bar{y}_1, \ldots, \bar{y}_{k-1}}\, \mathcal{S}'\big(d(\bar{y}_1, \ldots, \bar{y}_{k-1})\big) = E\big(m(\eta_k) \,|\, \varkappa_k = 1\big).$$

Since $\sum_{k=1}^{\infty} \mathbf{P}\{\varkappa_k = 1\} = E\tau$ and $\sum_{k=1}^{\infty} \mathbf{P}\{\varkappa_k = 1\} M_k = E\big(\sum_{k=1}^{\tau} m(\eta_k)\big) = E(\tau m(\eta))$ (see (53)), using the concavity of $C_M(\Pi_1)$ again we arrive at

$$I(\xi ; \bar{\eta}) \leqq H(\tau) + (E\tau)\, C_{\frac{E(\tau m(\eta))}{E\tau}}(\Pi_1). \tag{66}$$

Under the constraints $E\tau = t$ and $\dfrac{E(\tau m(\eta))}{E\tau} \leqq M$, the right hand side of (66) is $< \log_2(et) + tC_M(\Pi_1)$, completing the proof of Theorem 4.

References

1. *Csiszár, I., Katona, G., Tusnády, G.*, Information sources with different cost scales and the principle of conservation of entropy. Z. Wahrscheinlichkeitstheorie verw. Geb. **12** (1969) 185–222.
2. *Чисар И.* О каналах без шума. Проблемы передачи информации **6** (1970) 3–15.
3. *Добрушин Р. Л.* Общая формулировка основной теоремы Шеннона в теории информации. Успехи мат. наук **14** (1959) 3–104.
4. *Gallager, R. G.*, Information Theory and Reliable Communication, Wiley, 1968.
5. *Perez, A.*, Notions généralisées d'incertitude, d'entropie et d'information du point de vue de la théorie de martingales. Trans. First Prague Conference on Information Theory, etc., Prague, 1957, 183–208.
6. *Пинскер М. С.* Информация и информационная устойчивость случайных величин и процессов. Изд. АН СССР, Москва, 1960.

О проблеме емкости каналов с помехами при любых затратах на сигналы

И. ЧИСАР

(Будапешт)

Рассматриваются четыре типа емкости информации в обобщенных каналах с помехами при любых затратах на входные сигналы. Доказаны теоремы асимптотического равенства для входных сигналов с ограничениями и без ограничений. Исследуемые емкости относятся к передаче постоянных и изменяющихся затрат, особенно к определению затрат на длину передачи. Получены асимптотические выражения для средней емкости длины передачи. Представлены примеры каналов без памяти, включая и каналы с обратной связью.

I. Csiszár
Mathematical Research Institute
Hungarian Academy of Sciences
Budapest, V. Reáltanoda u. 13–15

Problems of Control and Information Theory, Vol. 2 (3—4), pp. 305 – 314 (1973)

BASIC CONSIDERATIONS CONCERNING FREQUENCY DICTIONARIES

L. OHERA

(Prague)

(Received February 12, 1973)

Preparing frequency dictionaries is one of basic tasks of quantitative linguistics. In this paper some problems concerning frequency dictionaries of large texts are discussed and as a result recommendations for the method of sampling, sample size and sample forming are given.

The fundamental task of quantitative linguistics (cf. e.g. [1 — 3]) is determining the occurrence frequencies of certain linguistic phenomena, graphemes, words, word types, etc., and investigating the relations that can be deduced from them. Systematically arranged occurrence frequencies of various tokens form frequency dictionaries, in which next to any individual token there is a number giving the occurrence (or its estimation) of the token in a given text.

Preparing a frequency dictionary is a costly business; great care should, therefore, be devoted to preliminary considerations; it is necessary to set the aim (e.g. to determine the quality of the frequency dictionary — reliability and precision of the resulting data) and after that to take this aim into consideration for deciding on the sampling method and the sample size.

Although special features of individual frequency dictionaries (e.g. those of graphemes, words, word types, etc.) greatly diverse, the basic considerations, dealt with in this paper, remain unchanged.

Frequency dictionary

Let be given: $\mathfrak{T}$ class of linguistic texts, $\mathfrak{M} \equiv \{\alpha^1, \ldots, \alpha^M\}$ finite set of token the occurrence frequencies of which are to be investigated, and G the set of rules defining unambiguous transformation of any text $\tau \in \mathfrak{T}$ into form $T = \alpha_1 \ldots \alpha_K$, where $\alpha_1, \ldots, \alpha_K \in \mathfrak{M}$.

Frequency dictionary $F(\tau, \mathfrak{M}, G)$ of text $\tau \in T$ based on token vocabulary $\mathfrak{M}$ and set of transformation rules G is defined as the set of doubles

$$F(\tau, \mathfrak{M}, G) = \{(\alpha, N_T(\alpha))\}_{\alpha \in \mathfrak{M}},$$

where $N_T(\alpha)$ is a nonnegative number representing the number of occurrences of token (symbol) α in text $T = G(\tau)$.

For practical purposes subset

$$F^*(\tau, \mathfrak{M}, G) = \{(\alpha, N_T(\alpha))\}_{\alpha \in \mathfrak{M} : N_T(\alpha) > 0}$$

of the frequency dictionary is quite commonly used, while the following equality holds:

$$F(\tau, \mathfrak{M}, G) = F^*(\tau, \mathfrak{M}, G) \cup \{(\alpha, 0)\}_{\alpha \in \mathfrak{M} : N_T(\alpha) = 0} .$$

For comparing frequency dictionaries of various texts of different lengths $K > 0$ the concept of relative frequency dictionary may be introduced. Relative frequency dictionary $f(\tau, \mathfrak{M}, G)$ is defined in the following way:

$$f(\tau, \mathfrak{M}, G) = \{(\alpha, n_T(\alpha))\}_{\alpha \in \mathfrak{M}},$$

where

$$n_T(\alpha) = \frac{N_T(\alpha)}{K}.$$

Frequency dictionary of a large text

Since for large texts a direct determination of $f(\tau, \mathfrak{M}, G)$ is usually not feasible, the following method has to be considered. For $0 < k \leq K$ a sample $t = \alpha_{i_1} \ldots \alpha_{i_k}$ from the transformed text T is drawn and set of doubles

$$\varphi_t(\tau, \mathfrak{M}, G) = \left\{\left(\alpha, \frac{N_t(\alpha)}{k}\right)\right\}_{\alpha \in \mathfrak{M}}$$

is formed, where $N_t(\alpha)$ is the nonnegative number of occurrences of token α in sample t; after that an estimation of $f(\tau, \mathfrak{M}, G)$ based on this set of doubles is made.

Before proceeding to make such estimations the inverse problem will be briefly discussed, namely: we shall investigate what can be said about $\varphi_t(\tau, \mathfrak{M}, G)$ of a sample $t \subset T$ provided that $f(\tau, \mathfrak{M}, G)$ is known. Knowing nothing about the system of sampling it can be said about $n_t(\alpha) = \frac{N_t(\alpha)}{k}$, $\alpha \in \mathfrak{M}$ only

$$\max\left(0, 1 - \frac{K}{k}(1 - n_T(\alpha))\right) \leq n_t(\alpha) \leq \min\left(1, \frac{K}{k} n_T(\alpha)\right), \qquad \sum_{\alpha \in \mathfrak{W}} n_t(\alpha) = 1 .$$

Two fundamental methods of the forming of sample t will be described now.

Method A. Sample t is a realization of a random sample of size k obtained by successive drawing symbols from text T without replacements.

Note: A realization of the random sample of size k from text T is obtained in such a way that one symbol $\alpha \in \mathfrak{M}$ is at random drawn from the text, each

symbol having equal probability to be drawn. Denoting α_{i_1} the drawn symbol, a new text is formed in such a way that in the former text symbol α_{i_1} is omitted, then again a symbol α_{i_2} at random drawn. Omitting the drawn symbol, α_{i_2}, a new text is formed and in such a way we proceed until altogether k symbols $\alpha_{i_1}, \ldots, \alpha_{i_k}$ have been drawn. Sample $t = \alpha_{i_1} \ldots \alpha_{i_k}$ is said to be formed by sampling without replacement. Note that by the term sampling with replacement is meant repeating the random choice without omitting chosen symbols. In such a case some of symbols of the text may be selected more than once.

Corresponding $N_t(\alpha)$ is random variable with hypergeometric probability distribution, i.e.

$$P[N_t(\alpha) = x] = \frac{\binom{Kn_T(\alpha)}{x}\binom{K(1 - n_T(\alpha))}{k-x}}{\binom{K}{k}},$$

$$x = \max(0,\ k - K + Kn_T(\alpha)), \ldots, \min(Kn_T(\alpha),\ k)$$

$$P[N_t(\alpha) = x] = 0 \text{ otherwise.}$$

For mathematical expectation $E[n_t(\alpha)]$ and variance $D[n_t(\alpha)]$ of random variable $n_t(\alpha)$

$$E[n_t(\alpha)] = n_T(\alpha)$$

$$D[n_t(a)] = \frac{1}{k} n_T(\alpha)\left(1 - n_T(\alpha)\right)\left(1 - \frac{k-1}{k-1}\right)$$

hold.

For $k \ll K$ (common and therefore very important condition) $N_t(\alpha)$ can be approximated by random variable with binomial distribution $\mathrm{Bi}(k, n_T(\alpha))$. In special cases, e.g. for $1 \ll k \ll K$ and $n_T(\alpha) \ll 1$ approximation by Poisson's distribution $\mathrm{Po}(kn_T(\alpha))$, for $1 \ll k \ll K$ and

$$k \gg \frac{1}{n_T(\alpha)\left(1 - n_T(\alpha)\right)}$$

approximation by normal distribution

$$\mathrm{No}(kn_T(\alpha),\ kn_T(\alpha)\ (1 - n_T(\alpha)))$$

may be used. In both cases

$$E[n_t(\alpha)] = n_T(\alpha),$$

for normal distribution

$$D[n_t(\alpha)] = \frac{n_T(\alpha)\left(1 - n_T(\alpha)\right)}{k}$$

and for Poisson's distribution

$$D[n_t(\alpha)] = \frac{n_T(\alpha)}{k}.$$

Method B. Sample t is formed in the following way. Let $K = Sr$, where S and $r \ll K$ are natural numbers. Text T is divided into S parts $T_1, \ldots, T_S$ so that

$$T_i \equiv \alpha_{(i-1)r+1}\ \alpha_{(i-1)r+2} \ldots \alpha_{ir}, \qquad i = 1, \ldots, S$$

and for chosen natural number $s < S$ sampling without replacement yielding sample $t = T_{i_1} \ldots T_{i_s}$ is performed.

If $s \ll S$, sampling without replacement may be approximated by sampling with replacement. Moreover, if $s \gg 1$, $n_t(\alpha)$ may be approximated by normally distributed random variable ξ having mathematical expectation

$$E[\xi] = n_T(\alpha)$$

and variance

$$D[\xi] = \frac{\sigma(\alpha)}{sr^2},$$

where $\sigma(\alpha)$ is the variance of random variable $\mu(\alpha)$ denoting the number of occurrences of token α in randomly chosen element of set $\{T_i\}_{i=1,\ldots,S}$.

We return now to the main problem encountered when constructing frequency dictionary: the estimation of relative frequency dictionary $f(\tau, \mathfrak{M}, G)$ based on the knowledge of $q_t(\tau, \mathfrak{M}, G)$.

For both methods of sampling $n_t(\alpha)$ is the best unbiased estimation of $n_T(\alpha)$. Since frequency dictionary $f(\tau, \mathfrak{M}, G)$ is intended to be used as one of quantitative characteristics of given text τ, the precision of the estimation for individual $n_T(\alpha)$ is required to be known, i.e., for example, to know the limits inside which lies $n_T(\alpha)$ with given reliability. Henceforth, confidence intervals will be used that are defined in the following way: interval $(l_t(\alpha), u_T(\alpha))$ is called a $100(1 - \beta)\,\%$ confidence interval for parameter $n_t(\alpha)$ of a given class of distribution if the relation

$$P[n_T(\alpha) \in (l_T(\alpha), u_T(\alpha))] = 1 - \beta$$

is valid for any value of parameter $n_T(\alpha)$ of parametric space.

In this paper two-sided confidence intervals will be used defined so that

$$P[N_t(\alpha) \geqq y_1] = \frac{1}{2}\beta$$

and

$$P[N_t(\alpha) \leqq y_2] = \frac{1}{2}\beta\,,$$

where y_1 is the random variable with a distribution of the given class and parameter $u_T(\alpha)$, while y_2 is the random variable with a distribution of the given class and parameter $l_T(\alpha)$.

Using Method A and *if* $k \ll K$ the following relations hold for lower and upper limits of confidence interval:

$$\sum_{i=0}^{N_t(\alpha)} \binom{k}{i} [u_T(\alpha)]^i [1 - u_T(\alpha)]^{k-i} = \frac{1}{2}\beta\,,$$

$$\sum_{i=N_t(\alpha)}^{k} \binom{k}{i} [l_T(\alpha)]^i [1 - l_T(\alpha)]^{k-i} = \frac{1}{2}\beta.$$

Hence

$$u_T(\alpha) = \frac{[N_t(\alpha) + 1]\, F_{1-\beta/2}\left[2\,[N_t(\alpha) + 1], 2\,[k - N_t(\alpha)]\right]}{k - N_t(\alpha) + [N_t(\alpha) + 1]\, F_{1-\beta/2}\left[2\,[k - N_t(\alpha) + 1], 2[k - N_t(\alpha)]\right]} \quad (1)$$

$$l_T(\alpha) = \frac{N_t(\alpha)}{N_t(\alpha) + [k - N_t(\alpha) + 1] F_{1-\beta/2}\left[2\,[k - N_t(\alpha) + 1], 2\,N_t(\alpha)\right]}\,, \quad (2)$$

($F_\gamma[\delta_1, \delta_2]$ is the γth quantile of Fisher's F distribution with δ_1 and δ_2 degrees of freedom).

For $1 \ll k$ and $n_t(\alpha) \ll 1$, $l_T(\alpha)$ and $u_T(\alpha)$ can be found by solving equations

$$\sum_{i=0}^{N_t(\alpha)} e^{-ku_T(\alpha)} \frac{[ku_T(\alpha)]^i}{i!} = \frac{1}{2}\beta$$

and

$$\sum_{i=N_t(\alpha)}^{\infty} e^{-kl_T(\alpha)} \frac{[kl_T(\alpha)]^i}{i!} = \frac{1}{2}\beta.$$

Hence,

$$l_T(\alpha) = \frac{1}{2k}\chi^2_{\beta/2}\,[2\,N_t(\alpha)]$$

$$u_T(\alpha) = \frac{1}{2k}\chi^2_{\gamma-\beta/2}\left[\,2\,[N_t(\alpha) + 1]\right],$$

where $\chi^2_\gamma[\delta]$ is the γth quantile of chi-square distribution with δ degrees of freedom.

Note: The following elementary results have been used:

a) given random variable ξ having binomial distribution Bi(k, λ) and random variable η having Fisher's distribution $F(\nu_1, \nu_2)$ where $\nu_1 = 2(k - x)$ and $\nu_2 = 2(x + 1)$ are degrees of freedom, the following equation holds:

$$P[\xi \leqq x] = P\left[\eta < \frac{x+1}{k-x}\,\frac{1-\lambda}{\lambda}\right];$$

b) given random variable ξ having Poisson's distribution Po(λ) and random variable η having χ^2 distribution with $2(x + 1)$ degrees of freedom, the following equation holds:

$$P[\xi \leqq x] = P[\eta > 2\lambda].$$

For $1 \ll k \ll K$ and

$$k \gg \frac{1}{n_T(\alpha)\left(1 - n_T(\alpha)\right)}$$

approximate relations

$$l_T(\alpha) = n_t(\alpha) - \mathrm{No}_{1-\beta/2}[0,1]\sqrt{\frac{n_t(\alpha)\left(1 - n_t(\alpha)\right)}{k}}$$

$$u_T(\alpha) = n_t(\alpha) + \mathrm{No}_{1-\beta/2}[0,1]\sqrt{\frac{n_t(\alpha)\left(1 - n_t(\alpha)\right)}{k}}$$

follow, where $\mathrm{No}_\gamma[0,1]$ is the γth quantile of the standardized normal distribution.

Using Method B and sufficiently large s because of the approximation by normal distribution $l_T(\alpha)$ and $u_T(\alpha)$ can be found out from the following approximate relations

$$l_T(\alpha) = n_t(\alpha) - \mathrm{No}_{1-\beta/2}[0,1]\sqrt{\frac{\sigma(\alpha)}{sr^2}}$$

$$u_T(\alpha) = n_t(\alpha) + \mathrm{No}_{1-\beta/2}[0,1]\sqrt{\frac{\sigma(\alpha)}{sr^2}},$$

or (taking into account that $k = sr$),

$$l_T(\alpha) = n_t(\alpha) - \mathrm{No}_{1-\beta/2}[0,1]\sqrt{\frac{s\sigma(\alpha)}{k}} \tag{3}$$

$$u_T(\alpha) = n_t(\alpha) + \mathrm{No}_{1-\beta/2}[0,1]\sqrt{\frac{s\sigma(\alpha)}{k}}. \tag{4}$$

Sample size

Question concerning the size k of sample t for determining relative frequency dictionary $f(\tau, \mathfrak{M}, G)$ cannot be answered unless the decision on the quality of the frequency dictionary (or selected doubles $(\alpha, n_T(\alpha)) \in f(\tau, \mathfrak{M}, G)$) for required confidence level as well as the method of sampling are agreed upon.

The following problem can be given as an example of the correct question concerning the sample size: given $\mathfrak{M}^* \subset \mathfrak{M}$, what should the size of a sample chosen by Method A (or B) be so that for all $\alpha \in \mathfrak{M}^*$ relative error

$$\varepsilon = \frac{u_T(\alpha) - l_T(\alpha)}{u_T(\alpha) + l_T(\alpha)}$$

determined for $100(1 - \beta)\%$ confidence interval $(l_T(\alpha), u_T(\alpha))$ does not exceed a given value $\varepsilon_{\max}$.

Using sampling Method A we determine $N_t^{\text{crit}}(\alpha)$ by putting $u_T(\alpha)$ and $l_T(\alpha)$ from (1), (2) and solving for $N_t(\alpha)$. Size k of sample t has to satisfy

$$N_t(\alpha) \geqq N_t^{\text{crit}}(\alpha), \quad \alpha \in \mathfrak{M}^*.$$

If the conditions for approximation by normal distribution are satisfied and if $n_t(\alpha) \ll 1$, then

$$\varepsilon = \text{No}_{1-\beta/2}[0,1] \frac{1}{\sqrt{N_t(\alpha)}}.$$

Hence, for $N_t(\alpha)$, $\alpha \in \mathfrak{M}^*$

$$N_t(\alpha) \geqq \left[\frac{\text{No}_{1-\beta/2}[0,1]}{\varepsilon_{\max}}\right]^2. \tag{5}$$

Using sampling Method B and for $1 \ll s \ll S$, according to (3) and (4),

$$\varepsilon = \text{No}_{1-\beta/2}[0,1] \frac{\sqrt{s\sigma(\alpha)}}{N_t(\alpha)},$$

the sample size, therefore, is to be chosen so that $N_t(\alpha)$ for all $\alpha \in \mathfrak{M}^*$ satisfy the following inequality

$$N_t(\alpha) \geqq \frac{\text{No}_{1-\beta/2}[0,1]}{\varepsilon_{\max}} \sqrt{s\sigma(\alpha)}. \tag{6}$$

Sampling method

Assuming equal cost of the processing of the sample chosen by Method A and that chosen by Method B (both samples consisting of k symbols) we certainly decide in favour of the method whose expected value of $N_t(\alpha)$ for chosen

token α is for the same error limiting $\varepsilon_{\max}$ less, according to (5) and (6), we decide according to the relation between

$$\left[\frac{\mathrm{No}_{1-\beta/2}[0,1]}{\varepsilon_{\max}}\right]^2$$

and

$$\frac{\mathrm{No}_{1-\beta/2}[0,1]}{\varepsilon_{\max}}\sqrt{s\sigma(\alpha)}\,;$$

if $E\,[N_t(\alpha)] = kn_T(\alpha) < s\sigma\,(\alpha)$, we prefer Method A and vice versa. For illustration two extreme examples are discussed.

Example: Let $\mathfrak{M} = \{\alpha, \beta\}$ and transformed text (of size K)

$$T \equiv \alpha\beta\alpha\beta\alpha\beta. \ldots \alpha\ \beta.$$

Using Method B the following very course estimation holds:

$$s\sigma(\alpha) \leq s;$$

for $r \geq 2$ inequality

$$E[N_t(\alpha)] \gg s\sigma(\alpha)$$

holds. Since for $r = 1$ Method B coincides with Method A, the last inequality holds for any natural r.

Example: Let again $\mathfrak{M} = \{\alpha, \beta\}$ and transformed text

$$T \equiv \alpha\alpha \ldots \alpha\beta\beta \ldots \beta$$

consists of $K/2$ symbols α followed by $K/2$ symbols β.
Using method B for $K \gg sr$

$$s\sigma(\alpha) \doteq \frac{1}{4}sr^2.$$

Hence

$$E[N_t(\alpha)] \leq s\sigma(\alpha).$$

The value of $\sigma(\alpha)$ is not known a priori, yet can be estimated

$$\sigma(\alpha) \doteq \frac{1}{s-1}\sum_{i=1}^{s}\left[\mu_i(\alpha) - \frac{1}{s}\sum_{j=1}^{s}\mu_j(\alpha)\right]^2,$$

where $\mu_i(\alpha)$, $i = 1, \ldots, s$ are realization of random variable $\mu(\alpha)$. This estimation preceeding the actual work on frequency dictionary uses a small test sample, by means of which also $n_T(\alpha)$ can be estimated.

It was assumed in comparing the two methods that the cost of sample processing (for equal sample size) is the same for both methods. For manual excerption with subsequent counting of cards, which is even at present a commonly used method, it is roughly so. Using automatic reading devices and transforming the text by electronic computers, however, the cost of Method B will be lower than that of Method A (for equal sample size). Denoting v_A cost and $A_{N_t}(\alpha)$ minimal required number of occurrences of any token $\alpha \in \mathfrak{M}^*$ for Method A and denoting similarly the variables for Method B we decide between the two methods according to the relation between

$$v_A \cdot E[{}^A N_t(\alpha)]$$

and

$$v_B \cdot E[{}^B N_t(\alpha)];$$

using (5) and (6) we come to the conclusion that we shall prefer Method A if

$$\left[\frac{v_A}{v_B}\right]^2 < \frac{s\sigma(\alpha)}{kn_T(\alpha)}$$

and Method B if inverse relation is valid.

Sample forming

So far it was assumed that sampling by Method A is performed in such a way that after the whole text τ having been transformed into form

$$T \equiv \alpha_1 \ldots \alpha_K$$

subsequent random choices without (or with) replacement take place. For reproducibility the using of pseudorandom numbers generated with equal probabilities from population $\{1, \ldots, K\}$ (e.g. by means of an electronic computer) may be recommended. Another way is very easily performed: a fixed system of selecting is chosen, e.g. sample t can be defined for $b < K/k$

$$t \equiv \alpha_b \alpha_{b+\left[\frac{K-b}{k-1}\right]} \alpha_{b+\left[\frac{K-b}{k-1}\right]2} \cdots \alpha_{b+\left[\frac{K-b}{k-1}\right](k-1)}$$

or

$$t \equiv \alpha_b \alpha_{b+\xi_1} \alpha_{b+\xi_1+\xi_2} \cdots \alpha_{b+\sum_{i=1}^{k-1}\xi_i},$$

where ξ_i, $i = 1, \ldots, k-1$ are realizations of random variable ξ having mathematical expectation $\left[\frac{K-b}{k-1}\right]$ and variance $\sigma_\xi \ll K/k$ and such that

$$\sum_{i=1}^{k-1} \xi_i \leqq K - b$$

([a] denotes the integer part of a). If $k \ll K$, the contextual dependence between two following symbols is sufficiently weak and such a sample can be considered to be a realization of the required sampling. Even using sampling by Method B similar method can be used. In practical problems encountered in linguistics when preparing frequency dictionaries the transformation rules are usually relatively simple and make possible the transformation only of those parts of the text that are important for the chosen system of sampling.

References

1. *Herdan, G.*, The Calculus of Linguistic Observations. Hague, 1962.
2. *Herdan, G.*, Quantitative Linguistics. London, 1964.
3. *Muller, Ch.*, Initiation à la statistique linguistique. Paris, 1968.

Общие методы составления частотных словарей

Л. ОХЕРА
(Прага)

Основной задачей квантитативной лингвистики является определение частот появления некоторых языковых явлений, слов, частей речи и т. д. и исследование зависимостей появления этих частот. Наглядным выражением частот появления служат частотные словари, в которых наряду с отдельными изучаемыми явлениями находятся частоты их появления (или их оценки) в данном тексте.

После формулировки задачи в статье исследована общая теория составления частотного словаря большого текста, когда необходимо пользоваться статистическими методами. Изучены и сравнены по их эффективности два основных типа выборки и даны некоторые рекомендации для объема, метода и техники выборки.

L. Ohera
Faculty of Nuclear Science and Physical Engineering
Czech Technical University
Břehová 1, Prague 1
Czechoslovakia

MTA Könyvtára
Periodika 3112 /19 74 -ből

BOOK REVIEW

J. Ackermann: Abtastregelung (Sampled-data control systems.) Springer Verlag, Berlin—Heidelberg—New York, 1972 (391 pages, 86 figures, 1 table; in German)

When the first books on sampled-data systems appeared in the fifties frequency-response techniques and z-transform methods were dominating. Some years after the a time-domain technique: the state-space method came into fashion. The latter overshadowed the difference between continuous-time and discrete-time systems. For a long time, however, discrete-time systems were treated as special cases of continuous-time systems.

With the advent and broadening field of application of digital computers the importance of discrete-time systems in general and specifically discrete-time control systems is more and more increasing. Under such circumstances it became meaningful to consider discrete-time systems self-contained.

The book reviewed has the task to give a comprehensive and direct introduction to sampled-data systems and to show elaborated recent techniques for the analysis and synthesis of complicated discrete-time feedback control systems.

The book is formally divided into thirteen chapters, the first being a short introduction whereas the last two are merely contents and index. Chapter Two is a concise review of the mathematical methods applied to the analysis of discrete-time systems. Chapter Three reviews the computation rules of z-transforms. Here besides common z-transforms also the so-called complete (vollständige) z-transforms are introduced by the author which are, however, the well-known advanced z-transforms in essence. Chapter Four develops the state-space method for linear, lumped-parameter, discrete-time systems. Chapter Five shows the linear transformation methods and introduces some canonical forms, for example, the so called observability and controllability canonical forms. Chapter Six serves as a brief introduction to multivariable (multiport) systems. Chapter Seven is devoted to identification and model building of discrete-time systems based on input-output measurements. Chapter Eight discusses the stability problems of discrete-time systems where numerous methods are presented. Chapter Nine shows some design methods for discrete-time control systems including also dead-beat systems. Chapter Ten develops synthesis methods for multivariable sampled-data control systems. The treatise is concerned with state-space methods. State variable estimation and state-variable feedback are also examined in some detail. Chapter Eleven is an appendix giving tables for Laplace and z-transforms.

The treatment is intelligible and consistent. Within the text many examples are elaborated which help to clarify the notions and concepts. Although the discrete-time systems are set in the focus of attention the connection with the continuous-time systems is often emphasized, especially the similarities and differences of the theorems on z-transforms and Laplace-transforms are pointed out.

In the opinion of the reviewer the most important feature of the book is that it gives selected topics on sophisticated multivariable discrete-time control systems.

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

PROBLEMS OF CONTROL AND INFORMATION THEORY, VOL. 2 (3—4)
ENGLISH TRANSLATION OF THE COMMUNICATIONS IN RUSSIAN

ENTROPY, INFORMATION AND WORK IN QUANTUM STATISTICS

V. V. MITYUGOV

(Gorki)

The paper discusses the role of the irreversible processes during decomposition of the signals. In the case of copying (duplication) of the information the entropy increases and at the same time the thermodynamical work decreases.

1. Introduction

Numerous studies into relations between physical and informational aspects of entropy [1—7] dating from Brillouin's and Gabor's works have elucidated the fact that entropy, defined as the functional $H_{cl} = -\sum_i p_i \ln p_i$ of physical system state probability distribution p_i, is at the same time a measure of system micro-state uncertainty, a characteristic of its non-equilibrium degree, and a measure of amount of information which may be transmitted under given macro-restrictions if the system is used as a physical information carrier. By varying the functional under a given mean energy, its maximum may be shown to be reached over the equilibrium distribution and to have the sense of the thermodynamical entropy.

Thus, consistence of various aspects of entropy now seems evident for a wide variety of problems where the state of a physical system may be described by probability distribution p_i, i.e. for the classical statistics and those quantum theory problems, where non-commutativity of various variables needs not to be taken into consideration, the quantum nature of the matter manifesting itself only through energy spectrum discreteness.

But for the consistently quantum description where density operator ϱ replaces distribution p_i, and the above-mentioned functional H_{cl} gives place to its natural generalization $H = -S_p \hat{\varrho} \ln \hat{\varrho}$ invariant to unitary transformations, the situation is not so clear. Von Neumann [8] and Elsasser [9] studied various aspects of entropy defined through the density operator. Comparatively recent papers [5, 6, 7] have elucidated the role of the functional in physical information theory problems. Moreover, the value of H, as well as that of H_{cl}, reaches its maximum for a given mean energy $E = S_p \hat{\varrho} \hat{E}$ on the Gibbs

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

distribution, thus enabling one to establish links between density matrix formalism and equilibrium thermodynamics [10]. Nevertheless, in discussions of non-equilibrium systems H is often refused to be identified with the physical entropy, thus preserving only its informational aspect.

The main reason for this approach is that H of a closed physical system does not change in time, i.e. is not subjected to the H-theorem of the physical kinetics. On the other hand, if a system may be broken up into a large number N of subsystems, the value of $-\sum_{\lambda=1}^{N} Sp\,\hat{\varrho}^{(\lambda)} \ln \varrho^{(\lambda)}$ may increase in time [4], thus making identification just of this variable with the physical entropy more reasonable, though this does not lead to an evident and united concept because in doing so it is not clear what should be referred to as the "true" entropy of a complex physical system.

All the same, the mentioned questions should be elucidated in order to give the correct understanding of the irreversible process nature, and to enable one to study the losses of information and work due to increase of physical entropy. The maximal work obtainable owing to system non-equilibrium rather than the informational aspect of entropy) should be taken as a universality criterion for the mentioned entropy definitions, because this question both historically, and logically is the corner-stone of the thermodynamics.

2. Transfer of work and information under interactions

It seems intuitively clear that interaction of physical systems is accompanied by exchanges of energy and information. We are going to make a more strict analysis of these considerations based on the definitions of mean energy, work and information accepted in the statistical mechanics and quantum theory of signal transmission. In doing so, only collision-type interactions will be considered where the totals of all additive movement integrals of system do not change. First, we do so because the processes of dispersion of light, relaxation at collision of particles, etc. which we will be mostly interested in, are just of this kind. Second (and this is more important from the principal viewpoint), even the problems including "switch-on" or "switch-off" of interaction able to change the value of the total energy, may be reduced to collision ones by consistently quantum description of the physical factor responsible for such a "switching".

Consider the process of information and work transfer under interaction of two systems each of which may have any number of degrees of freedom and be divided into subsystems. Let, prior to the collision, their states be statisti-

cally independent and described by density operators $\hat{\varrho}^{(1)}$ and $\hat{\varrho}^{(2)}$ with energies and entropies, respectively,

$$E_\lambda = Sp\,\hat{\varrho}^{(\lambda)} E_\lambda\,;\; H_\lambda = -\,Sp\,\hat{\varrho}^{(\lambda)} \ln \hat{\varrho}^{(\lambda)};\;\; \lambda = 1, 2 \tag{2.1}$$

where E_λ are Hamiltonians of systems without interaction. In the course of time the density operator $\varrho(t)$ of a complex system changes in accordance with

$$\hat{\varrho}_{(t)} = \hat{U}_{(t)}^{-1}\,\hat{\varrho}^{(1)}\,\hat{\varrho}^{(2)}\,U_{(t)} \tag{2.2}$$

where $U_{(t)}$ is an evolutionary unitary operator. For example, in Schrödinger's representation, given interaction $\hat{V}_{(t)}$, it is expressed through the complete energy operator $\hat{E} = \hat{E}_1 + \hat{E}_2 + \hat{V}$ as $\hat{U} = \exp\,(-i\hbar^{-1}\hat{E}t)$. Generally speaking, system energies and entropies also change:

$$E_{\lambda(t)} = Sp\,\hat{\varrho}_{(t)}\,\hat{E}_\lambda\,;\; H_{\lambda(t)} = -\,Sp\,\hat{\varrho}_{(t)}^{(\lambda)} \ln \hat{\varrho}_{(t)}^{(\lambda)} \tag{2.3}$$

where $\hat{\varrho}_{(t)}^{(1)} = Spy\,\hat{\varrho}_{(t)}\,;\; \hat{\varrho}_{(t)}^{(2)} = Spx\,\hat{\varrho}_{(t)}$, and by x, y complete sets of variables, respectively of the first and second systems, are understood. After the interaction the following relations hold:

$$E_{1(t)} + E_{2(t)} = E = E_1 + E_2\,;\; H_{1(t)} + H_{2(t)} \geq H_1 + H_2\,. \tag{2.4}$$

The latter inequality, which is the quantum generalization of the well known relation in the Shannon theory [11], is readily inferred by means of the Klein lemma (see [12]).

For the sake of generality, take that system 1 is the source of information and work for system 2 receiving them under interaction.

Consider first exchange of work between subsystems 1 and 2 which will be referred to below as "source" and "adopter",* respectively. To this end, suppose that at $t = 0$ adopter's state is characterized by energy E_2 and zero entropy, and that by work R, adopted during interaction with the source, energy increase $E_{2(t)} - E_2$ is meant which is not accompanied by increase of entropy. If $H^{(1)}_{\max}(E_1)$ is the maximal source entropy for the given mean energy, then it may be readily shown through (2.4) that

$$R \leq E_1 - E^{(1)}_{\min}(H_1) \tag{2.5}$$

where E_1 and H_1 are source mean energy and state entropy at $t = 0$, and $E^{(1)}_{\min(H_1)}$** is the inverse function of $H^{(1)}_{\max(E)}$. It is natural that $R \leq 0$ if the

* We refrain from the term "receiver" since in the quantum communication theory it stands for a measuring instrument in Von Neumann sense.

** In systems with finite number of levels this function may turn out to be two-valued; then by $E_{\min(H)}$ the least value corresponding to thermal equilibrium should be understood.

density operator corresponds to the equilibrium case:

$$\varrho_{E_1 E_1'} = Z^{-1} \exp(-E_1/kT)\, \delta_{E_1 E_1'}.$$

In the case when the source is a classical homogeneous body whose state non-equilibrium is due to deviation of temperature in a small volume v from the mean value T, (2.5) leads to the well known formula

$$R \leq v c_v kT \ln(1 + \Delta T/T). \tag{2.6}$$

Continuing with similar examples, one may easily show that the operator definition of entropy based on (2.4) leads naturally to the known thermodynamical theorem on maximal work. Note that feasibility of equality in (2.5) is a very fine question which seemingly cannot be solved without more concrete data on system energy spectra and their interaction. But we shall disengage ourselves from this fact and introduce, for the sake of convenience, a value $R_{\max}$ corresponding to the equality sign in (2.5).

Infer now a useful consequence of (2.5). Consider a complex system, consisting of source and adopter, as the source of work which can be extracted by an external device. If systems 1 and 2 did not interact in the past and if total entropy is $H = H_1 + H_2$, the maximal value of this work is $R_{\max} = E - E_{\min(H)}$ where $E_{\min(H)}$ is the equilibrium energy of the complex system corresponding to the entropy $H = H_1 + H_2$. Moreover, one may speak about amounts of work R_λ with respect to the external device contained separately in each system. Then it may be easily shown through (2.5) that

$$R_{1\max} + R_{2\max} = R_{\max} - E^{(1)}_{\min(H_1(t))} - E^{(2)}_{\min(H_2(t))} + E_{\min(H)}. \tag{2.7}$$

This relation is used below in discussions of losses of work and information transfer in irreversible processes.

Embark now on the information exchange under system interaction. Let source states described by density operators $\hat{\varrho}_i^{(1)}$ be used as signals sent with probabilities p_i. Then Shannon's amount of information which may be obtained through measurements of the complete set x of source variables at $t = 0$ is as follows:

$$I_x = -\sum_{i,x} p(x \mid i)\, p_i \ln \sum_i p(x \mid i)\, p_i + \sum_i p_i \sum_x p(x \mid i) \ln p(x \mid i) \tag{2.8}$$

where $p(x|i)$ are diagonal elements of $\hat{\varrho}_i^{(1)}$ in x-representation. $\operatorname{Sup}_x I_x$ over all possible complete sets x may be referred to as the "amount of information I_1 in source states". Measurement of adopter variables y together with x at $t = 0$ does not, naturally, give any additional information about the signal.

The situation may change as the result of interaction between systems. Indeed, at $t > 0$ both density operator of complex system $\hat{\varrho}_{i(t)} = \hat{U}^{-1}_{(t)}\, \hat{\varrho}_i^{(1)}\, \hat{\varrho}^{(2)}\, \hat{U}_{(t)}$

and operator of their ensemble $\hat{\varrho}_{(t)} = \sum_i p_i \hat{\varrho}_{i(t)}$ do not separate, generally speaking, into the product $\hat{\varrho}^{(1)}_{(t)} \hat{\varrho}^{(2)}_{(t)}$. Therefore any dynamic variables x and y characterizing source and adopter states will be statistically related, and their joint probability distribution may be found through computation of diagonal elements $\varrho_{xy,xy(t)}$ of the overall system density operator in a corresponding representation. It should be noted that the correlation entropy for x and y

$$H_{\text{rel}} = (x, y) = -\sum_x \varrho^{(1)}_{xx} \ln \varrho^{(1)}_{xx} - \sum_y \varrho^{(2)}_{yy} \ln \varrho^{(2)}_{yy} + \sum_{x,y} \varrho_{xy,xy} \ln \varrho_{xy,xy} \qquad (2.9)$$

characterizes the degree of statistical irreversibility of source-adopter interaction, but it by no means measures the transmitted information. Indeed, even at $H_1 = 0$ when the source is in a definite quantum state and does not send any messages, $H_{\text{rel}}(x, y)$ may differ from zero in consequence of interaction.

To compute the amount of information $I_y(t)$ extractable by measurements of adopter variables y, note that under transmission of ith signal followed by interaction with source, adopter is in the state having density operator $\varrho^{(2)}_i(t) = Sp_x \hat{\varrho}_i(t)$, therefore

$$I_y(t) = H(y, t) - \bar{H}_i(y, t) =$$

$$= -\sum_y \varrho^{(2)}_{yy}(t) \ln \varrho^{(2)}_{yy}(t) + \sum_i p_i \sum_y \varrho^{(2)}_{iyy}(t) \ln \varrho^{(2)}_{iyy}(t). \qquad (2.10)$$

Just in the same manner, the amount of information $I_x(t)$, obtained through measurement of source variables at the instant t, is expressed through the elements of matrices $\varrho^{(1)}_{ix,x}(t)$.

Let us see the regularities followed by redistribution of information under interaction between source and adopter. For this purpose sum $I_x(t)$ and $I_y(t)$ and make use of the evident inequality $H_i(x, t) + H_i(y, t) \geq H_i(x, y, t)$, thus obtaining

$$I_x(t) + I_y(t) \leq H_1(x, t) + H_2(y, t) - \bar{H}_i(x, y, t). \qquad (2.11)$$

To estimate $\bar{H}_i(x, y, t)$ note that x, y is not generally that set of variables where $\text{Sup}_{x,y} I_{xy} = I^*$ is reached, therefore $I_{xy}(t) = H(x, y, t) - \bar{H}_t(x, t) \leq I$, thus resulting in

$$I_1(t) + I_2(t) \leq I + H(x, t) + H(y, t) - H(x, y, t). \qquad (2.12)$$

The inequality (2.12) does not explicitly contain any characteristics of source states used as signals, but assumes that sets of variables x and y (whose mea-

* It is easy to show that $I(x, y, t) \leq H_1$. However, existence of, at first sight, evident inequality $I_{xy}(t) \leq \text{Sup}_x I_x(0)$ is still open to question, though its decision would be of principal importance for the indirect measurement theory.

surement is necessary for the information extraction procedure) are given. To do away with this restriction, that would be convenient in discussions of general regularities governing information transfer under interactions, make use of the statement below.

Theorem. Correlation entropy of relation between any variables x and y of various systems — $H_{\mathrm{rel}}(x, y) = H(x) + H(y) - H(x, y)$ — does not exceed $H_1 + H_2 - H$.

Refer to the proof of the Appendix.

Using this inequality together with (2.12) find

$$I_{1(t)} + I_{2(t)} \leq I + H_1(t) + H_2(t) - H\,. \tag{2.13}$$

To analyze the processes of energy, work, and information transfer under interactions, make use of (2.13) and (2.7). Consider the case of $H_2 = 0$. Physically this means that the device intended to accumulate the useful work or information is protected somehow against foreign noises and fluctuations.

Note first of all that adoptions of work and information from the source are, in a sense, mutually exclusive. Indeed, in the case of adoption of the maximal work corresponding to the equality sign in (2.5), by the end of interaction adopter is in the state $H_2(t) = 0$. Together with (2.10), this leads to $I_y(t) = 0$ since in this case the operators $\varrho_i^{(2)}(t)$ are independent of the signals sent.

Consider now an intermediate case of $H_1(t), H_2(t) \neq 0$. Refer to this end to (2.7) and (2.13). First, if interaction satisfies the condition of statistical reversibility $H_1(t) + H_2(t) = H$, it follows from (2.13) that $I_1(t) + I_2(t) \leq I$, i.e. sum of amounts of information extracted from the parts cannot exceed the total information contained in the system. To analyze (2.7), it is convenient to introduce "effective" temperatures $T_{1(t)}^{\mathrm{eff}}$, $T_{2(t)}^{\mathrm{eff}}$, T^{eff} related unambiguously to the entropies $H_1(t)$, $H_2(t)$ and H by

$$H_{\lambda(t)} = \int_0^{T_\lambda^{\mathrm{eff}}} \frac{C_\lambda(T')}{T'}\,dT'\,; \qquad H(T) = \int_0^{T^{\mathrm{eff}}} \frac{(C_1(T') + C_2(T'))}{T'}\,dT'\,, \tag{2.14}$$

where C_1 and C_2 are, respectively, source and adopter thermal capacities. If $T_1^{\mathrm{eff}} \neq T_2^{\mathrm{eff}}$, then $E_{\min(H)} < E_{\min(H_1)}^{(1)} + E_{\min(H_2)}^{(2)}$ if for no other reason than that the establishment of equilibrium in overall system is related to the equalization of the temperatures T_1^{eff} and T_2^{eff}. If $T_1^{\mathrm{eff}} = T_2^{\mathrm{eff}}$ and the reversibility condition is met, then instead of (2.7) obtain that

$$R_{1\max} + R_{2\max} = R_{\max}\,.$$

Analyse now behaviour of amounts of work and information in irreversible processes. As is seen from (2.13), the possibility of reproduction (i.e. $I_1(t) + I_2(t) > I$) leads inevitably to the increase of entropy and takes place only under irreversible source-adopter interaction. On the contrary, in this case the total amount of work $R_{1\max} + R_{2\max}$ contained separately in each part drops down because if $H_1(t) + H_2(t) > H$, relation $E_{\min(H)} < E^{(1)}_{\min(H_1)} + E^{(2)}_{\min(H_2)}$ holds even for $T_1^{\text{eff}} = T_2^{\text{eff}}$. Establish the relation between losses of total work and information. Introducing notation $\Delta I = I_1 + I_2 - I$, differentiating (2.7) and (2.13) and noting that $(\partial E_{\min}/\partial H) = T^{\text{eff}}$, find that

$$\begin{aligned} -d(\Delta R) &= T_1^{\text{eff}}\, dH_1 + T_2^{\text{eff}}\, dH_2 \\ d(\Delta I) &\leq dH_1 + dH_2 . \end{aligned} \tag{2.15}$$

If $T_1^{\text{eff}} = T_2^{\text{eff}}$, it follows from (2.15) that

$$-d(\Delta R) \geq T^{\text{eff}}\, d(\Delta I) . \tag{2.16}$$

This suggests that reproduction of information involves increase of entropy and inevitable loss in the total amount of useful work contained separately in each of the interacting systems.

3. Reproduction of information and loss of work in multichannel systems

An important problem which attracted repeated attention not only of physicists and mathematicians, but also of philosophers, is whether the loss of thermodynamical work under relaxation owing to collision of particles cannot be eliminated in principle, or this is due to our inability to compute phase trajectories of all particles for given initial conditions. Without claiming to have found the final solution of this problem, we are going only to discuss it in connection with the problem of statistical irreversibility accompanying signal ramification process.

Consider a complex physical system consisting of N subsystems with states statistically independent at $t = 0$ and characterized by mean energies E_λ and entropies H_λ $(1 \leq \lambda \leq N)$. As the result of collisions between subsystems, the values of total energy $E = \sum_\lambda E_\lambda$ and entropy $H = \sum_\lambda H_\lambda$, do not change, but, generally speaking, the sum $\sum_\lambda H_\lambda(t)$ increases and after long enough intermixing for $N \gg 1$ may, in the limit, reach $H_{\max}(E)$ corresponding to statistical equilibrium.

Try to make clear whether during relaxation (resulting from internal collisions in the system) the useful work is lost, and which of values $-H$ or

$\sum_\lambda H_\lambda(t)$ — is here of major importance. Generally speaking, at no stage of system evolution does (2.4) prevent form obtaining the work $R_{\max} = E - E_{\min}(H)$ and, at least in principle, for each finite N a complex enough method of influence upon the system may be worked out approaching arbitrarily this effectiveness limit.

It should be kept in mind, however, that all conclusions of the statistical thermodynamics concerning maximal work are obtained for devices influencing upon the system macroscopically, and therefore are based on the treatment of "one-particle" distribution functions or density matrices ignoring correlation between subsystems. This class comprises all systems and apparatuses used presently for extraction of useful work from non-equilibrium media, such as the Carnot machine piston pressing the working volume of gas, laser using spectral properties of pumping light to generate a monochromatic light but ignoring spectrum correlations, etc. With such an approach, it becomes inessential whether relaxation is due to internal collisions in the system, or to its interaction with an appropriate thermostat. Therefore work $R^{(T)}$ extracted by devices of this type is associated only with non-equilibrium of one-particle density matrices and is limited by the following inequality:

$$R_{(t)}^{(T)} \leq E - E_{\min}\left(\sum_\lambda H_\lambda(t)\right). \tag{3.1}$$

In the case where at $t = 0$ states of one or several systems contain information about the signal, after a sufficiently long period it will be distributed among all parts owing to internal interactions in the system, i.e. the signal will ramify. Amounts of information $I_\lambda(t)$ in each subsystem are in this case related by the following inequality which is direct generalization of (2.13):

$$\sum_\lambda I_\lambda(t) \leq I - H + \sum_\lambda H_\lambda(t) \tag{3.2}$$

where I as before is Sup of amount of information over complete sets of variables of the overall system.

Of course, the set of N subsystems under consideration should not be understood only as a set of particles. This example is but the most traditional in discussions of relaxation processes. In the theory of information transmission it is more natural to identify the subsystems with multichannel network branches whose interaction gives rise to transfer of information and energy under signal ramification.

Consider the case where at $t = 0$ only one of the subsystems carries information (assign to it No. 1). This subsystem corresponds to the channel feeding information into the system. See how much information each of sys-

tem branches may contain at an arbitrary instant $t > 0$ under various assumptions.

First, let initial ensemble of signals be in statistical equilibrium ($E_1 = E^{(1)}_{\min(H_1)}$), and the remaining subsystems be at the initial moment in the states with zero energies and entropies (i.e. without self-noise). Taking into consideration that $I \leq H_1(0)$ (see [18]), one can readily obtain for this case through (3.2):

$$\sum_{\lambda} I_\lambda(t) \leq H_{\max(E_1)} \tag{3.3}$$

where $H_{\max(E_1)}$ is the equilibrium entropy of the overall system corresponding to the energy E_1. Thus, in the absence of additional noise in multichannel system, (3.2) does not, basically, introduce new restrictions in addition to the evident $I \leq H$. For example, if all N branches are linear oscillators of the same frequency ω, then in the quasiclassical limit $E_1 \gg \hbar\omega N$ we have that

$$H^{(1)}_{\max(E)} = \ln e\,E - \ln \hbar\omega\,, \quad H_{\max(E)} = N(\ln eE - \ln \hbar\omega\, N)$$

and instead of (3.3) obtain

$$N^{-1} \sum_{\lambda} I_\lambda(t) \leq \ln e\,E/\hbar\omega - \ln N \tag{3.4}$$

i.e. under any finite N asymptotically complete reproduction is possible where, basically, each branch may contain practically complete information about the signal.

On the other hand, for the quantum case $E_1 < \hbar\omega$ we have $H^{(\lambda)}_{\max}(E_\lambda) = (E_\lambda/\hbar\omega) \ln \hbar\omega e/E_\lambda$, therefore $\sum_{\lambda} I_\lambda \leq \frac{E_1}{\hbar\omega} \ln \frac{\hbar\omega\, Ne}{E_1}$, and amount of mean information in each branch does not exceed $N^{-1}\left(\frac{E_1}{\hbar\omega} \ln \frac{\hbar\omega e}{E_1} + \ln N\right)$. This indicates that losses of information under signal ramification are inevitable even in the absence of additional noise. Note that if in the initial state is given the same entropy and information, but greater energy, then (3.3) will not impede the decrease of losses. This leads to the conclusion of expediency to amplify preliminarily the quantum signal or confer to it nonthermal statistical character.

But if in each branch of a multichannel system there are statistically signal-like thermal noises, all effective temperatures $T^{\text{eff}}_{\lambda(0)}$ are equal to each other, interaction is thermodynamically reversible, and information reproduction is at all impossible because $\sum_{\lambda} I_\lambda(t) \leq I$.

We shall illustrate the discussed relations by an example of a physical communication system using coherent signals. Let the transmitter create field

oscillator states $<\alpha_\lambda|$ within the frequency band $(\omega_0 - \Delta/2,\ \omega_0 + \Delta/2)$ and send them with probabilities $p(A_\lambda, \Phi_\lambda) = D^{-1} A_\lambda \exp(-A_\lambda^2/2D)$, where

$$D = 2\pi W/\hbar\omega_0 \Delta,\ A = |\alpha_\lambda|,\ \Phi_\lambda = i \ln(\alpha_\lambda/|\alpha_\lambda|),$$

and W is the total signal power. The transmission rate of information extracted under simultaneous measurement of canonical coordinates and signal pulses e.g. by the method of beam splitting [14]) is $\frac{\Delta}{2\pi} \ln\left(1 + \frac{2\pi W}{\hbar\omega_0 \Delta}\right)$. If the flow of transmitter emission ramifies into two similar noiseless channels, in each of them again there will be coherent signals [15] with amplitudes $A_{\lambda\nu} = A_\lambda/\sqrt{2}$ $(\nu = 1, 2)$ and phases $\Phi_{\lambda\nu} = \Phi_\lambda$ distributed in accordance with $p(A_{\lambda\nu}, \Phi_{\lambda\nu}) = (D/2)^{-1} A_{\lambda\nu} \exp(A_{\lambda\nu}^2/D)$. The information flow therefore may be obtained through the same formula and is $\frac{\Delta}{2\pi} \ln(1 + \pi W/\hbar\omega_0 \Delta)$. But if the channel extracting half of the signal power contains statistically signal-like thermal noise, the information flows r, r_ν in the original channel and output branches, respectively, are defined (see [14]) as follows:

$$r = \frac{\Delta}{2\pi} \ln(1 + 2\pi W/\hbar\omega_0 \Delta);\ r_\nu = \frac{\Delta}{2\pi} \ln\left(\frac{2\hbar\omega_\nu \Delta + 2\pi W}{\hbar\omega_0 A + 2\pi W}\right) \tag{3.5}$$

thus $r_1 + r_2 < r$. It is easy to see that this example demonstrates all discussed above peculiarities of information reproduction under signal ramification that follows from (3.20).

In conclusion we shall dwell briefly on the interrelation between losses of thermodynamical work and information reproduction under interaction of subsystems. Differentiating (3.1) and (3.2), and introducing effective temperature $T^{\text{eff}} = (dH_{\max}/dE)_{Z_\lambda N_\lambda}$ as before, we find the following relation:

$$-dR_{\max}^{(T)} \geq T^{\text{eff}} \propto \left(\sum_\lambda I_\lambda - I\right) \tag{3.6}$$

which at $N = 2$ boils down to (2.16) because for $T_1^{\text{eff}} = T_2^{\text{eff}}$ $R_{\max}^{(T)}$ and $R_{1\max} + R_{2\max}$ of (2.7) coincide. Thus, data reproduction under interaction of subsystems is inevitably accompanied by a thermodynamically irreversible process and by a loss of useful work $R^{(T)}$.

Appendix

Let us prove the stated above theorem. Denote by $\varrho_{nm,n'm'}$ the elements of overall system density matrix in a representation diagonalizing $\hat{\varrho}^{(1)}$ and $\hat{\varrho}^{(2)}$, by ϱ_i its eigenvalue, and by $\varrho_n^{(1)}$ and $\varrho_m^{(2)}$, respectively, eigenvalues of subsystem density matrices. If $V_{nx}^{(1)}$ and $V_{my}^{(2)}$ are unitary matrices relating "n, m"

and "x, y" representations, the correlation entropy for any variables x and y will have the following form

$$H_{\text{rel}}(x, y) = H(x) + H(y) - H(x, y) = -\sum_x p(x) \ln p(x) - \\ - \sum_y p(y) \ln p(y) + \sum_{x,y} p(x, y) \ln p(x, y) \qquad \text{(A. 1)}$$

where

$$p(x) = \sum_n \varrho_n^{(1)} |V_{nx}^{(1)}|^2; \; p_y = \sum_m \varrho_m^{(2)} |V_{my}^{(2)}|^2; \; p_{xy} = \Sigma V_{xn}^{(1)} V_{ym}^{(2)} \varrho_{nm,n'n} V_{xn'}^{*(1)} V_{ym'}^{*(2)}.$$

Introduce also $\widetilde{H}_{\text{rel}}(x, y)$ differing from (A.1) in that

$$\widetilde{H}(xy) = -\sum_{x,y,n,m} \varrho_{nm,nm} |V_{nx}^{(1)} V_{my}^{(2)}|^2 \times \ln \sum_{m,n} \varrho_{nm,nm} |V_{nx}^{(1)} V_{my}^{(2)}|^2$$

is substituted for $H(x, y)$.

Show that $\widetilde{H}_{\text{rel}}(x, y) \leq H_{\text{rel}}(n, m) = -\sum_n \varrho_n^{(1)} \ln \varrho_n^{(1)} - \sum_m \varrho_m^{(2)} \ln \varrho_m^{(2)} + \\ + \sum \varrho_{nm,mm} \ln \varrho_{nm,nm}$. By writing joint distribution of probabilities of four random variables $\tilde{p}(x, n, m, y) = |V_{nx}^{(1)} V_{my}^{(2)}|^2 \varrho_{nm,nm}$, one readily verifies that they constitute a Markov chain where x and y are extreme elements. Thus $\widetilde{H}_{\text{rel}}(x, y) \leq H_{\text{rel}}(n, m)$.

Before returning to discussion of related systems, consider the following abstract scheme. Let ϱ_i be eigenvalues of the density matrix of a physical system, and $\hat{U}_{ik}$ and $\hat{V}_{kz}$ be unitary matrices whose product $\hat{U}\hat{V} = S$ is fixed. Introduce entropies

$$H(k) = -\sum_k p(k) \ln p(k); \; p(k) = \sum_i \varrho_i (U_{ik})^2$$

$$H(z) = -\sum_z p(z) \ln p(z); \quad p(z) = \sum_i \varrho_i |S_{iz}|^2 \qquad \text{(A.2)}$$

$$\widetilde{H}(z) = -\sum_z \tilde{p}(z) \ln \tilde{p}(z); \; \tilde{p}(z) = \sum_{i,k} \varrho_i |U_{ik} V_{kz}|^2$$

$$H = -\sum_i \varrho_i \ln \varrho_i$$

and compare $H(z) - H$, $\widetilde{H}(z) - H(k)$ assuming that $H \leq H(k) \leq H(z)$, and unitary operators $\hat{U}$ and $\hat{V}$ belong to the continuous sequence $\hat{U}_{(\beta)}$, $\hat{V}_{(\beta)}$ ($0 \leq \beta \leq 1$, $\hat{U}_{(0)} = \hat{1}$, $\hat{V}_{(1)} = \hat{1}$) for which $\partial H(k, \beta) \mid \partial\beta > 0$. It is easy to see that $H(z) - H$ is β-independent, whereas $\widetilde{H}(z) - H(k)$ decreases from $H(z) - \\ - H$ to zero as β changes from 0 to 1. By means of direct, but somewhat tire-

some varying, one may show that under our assumptions this decrease is monotonic, thus the following inequality holds

$$\tilde{H}(z) - H(k) \leq H(z) - H. \tag{A.3}$$

Return now to the entropy of relation between two systems. It follows from (A.3) that

$$\tilde{H}(x, y) - H(n, m) \leq H(x, y) - H. \tag{A.4}$$

Indeed, the unitary matrix $U_{inm,jn'm'}$ reducing $\varrho_{nm,n'm'}$ to the diagonal form is defined to the interchange of eigenvalues ϱ_i, and may be always chosen to have a form enabling one to deform it continuously to the identity matrix with monotonic change of $H(n, m)$. Similar reasoning holds for the matrices $V^{(1)}_{nx}$, $V^{(2)}_{my}$ diagonalizing $\varrho^{(1)}_{xx'}$ and $\varrho^{(2)}_{yy'}$, and, consequently, for their direct product effecting transition from "xy-representation" to "nm-representation". Thus (A.3) is satisfied. Combining (A.4) and $\tilde{H}_{\text{rel}}(x, y) < H_{\text{rel}}(n, m)$, and taking into consideration that $\varrho^{(1)}_{nn'} = \varrho^{(1)}_{n}\, \delta_{nn'}$, $\varrho^{(2)}_{mm'} = \varrho^{(2)}_{m}\, \delta_{mm'}$, obtain

$$H_{\text{rel}}(x, y) \leq H_1 + H_2 - \sum_i \varrho_i \ln \varrho_i \,. \tag{A.5}$$

Q.E.D.

References

1. *Gabor, D.*, Phyl. Mag., **41**, 7, 1950.
2. *Brillouin, Leon*, "Science and Information Theory", Academic Press Inc., Publishers New York, 1956.
3. *Szilard, L.*, Z. Phys., B. 53, 1929.
4. *Файн, В. М., Ханин, Я. И.*, Квантовая радиофизика. Изд. «Сов. радио». М.
5. *Лебедев, Д. С., Левитин, Л. Б.*, Сб. Теория передачи информации, Наука, М., 1964.
6. *Боровицкий, С. И., Митюгов, В. В.*, Изв. ВУЗов — Радиофизика, 7, 854, 1964.
7. *Gordon, J. P.*, Proc. IRE, **50**, 9, 1962.
8. *Von Neumann, J. V.*, "Mathematische Grundlagen der Quantenmechanik", Springer Verlag, 1932 (or "Mathematical Foundations of Quantum Mecanics" Princeton Univ. Press. 1955)
9. *Elsasser, N. M.*, Phys. Rev., **52**, 987, 1937.
10. *Ingarden, R. S.*, Acta Physica Polonica (Preprint No 60).
11. *Shannon, C. E.*, A mathematical theory of communication. Bell System Tech. J., **27** (1948), *3*, 379–423; **27** (1948), *4*, 623–656.
12. *Стратонович, Р. Л.*, Изв. ВУЗов — Радиофизика, **8**, 116, 1965.
13. *Левитин, Л. Б.*, Труды IV конференции по теории передачи информации, секция П. М. — Ташкент, 1969.
14. *Левитин, Л. Б., Митюгов, В. В.*, Тезисы докладов I Всесоюзной конференции по проблемам передачи информации лазерным излучением. г. Киев, 1969.
15. *Glauber, R. J.*, see in "Optique et Electronique Quantique. Quantum Optics and Electronics" Gordon and Breach, Science Publishers, New York–London–Paris, 1965.

REALIZATION OF PARTIALLY SYMMETRIC FUNCTIONS BY NETWORKS OF FUNCTIONAL ELEMENTS

A. V. MARKOVSKII, L. B. SHIPILINA

(Moscow)

Realization of partially-symmetric functions by networks of functional elements is discussed. An asymptotic expression of Shannon's function is derived. Two specific kinds of partially symmetric functions are studied which result in simpler, as opposed to the general case, realization. Asymptotic Shannon's functions are given for them.

The present paper is one more example of the local coding principle by O. B. Lupanov [1], and it is based essentially on his concepts and notations.

A Boolean function $f(\mathbf{x})$ is referred to as partially symmetric (PSF) [2, 3] if there is a partition of the set of its arguments $\mathbf{x} = (x_1, \ldots, x_n)$ into non-empty subsets $\mathbf{x}_i = (x_{i1}, \ldots, x_{ini})$; $i = 1, 2, \ldots, k$, $n = \sum_{k=1}^{k} n_i$ (called "Symmetry groups") such that $f(\mathbf{x})$ does not change under an arbitrary permutation of arguments inside any of the groups $\mathbf{x}_i$. Without loss of generality we shall take that

$$\mathbf{x}_2 = (x_{n_1+1}, \ldots, x_{n_2+n_2}), \ldots, \mathbf{x}_k = (x_{n_1+n_2+\ldots+n_{k-1}+1}, \ldots, x_{n_1+\ldots+n_k}).$$

The numbers $n_1, \ldots, n_k$ will be referred to as parameters of the PSF $f(\mathbf{x})$ which will be denoted as $f_{n_1,\ldots,n_k}(\mathbf{x})$.

It is clear that by taking $n_i = 1$, $i = 1, 2, \ldots k$, $k = n$, any Boolean function may be regarded as partially symmetric. Symmetric functions are another specific kind of PSF [4].

Consider implementation of PSF by networks of functional elements over a finite basis [5]. Complexity $Z(\varrho)$ of a network S over a given basis is understood as the sum of weights (prices) of constituing elements. By complexity $Z(F^m)$ of a system $F^m(\mathbf{x})$ of m PSF $f_1(\mathbf{x}), \ldots f_m(\mathbf{x})$, the minimum of complexities of networks implementing $F^m(\mathbf{x})$ over the given basis is meant.

Let $\mathfrak{F}^m_{n_1,\ldots,n_k}$ be the class of all PSF systems $F^m_{n_1,\ldots,n_k}(\mathbf{x})$ having parameters $n_1, \ldots, n_k$. The function $Z(n_1, \ldots, n_{k,m}) = \max Z(F^m_{n_1,\ldots,n_k}(\mathbf{x}))$ over all $F^m_{n_1,\ldots,n_k}(\mathbf{x}) \in \mathfrak{F}^m_{n_1,\ldots,n_k}$ is referred to as Shannon's function for the class $\mathfrak{F}^m_{n_1,\ldots,n_k}$.

Let $H(n_1, \ldots, n_k) = \prod_{i=1}^{k} (n_i + 1)$. The following theorem on behaviour of the function $Z(n_1, \ldots, n_k, m)$ holds.

Theorem 1. For any sequence $\mathfrak{F}^{m_1}_{n_{11}, \ldots, n_{k_1}}$, $\mathfrak{F}^{m_2}_{n_{12}, \ldots, n_{k_2}}, \ldots,$ $\mathfrak{F}^{m_t}_{n_{1t}, \ldots, n_k}$ of PSF classes such that $n = \sum_{i=1}^{k} n_i \to \infty$,

$$\frac{\log m}{H} \overset{1}{\to} 0 \tag{1}$$

and

$$\frac{n \log mH}{mH} \to 0 \tag{2}$$

the following holds:

$$Z(n_1, \ldots, n_k, m) \overset{2}{\sim} \varrho \frac{mH(n_1 \ldots, n_k)}{\log mH(n_1, \ldots, n_k)} \tag{3}$$

where ϱ is the minimum of basis element reduced weights.

The reduced weight is defined only for elements ε having at least two inputs, and is equal to $\frac{p(\varepsilon)}{i_\varepsilon - 1}$ where $p(\varepsilon)$ and i_ε are, respectively, weight and number of inputs of the element ε [5].

Proof. Let $\mathfrak{F}^{m}_{n_1, \ldots, n_k}(\mathbf{x})$ be a system of m PSF $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ having parameters $n_1, \ldots, n_k$. Consider the union of all possible sets $\mathbf{b} = (b_1, \ldots, b_k)$ of integers b_i where $0 \leq b_i \leq n_i$; $i = 1, 2, \ldots k$.

The number of such sets is $H(n_2, \ldots, n_k)$.

Order lexicographically the union of the sets $\mathbf{b}$ from right to left, and number them by natural numbers 0 through $H - 1$. Denote the number of a set $\mathbf{b}$ by $B(\mathbf{b})$:

$$\begin{aligned} B(\mathbf{b}) = {} & (n_1 + 1)(n_2 + 1) \cdot \ldots \cdot (n_{k-1} + 1) b_k + \ldots + \\ & (n_1 + 1)(n_2 + 1) \cdot \ldots \cdot (n_{i-1} + 1) b_i + \ldots + \\ & (n_1 + 1)(n_2 + 1) b_3 + (n_1 + 1) b_2 + b_1. \end{aligned} \tag{4}$$

Let $0 \leq B \leq H - 1$. Denote by $\mathbf{b}(B) = \{b_1(B), \ldots, b_i(B), \ldots, b_k(B)\}$ the set $\mathbf{b}$ having the number B. Take $H_i = \prod_{j=2}^{i} (n_{j-1} + 1)$ for $i = 2, 3, \ldots, k$ and $H_1 = 1$; then $B(\mathbf{b}) = \sum_{i=1}^{k} H_i b_i$.

[1] Indices are omitted. The log is taken to the base 2.

[2] $a \sim b$ means that $\frac{a}{b} \to 1$ for $a, b \to \infty$.

The system $F^m_{n_1,\ldots,n_k}(\mathbf{x})$ is defined completely by the sequence $\Pi = \tilde{\pi}_0, \tilde{\pi}_1, \ldots, \tilde{\pi}_{H-1}$ of the length mH where $\tilde{\pi} = f_1(\mathbf{b}(B)), f_2(\mathbf{b}(B)), \ldots, f_m(\mathbf{b}(B))$ is a set of PSF values $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$ over (any) set of argument values containing in the group $\mathbf{x}_i$ $(i = 1, 2, \ldots, k)$ $b_i(B)$ ones and $n_i - b_i(B)$ zeroes. It is obvious that the sequence Π may be constructed unambiguously through the system $\mathfrak{F}^m_{n_1,\ldots,n_k}(\mathbf{x})$, various sequences Π corresponding to various systems from $\mathfrak{F}^m_{n_1,\ldots,n_k}$. Thus the sequence Π may be used as the code of a system $F^m_{n_1,\ldots,n_k}(\mathbf{x})$ in the class $\mathfrak{F}^m_{n_1,\ldots,n_k}$. Hence, the class $\mathfrak{F}^m_{n_1,\ldots,n_k}$ contains 2^{mH} systems $F^m_{n_1,\ldots,n_k}(\mathbf{x})$. By the virtue of (1) and (2), and Theorem D1[1] we obtain immediately the lower bound of cardinality for $Z(n_1, \ldots, n_k, m)$

$$Z(n_1 \ldots n_k, m) \overset{3}{\gtrsim} \varrho \frac{mH(n_1, \ldots, n_k)}{\log m\, H(n_1, \ldots, n_k)}. \tag{5}$$

The upper bound for $Z(n_1, \ldots, n_k, m)$ is obtained by means of the local coding principle [1]. Let $F^m_{n_1,\ldots,n_k}(\mathbf{x})$ be an arbitrary system from $\mathfrak{F}^m_{n_1,\ldots,n_k}$, and $\Pi = \tilde{\pi}_0, \tilde{\pi} \ldots, \tilde{\pi}_{H1}$ be the code of the system $F^m_{n_1,\ldots,n_k}(\mathbf{x})$. Let $\tilde{\sigma}$ be an arbitrary set of values of the variables $x_1, \ldots, x_n$, and $\tilde{\sigma}_i$; $i = 1, 2, \ldots, k$ be a subset of the set $\tilde{\sigma}$ comprising values of the variables of the group $\mathbf{x}_i$ and let b_i be the number of ones in $\tilde{\sigma}_i$. The set $F^m_{n_1,\ldots,n_k}(\tilde{\sigma}) = \{f_1(\tilde{\sigma}), \ldots, f_m(\tilde{\sigma})\}$ will be computed as follows:

1) compute b_i for each $\tilde{\sigma}_i$,
2) for the set $\mathbf{b} = (b_1, \ldots, b_k)$ compute $B(\mathbf{b})$ through (4), and
3) separate from Π a part $\tilde{\pi}_B$ coinciding with $F^m_{n_1,\ldots,n_k}(\tilde{\sigma})$.

Denote by $\tilde{\beta}_i$, $\tilde{\nu}_i$ and $\tilde{B}(\mathbf{b})$ binary sets of lengths $]\log(n_i + 1)[$, $]\sum_{i=1}^{k} \log(n_i + 1)[$, and $]\log H[$ representing, respectively, the numbers b_i, H_i, and $B(\mathbf{b})$.

A network S for $F^m_{n_1,\ldots,n_k}$ is constructed from the following blocks (Fig. 1):

1) blocks Б_i, $i = 1, 2, \ldots, k$ compute the number of ones (b_i) in the sets $\tilde{\sigma}_i$. Block Б_i has n_i inputs and $]\log(n_i + 1)[$ outputs. According to Theorem 3.1 of [1] obtain

$$\sum_{i=1}^{k} Z(\sigma_i) = O\left(\sum_{i=1}^{k} n_i\right) = O(n)^4. \tag{6}$$

[3] $a \gtrsim b$ means that $\lim \frac{a}{b} \geq 1$ for $a, b \to \infty$.

[4] $a = O(b)$ means that $\frac{a}{b} \leq k$ for $a, b \to \infty$ where k is constant.

2) Block B computes $B(\mathbf{b})$ through the set $\mathbf{b} = (b_1, \ldots, b_k)$. It has $\sum_{i=1}^{k}]\log(n^i \neq 1)[$ inputs and $]\log H[=] \sum_{i=1}^{k} \log(n_i + 1)[$ outputs. Block B multiplies binary numbers $\tilde{p}_i$ and $\tilde{v}_i$, and sums subsequently the results. In doing so k multiplication and k addition networks for two binary numbers of at most $]\log H[$ bits are required. The multiplication network may be chosen so that its complexity does not exceed $C_1 \log^2 H$. (It may be based, for instance, on the conventional algorithm for multiplication of binary numbers). The addition network may be chosen so that its complexity does not exceed $C_2 \log H$[1], C_1, C_2 = const. Hence, it follows that

$$Z(\mathrm{B}) = O(k \log^2 H).$$

3) Block C defines the part $\tilde{\pi}_B$ of the code Π through $B(\mathbf{b})$. It has $]\log H]$ inputs feeding in the number $\tilde{B}(\mathbf{b})$, and m outputs feeding out the set $\pi_B = F^m_{n_1,\ldots,n_k}(\tilde{\sigma})$ of values of the functions $f_1, \ldots, f_m$ over the set $\tilde{\sigma}$.

Note that though the number of al possible value sets at $]\log H[$ inputs of block C is $2^{]\log H[}$, in practice only H first[5] sets appear at block C inputs. Therefore block C may be regarded as realizing a system of functions $\varphi_1, \ldots, \varphi_m$ of $]\log H[$ arguments that are defined arbitrarily over H first sets of arguments and equal to zero over the remaining ones. According to Theorem D8 [1], the network implementing block C may be chosen to have complexity

$$Z(c) \lesssim \varrho \frac{mH}{\log mH} \tag{8}$$

obtain through (6), (7) and (8)

$$Z(\varrho) \lesssim \varrho \frac{mH}{\log mH} + \sigma(n) + O(k \log^2 H). \tag{9}$$

By means of the obvious relation

$$H(n_1, \ldots, n_k) \geq 2^k$$

obtain

$$\frac{k \log^3 H}{mH} = \frac{1}{m} \cdot \frac{K}{H^{1/2}} \cdot \frac{\log^3 H}{H^{1/2}} \leq \frac{1}{m} \cdot \frac{k}{\sqrt{2^k}} \cdot \frac{\log^3 H}{\sqrt{H}} \to 0.$$

Then through (1), (2), and (9) obtain

$$Z(n_1, \ldots, n_k, m) \leq Z(F^m_{n_1,\ldots,n_k}) \leq Z(\varrho) \lesssim \varrho \frac{mH}{\log mH} \tag{10}$$

and through (10) and (5) obtain (3). Q.E.D.

[5] The sets are considered as ordered in the order of increase of numbers whose binary "images" they are.

Note that for a system of m symmetric functions (in this case $H(n) = n + 1 \sim n$) (3) turns into

$$Z(n) \sim \varrho \frac{mn}{\log mn}$$

(see Theorem 3.2 [1]), and for a system of arbitrary functions (in this case $k = n$, $n_1 = n_2 = \ldots = n_k = 1$, $H(n_1, \ldots, n_k) = 2^n$) it turns into

$$Z(n) \sim \varrho \frac{m\, 2^n}{n + \log m}$$

(see Theorem D10 [1]).

Consider two specific cases where implementation of PSF is simpler than in the general case.

1. Let $F^{(1)m}_{n_1, \ldots, n_k} = \{f_1, \ldots, f_m\}$ be a PSF system, each of them being representable as

$$f_i = \varphi_i(\varrho_1(\mathbf{x}_1), \varrho_2(\mathbf{x}_2), \ldots, \varrho_k(\mathbf{x}_k)), \qquad i = 1, 2, \ldots, m,$$

where $\varrho_i, \ldots, \varrho_k$ are symmetric functions, φ is a Boolean function of k arguments, and the sets $\mathbf{x}_i$, $i = 1, 2, \ldots, m$, comprise n_i arguments. Denote the set of all systems $F^{(1)m}_{n_1, \ldots, n}(\mathbf{x})$ by $\mathfrak{F}^{(1)m}_{n_1, \ldots, n_k}$ and its Shannon's function by $Z^{(1)m}_{(n_1, \ldots n_k)})$.

Theorem 2. If for $n \to \infty$, $k \to \infty$

$$\log \frac{m}{2^k} \to 0 \tag{11}$$

and

$$\frac{k \cdot n \cdot \log m}{m\, 2^k} \to 0 \tag{12}$$

then

$$Z^{(1)}_{(n_1, \ldots, n_k, m)} \sim \varrho \frac{m\, 2^k}{k + \log m}\,. \tag{13}$$

Proof. Denote the number of systems $F^{(1)m}_{n_1, \ldots, n_k}$ by $N^{(1)}_{(n_1, \ldots, n_k, m)}$.

It is evident that

$$N^{(1)}_{(n_1, \ldots, n_k, m)} = 2^{n_1+1} \cdot \ldots \cdot 2^{n_k+1} \cdot 2^{m2^k} = 2^{n+2} \cdot 2^{m2^k}.$$

The first factor gives the number of possible choices of the functions $\varrho_1, \ldots, \varrho_k$; the second gives the number of choices of the functions $\varphi_1, \ldots, \varphi_m$. Denote the binary logarithm of a number $N^{(1)}_{(n_1, \ldots, n_k, m)}$ by $H^{(1)}_{(n_1, \ldots, n_k, m)}$. Then, by the virtue of (12),

$$H^{(1)}_{(n_1, \ldots, n_k, m)} = n + k + m \cdot 2^k \sim m2^k\,. \tag{14}$$

Through (14) and Theorem D1 [1], and by the virtue of (11) and (12) obtain

$$Z^{(1)}_{(n_1,\ldots,n_k)} \geq \varrho \frac{H^{(1)}_{(n_1,\ldots,n_k,m)}}{\log H^{(1)}_{(n_1,\ldots,n_k,m)}} \sim \varrho \frac{m2^k}{\log m2^k}. \tag{15}$$

Figure 2 shows a network S implementing an arbitrary system $F^{(1)m}_{n_1,\ldots n_k}$. Blocks $\varrho_1, \ldots, \varrho_k$ realize symmetric functions $\varrho_1(\mathbf{x})_1, \ldots, \varrho_k(\mathbf{x}_k)$. By Theorem 3.1 [1] obtain

$$Z(\varrho_i) = O(n_i). \tag{16}$$

Block D has k inputs and m outputs, and realizes functions $\varphi_1, \ldots, \varphi_m$. By the virtue of (11) and Theorem D8 [1] obtain

$$Z(D) \leq \varrho \frac{m2^k}{\log m2^k}.$$

This and (16) result in

$$Z^{(4)}_{(n_1,\ldots,n_k)} \leq Z(s) \lesssim \varrho \frac{m2^k}{\log m2^k} + O(n_1) + \ldots + O(n_k) = \frac{\varrho\, m2^k}{\log m2^k} + O(n).$$

By the virtue of (12) and (15) obtain (13). Q.E.D.

2. Let $F^{(2)m}_{n_1,\ldots,n_k} = \{f_1, \ldots, f_m\}$ be a PSF system, each of them being representable as

$$f_j = C_j(S_1(\mathbf{x}), \varrho_2(\mathbf{x}_2), \ldots, \varrho_k(\mathbf{x}_k)), \qquad j = 1, 2, \ldots, m,$$

where $\varrho_1, \ldots, \varrho_k$; $C_1, \ldots, C_m$ are symmetric functions, and the sets $\mathbf{x}_i$ comprise n_i arguments.

Denote by $\mathfrak{F}^{(2)m}_{n_1,\ldots n_k}$ the set of all systems $F^{(2)m}_{n_1,\ldots,n_k}(\mathbf{x})$, by $N^{(2)}_{(n_1,\ldots,n_k,m)}$ the cardinality of the set $\mathfrak{F}^{(2)m}_{n_1,\ldots,n_k}$, and by $Z^{(2)}_{(n_1,\ldots,n_k,m)}$ the Shannon function for the class $\mathfrak{F}^{(2)m}_{n_1,\ldots n_k}$. Let $H^{(2)}_{(n_1,\ldots,n_k,m)} = \log N^{(2)}_{(n_1,\ldots,n_k,m)}$.

Theorem 3. If $n \to \infty$, $k \to \infty$ and

$$\frac{\log k}{m} \to 0 \tag{17}$$

$$\frac{\log m}{k} \to 0 \tag{18}$$

$$\frac{n \log km}{km} \to 0 \tag{19}$$

then

$$Z^{(2)}_{(n_1,\ldots,n_k,m)} \sim \varrho \frac{km}{\log km}. \tag{20}$$

Proof. It is evident that

$$N^{(2)}_{(n_1,\ldots,n_k,m)} = 2^{n+k} \cdot 2^{m(k+1)} = 2^{n+k+m+mk}.$$

Then, by the virtue (19), $H_{(n_1,\ldots,n_k,m)} \sim m_k$. Hence, by Theorem D1 [1] the lower bound is:

$$Z^2_{(n_1,\ldots,n_k,m)} \geq \varrho \frac{mk}{\log mk}. \tag{21}$$

Figure 2 shows a network implementing an arbitrary system $F^{(2)m}_{n_1,\ldots,n_k}(\mathbf{x})$. Block $S_1, \ldots, \varrho_k$ are the same as above. Block D has k inputs and m outputs, and realizes functions $C_1, \ldots, C_m$. By the virtue of (17) and (18) and in accordance with Theorem 3.2 [1] obtain

$$Z(D) \leq \varrho \frac{mk}{\log mk}.$$

This and (16) result in

$$Z^2_{(n_1,\ldots,n_k,m)} \leq Z(\varrho) \leq \varrho \frac{mk}{\log mk} + O(n).$$

By the virtue of (19) and (21) obtain (20). Q.E.D.

References

1. *Лупанов, О. Б.:* Об одном подходе к синтезу управляющих систем — принципе локального кодирования. Сб. «Проблемы кибернетики», вып. **14,** 1965, 31—110.
2. *Mukhopadhyay, A.:* Detection of Total or Partial Symmetry of a Switching Function with the Use of Decomposition Charts. IEEE Trans. on Electr. Comp., vol. **EC—12,** No. *5,* October 1963, pp. 553—561.
3. *Dietmeyer, D. L., Schneider, P. R.:* Identification of Symmetry, Redundancy and Equivalence of Boolean Functions. IEEE Trans. on Electr. Comp., vol. **EC—16,** No. *6,* December 1967, pp. 804—817.
4. *Caldwell, S. H.:* Switching Circuits and Logical Design, Wiley, New York, 1958.
5. *Лупанов, О. Б.:* Об одном классе схем из функциональных элементов. Сб. «Проблемы кибернетики», вып. **7,** 1962, 61—114.

MAGYAR TUDOMÁNYOS AKADÉMIA KÖNYVTÁRA

NOTE TO CONTRIBUTORS

Two copies of the manuscripts (each duly completed by figures, tables and references) are to be sent either to

E. D. Teryaev coordinating editor

Department of Mechanics and Control Processes

Academy of Sciences of the USSR

Leninsky Prospect 14, Moscow V-71, USSR

or to

J. Kocsis coordinating editor

Department of Automation

Technical University

H–1111 Budapest XI., Goldman György tér 3, Hungary

The authors are requested to retain a third copy of the submitted typescript to be able to check the proofs against it

The papers, preferably in English or Russian, should be typed double-spaced on one side of good-quality paper, with wide margins (c. 4–5 cm) should carry the title of the contribution, the author(s)' name, and the name of the country. At the end of the typescript the name of that author who manages the proof-reading should also be given.

An abstract of about 50 to 100 words should head the paper.

The authors are encouraged to use the following headings: Introduction, (outlining the problem), Methods and results, Discussion, Conclusions, References. The entire material should not exceed 15 pages including tables and references. The proper location of the tables and figures must be indicated on the margins. Mathematical notations should follow up-to-date usage.

The summary – possibly in Russian if the paper is in English and *vice versa* – should contain a brief account of the proposition and indications of the formulas used and figures shown in the paper. The summary is not supposed to exceed 10–15 per cent of the paper.

The authors will be sent sheet-proofs which they are to return by next mail to the sender Regional Editorial Board.

Authors are entitled to 100 reprints free of charge. Rejected manuscripts will be returned to the authors.

К СВЕДЕНИЮ АВТОРОВ

Рукописи в двух экземплярах (каждый из которы должен содержать рисунки, таблицы и литературу направляются

Е. Д. Теряеву — Научный секретарь журнала

Отделение механики и процессов управления

Академия Наук СССР

Ленинский Проспект, 14, Москва В-71, СССР

или

Я. Кочшшу — Научный секретарь журнала

Кафедра Автоматизации
Будапештского Технического
Университета, Будапешт XI, площадь
H–1111 Голдманн Д., 3, Венгрия

Авторам рекомендуется оставлять у себя копию всех представленных ими материалов для справок при корректуре.

Статьи, желательно на русском или английском языках, отпечатанные на бумаге хорошего качества, с промежутком в два интервала и широкими (4–5 см) полями должны содержать наименование статьи, фамилию автора (авторов), название страны. В конце статьи необходимо также указать фамилию автора, ответственного за корректуру гранок.

Статье должна предшествовать аннотация объемом до 50–100 слов.

Авторы при написании статьи должны придерживаться следующей формы: введение (постановка задачи), основное содержание и результаты, обсуждение, выводы и литература. Объем статьи не должен превышать 15 печатных страниц, включая таблицы и ссылки. Последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями.

К статье обязательно должно быть приложено резюме-реферат. Резюме — на русском языке, если статья написана на английском, и наоборот — должно содержать краткое изложение текста статьи со ссылками на необходимые формулы и графики, имеющиеся в основном тексте. Объем разюме не должен превышать 10–15 % объема статьи.

Авторам высылаются гранки статьи, которые они должны незамедлительно возвратить в Региональную секцию Редколлегии журнала.

Авторам обеспечивается бесплатно 100 оттисков их статей. Рукописи непринятых статей возвращаются авторам.

CONTENTS • СОДЕРЖАНИЕ

Yu, P. L.—Leitmann, G.: Compromise solutions, domination structures and Salukvadze's solution (*Йу П. Л.—Лейтманн Г:* Компромиссные решения, доминирующие структуры и решение Салуквадзе) 183

Willems, J. L.: Mean square stability criteria for linear white noise stochastic systems (*Виллемс Й. Л:.* Квадратичные критерии устойчивости для линейных стохастических систем с белыми шумами) 199

Apokorin, D. S.: Asymptotic properties of Markov chain parameter estimates (*Апокорин Д. С.:* Оценка параметра марковской цепи) 219

Vittikh, V. A.: Synthesis of algorithms for data compression (*Виттих В. А.:* Синтез алгоритмов сжатия данных) 235

Митюгов В. В.: Энтропия, информация и работа в квантовой статистике (*Mityugov, V. V.:* Entropy, information and work in quantum statistics) 243

Bunich, A. L.—Raibman, N. S.: Dispersional identification equation of nonlinear plants (*Бунич А. Л.—Райбман Н. С.:* Дисперсионное уравнение идентификации нелинейных объектов) 257

Марковский А. В.—Шипилина Л. Б.: Реализация частично-симметрических функций схемами из функциональных элементов (*Markovskii, A. V.—Shipilina, L. B.:* Realization of partially symmetric functions by networks of functional elements) 273

Csiszár, I.: On the capacity of noisy channels with arbitrary signal costs (*Чисар И.:* О проблеме емкости каналов с помехами при любых затратах на сигналы) 283

Ohera, L.: Basic considerations concerning frequency dictionaries (*Охера Л.:* Общие методы составления частотных словарей) 305

Book review 315