

316920

VOL. 20 • NUMBER 1
TOM 20 • НОМЕР 1

20
1991

9

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

12

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
УПРАВЛЕНИЯ И
ИНФОРМАЦИИ
ТЕОРИИ



АКАДЕМИЯ НАУК С С С Р **1991**
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

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

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

- theory of control processes
- theory of adaptive systems
- theory of estimation and identification
- theory of controlling robot-technologic and flexible manufacturing systems
- information theory
- information-theoretic aspects of multiple access networks.

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

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

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

- теория процессов управления;
- теория адаптивных систем;
- теория оценивания и идентификации;
- теория управления робототехническими и гибкими производственными системами;
- теория информации;
- теория информации в области сетей с множественным доступом.

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

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czech and Slovak Federal Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia:

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

For all other countries:

PERGAMON PRESS PLC Headington Hill Hall, Oxford OX3 OBW, England

or

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

1991 Subscription Rate DM 627,— per annum including postage and insurance.

SUBJECT INDEX

- Carbonez, A., Györfi, L., van der Meulen, E. C.*: Nonparametric entropy estimation based on randomly censored data. **20**, *6*, pp. 441–451
- Chentsov, A. G.*: On the construction of solution to nonregular problems of optimal control. **20**, *2*, pp. 129–143
- Chernyak, A. I., Sztrik, J.*: Asymptotic behaviour of a complex renewable standby system with fast repair. **20**, *1*, pp. 37–44
- Emelyanov, S. V., Zhivoglyadov, P. V., Korovin, S. K.*: Analysis of admissible perturbations and stabilization of uncertain discrete-time plants. **20**, *5*, pp. 353–371
- En-hui Yang*: Universal almost sure data compression for abstract alphabets and arbitrary fidelity criterions. **20**, *6*, pp. 397–408
- Faragó, A., Linder, T., Lugosi, G.*: Nearest neighbor search and classification in $O(1)$ time. **20**, *6*, pp. 383–395
- Ferrante, M.*: On finite dimensional filtering in discrete time. **20**, *4*, pp. 257–265
- Gabasov, R., Kirillova, F. M.*: Optimization of dynamical systems with identification of input perturbations. **20**, *3*, pp. 233–244
- Gabasov, R., Kirillova, F. M., Gaishun, P. V., Prischepova, S. V.*: Synthesis of optimal controls on nonexact measurements of output signals. **20**, *6*, pp. 409–427
- Haroutunian, E. A., Maroutian, R. Sh.*: (E, Δ) -achievable rates for multiple descriptions of random varying source. **20**, *2*, pp. 165–178
- Hulkó, G.*: Lumped input and distributed output systems at the control of distributed parameter systems. **20**, *2*, pp. 113–128
- Ishii, H., Menaldi, J-L., Zaremba, L.*: Viscosity solutions of the Bellman equation on an attainable set. **20**, *5*, pp. 317–328
- Korbicz, J., Podladchikov, V., Bidyuk, P.*: Suboptimal control algorithm for discrete systems. **20**, *4*, pp. 281–290
- Korovin, S. K., Nikitina, M. G., Nikitin, S. V.*: Infinite-dimensional systems: Approximate controllability and observability. Part I. **20**, *1*, pp. 59–76
- Korovin, S. K., Nikitina, M. G., Nikitin, S. V.*: Infinite-dimensional systems: Design of Sakawa controllers. Part II. **20**, *2*, pp. 97–111
- Kramosil, I.*: Definition and recognition of classical sets by the rough ones. **20**, *2*, pp. 77–95
- Krasovskii, A. A.*: Optimization and stochastic dynamics in the state space. **20**, *1*, pp. 45–57
- Kurzanski, A. B., Pschenichnyi, B. N., Pokotilo, V. G.*: Optimal inputs for guaranteed identification. **20**, *1*, pp. 12–23

- Lemos, J. M.*: Long-range adaptive control of ARMAX plants with accessible disturbances. **20**, 2, pp. 145-164
- Linder, T.*: On asymptotically optimal companding quantization. **20**, 6, pp. 475-484
- Lugosi, G.*: Pattern classification from distorted sample. **20**, 6, pp. 465-473
- Malanowski, K.*: Stability and sensitivity analysis of discrete optimal control problems. **20**, 3, pp. 187-200
- Martos, B.*: Viable control trajectories in linear systems. **20**, 4, pp. 267-280
- Mikleš, J., Mészáros, A.*: A decoupling pole-placement controller for a class of multivariable systems. **20**, 4, pp. 291-298
- Morvai, G.*: Empirical log-optimal portfolio selection. **20**, 6, pp. 453-463
- Nguyen Van Su*: Null-controllability of infinite-dimensional discrete-time system with restrained control. **20**, 3, pp. 215-232
- Otáhal, A.*: Parameter estimation for nearest neighbor Gaussian random fields in the plane. **20**, 6, pp. 429-439
- Papageorgiou, N. S.*: Relaxability and well-posedness for infinite dimensional optimal control problems. **20**, 3, pp. 201-214
- Papageorgiou, N. S.*: On the optimal control and relaxation of finite dimensional systems driven by maximal monotone differential inclusions. **20**, 4, pp. 245-255
- Rosinová, D.*: On decentralized stabilization of large-scale linear discrete systems. **20**, 5, pp. 329-339
- Shiryayev, V. I.*: Minimax filtering in real time of multistage systems. **20**, 5, pp. 309-316
- Smagina, Ye. M.*: A method of desinging of observable output ensuring given zeros location. **20**, 5, pp. 299-307
- Studniarski, M.*: The discrete maximum principle as a sufficient optimality condition. **20**, 3, pp. 179-186
- Taras'ev, A. M.*: The function of an optimal guaranteed result of control problems with a vector criterion. **20**, 1, pp. 25-36
- Timofeev, A. V.*: Non-asymptotic solution of confidence estimation parameter task of a non-linear regression by means of sequential analysis. **20**, 5, pp. 341-351
- Vaněček, A.*: Strongly nonlinear and other control systems. **20**, 1, pp. 3-12
- Vesely, V., Barč, V., Hindi, K. S.*: A decentralized control scheme for continuous-time systems through partial aggregation. **20**, 6, pp. 373-381

AUTHOR INDEX

- Barč, V. 20, 6, pp. 373-381
 Bidyuk, P. 20, 4, pp. 281-290
 Carbonez, A. 20, 6, pp. 441-451
 Chentsov, A. G. 20, 2, pp. 129-143
 Chernyak, A. I. 20, 1, pp. 37-44
 Emelyanov, S. V. 20, 5, pp. 353-371
 En-hui Yang 20, 6, pp. 397-408
 Faragó, A. 20, 6, pp. 383-395
 Ferrante, M. 20, 4, pp. 257-265
 Gabasov, R. 20, 3, pp. 233-244;
 20, 6, pp. 409-427
 Gaishun, P. V. 20, 6, pp. 409-427
 Györfi, L. 20, 6, pp. 441-451
 Haroutunian, E. A. 20, 2, pp. 165-178
 Hindi, K. S. 20, 6, pp. 373-381
 Hulkó, G. 20, 2, pp. 113-128
 Ishii, H. 20, 5, pp. 317-328
 Kirillova, F. M. 20, 3, pp. 233-244;
 20, 6, pp. 409-427
 Korbicz, J. 20, 4, pp. 281-290
 Korovin, S. K. 20, 1, pp. 59-76;
 20, 2, pp. 97-111; 20, 5, pp. 353-371
 Kramosil, I. 20, 2, pp. 77-95
 Krasovskii, A. A. 20, 1, pp. 45-57
 Kurzanski, A. B. 20, 1, pp. 12-23
 Lemos, J. M. 20, 2, pp. 145-164
 Linder, T. 20, 6, pp. 383-395;
 20, 6, pp. 475-484
 Lugosi, G. 20, 6, pp. 383-395;
 20, 6, pp. 465-473
 Malanowski, K. 20, 3, pp. 187-200
 Maroutian, R. Sh. 20, 2, pp. 165-178
 Martos, B. 20, 4, pp. 267-280
 Menaldi, J.-L. 20, 5, pp. 317-328
 Mészáros, A. 20, 4, pp. 291-298
 van der Meulen, E. C. 20, 6,
 pp. 441-451
 Mikleš, J. 20, 4, pp. 291-298
 Morvai, G. 20, 6, pp. 453-463
 Nguyen Van Su 20, 3, pp. 215-232
 Nikitin, S. V. 20, 1, pp. 59-76;
 20, 2, pp. 97-111
 Nikitina, M. G. 20, 1, pp. 59-76;
 20, 2, pp. 97-111
 Otáhal, A. 20, 6, pp. 429-439
 Papageorgiou, N. S. 20, 3, pp. 201-214;
 20, 4, pp. 245-255
 Podladchikov, V. 20, 4, pp. 281-290
 Pokotilo, V. G. 20, 1, pp. 12-23
 Prischepova, S. V. 20, 6, pp. 409-427
 Pschenichnyi, B. N. 20, 1, pp. 12-23
 Rosinová, D. 20, 5, pp. 329-339
 Shiryaev, V. I. 20, 5, pp. 309-316
 Smagina, Ye. M. 20, 5, pp. 299-307
 Studnarski, M. 20, 3, pp. 179-186
 Sztrik, J. 20, 1, pp. 37-44
 Taras'ev, A. M. 20, 1, pp. 25-36
 Timofeev, A. V. 20, 5, pp. 341-351
 Vaněček, A. 20, 1, pp. 3-12
 Veselý, V. 20, 6, pp. 373-381
 Zaremba, L. 20, 5, pp. 317-328
 Zhivoglyadov, P. V. 20, 5, pp. 353-371

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITOR

N. N. KRASOVSKII (USSR)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

K. V. FROLOV

A. B. KURZHANSKI

I. A. OVSEEVICH

E. D. TERYAEV

R. Z. KHASMINSKI

HUNGARY

T. VÁMOS

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

L. GYÖRFI

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJČ

I. VAJDA

РЕДАКТОР ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

К. В. ФРОЛОВ

А. Б. КУРЖАНСКИЙ

И. А. ОВСЕВИЧ

Е. Д. ТЕРЯЕВ

Р. З. ХАСЬМИНСКИЙ

ВНР

Т. ВАМОШ

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

Л. КЕВИЦКИ

Л. ДЪЕРФИ

Я. КОЧИШ

ЧССР

Й. БЕНЕШ

В. СТРЕЙЦ

И. ВАЙДА



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADEMIA
KÖNYVTÁRA

STRONGLY NONLINEAR AND OTHER CONTROL SYSTEMS

A. VANĚČEK

(Prague)

(Received December 22, 1989)

The control systems are divided into two classes: the strongly nonlinear systems and the other systems. Firstly, the control systems are embedded into dynamical systems with parameters. To analyse the nonlinear control systems, the invariant manifolds are used. The fractal dimension of the systems' attractors as opposed to the entire dimension is used for the systems division into two classes. As applications, vital and anti-vital goals of the control systems are stated.

1. Introduction

At a seminar Theory of Strongly Nonlinear Processes (Prague), it only gradually became clear what is the main topic of this seminar attended mainly by physicists — it was the turbulence. The turbulence is, mainly in engineering, considered as a negative phenomenon, as it makes noise and dangerous strain. The turbulence is a chaotic process with negative evaluation, at least in engineering. At the opposite, in the physics of atmosphere, the turbulence is considered as a positive phenomenon — it allows the mixing of air stratas. Only of minor importance is that its chaotic behaviour makes impossible, due to inexact knowledge of the initial conditions, the weather prediction for more than a few days. But the negative evaluation of chaos in general, based on its abasement of human pride in that it makes impossible the long term scientific prediction, in the evaluation of chaos entirely dominates. After several signs of the possible turns of the evaluation of chaos, the real change of the evaluation paradigma was caused by Goldberger [7, 22]. On the basis of the spectrum of ECG during a heart failure, Goldberger observed that there occurs some sort of pathological periodicity, the spectrum being concentrated to some narrow frequency band. On the opposite, Goldberger observed, on the basis of the spectrum of ECG of a healthy heart, the wide-band spectrum of the type $1/f$ — it is such what Mandelbrot connected with fractals.

The author's field of interest are the controlled systems — the dynamical systems with parameters, parametrized in such a way to behave properly, especially

to reach the given asymptotic goals. The author's interest in the synthesis of the chaos was awakened only by the new paradigm — the evaluation of the chaos as a positive phenomenon. The Dynamical Systems theorists are concerned only with the analysis. In the entire Control Theory, systems to be synthesized are always the ones with point attracting, equilibrium state. On the opposite, the newly proposed control synthesis should change the system parameters in such a way, that the equilibrium states are the bounded, persisting, non-periodic cycles. According to the new paradigm, these are connected with the 'health', at the difference to the periodic cycles connected, according to the new paradigm, with the 'illness', and at the difference to the point attractor connected with the 'death'.

2. Control systems as dynamical systems with parameters

DEFINITION 1. Dynamical systems with parameters are defined as systems of differential equations

$$dx/dt = f(x, K)$$

where the time $t \in \mathbb{R}$, the state vector $x \in \mathbb{R}^n$, the vector field $f \in \text{Lip}$, the parameter matrix K of the time invariant parameters, $dK/dt = 0$.

Fact. The control systems can be described as dynamical systems with parameters.

Method. The consistent usage of the state description of control systems.

Example 1. Linear control system with right state feedback:

$$dx/dt = (F + GK)x.$$

The parameter of this linear dynamical system (being linear in the state x) is matrix K . The control is introduced because

$$e^{Ft}$$

is not sufficiently damped and may be even unstable. The state feedback parameter K is introduced in such a way to make

$$e^{(F+GK)t}$$

properly and quickly damped.

Example 2. Nonlinear control system with both left and right state feedback:

$$\begin{bmatrix} dy/dt \\ dz/dt \end{bmatrix} = \begin{bmatrix} f(y) + g(K_r z) \\ f(z) + g(K_r z) + K_l(h(y) - h(z)) \end{bmatrix}$$

or

$$dx/dt = F(x, K)$$

where $F \in \text{Lip}$, $F(0,0) = 0$, $K = [K'K_r]$, $x_1 = y$, $x_2 = z$. Traditionally, we are introducing the control because

$$\lim_{N \rightarrow \infty} [I + F(\cdot, 0)t/N]^N$$

is not sufficiently damped and may be even unstable. The state feedback parameter K is introduced in such a way to make

$$\lim_{N \rightarrow \infty} [I + F(\cdot, K)t/N]^N$$

properly and quickly damped. (Here we have used the nonlinear response written in the closed form with the help of the limit form of the Euler integration formula, see Arnold [2].)

Note. The embedding of control systems into dynamical systems with parameters entails that for a description of the systems, the states are sufficient and the inputs and outputs are not needed for this purpose. So, we are leaving the State Theory according to Kalman and Zadeh which is a hybrid theory, mixing the internal and external description. (Here we are speaking about control in the narrow sense: in the case we need also to follow, we model the object to be followed by some other dynamical system as was observed early by Luenberger.) From the system theory we are eliminating its basic problem, i.e. the problem of the minimal realization of the internal description from idealized external description which is the prototype of the identification problem. The control system is originated by the connection of the controlled plant and the regulator. The controlled plant is originated from the connection of elementary blocks. The only measurements we need for such a modelling are scalar, static ones. This approach to the modelling is perhaps the only one which has been successfully tested in physics. According to Tonti [20], the fundament of this approach, i.e. the fundament of the internal modelling, is a cohomology of a cell complex. This cell complex we are always building from the boundary elements: the edges from the vertices, the faces from the edges, the space cubes from the faces, the time-space hypercubes from the cubes, and the cell complex from hypercubes. In such a way we obtain some nonlinear partial differential equations. For the dynamical systems, the cell complex is two-dimensional and the differential equations are the ordinary ones. For example for a dynamical electrical circuit, we are connecting such a circuit from *RLC* elements and the only measurements for modelling are the static scalar ones of usually linear *LC* elements and of generally nonlinear characteristic of *R*. Finally, let us note that according to Vinogradov [19] the conservation laws of physics are just the group cohomologies and that there exists the so-called cohomological physics (Stasheff [17]).

3. The invariant sets — The basis of control systems analysis

DEFINITION 2 [4]. The invariant set M of the dynamical system is a set of elements of which are the whole trajectories, such that M is the solution of the equation:

$$\varphi(M) = M, \quad t \in \mathbb{R}$$

where φ is the mapping of M by the trajectories.

Example 3. For $n = 2$ the stationary point, the trajectory, the plane filled by trajectories are the examples of invariant sets of dimensions 0, 1, 2, respectively.

Centre Manifold Theorem (Kelley [10]). Let us suppose the dynamical system

$$dx/dt = f(x), \quad x \in \mathbb{R}^n$$

where $f(x)$ of smoothness class C^r ($r \geq 2$) is zero at the origin: $f(0) = 0$, and its linear part is Fx . Then the invariant set of the dynamical system can be resolved into the three locally disjoint (up to the origin) manifolds

$$W^s, \quad W^a, \quad W^c$$

of smoothness classes C^r, C^r, C^{r-1} , respectively. The invariant set of the linearized dynamical system $dx/dt = Fx$ can be resolved into three linear spaces

$$T^s, \quad T^a, \quad T^c.$$

Spaces T^i ($i = s, a, c$) are the tangent spaces to the manifolds W^i in the origin. The manifold W^s is stable, the manifold W^a is anti-stable (i.e. stable for $t \rightarrow -\infty$). The asymptotic behaviour on the central manifold W^c is determined by the higher than the linear part Fx order part of $f(x)$.

Applications of the Centre Manifold Theorem. We shall apply the theorem for the control systems, limiting ourselves to the hyperbolic systems, i.e. systems without centre manifold. Their stability, unstability and anti-stability is locally determined by the eigenvalues of their linearization. Further, we shall suppose the semi-simplicity of the linearized system, i.e. the eigenvectors of the linearized system matrix F are forming the basis. We shall search for those initial segments of the n solutions of the $dx/dt = Fx + o(x)$ equation which are naturally connected with those n solutions of the $dx/dt = Fx$ equation which are the most elementary invariant sets of $dx/dt = Fx$, the eigenvectors of both F and e^{Ft} . Using the Picard integration method (for the method see e.g. [1]), we obtain the n curved segments: we shall call them the initial segments of the eigencurves or just eigencurves. The Picard iterative construction of the solution of nonlinear differential equation:

$$x_{k+1}(t) = x_k(t) + \int_{t_0}^t f(x_k(\tau)) d\tau$$

will give us the eigencurve v obeying the eigen-equation

$$f(\cdot)v = s(\cdot)v$$

which is the generalization of the classical eigen-equation $Fv = sv$. The situation is illustrated on the upper part of the figure. The field $s(\cdot)$ corresponding to the eigencurve v is the restriction of the vector field $f(\cdot)$ on v . The gradient of the eigencurve at the equilibrium point is the eigenvalue. Similarly, for the eigenplane on which all the oscillating solutions of the equation $dx/dt = Fx$ lay for a couple of complex conjugate eigenvalues. Now, for the Picard construction of the eigensurface, we need one parametrical set of solutions lying on the eigenplane as a set of initial iterations. At the end of iterations, the partial derivatives of the eigensurface in the origin are the real and imaginary parts of complex conjugate eigenvalues (see the lower part of the figure).

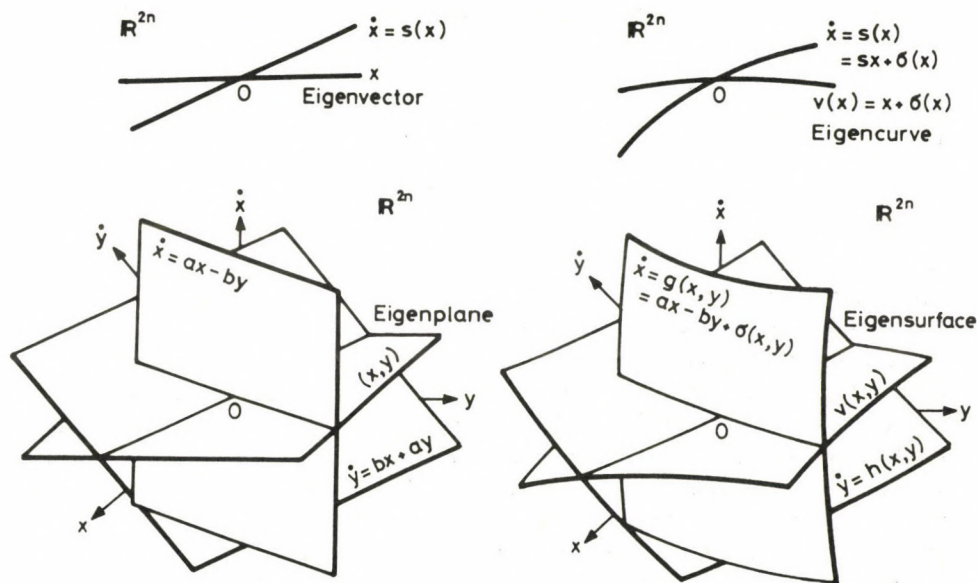


Fig. 1

From the Linear Control Theory (see e.g. Kailath [10]), we know that the fundamental conditions for the eigenstructure changeability are given by the conditions of Popov [15] and Moore [12]. Here we shall generalize them for nonlinear control systems. For the *right state feedback*:

$$dx/dt = f(x) + g(K(x)); \quad f, g \in \text{Lip}, \quad g(0) = 0,$$

we have the eigen-equation $[f(\cdot) + g(K(\cdot))]v = s(\cdot)v$. Equivalently, $[Is(\cdot) - f(\cdot)]v = g(K(\cdot))v$ and $v = [Is(\cdot) - f(\cdot)]^{-1}g(K(\cdot))v$. From the last equation, we have the *eigenfunction changeability from right*:

$$v \in \text{Im}[(Is - f)^{-1}g].$$

For the *left state feedback*:

$$dx/dt = f(x) + Lh(x); \quad f, h \in \text{Lip}, \quad h(0) = 0,$$

we have the eigen-equation $[f(\cdot) + Lh(\cdot)]v = s(\cdot)v$ from which follows the *right eigencurve changeability*:

$$v \notin \text{Ker } h.$$

We shall introduce the left eigencurve w of f as the right eigencurve of transposed f , i.e. of f' . So, we shall obtain the *eigenfunction changeability from left*:

$$w \in \text{Im}[(Is - f')^{-1}h']$$

and, finally, the *left eigencurve changeability*:

$$w \notin \text{Ker } g'.$$

For $f(x) = Fx$, $g(x) = Gx$, $h(x) = Hx$, $s(x) = sx$, $f(v) = Fv$, $f'(w) = F'w$, we have the Popov-Moore conditions.

Up to now, our results, based on the Centre Manifold Theorem, has been only of local nature — even if the vicinity of the origin may be very broad. Now, we shall sketch the globalization of the local results but again for hyperbolic systems (i.e. with no centre manifold). The common point of both the stable and anti-stable manifolds and of the stable and anti-stable spaces is the equilibrium state. The nonlinear systems have generally more than one equilibrium state, i.e. more than one solutions x of $dx/dt = 0$ or $f(x) = 0$. The global behaviour is determined by patching the neighbouring manifolds containing various equilibrium states. Invariant manifolds of nonlinear systems with more than one equilibrium states are the manifolds with boundaries. The nonlinear systems which will interest us most — the strongly nonlinear systems — will be the ones with several (at least two) equilibrium states.

4. The attractors — The basis of control systems goals

DEFINITION 3. The attractor of the dynamical system is such an invariant set of the system which is compact and stable, i.e. all trajectories from some vicinity

of the attractor converge to the attractor for $t \rightarrow +\infty$. The repeller of the system is such an invariant set which is anti-stable.

Example 4. The attractors of integer dimension D are the sink, the cycle, and the torus with $D = 0, 1, 2$, respectively.

Review of Lyapunov exponents and the fractal dimension of attractors [3]. The Lyapunov exponent LE is a generalization of the real part of the eigenvalue for non-stationary linear systems which was obtained as a time-mean of the eigenvalue. The non-stationary linear system $dx/dt = A(t)x$ we shall obtain as the linearization of the nonlinear system in the vicinity of its solution. The i -th solution $x_i(t)$ has i -th LE

$$\lambda_i = \lim_{T \rightarrow \infty} \sup \frac{1}{T} \ln |x_i(t)|$$

where $i = 1, \dots, n$. After the diagonalization, the i -th solution is

$$y_i(t) = \exp \left(\int_0^t d_i(t) dt \right)$$

and its LE is the real part of the mean value of the eigenvalue. To derive the dimension of the attractor using LE, let us integrate the original differential equation on the attractor and at the same time let us integrate its non-stationary linearization. For simplicity, let us assume $n = 3$ and the ordering of LE's: $\lambda_1 > 0$, $\lambda_2 \geq 0$, $\lambda_3 < 0$. In the vicinity of the trajectory of the attractor, let us introduce the cube with edge ε . In the proper coordinates, the i -th edge of the cube is in the mean evolving as $\varepsilon \exp(\lambda_i t)$. The number of the cubes with the edge $\varepsilon \exp(\lambda_3 t)$ which are needed to cover the attractor is

$$N(t) = \prod_{i=1}^3 \varepsilon \exp(\lambda_i t) / \varepsilon \exp(\lambda_3 t) = \exp((\lambda_1 + \lambda_2 - 2\lambda_3)t).$$

The Lyapunov dimension of the attractor is

$$D_L = - \lim_{t \rightarrow \infty} \frac{\ln N(t)}{\ln \exp(\lambda_3 t)} = 2 + \frac{\lambda_1 + \lambda_2}{|\lambda_3|}.$$

In fact, by the theorem of Haken [3], $\lambda_2 = 0$.

Example 6. The Lorentz attractor [1, 2, 16, 18], is an attractor of the set of three Lorentz bilinear differential equations for some fixed values of the three parameters. It is some non-periodic, bounded, non-vanishing trajectory with axis symmetry. Its Lyapunov exponents are $\lambda_1 = 1.37$, $\lambda_2 = 0$, $\lambda_3 = -22.37$. $D_L = 2 + (1.37 + 0)/22.37 = 2.06$. The repeller is ∞ .

Example 7. The double scroll attractor of Chua–Matsumoto–Komuro [3, 11] is an attractor of the three differential equations, the first with slight nonlinearity,

the second and the third one being linear. Again, it is some non-periodic, bounded, non-vanishing trajectory, now with centre-symmetry. Its Lyapunov exponents are $\lambda_1 = 0.23$, $\lambda_2 = 0$, and $\lambda_3 = -1.78$ and its dimension is $D_L = 2.13$.

Comment on Examples 6, 7. Both dynamical systems have three equilibrium points, each of these six equilibrium points are hyperbolic, and each six points have both stable and anti-stable eigenvalues repell, after some time, the trajectories from the vicinity of the equilibrium point, the stable eigenvalues attract, for some time, the trajectories in the vicinity of their equilibrium points. After that time, the anti-stable eigenvalues again repell, etc. Generalizing both examples, we write the

Scenario of the synthesis of control systems with fractal dimension of attractors or the chaotic systems or, by definition, the strongly nonlinear systems:

A. Introduce the state space \mathbb{R}^n with the co-dimension of the state trajectories at least 2, i.e. the state space of dimension at least $n = 3$.

B. Introduce the vector field of at least 2 equilibrium points.

C. Parametrize the vector field in the vicinity of each of the equilibrium points to obtain both stable eigenvalues and anti-stable eigenvalues of the linearized systems near each of the equilibrium points, i.e. make some generalized shift of eigenvalues — the shiftability conditions are the generalized Popov conditions.

D. If needed, use the parametrization, moreover, even to generalized eigenvectors adjustment — the adjustability conditions are the generalized Moore conditions.

Systems in which such parametrizations, leading to non-vanishing, persisting, and non-periodic attractor trajectories are impossible, we shall call weakly nonlinear systems, the special weakly nonlinear systems being the linear systems.

Note on the Scenario. Our Scenario is in concordance with the known analysis of the birth of bounded, non-vanishing and non-periodic trajectories. This analysis is based on homo- and heteroclinic trajectories and on the Smale horseshoe. The heteroclinic trajectory is a loop containing at least two equilibrium points. Homeoclinic trajectory is a loop containing just one equilibrium point; if it lays also in the vicinity of the other equilibrium state, it is near to our scenario. The Smale horseshoe is the special case of the Poincaré mapping which is based on the expansion (anti-stability), contraction (stability) and on the folding or the transition between the areas of two equilibrium points.

5. Conclusion

The positive properties of the Strongly Nonlinear Systems:

– Qualitatively higher possibilities of systems “far from equilibrium” (Prigogine); in our interpretation the systems with the trajectories in the areas of several equilibrium points.

– The making possible of mixing and in this way enlarging the capacity of absorption of the incoming thermal or kinetic energy (Ottino [13]) and with this connected functionality and adaptivity (Garfinkel [5]).

– The sensitivity on initial conditions and at the same time high structural stability (Paluš *et al.* [14]).

– Convergence to the attractor and at the same time divergence within the attractor.

– Healthy, at the difference to epileptic state of the brain (Haken [8], Freeman [6]).

– Healthy state of the heart (Goldberger [7]).

These properties motivate the synthesis of the control of strongly nonlinear systems as the stabilization on the chaotic or fractal attractor. At the difference to the much publicized analysis of specific cases and to the synthesis of weakly nonlinear and linear control systems, we do not know much more than we presented in our Scenario, and the germ of theory based on the generalization of linear control synthesis based on the generalized eigenstructure.

References

1. Abraham, R. H., Shaw, C. D., Dynamics — The Geometry of Behavior. Pt. 3: Global Behavior. Aerial Press, Santa Cruz, 1981.
2. Arnold, V. I., Ordinary Differential Equations (in Russian). Nauka, Moscow 1975. (Also in English translation.)
3. Chaotic Systems. Special Issue of Proceedings IEEE 75, No. 3, 1987.
4. Dynamical Systems 1. (In Russian.) VINITI, Moscow 1985. (Also in English translation.)
5. Garfinkel, A., The Virtues of Chaos. Beh. Brain Sci. 10, pp. 178–179, 1987.
6. Freeman, W. J., Strange Attractors that Govern Mammalian Brain Dynamics Shown by Trajectories of EEG Potential. IEEE Trans. CAS-35, pp. 781–783, 1988.
7. Goldberger, A.L., Nonlinear Dynamics, Fractals, Cardiac Physiology and Sudden Death. Temporal Disorder in Human Oscillatory Systems, Springer-Verlag, Berlin, 1987.
8. Haken, H., Synergetik — Selbstorganisationsvorgänge in Physik, Chemie und Biologie. A. v. Humboldt-Stiftung Mitteilungen No. 43, pp. 12–23, 1984.
9. Kailath, T., Linear Systems. Prentice Hall, Englewood Cliffs, 1980.
10. Kelley, A., The Stable, Center Stable, Center, Center-Unstable, and Unstable Manifold. J. Diff. Eq. 3, pp. 546–570, 1967. (Reprinted in: Abraham, R., Robbin, J., Transverse Mappings and Flows. Benjamin, New York, 1967.)
11. Matsumoto, T., Chua, L. O., Komuro, M., The Double Scroll. IEEE Trans. CAS-32, pp. 789–818, 1985.
12. Moore, B.C., On the Flexibility Offered by State Feedback in Multivariable Systems beyond Closed-Loop Eigenvalue Assignment. IEEE Trans. AC-21, pp. 689–692, 1976.
13. Ottino, J. M., The Mixing of Fluids. Sci. Amer. pp. 56–67, 1989.
14. Paluš, M., Dvořák, I., Šiška, J., Deterministic Chaos and Living Systems. Physics and Mathematics, pp. 98–113. (In Czech.) JČSMF, Prague 1987.

15. Popov, V. M., Hyperstability of Control Systems. (Russian translation from Roumanian.) Nauka, Moscow, 1970. (Also in English translation.)
16. Sparrow, C., The Lorentz Equations: Bifurcations, Chaos, and Strange Attractors. Springer-Verlag, New York, 1982.
17. Stasheff, J., Cohomological Physics. Algebraic Topology, Rational Homotopy. Lecture Notes Math. 1318, pp. 228-237. Springer-Verlag, Berlin, 1988.
18. Strange Attractors. (Russian translation from English.) Mir, Moscow, 1982.
19. Symmetries in Partial Differential Equations. Spec. Issue of Math. Applicandae, Nos 1, 2, 1989.
20. Tonti, E., The Algebraic-Topological Structure of Physical Theories. Proc. Symp. Symmetry, Similarity and Group Theoretic Methods in Mechanics, held at the University of Calgary, 1974.
21. Vaněček, A., Control Systems under Path Integrals. IFAC Congress, München, Preprints 9, pp. 22-27, Düsseldorf, 1987.
22. Why a Steady Heart May Not be Healthy? (Review of a message given by A. Goldberger of the Harvard Medical School to the American Association for the Advancement of Science.) New Scientist, 21 January 1989, p. 31.

Строго нелинейные и другие системы управления

А. ВАНЕЧЕК

(Прага)

Системы управления можно разбить на два класса: строго нелинейные и другие. Отправным пунктом является вложение систем управления в множество динамических систем с параметрами. Для анализа нелинейных систем применены инвариантные многообразия. Фрактальная размерность аттракторов системы в отличие от целочисленной размерности использована для классификации принадлежности систем к двум классам. В качестве примера применения формулируются задачи систем, целью которых является или выживание, или успокоение процесса.

Antonín Vaněček
Institute of Information Theory and Automation
Czechoslovak Academy of Sciences
CS 182 08 Prague 8, Czechoslovakia

OPTIMAL INPUTS FOR GUARANTEED IDENTIFICATION

A. B. KURZHANSKI, B. N. PSCHENICHNYI, V. G. POKOTILO

(Lazenburg, Kiev)

(Received December 19, 1989)

This paper deals with the problem of identifying a finite dimensional vector parameter on the basis of observations that are generated by an infinite dimensional input and corrupted by an unknown but bounded "noise". The specific problem solved here is one of selecting an optimal input that would ensure the smallest worst-case error for the identification procedure. This is taken as the diameter of the smallest ball that would contain the set of states consistent with the measurement and the given constraints on the unknowns. The paper continues the investigation of [1-8].

1. Assume the following notations: H stands for a Hilbert space, \mathbb{R}^n for the n -dimensional Euclidean space, the respective inner products for those spaces being $\langle \cdot, \cdot \rangle$ and (\cdot, \cdot) and the norms being $\|\cdot\|$ and $|\cdot|$.

The problem under discussion is as follows. Consider a system

$$y = \sum_{i=1}^m z_i a_i + \zeta \quad (1)$$

where

$$y, a_i, \zeta \in H, \quad z_i \in \mathbb{R} \quad (i = 1, \dots, m).$$

With y, a_i given, one is to identify the unknown vector $z = (z_1, \dots, z_m)$ under the restriction $\|\zeta\| \leq 1$.

Here y is the available measurement, a_i are the given inputs, ζ is the unknown but bounded disturbance. We further assume the elements a_i to be linearly independent.

Also denote H^m the Hilbert space of columns so that $x \in H^m$ if

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \quad x_i \in H.$$

If C is a matrix of dimension $k \times m$ with elements $C_{ij} \in H$, then Cx is a column with k elements

$$\left\{ \sum_{j=1}^m C_{ij} x_j \right\}, \quad i = 1, \dots, k,$$

so that

$$Cx \in H^k.$$

Let the asterisk indicate the transpose for a vector or a matrix. Then for $\psi_i \in \mathbb{R}$, $a_i \in H$ we will have $\psi_a^* = \sum_{i=1}^m \psi_i a_i$.

The operations on matrices whose elements belong to H are performed according to the standard rules of "ordinary" matrix calculations except that the products of the respective elements are taken as scalar products in H , e.g.

$$a^* a = \sum_{i=1}^m \langle a_i, a_i \rangle = \sum_{i=1}^m \|a_i\|^2$$

$$aa^* = \begin{pmatrix} \langle a_1, a_1 \rangle & \dots & \langle a_1, a_m \rangle \\ \dots & \dots & \dots \\ \langle a_m, a_1 \rangle & \dots & \langle a_m, a_m \rangle \end{pmatrix}.$$

Finally, assume

$$z = \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix} \in \mathbb{R}^m.$$

Formula (1) may now be rewritten as

$$y = z^* a + \zeta. \quad (2)$$

2. Given y, a , let us find the set of states of system (2) consistent with the constraint $\|\zeta\| \leq 1$:

$$z(y) = \{z : \exists \zeta \in H, \|\zeta\| \leq 1, y = z^* a + \zeta\}.$$

From (2) it follows that

$$\|\zeta\|^2 = yy - 2a^* z + z^* aa^* z \leq 1$$

or, taking,

$$p = Aq, \quad p = ay, \quad A = aa^*$$

that

$$(z - q)^* A(z - q) \leq 1 - h^2(y), \tag{3}$$

where $h^2(y) = yy - q^*p$ and, obviously, $0 \leq h^2(y) \leq 1$.

Inequality (3) describes an *ellipsoid* $E(q, A)$ in \mathbb{R}^n whose matrix A and center q depends upon the measurement y . The *diameter* of this ellipsoid is defined as twice the radius $r(y)$ of the smallest ball that includes it.

According to a well-known property of the eigenvalues of a positive definite quadratic form, we have, [9]

$$r(y) = (1 - h^2(y))\lambda^{-1}(a),$$

where $\lambda(a)$ is the *smallest eigenvalue* of the form

$$x'aa^*x.$$

It is clear that the diameter $d(y) = 2r(y)$ will be the largest iff $h^2(y) = 0$ which happens if and only if $y = 0$ (the "worst-case" realization).

Our objective will now be to select the input a in such a way that the "worst-case" diameter $d(0)$ would be as small as possible.

Hence, we are to minimize $\lambda^{-1}(a)$ — the inverse of the smallest eigenvalue $\lambda(a)$ of the matrix $A = aa^*$ of the ellipsoid

$$E(0, A) = \mathcal{E}(A) = \{x : x'Ax \leq 1\}$$

(the location of the center does not matter and may be taken to be the origin).

As $A = aa^*$ is invertible, the minimization of $\lambda^{-1}(a)$ is equivalent to the maximization of $\lambda(a)$. The procedure makes sense (the solution remains in H^m) once the admissible values of a are bounded by a certain set \mathcal{M} .

The problem to be discussed is, therefore, as follows: *specify an element $a \in \mathcal{M}$ such that $\lambda(a)$ would attain its maximal value.*

Remark 2.1. The center $q = A^{-1}ay$ could be presented as by where $b = A^{-1}a$ is a vector biorthogonal to a , i.e.

$$ab^* = ba^* = A^{-1}aa^* = I_m,$$

where I_m is an m -dimensional unit matrix.

3. According to the theory of necessary conditions of optimality let us first investigate the local behaviour of $\lambda(a)$ by calculating the directional derivative

$$\lambda'(a, \bar{a}) \equiv \lim_{\gamma \downarrow 0} \frac{\lambda(a + \gamma\bar{a}) - \lambda(a)}{\gamma}, \quad \bar{a} \in H^m.$$

Due to the extremal properties of the eigenvalues of A we have

$$\begin{aligned} \lambda(a) &= \min\{(\psi, A\psi) \mid |\psi| = 1\} = \\ &= \min\{(\psi^*a, \psi^*a) \mid |\psi| = 1\}. \end{aligned} \tag{4}$$

Denote

$$\Psi(a) = \{\psi \in \mathbb{R}^m : (\psi, A\psi) = \lambda(a), |\psi| = 1\}.$$

Clearly, $\Psi(a)$ is the set of normalized eigenvectors corresponding to the minimal eigenvalue $\lambda(a)$ of A .

Since

$$\begin{aligned} \frac{d}{d\gamma} \langle \psi^*(a + \gamma\bar{a}), \psi^*(a + \gamma\bar{a}) \rangle_{\gamma=0} &= \\ &= 2 \langle \psi^*\bar{a}, \psi^*a \rangle = 2a^*\psi \cdot \psi^*\bar{a}, \end{aligned}$$

it follows from [10] that

$$\lambda'(a, \bar{a}) = \min\{2a^*\psi \cdot \psi^*\bar{a} \mid \psi \in \Psi(a)\}. \quad (5)$$

Denote

$$\begin{aligned} \partial\lambda(a) &= \text{co} \{2a^*\psi\psi^* : \psi \in \Psi(a)\} = \\ &= 2a^*\text{co} \{\psi\psi^* : \psi \in \Psi(a)\} \end{aligned} \quad (6)$$

Relations (5), (6) yield

THEOREM 1. The following formula is true

$$\lambda'(a, \bar{a}) = \min\{w\bar{a} \mid w \in \partial\lambda(a)\}. \quad (7)$$

Let us discuss the latter relation in more detail.

According to the terminology of convex analysis [11, 12] the set $\partial\lambda(a)$ is defined as the *subdifferential* (of function $\lambda(a)$ at point a) and its elements as the respective *subgradients*. The finite dimensionality of $\partial\lambda(a)$ also implies that $\partial\lambda(a)$ is a convex compact set.

Following [11, 12] it is possible to indicate that if an $m \times m$ -dimensional matrix

$$\Gamma \in \text{co} \{\psi\psi^* : \psi \in \Psi(a)\},$$

then there exists an integer $k \leq m^2 + 1$ such that

$$\Gamma = \sum_{j=1}^k \gamma_j \psi_j \psi_j^*, \quad \psi_j \in \Psi(a); \quad j = 1, \dots, k, \quad (8)$$

$$\sum_{j=1}^k \gamma_j = 1; \quad \gamma_j \geq 0, \quad j = 1, \dots, k.$$

Therefore, all the elements of $\partial\lambda(a)$ turn to have the form $2a^*\Gamma$ where Γ is given by relation (8).

4. Let us now proceed with the necessary conditions of optimality for the basic problem which is to maximize $\lambda(a)$ under the restriction $a \in \mathcal{M}$. For doing this we will need the notion of *tangent cone* [12].

Recall that a *tangent cone* $K(a)$ to set \mathcal{M} at point a is a convex cone such that $\bar{a} \in K_{\mathcal{M}}(a)$ yields the existence of a function $\psi(\sigma) : [0, 1] \rightarrow H^m$ that ensures for a sufficiently small $\varepsilon > 0$ the inclusion

$$a + \sigma \bar{a} + \psi(\sigma) \in \mathcal{M}; \quad \sigma < \varepsilon;$$

and

$$\lim_{\sigma \rightarrow 0} \psi(\sigma) \sigma^{-1} = 0.$$

With M convex

$$K_{\mathcal{M}}(a) = \{\bar{a} \in H^m : \bar{a} = \gamma(w - a), \gamma > 0, w \in \mathcal{M}\}.$$

Denote $K_{\mathcal{M}}^*(a)$ to be the *adjoint cone* for $L_{\mathcal{M}}(a)$ so that

$$K_{\mathcal{M}}^* = \{w^* \in H^{m^*} : w^* \bar{a} \geq 0, \forall \bar{a} \in K_{\mathcal{M}}(a)\},$$

$$w^* = (w_1, \dots, w_m), \quad w_i \in H.$$

THEOREM 2. Once the element a delivers the maximum of function $\lambda(a)$ on the set \mathcal{M} there exists an array of values $\gamma_j > 0$, $j \leq k \leq m^2 + 1$ and normalized eigenvectors ψ_j of the matrix $A = aa^* : |\psi_j| = 1$, $A\psi_j = \lambda(a)\psi_j$, that

$$-a^* \Gamma \in K_{\mathcal{M}}^*(a), \quad \Gamma = \sum_{j=1}^k \gamma_j \psi_j \psi_j^*. \quad (9)$$

Proof. According to the theory of necessary conditions of optimality at point a , one must have [10, 12]

$$(-\partial\lambda(a)) \cap K_{\mathcal{M}}^*(a) \neq \emptyset.$$

But the elements of $\partial\lambda(a)$ are of the form $2a^* \Gamma$, the structure of Γ being defined by (8). As $K_{\mathcal{M}}^*(a)$ is a cone, its elements could be multiplied by any positive constant with the resulting element still in $K_{\mathcal{M}}^*(a)$. The multiplier 2 and the normalizing relation for the sum of γ_j 's being equal to unity may, therefore, be substituted by the requirement that $\gamma_j \geq 0$ for all $j = 1, \dots, k$.

Consider some specific properties of the matrix

$$\Gamma = \sum_{j=1}^k \gamma_j \psi_j \psi_j^*$$

that may facilitate the further analysis:

- (a) Matrix Γ is symmetric and positive definite. Indeed, once $W \in \mathbb{R}^m$ we have

$$(W, \Gamma W) = \sum_{j=1}^k \gamma_j W^* \psi_j \psi_j^* W = \sum_{j=1}^k \gamma_j (W^* \psi_j)^2 \geq 0.$$

- (b) For each column Γ_i , $i = 1, \dots, m$, of the matrix Γ ($\Gamma = \Gamma_1, \dots, \Gamma_m$) we have $A\Gamma_1 = \lambda(a)\Gamma_1$. By direct calculation

$$A\Gamma = \sum_{j=1}^k \gamma_j A\psi_j \psi_j^* = \lambda(a) \sum_{j=1}^k \gamma_j \psi_j \psi_j^* = \lambda(a)\Gamma$$

and further, due to the rules of matrix multiplication

$$A\Gamma = (A\Gamma_1, \dots, A\Gamma_m) = \lambda(a)(\Gamma_1, \dots, \Gamma_m)$$

which proves the assertion.

- (c) If there exists an eigenvector ψ such that

$$A\psi = \lambda\psi, \quad \lambda > \lambda(a),$$

then $\Gamma\psi = 0$ (matrix Γ is degenerate). Under the conditions of the above $(\psi_j, \psi) = 0$. Therefore,

$$\Gamma\psi = \sum_{j=1}^k \gamma_j \psi_j (\psi_j^*, \psi) = 0.$$

Let us now specify some particular cases.

5. Suppose

$$\mathcal{M} = \{a \in H^m : f(a) \leq 0\}$$

with $f(a) \equiv f(a_1, \dots, a_m)$ assumed to be a smooth function with a non-degenerate gradient

$$f'(a) = (f'_1(a), \dots, f'_m(a))$$

($f'_i(a)$ stands for the partial derivative of f in a_i).

As it is well known [12] in this case

$$K_m^*(a) = \{-\sigma f'(a) : \sigma \geq 0, \sigma f(a) = 0\}.$$

On the other hand, matrix Γ is non-zero as for example

$$(\psi_1, \Gamma\psi_1) = \sum_{j=1}^k \gamma_j (\psi_1^* \psi_j)^2 \geq \gamma_1 |\psi_1|^2 = \gamma_1 > 0.$$

Therefore, at least one of its vector columns is non-zero, for example, $\Gamma_1 \neq 0$. The necessary condition for the case under consideration yields

$$a^* \Gamma = \sigma f'(a), \quad \sigma \geq 0. \tag{10}$$

If $\sigma = 0$, then $a^* \Gamma = 0$, hence $a^* \Gamma_1 = 0$, $\Gamma_1 \neq 0$, i.e. the a_i 's are linearly dependent which contradicts the condition that $\lambda(a) > 0$.

We have just proved

Corollary 1. If $\mathcal{M} = \{a \in H^m : f(a) \leq 0\}$ then the maximizing point for $\lambda(a)$, $a \in \mathcal{M}$, satisfies the relations

$$a^* \Gamma = \sigma f'(a), \quad \sigma > 0, \quad f(a) = 0,$$

$$\Gamma = \sum_{j=1}^m \gamma_j \psi_j \psi_j^*, \quad \gamma_j > 0, \quad a a^* \psi_j = \lambda(a) \psi_j, \quad |\psi_j| = 1.$$

Particularly, if

$$f(a) = \sum_{i=1}^m \|a_i\|^2 - 1 = a^* a - 1$$

then

$$f'(a) = 2(a_1, \dots, a_m) = 2a^*$$

and the necessary condition yields

$$a^* \Gamma = 2\sigma a^*, \quad \sigma > 0, \quad a^* a = 1.$$

If matrix Γ would be degenerate we would have $\Gamma \psi = 0$ for a certain $\psi \in \mathbb{R}^m$, $|\psi| = 1$. Therefore,

$$a^* \Gamma \psi = 2\sigma a^* \psi = 0,$$

i.e. $a^* \psi = 0$, a_1, \dots, a_m would be linearly dependent and $\lambda(a) = 0$ which contradicts the maximality of $\lambda(a) > 0$. Matrix Γ is, therefore, non-degenerate.

From the representation (8) of matrix Γ it follows that it may be non-degenerate only if among the vectors ψ_j , $j = 1, \dots, k$, there exists a subset of m linearly independent vectors. In this case all of the latter eigenvectors of A would correspond to $\lambda(a)$. This is possible only if

$$A = a \cdot a^* = \lambda(a) I_m,$$

i.e.

$$(a_i, a_j) = 0, \quad i \neq j, \quad \|a_i\|^2 = \lambda(a).$$

Hence, the solution to the basic problem results in an array of orthogonal vectors a_i with equal norms.

Since

$$\sum_{i=1}^m \|a_i\|^2 = m\lambda(a) = 1,$$

we have

$$\lambda(a) = m^{-1}.$$

6. Consider a specific problem of controlling the observation process when

$$a \in H^m, \quad H = L^2[0, T].$$

The set \mathcal{M} is the set of solutions to the m -dimensional differential system

$$\dot{a} = Ca + Bu, \quad t \in [0, T], \quad a[0] = a_0, \quad (11)$$

with control $u(t)$ selected from a convex set U of functions that ensure the existence of solutions to (11).

On the interval $[0, T]$ we are, therefore, considering the measured signal

$$y(t) = a^*(t)z + \zeta(t), \quad \zeta(\cdot) \in L^2[0, T],$$

$$\int_0^T \zeta^2(t) dt \leq 1.$$

The optimal control problem now consists in the selection of a control $u(\cdot) \in U$ that would maximize the minimal eigenvalue of the matrix A with elements

$$\int_0^T a_i(t)a_j(t) dt.$$

Once $u_0(t)$ is the optimal control and $a^0(t)$ the respective solution to system (11), the adjoint cone would be determined as

$$K_M^*(\cdot) = \left\{ \psi^*(\cdot) : \int_0^T \psi^*(t)(a(t) - a^0(t)) dt \geq 0 \right\}, \quad (12)$$

where the inequality should be fulfilled for all the solutions $a(t)$ to equation (11) generated by all the controls $u(\cdot) \in U$.

Moreover,

$$\psi(t) = \begin{pmatrix} \psi_1(t) \\ \vdots \\ \psi_m(t) \end{pmatrix}, \quad \psi_i(\cdot) \in L^2[0, T]; \quad i = 1, \dots, m.$$

Since

$$a(t) = (\exp Ct)a(0) + \int_0^t (\exp C(t-\tau))Bu(\tau)d\tau,$$

this may be substituted into the inequality which yields (12). After an obvious calculation this yields

$$\int_0^T \left(\int_{\tau}^T \psi^*(\sigma)(\exp C(t-\tau)) dt \right) B(u(\tau) - u_0(\tau))d\tau \geq 0, \quad (13)$$

$$u(\tau) \in U.$$

Denoting

$$\psi^*(\tau) = - \int_{\tau}^T \psi^*(t)(\exp C(t-\tau)) dt \quad (14)$$

we come to

THEOREM 3. The inclusion $\psi^*(\cdot) \in K_M^*(a^0(\cdot))$ holds if and only if the inequality

$$\int_0^T \psi^*(\tau)B(u(\tau) - u_0(\tau))d\tau \leq 0 \quad (15)$$

is true for any $u(\cdot) \in U$.

Passing to the necessary conditions of optimality we have to check the condition of Theorem 2 which is

$$-a^{0*}(t)\Gamma = \psi^*(t), \quad \psi^*(\cdot) \in K_M^*(a^0(\cdot)).$$

Combining this with (14) we come to the relation

$$\psi^*(\tau) = \int_{\tau}^T a^{0*}(t)\Gamma(\exp C(t-\tau)) dt \quad (16)$$

which should be coupled with inequality (15).

The principal result now sounds as follows.

THEOREM 4. In order that the control $u \in U$ and the respective trajectory $a_0(t)$, $t \in [0, T]$ would determine the maximum for the minimal eigenvalue of the matrix

$$A = \left\{ \int_0^T a_i^0(t)a_j^0(t) dt \right\}$$

it is necessary that one could indicate such numbers $\gamma_j > 0$ and such eigenvectors ψ_j of the matrix A ($i = 1, \dots, k$) that the following relations would be true:

1. $\dot{a}^0(t) = Ca^0(t) + Bu_0(t)$, $t \in [0, T]$,
2. $\dot{\psi}^*(\tau) = -a^*(\tau)\Gamma - \psi^*(\tau)C$, $\tau \in [0, T]$, $\psi^*(T) = 0$,
3. $\Gamma = \sum_{j=1}^k \gamma_j \psi_j$,
4. $\int_0^T \psi^*(t)Bu(t) dt \leq \int_0^T \psi^*(t)Bu_0(t) dt$; $u(\cdot) \in U$.

The proof follows from above having in view that relation (2) is obtained by a direct differentiation of (16) in τ .

References

1. Krasovskii, N. N., On the theory of controllability and observability of linear dynamic systems. Prikl. Mat. Mech. **28**, 1 (1964) (in Russian).
2. Kurzhanski, A. B., Control and Observation Under Uncertainty. Nauka, Moscow, 1977.
3. Pschenichnyi, B. N., Pokotilo, V. G., The Minmax Approach to the Estimation of the Parameters of Linear Regression. Izv. Akad. Nauk SSR, Tech. Cybernetics, **2** (1983) (translated as "Engineering Cybernetics").
4. Witsenhausen, H. S., Sets of possible states of linear systems given perturbed observations. IEEE Trans. Automat. Control **AC-3** (1968).
5. Schweppe, F. C., Uncertain Dynamic Systems. Prentice Hall, 1973.
6. Koscheev, A. S., Kurzhanskii, A. B., Adaptive estimation of multistage systems under uncertainty. Izv. Akad. Nauk SSR, Tech. Cybernetics, **2** (1983) (translated as "Engineering Cybernetics").
7. Fogel, E., System identification via membership set constraints with energy constrained noise. IEEE Trans. Automat. Control **AC-24** (1979).
8. Kurzhanskii, A. B., Identification — a theory of guaranteed estimates. In: "From Data to Model", ed. J.C. Willems, Springer-Verlag, 1989.
9. Gantmacher, F. R., The Theory of Matrices. Nauka, Moscow, 2nd edition, 1986.
10. Pschenichnyi, B. N., The Necessary Conditions of Extremum. Becker, New York, 1971.
11. Rockafellar, R. T., Convex Analysis. Princeton University Press, 1970.
12. Pschenichnyi, B. N., Convex Analysis and Extremal Problems. Nauka, Moscow, 1980.

Оптимальные входы в задаче гарантированной идентификации

А. Б. КУРЖАНСКИЙ, Б. Н. ПШЕНИЧНЫЙ, В. Г. ПОКОТИЛО

(Лаксенбург, Киев)

В данной статье рассматривается задача об идентификации конечномерного векторного параметра на основе наблюдений, порожденных бесконечномерным входом, в

условиях неопределенных помех. Предполагается, что информация о помехах исчерпывается заданием априорного ограничения на их реализации. Специфика задачи, изложенной в данной работе, состоит в выборе оптимального входа, который бы обеспечил наименьшее значение гарантированной ошибки процесса идентификации. При этом последняя определяется как диаметр наименьшего шара, содержащего область идентифицируемых параметров, совместимых с результатами наблюдений. Работа продолжает исследования [1-8].

THE FUNCTION OF AN OPTIMAL GUARANTEED RESULT OF CONTROL PROBLEMS WITH A VECTOR CRITERION

A. M. TARAS'EV

(*Sverdlovsk*)

(Received January 24, 1990)

A control system whose dynamics is subject to uncertain disturbances is considered. Quality of trajectories of the system is evaluated by a terminal vector criterion. The vector multi-valued function of optimal guaranteed result of the given problem is defined. Properties of this function are examined. Necessary and sufficient conditions for a vector multi-valued function to be the function of the optimal guaranteed result are given.

1. Introduction

In this paper we consider a control system whose dynamics is described by an ordinary differential equation. It is supposed that the right-hand side of the system depends not only on the control but also on uncontrollable disturbances. The control is formed as a function of position. On the motions of the system a vector functional is defined. Quality of the control is evaluated by the vector which componentwise majorizes the values of the vector functional calculated on the motions corresponding to this control and an arbitrary disturbance. This estimating vector is guaranteed by the considered control. Therefore, it is called the guaranteed result.

Such a problem statement can arise in applications when quality of the process is evaluated by several criteria. Besides, each criterion is important for evaluation and should not be worsened at the expense of improving of others. For example, in the aircraft landing problem the role of such criteria can be assigned to the lateral and vertical deviations from the glide path, the lateral and vertical components of the velocity vector at the moment of landing, etc.

The optimal guaranteed result is defined as the set of all Pareto minimums among guaranteed results. The multi-valued function that associates with initial positions the corresponding optimal guaranteed result is called the function of optimal guaranteed result (FOGR). Properties of the vector multi-valued FOGR are analysed in the present paper. Investigations are developed within the framework

of the approach proposed in [1, 2]. The properties of stability are formulated for the FOGR. The infinitesimal form of the stability properties is studied. We formulate the necessary and sufficient conditions for a vector multi-valued function to be the FOGR.

The definition of the FOGR accepted in the present paper is similar to the definition of the optimal guaranteed estimate proposed in [3]. The above mentioned paper deals with multi-criteria problems of guaranteed control within the framework of the first direct Pontryagin method.

It should be mentioned that there are other approaches to the analysis of multi-criteria control problems under indeterminacy. The definition of equilibrium as generalization of the concepts of Pareto optimality and Nash equilibrium lies on the basis of one of them [4]. Note that in the present paper, in contrast to [4], we consider the control problem in which the optimal guaranteed result is ensured by one participant independently of a disturbance realization. In [4], on the contrary, all participants are equivalent. This is expressed in the symmetric definition of equilibrium for them.

2. Statement of the problem, main definitions

We consider the control system whose dynamics is described by the general differential equation

$$\dot{x} = f(t, x, u, v). \quad (1)$$

Here $t \in [t_0, \vartheta] = T$, $x \in \mathbb{R}^n$ is an n -dimensional phase vector, $u \in P \subset \mathbb{R}^p$ is a p -dimensional vector of control from the compact set P , $v \in Q \subset \mathbb{R}^q$ is a q -dimensional vector of disturbances from the compact set Q .

It is supposed that the function $f : T \times \mathbb{R}^n \times P \times Q \rightarrow \mathbb{R}^n$ is continuous of all arguments, locally Lipschitz continuous with respect to x , and satisfies the condition of extendability of solutions of system (1).

Assume that the vector cost functional is determined on the motions $x(\cdot)$ of system (1) by the relation

$$J(x(\cdot)) = \sigma(x(\vartheta)) = (\sigma_1(x(\vartheta)), \dots, \sigma_m(x(\vartheta))). \quad (2)$$

Here $\sigma_i : \mathbb{R}^n \rightarrow \mathbb{R}^1$, $i = 1, \dots, m$ are continuous functions.

Let us fix an arbitrary position $(t_*, x_*) \in T \times \mathbb{R}^n$. Let $U = U(\cdot) : T \rightarrow \mathbb{R}^n \rightarrow P$ be a positional control strategy. The set of the motions generated by the strategy U from the position (t_*, x_*) is defined as the set of all limits of the Euler splines constructed by virtue of U (see [1], p. 32). This set is a compactum in the space of continuous functions. Denote it by the symbol $X(t_*, x_*, U)$.

The aim of control is to "minimize" the cost functional (2) in the sense stated below.

Let us introduce an order ratio on m -dimensional vectors. For $a = (a_1, \dots, a_m)$, $b = (b_1, \dots, b_m)$ we shall assume

$$\begin{aligned} a &\leq b \text{ if } a_i \leq b_i, \quad i = 1, \dots, m, \\ a &\not\leq b \text{ if for some } j \in \{1, \dots, m\}, \quad a_j > b_j. \end{aligned}$$

Let us pass to the definition of the vector guaranteed result. Now, we shall introduce notations. For a position $(t_*, x_*) \in T \times \mathbb{R}^n$ and strategy U assume

$$\Sigma(t_*, x_*, U) = \{s \in \mathbb{R}^n : s = \sigma(x(\vartheta)), x(\cdot) \in X(t_*, x_*, U)\},$$

$$\Sigma_{\max}(t_*, x_*, U) = \{s^0 \in \mathbb{R}^m : s \leq s^0 \text{ for all } s \in \Sigma(t_*, x_*, U)\},$$

$$\Sigma_{\max}(t_*, x_*) = \bigcup_U \Sigma_{\max}(t_*, x_*, U),$$

$$\Sigma_{\max} = \{(t, x, s) \in T \times \mathbb{R}^n \times \mathbb{R}^m : s \in \Sigma_{\max}(t, x)\}. \quad (3)$$

Thus, $\Sigma_{\max}(t_*, x_*, U)$ is the set of vector results guaranteed by a strategy U at a position (t_*, x_*) by all components simultaneously. The set $\Sigma_{\max}(t_*, x_*)$ is the collection of all guaranteed results at a position (t_*, x_*) .

DEFINITION 1. The set of Pareto minimums from the set of guaranteed results $\Sigma_{\max}(t_*, x_*)$ we shall call the optimal guaranteed result, i.e.

$$c(t_*, x_*) = \{s^0 \in \Sigma_{\max}(t_*, x_*) : s \not\leq s^0 \text{ for all } s \in \Sigma_{\max}(t_*, x_*) \setminus \{s^0\}\}.$$

DEFINITION 2. The multi-valued mapping $(t, x) \rightarrow c(t, x)$ that associates with initial position $(t, x) \in T \times \mathbb{R}^n$ the corresponding optimal guaranteed result $c(t, x)$ is called the function of the optimal guaranteed result (FOGR).

3. Properties of the FOGR

With a view to investigate the properties of FOGR we shall consider the auxiliary augmented system

$$\begin{aligned} \dot{x} &= f(t, x, u, v), \quad t \in T, \quad x \in \mathbb{R}^n, \quad u \in P, \quad v \in Q \\ \dot{s}_1 &= 0 \\ &\vdots \\ \dot{s}_m &= 0. \end{aligned} \quad (4)$$

We pose for it the problem of pursuit with the target set

$$M = \{(x, s) \in \mathbb{R}^n \times \mathbb{R}^m : \sigma(x) \leq s\} \quad (5)$$

at the given instant ϑ .

In the followings, constructions of the positional differential games theory [1] are used.

Let $(t_*, x_*, s_*) \in T \times \mathbb{R}^n \times \mathbb{R}^m$, $U_r : T \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow P$ be a positional strategy of control for system (4), $Y(t_*, x_*, s_*, U_r)$ the set of the motions constructed by passing to the limit from the Euler splines of system (4) corresponding to U_r . The aim of the control U_r is to lead the motions of the corresponding set $Y(t_*, x_*, s_*, U_r)$ to the target set M at the instant ϑ .

In the auxiliary differential game arbitrary information may be used by forming the vector of disturbances v , i.e. it may turn out to be very unfavourable. For the theorem on alternative to be true ([1], p. 367) it is sufficient to assume that the vector of disturbances is formed as the function of positions $(t, x) \in T \times \mathbb{R}^n$ and vectors of control $u \in P$. Such a function $V_u : T \times \mathbb{R}^n \times P \rightarrow Q$ is called a counter-strategy. The set of motions generated by a counter-strategy V_u is defined by passing to the limit from the Euler splines constructed according to V_u . This set is a compactum in the space of continuous functions. Let $(t_*, x_*, s_*) \in T \times \mathbb{R}^n \times \mathbb{R}^m$, $V_u^r : T \times \mathbb{R}^n \times \mathbb{R}^m \times P \rightarrow Q$ be some counter-strategy in system (4), $Y(t_*, x_*, s_*, V_u^r)$ be the set of motions constructed by passing to the limit from the Euler splines of system (4).

Remark 1. The specific character of system (4) implies that its last m coordinates do not change and are identically equal to the vector s_* . From here it follows, firstly, that the positional strategy $U_r : T \times \mathbb{R}^n \times \mathbb{R}^m \rightarrow P$ really depends on the first $n + 1$ coordinates $(t, x) \in T \times \mathbb{R}^n$ only, and the counter-strategy $V_u^r : T \times \mathbb{R}^n \times \mathbb{R}^m \times P \rightarrow Q$ depends only on positions $(t, x) \in T \times \mathbb{R}^n$ and vectors of control $u \in P$. Therefore, we may assume $U_r \equiv U : T \times \mathbb{R}^n \rightarrow P$, $V_u^r \equiv V_u : T \times \mathbb{R}^n \times P \rightarrow Q$. Secondly, the following relations are true

$$Y(t_*, x_*, s_*, U_r) = Y(t_*, x_*, s_*, U) = \{(x(\cdot), s_*) : x(\cdot) \in X(t_*, x_*, U)\}, \quad (6)$$

where $X(t_*, x_*, U)$ is the set of motions of system (1) corresponding to the positional strategy U

$$Y(t_*, x_*, s_*, V_u^r) = Y(t_*, x_*, s_*, V_u) = \{(x(\cdot), s_*) : x(\cdot) \in X(t_*, x_*, V_u)\}, \quad (7)$$

where $X(t_*, x_*, V_u)$ is the set of motions of system (1) corresponding to the counter-strategy V_u .

The theorem on alternative asserts that for problem (4), (5) there exists a closed set $W_u \subset T \times \mathbb{R}^n \times \mathbb{R}^m$ called the positional absorption set (the maximal u -stable bridge) with the following properties. If a position (t_*, x_*, s_*) belongs to

W_u then there exists a positional control strategy $U_r \equiv U : T \times \mathbb{R}^n \rightarrow P$ such that for any motion $y(\cdot) \in Y(t_*, x_*, s_*, U_r)$ the inclusion $y(\vartheta) \in M$ is true. If a position (t_*, x_*, s_*) does not belong to W_u then there exists a counter-strategy $V_u^r \equiv V_u : T \times \mathbb{R}^n \times P \rightarrow Q$ such that for any motion $y(\cdot) \in Y(t_*, x_*, s_*, V_u^r)$ the relation $y(\vartheta) \notin M$ is true.

The following statement may be proved with the help of the theorem on alternative.

Lemma 1. The set of vectors Σ_{\max} determined by (3) and the positional absorption set W_u of the augmented problem (4), (5) coincide.

The theorem on alternative, Lemma 1 and Remark 1 imply the following result.

THEOREM 1. For the FOGR $(t, x) \rightarrow c(t, x)$ at any position $(t_*, x_*) \in T \times \mathbb{R}^n$ alternative takes place

1) for any optimal guaranteed vector $s^0 \in c(t_*, x_*)$ there exists a positional strategy $U : T \times \mathbb{R}^n \rightarrow P$ such that the vector inequality $\sigma(x(\vartheta)) \leq s^0$ is true for all motions $x(\cdot) \in X(t_*, x_*, U)$;

2) for all vectors s which are not guaranteed at a position (t_*, x_*) (i.e. $s^0 \not\leq s$ for all $s^0 \in c(t_*, x_*)$) there exists a counter-strategy $V_u : T \times \mathbb{R}^n \times P \rightarrow Q$ such that $\sigma(x(\vartheta)) \not\leq s$ for all $x(\cdot) \in X(t_*, x_*, V_u)$.

Let us formulate some properties of the vector multi-valued FOGR. Proofs of these properties are not complicated, therefore, will be omitted here.

Property 1. For any position $(t_*, x_*) \in T \times \mathbb{R}^n$ the set $\Sigma_{\max}(t_*, x_*)$ is completely determined by its Pareto points, namely, by the vector values of the FOGR $c(t_*, x_*)$, and the following equality takes place

$$\Sigma_{\max}(t_*, x_*) = \{s \in \mathbb{R}^m : s^0 \leq s, s^0 \in c(t_*, x_*)\}. \quad (8)$$

Property 2. Let $\omega_i : T \times \mathbb{R}^n \rightarrow \mathbb{R}^1$ be the function of the optimal guaranteed result of the control problem for system (1) with the scalar criterion

$$J(x(\cdot)) = \sigma_i(x(\vartheta)), \quad i = 1, \dots, m. \quad (9)$$

Here functions $\sigma_i : \mathbb{R}^n \rightarrow \mathbb{R}^1$, $i = 1, \dots, m$ are the components of the vector function $\sigma : \mathbb{R}^n \rightarrow \mathbb{R}^m$ from the functional (2).

Then, for any position $(t_*, x_*) \in T \times \mathbb{R}^n$, and vector $s \in c(t_*, x_*)$ the vector inequality $\omega_0(t_*, x_*) \leq s$ is true. Here $\omega_0(t_*, x_*) = (\omega_1(t_*, x_*), \dots, \omega_m(t_*, x_*))$.

Property 3. For any position $(t_*, x_*) \in T \times \mathbb{R}^n$ and index $i \in \{1, \dots, m\}$ there exist numbers s_j^0 , $j = 1, \dots, m$, $j \neq i$ such that $s^0 = (s_1^0, \dots, s_{i-1}^0, \omega_i(t_*, x_*), s_{i+1}^0, \dots, s_m^0)$ is a vector of the optimal guaranteed result, i.e. $s^0 \in c(t_*, x_*)$.

Property 4. For all positions $(t_*, x_*) \in T \times \mathbb{R}^n$ the optimal guaranteed result $c(t_*, x_*)$ is a bounded set.

Property 5. The epigraph $\text{epi } c = \{(t, x, s) \in T \times \mathbb{R}^n \times \mathbb{R}^m : s^0 \leq s, s^0 \in c(t, x)\} = \Sigma_{\max}$ of the FOGR $(t, x) \rightarrow c(t, x)$ is a closed set. In this sense the

function $(t, x) \rightarrow c(t, x)$ is lower semi-continuous, i.e. for any position $(t_*, x_*) \in T \times \mathbb{R}^n$ and sequences $\{(t_k, x_k) \in T \times \mathbb{R}^n\}$, $\{s_k \in c(t_k, x_k)\}$, $(t_k, x_k) \rightarrow (t_*, x_*)$, $s_k \rightarrow s_*$, for $k \rightarrow \infty$ there exists a vector $s^0 \in c(t_*, x_*)$ satisfying the vector inequality $s_0 \leq s_*$.

Remark 2. It follows from lemma 1 that for the construction of the vector multi-valued FOGR $(t, x) \rightarrow c(t, x)$ the algorithms and programs destined for solving guaranteed control problems of the form (4), (5) may be used. Such algorithms and programs, for example [5], have been developed in the Dynamic System Department of the Institute of Mathematics and Mechanics of the Urals Branch of the USSR Academy of Sciences.

The property of u -stability is one of the basic properties of the positional absorption set W_u . This property is the base of step-by-step back procedures for constructing the set W_u . If the program absorption set has the property of u -stability then it coincides with the positional absorption set. It is found that the property of u -stability may be formulated for the vector multi-valued functions including the FOGR $(t, x) \rightarrow c(t, x)$.

For this purpose we shall introduce the following notations. Denote by SC the class of vector multi-valued functions $(t, x) \rightarrow \omega(t, x)$ satisfying the following conditions:

- a) for all $(t, x) \in T \times \mathbb{R}^n$ the set $\omega(t, x) \subset \mathbb{R}^m$ is bounded;
- b) for all $(t, x) \in T \times \mathbb{R}^n$ the set $\omega(t, x)$ has the property of regularity: $s^{(1)} \not\leq s^{(2)}$ for $s^{(1)} \neq s^{(2)}$, $s^{(i)} \in \omega(t, x)$, $i = 1, 2$;
- c) the epigraph $W = \text{epi } \omega = \{(t, x, s) \in T \times \mathbb{R}^n \times \mathbb{R}^m : s^0 \leq s, s^0 \in \omega(t, x)\}$ is a closed set, i.e. a function $(t, x) \rightarrow \omega(t, x)$ is lower semi-continuous.

Note that the FOGR $(t, x) \rightarrow c(t, x)$ belongs to the class SC , i.e. $c \in SC$.

Suppose that the set $W = \text{epi } \omega \subset T \times \mathbb{R}^n \times \mathbb{R}^m$ has the property of u -stability for the augmented system (4). Let us remind the formulation of the property of u -stability for the set W : for any position $(t_*, x_*, s_*) \in W$ ($t_* < \vartheta$), moment $t \in (t_*, \vartheta]$, and unit vector $l \in S = \{r \in \mathbb{R}^n : \|r\| = 1\}$ there exists a solution $(x(\cdot), s(\cdot))$ of the differential inclusion

$$\begin{aligned} \dot{x}(\tau) &\in F(\tau, x(\tau), l) \\ \dot{s}(\tau) &= 0 \\ x(t_*) &= x_*, \quad s(t_*) = s_*, \quad \tau \in [t_*, t] \end{aligned} \tag{10}$$

such that $(t, x(t), s(t)) \in W$.

Here

$$F(\tau, y, l) = \Pi(\tau, y, l) \cap G(\tau, y), \tag{11}$$

$$\begin{aligned} G(\tau, y) &= \text{co}\{f \in \mathbb{R}^n : f = f(\tau, y, u, v), u \in P, v \in Q\}, \\ \Pi(\tau, y, l) &= \{r \in \mathbb{R}^n : \langle l, r \rangle \geq H(\tau, y, l)\}, \\ H(\tau, y, l) &= \min_{u \in P} \max_{v \in Q} \langle l, f(\tau, y, u, v) \rangle, \quad (\tau, y, l) \in T \times \mathbb{R}^n \times S. \end{aligned} \tag{12}$$

Taking into account that $s(\tau) \equiv s_*$ when $\tau \in [t_*, \vartheta]$ it is possible to express the property of u -stability of the set W in terms of a vector multi-valued function $(t, x) \rightarrow \omega(t, x)$, $\omega \in SC$.

DEFINITION 3. We shall say that a function $(t, x) \rightarrow \omega(t, x) : T \times \mathbb{R}^n \rightarrow 2^{\mathbb{R}^m}$, $\omega \in SC$ has the property of u -stability if, for any position $(t_*, x_*) \in T \times \mathbb{R}^n$ ($t_* < \vartheta$), vector $s_* \in \omega(t_*, x_*)$, moment $t \in (t_*, \vartheta]$, and vector $l \in S$ there exist a solution $x(\cdot)$ of the differential inclusion

$$\dot{x}(\tau) \in F(\tau, x(\tau), l), \quad x(t_*) = x_*, \quad \tau \in [t_*, t]$$

and a vector $s \in \omega(t, x(t))$ such that $s \leq s_*$.

Remark 3. Since the positional absorption set $W_u = \Sigma_{\max}$ of the auxiliary problem (4), (5) is the epigraph of the FOGR $(t, x) \rightarrow c(t, x)$, $c \in SC$ and the set W_u is a u -stable bridge (the maximal relative to the inclusion u -stable bridge) then the FOGR $(t, x) \rightarrow c(t, x)$ is a u -stable function in the sense of Definition 3.

4. Infinitesimal constructions

Stability properties may be defined by different equivalent ways. The infinitesimal form [6-8] is convenient for the property of u -stability. Infinitesimal constructions may be used also for the definition of the property of u -stability of vector multi-valued functions $\omega \in SC$.

Let $\bar{\mathbb{R}} = \mathbb{R}^1 \cup \{+\infty\} \cup \{-\infty\}$, $T^0 = [t_0, \vartheta]$.

Define the lower derivative of a vector multi-valued function $\omega \in SC$ at a position (t_*, x_*, s_*) ($(t_*, x_*) \in T^0 \times \mathbb{R}^n$, $s_* \in \omega(t_*, x_*)$) in a given direction $(1, h)$, $h \in \mathbb{R}^n$.

Let

$$D\omega(t_*, x_*, s_*) = \left\{ (h, d) \in \mathbb{R}^n \times \bar{\mathbb{R}}^m : h = \lim_{k \rightarrow \infty} \frac{x_k - x_*}{t_k - t_*}, \right. \\ \left. d = \lim_{k \rightarrow \infty} \frac{s_k - s_*}{t_k - t_*}, t_k \rightarrow t_*, t_k \in (t_*, \vartheta], \right. \\ \left. x_k \in \mathbb{R}^n, s_k \in \omega(t_k, x_k) \right\},$$

$$\partial\omega(t_*, x_*, s_*)|(h) = \{d \in (\bar{\mathbb{R}})^m : (h, d) \in D\omega(t_*, x_*, s_*)\}.$$

Remark 4. For all $(t_*, x_*) \in T^0 \times \mathbb{R}^n$, $s_* \in \omega(t_*, x_*)$, $h \in \mathbb{R}^n$ the sets $D\omega(t_*, x_*, s_*)$, $\partial\omega(t_*, x_*, s_*)|(h)$ are closed.

DEFINITION 4. The set of all Pareto minimal points from the set $\partial\omega(t_*, x_*, s_*)|(h)$ is called the lower derivative $\partial_- \omega(t_*, x_*, s_*)|(h)$ of a function

$(t, x) \rightarrow \omega(t, x), \omega \in SC$ at a position $(t_*, x_*, s_*) ((t_*, x_*) \in T^0 \times \mathbb{R}^n, s_* \in \omega(t_*, x_*))$ in a direction $(1, h), h \in \mathbb{R}^n$, i.e.

$$\begin{aligned} \partial_- \omega(t_*, x_*, s_*)|(h) &= \{d^0 \in \partial \omega(t_*, x_*, s_*)|(h) : \\ &d \not\leq d^0 \text{ for all } d \in \partial \omega(t_*, x_*, s_*)|(h) \setminus \{d^0\}\}. \end{aligned} \quad (13)$$

Now, remind [7] the definition of the derivative of a multi-valued mapping $t \rightarrow W(t) : T \rightarrow 2^{\mathbb{R}^{n+m}}$ at a point (t_*, x_*, s_*) , $(x_*, s_*) \in W(t_*)$. Here $W(t)$ is a closed set in $\mathbb{R}^n \times \mathbb{R}^m$ for all $t \in T$.

The derivative of a multi-valued mapping $t \rightarrow W(t)$ at a point (t_*, x_*, s_*) , $(x_*, s_*) \in W(t_*)$ is the set

$$\begin{aligned} DW(t_*, x_*, s_*) &= \left\{ (h, d) \in \mathbb{R}^n \times \mathbb{R}^m : h = \lim_{k \rightarrow \infty} \frac{x_k - x_*}{t_k - t_*}, d = \lim_{k \rightarrow \infty} \frac{\psi_k - s_*}{t_k - t_*}, \right. \\ &\left. t_k \rightarrow t_*, t_k \in (t_*, \vartheta], (x_k, \psi_k) \in W(t_k) \right\}. \end{aligned} \quad (14)$$

Let ω be a function of the class SC and the set $W = \{(t, x, s) \in T \times \mathbb{R}^n \times \mathbb{R}^m : s^0 \leq s, s^0 \in \omega(t, x)\}$ its epigraph. Define the multi-valued mapping $t \rightarrow W(t)$ by the formula

$$W(t) = \{(x, s) \in \mathbb{R}^n \times \mathbb{R}^m : (t, x, s) \in W\}. \quad (15)$$

It is possible to prove the following statement.

Lemma 2. Let $\omega \in SC$, W be the epigraph of the function ω , and the multi-valued mapping $t \rightarrow W(t)$ determined by (15), $(t_*, x_*) \in T^0 \times \mathbb{R}^n, s_* \in \omega(t_*, x_*)$. Then

$$DW(t_*, x_*, s_*) = \{(h, \alpha) : h \in \mathbb{R}^n, d \leq \alpha, d \in \partial_- \omega(t_*, x_*, s_*)|(h)\}. \quad (16)$$

Lemma 2 states, in fact, that the set $DW(t_*, x_*, s_*)$ coincides with the epigraph of the lower derivative $h \rightarrow \partial_- \omega(t_*, x_*, s_*)|(h)$. This means that statements operating with the notion of the derivative $DW(t_*, x_*, s_*)$ of a multi-valued mapping $t \rightarrow W(t)$ may be re-formulated for the lower derivative $h \rightarrow \partial_- \omega(t_*, x_*, s_*)|(h)$ of a function $(t, x) \rightarrow \omega(t, x), \omega \in SC$. Using this argument we shall transfer the infinitesimal formulation of the property of u -stability of a closed set in terms of the derivative of a multi-valued mapping to the lower derivative $h \rightarrow \partial_- \omega(t_*, x_*, s_*)|(h)$ of a function $\omega \in SC$.

Now, we shall cite [7] the infinitesimal form of the property of u -stability of a closed set $W \subset T \times \mathbb{R}^n \times \mathbb{R}^m$. A set W is a u -stable bridge if, for any position $(t_*, x_*, s_*) \in \partial W$, and $l \in S$

$$DW(t_*, x_*, s_*) \cap FR(t_*, x_*, l) \neq \emptyset. \quad (17)$$

Here

$$FR(t_*, x_*, l) = \{(f, 0) \in \mathbb{R}^n \times \mathbb{R}^m : f \in F(t_*, x_*, l), 0 \in \mathbb{R}^m\}, \quad (18)$$

∂W is the boundary of a set W .

Taking into account the inclusion $DW(t_*, x_*, s_*) \subseteq DW(t_*, x_*, s^*)$ for $s_* \leq s^*$ and formula (18) we may write the property of u -stability of a function $\omega \in SC$ in the following equivalent to Definition 3.

DEFINITION 5. A function $\omega \in SC$ has the property of u -stability if, for any position $(t_*, x_*) \in T^0 \times \mathbb{R}^n$, the vector value $s_* \in \omega(t_*, x_*)$, and vector $l \in S$ there exist a vector $f \in F(t_*, x_*, l)$ and a vector value of the lower derivative $d \in \partial_- \omega(t_*, x_*, s_*)(f)$ such that

$$d \leq 0. \quad (19)$$

Here $0 \in \mathbb{R}^m$ is the m -dimensional zero-vector.

Remark 5. The convex compact set $F(t_*, x_*, l)$ (11) appearing in relation (18) and Definition 5 may be replaced according to [7-9] by the half-space $\Pi(t_*, x_*, l)$ of (12), $(t_*, x_*, l) \in T \times \mathbb{R}^n \times S$ without losing the equivalence of definitions.

Remark 6. The property of u -stability of the FOGR $(t, x) \rightarrow c(t, x)$ (see Remark 3) may be written in the equivalent infinitesimal form (19).

5. Necessary and sufficient conditions

To formulate the necessary and sufficient conditions which the vector multi-valued FOGR $(t, x) \rightarrow c(t, x)$ must satisfy we introduce the notion of a v -stable function. Let us cite [1] the definition of the property of v -stability of a closed set W_v . A closed set $W_v \subset T \times \mathbb{R}^n \times \mathbb{R}^m$ is called v -stable for the augmented system (4) if, for any position $(t_*, x_*, s_*) \in W_v$, moment $t \in (t_*, \vartheta]$, and control vector $u \in P$ there exists a solution $x(\cdot)$ of the differential inclusion

$$\dot{x}(\tau) \in F(\tau, x(\tau), u), \quad x(t_*) = x_* \quad (20)$$

such that $(t, x(t), s_*) \in W_v$.

Here $F(\tau, y, u) = \text{co}\{f : f = f(\tau, y, u, v), v \in Q\}$.

DEFINITION 6. A vector multi-valued function $(t, x) \rightarrow \omega(t, x) : T \times \mathbb{R}^n \rightarrow 2^{\mathbb{R}^m}$, $\omega \in SC$ is called v -stable if, for any position (t_*, x_*, s_*) from the hypograph $\text{hypo } \omega = \{(t, x, s) : s_0 \not\leq s \text{ for all } s_0 \in \omega(t, x)\}$ ($(t_*, x_*, s_*) \in \text{hypo } \omega$) there exists a v -stable set W_v satisfying the following inclusions $(t_*, x_*, s_*) \in W_v \subset \text{hypo } \omega$.

Remark 7. The epigraph $W_u = \Sigma_{\max} = \text{epi } c$ of the FOGR $(t, x) \rightarrow c(t, x)$ is the positional absorption set of the augmented problem (4), (5). The complement

hypo $c = (T \times \mathbb{R}^n \times \mathbb{R}^m) \setminus \text{epi } c$ of the set $\text{epi } c$ represents the union of v -stable bridges [1]. Therefore, the FOGR $(t, x) \rightarrow c(t, x)$ is v -stable in the sense of Definition 6.

Remarks 3 and 7 indicate the necessity of the properties of u - and v -stability for the FOGR $(t, x) \rightarrow c(t, x)$. Let us formulate the statement about sufficiency of these conditions.

THEOREM 2. For a vector multi-valued function $(t, x) \rightarrow \omega(t, x) : T \times \mathbb{R}^n \rightarrow 2^{\mathbb{R}^m}$, $\omega \in SC$ to be the FOGR $(t, x) \rightarrow c(t, x)$ it is necessary and sufficient that the following conditions are satisfied

- 1) the boundary condition $\omega(\vartheta, x) = \sigma(x)$, $x \in \mathbb{R}^n$;
- 2) the property of u -stability in the sense of Definitions 3 or 5;
- 3) the property of v -stability in the sense of Definition 6.

This statement may be proved by using the theorem on alternative and properties of functions $\omega \in SC$.

Remark 8. Program constructions (maximin program functions) are often used for estimating the optimal guaranteed result in problems with a scalar criterion. Program constructions are called so because maximum and minimum operations determining them are fulfilled on the sets of program controls (depending on time) only. If, in general, verification of the v -stable property of a function $\omega \in SC$ in the sense of Definition 6 seems to be very hard then the maximin program function $(t, x) \rightarrow \text{pm}(t, x)$ is v -stable automatically, by definition. When, in addition, the maximin program function $(t, x) \rightarrow \text{pm}(t, x)$ is u -stable then it is said that it is regular [1, 2, 10] since it coincides, by Theorem 2, with the FOGR $(t, x) \rightarrow c(t, x)$. For obtaining convenient for testing conditions of regularity of the maximin program function $(t, x) \rightarrow \text{pm}(t, x)$ in linear problems with a vector criterion it is possible to use the infinitesimal form (19) of u -stability. These conditions of regularity are contained in [11]. Detailed proofs of the statements of the present paper can be found there.

6. Conclusions

The notion of the vector multi-valued function of optimal guaranteed result (FOGR) is introduced for a control problem with a vector criterion. It is defined as the aggregate of the best (minimal in the Pareto sense) points from the set of guaranteed (maximal by all components) vector results. This definition coincides with the known definition of the guaranteed result [1, 2] for a control problem with scalar criterion — a vector criterion of one component ($m = 1$).

Functional properties of the vector multi-valued FOGR are analysed in the present paper. The so-called stability properties are formulated. Necessary and sufficient conditions characterizing the FOGR are obtained. The infinitesimal form of the stability properties is studied. The method of construction of the FOGR is outlined.

Knowing the FOGR enables us to solve the problem. The method of extremal aiming to the epigraph of the FOGR [1, 2] can be used for constructing the solving control procedures in this case.

As for the transformation of the vector criterion to the scalar one the following should be stated. The scalarization of the vector criterion leads to the scalar control problem the solution of which depends on the coefficients of the scalar transformation. The choice of these coefficients a priori does not completely take into account the essence of the control problem. Meanwhile, the construction of the vector multi-valued FOGR provides the complete solution of the vector control problem, since all vector optimal guaranteed results and corresponding optimal solving strategies are determined. One can choose among them the results corresponding to the essence (dynamics and aims) of the control problem.

References

1. Krasovskii, N. N., Subbotin, A. I., Positional differential games. Moscow, Nauka, 1974, 456 pp. (in Russian).
2. Krasovskii, N. N., Subbotin, A. I., Game-theoretical control problems. New York, Springer, 1987, 517 pp.
3. Nikolskii, M. S., On guaranteed estimates in differential games with a vector quality criterion and fixed duration. *Izv. Akad. Nauk SSSR. Tekhn. kibernet.*, 1980, No. 2, pp. 37-43 (in Russian).
4. Gusev, M. I., Kurzhanskii, A. B., On equilibriums in multi-criterial game problems. *Dokl. Akad. Nauk SSSR*, 1976, vol. 229, No. 6, pp. 1295-1298 (in Russian).
5. Algorithms and programs of solution of linear differential games (Materials on math. software). Sverdlovsk, Urals Sci. Centre Akad. Nauk SSSR, 1984, 295 pp. (in Russian).
6. Subbotin, A. I., Generalization of the basic equation of the differential games theory. *Dokl. Akad. Nauk SSSR*, 1980, vol. 254, No. 2, pp. 293-297 (in Russian).
7. Guseinov, H. G., Subbotin, A. I., Ushakov, V. N., Derivatives for multi-valued mappings with application to game-theoretical problems of control. *Probl. Control and Inform. Theory*, 1985, vol. 14, No. 3, pp. 155-167.
8. Subbotin, A. I., Taras'ev, A. M., Conjugate derivatives of the value function of a differential game. *Dokl. Akad. Nauk SSSR*, 1985, vol. 283, No. 3, pp. 559-564 (in Russian).
9. Subbotin, A. I., Taras'ev, A. M., Stability properties of the value function of a differential game and viscosity solutions of Hamilton-Jacobi equations. *Probl. Control and Inform. Theory*, 1986, vol. 15, No. 6, pp. 451-463.
10. Subbotin, A. I., Chentsov, A. G., Optimization of guarantee in control problems. Moscow, Nauka, 1981, 288 pp. (in Russian).
11. Taras'ev, A. M., Differential games with a vector criterion. Sverdlovsk, Urals Branch Akad. Sci. USSR, Inst. Math. and Mech., 1989. 43 pp. Dep. in VINITI 11.07.89, No. 4608-B89 (in Russian).

Функция оптимального гарантированного результата в задачах управления с векторным критерием

А. М. ТАРАСЬЕВ

(Свердловск)

Основной чертой рассматриваемой в работе постановки задачи управления является применение векторного критерия для оценки качества процесса. При исследовании задачи введено понятие векторной многозначной функции оптимального гарантированного результата. Изучаются свойства этой функции. Рассматривается инфинитезимальная форма свойств стабильности. Приводятся необходимые и достаточные условия, которым должна удовлетворять векторная многозначная функция оптимального гарантированного результата.

А. М. Тарасьев

Институт математики и механики УрО АН СССР
СССР, 620219, Свердловск, ГСП-384,
ул. С. Ковалевской, 16.

ASYMPTOTIC BEHAVIOUR OF A COMPLEX RENEWABLE STANDBY SYSTEM WITH FAST REPAIR

A. I. CHERNYAK, J. SZTRIK

(Kiev) (Debrecen)

(Received December 21, 1989)

The present paper is concerned with an asymptotic analysis of a complex renewable standby system operating in random environments. Supposing "fast repair" it is shown that the time to the first system failure converges in distribution, under appropriate norming, to an exponentially distributed random variable.

1. Introduction

In this paper we deal with a special queueing problem which is of considerable importance in reliability theory. In many models of practical interest "small parameters" are usually present, e.g. the failure rate of the elements are much smaller than their repair rates. (This is termed in reliability theory as "fast repair".) This situation enables us to use approximate methods in reliability calculations. For good reviews and materials the interested reader is referred to, among others [3-8, 11-14, 16]. It is also well known that the great majority of problems can be treated by the help of Semi-Markov Processes (SMP), Semi-Regenerative Processes or, more generally, processes with an embedded point process (cf. Franken *et al.* [5]). For those models, mostly stationary reliability measures are obtained, and characteristics like time to the first system failure are difficult to obtain. Since the failure-free operation of the system corresponds to sojourn time problems we can use the results obtained for SMP. It is easy to see that in the case of "fast repair" the exit from a given subset of the state space of the underlying SMP is a "rare" event, that is, it occurs with a small probability. Thus, it is natural to investigate the asymptotic behavior of sojourn time in a given subset, provided that the probability of exit from it tends to zero (see Anisimov [1-2], Keilson [9], Korolyuk and Turbin [10]).

The aim of the present paper is to deal with an asymptotic analysis of a complex renewable standby system operating in random environments. Supposing "fast repair" it is shown that the time to the first system failure converges in

distribution, under appropriate norming, to an exponentially distributed random variable.

The main contribution of our paper is the following. The failure and repair intensities of the elements depend on the number of the failed elements and the state of the given random environment. As a result of this assumption, the corresponding subset of the limiting Markov process—constructed to this problem—is not a simple essential class of states. Hence, the “classical” methods cannot be applied. Using the results of Anisimov [1-2] the asymptotic exponentiality is proved.

2. The mathematical model

Let us consider a renewable system consisting of n_1 operating units, n_2 loaded standby units, n_3 lightly loaded standby units, and r repair crews. The operating elements are assumed to be embedded in a random environment governed by an irreducible, aperiodic Markov chain $(X_1(t), t \geq 0)$ with state space $\{1, \dots, r_1\}$ and with transition density matrix

$$\left\{ a_{i_1 j_1}^{(1)}, i_1, j_1 = \overline{1, r_1}, a_{i_1 i_1}^{(1)} = \sum_{j \neq i_1} a_{i_1 j}^{(1)} \right\}.$$

Whenever $X_1(t) = i_1$ and at time t there are s elements at the repair facility, the probability of failure of each operating unit in the interval $(t, t+h)$ is

$$\lambda(i_1, s)h + o(h), \quad i_1 = \overline{1, r_1}, \quad s = \overline{0, n_1 + n_2 + n_3 - 1}.$$

Similarly, the loaded standby units are supposed to be embedded in a random environment governed by an irreducible, aperiodic Markov chain $(X_2(t), t \geq 0)$ with state space $\{1, \dots, r_2\}$ and with transition density matrix

$$\left\{ a_{i_2 j_2}^{(2)}, i_2, j_2 = \overline{1, r_2}, a_{i_2 i_2}^{(2)} = \sum_{j \neq i_2} a_{i_2 j}^{(2)} \right\}.$$

Whenever $X_2(t) = i_2$ and at time t there are s elements at the repair facility, the probability of failure of each loaded standby unit in the interval $(t, t+h)$ is

$$\beta(i_2, s)h + o(h), \quad i_2 = \overline{1, r_2}, \quad s = \overline{0, n_1 + n_2 + n_3 - 1}.$$

Furthermore, the lightly loaded standby elements are also supposed to be embedded in a random environment governed by an irreducible, aperiodic Markov chain $(X_3(t), t \geq 0)$ with state space $\{1, \dots, r_3\}$ and with transition density matrix

$$\left\{ a_{i_3 j_3}^{(3)}, i_3, j_3 = \overline{1, r_3}, a_{i_3 i_3}^{(3)} = \sum_{j \neq i_3} a_{i_3 j}^{(3)} \right\}.$$

Whenever $X_3(t) = i_3$ and at time t there are s elements at the repair facility, the probability of failure of each lightly loaded standby unit in the interval $(t, t+h)$ is

$$\nu(i_3, s)h + o(h), \quad i_3 = \overline{1, r_3}, \quad s = \overline{0, n_1 + n_2 + n_3 - 1}.$$

When the elements fail they enter a repair facility and will be immediately served, unless all the repairmen are busy, otherwise they wait in a queue in the order of their breakdowns. The repair facility is supposed to be embedded in a random environment, governed by an irreducible, aperiodic Markov chain $(X_4(t), t \geq 0)$ with state space $\{1, \dots, r_4\}$ and with transition density matrix

$$\left\{ a_{i_4 j_4}^{(4)}, \quad i_4, j_4 = \overline{1, r_4}, \quad a_{i_4 i_4}^{(4)} = \sum_{j \neq i_4} a_{i_4 j}^{(4)} \right\}.$$

Whenever $X_4(t) = i_4$ and at time t there are s elements at the repair facility, the probability of repair of each unit under service in the interval $(t, t+h)$ is

$$\mu(i_4, s, \varepsilon)h + o(h), \quad i_4 = \overline{1, r_4}, \quad s = \overline{1, n_1 + n_2 + n_3}.$$

Each operating unit that fails is instantaneously replaced by a unit from the loaded standby; each unit that fails or that is put into operation from the loaded standby is immediately replaced by a unit from the light standby. Each unit after renewal is put into the light standby.

The environmental processes and all the random variables are assumed to be independent of each other.

Let us consider the system assuming "fast repair", that is, $\mu(i_4, s, \varepsilon) \rightarrow \infty$ as $\varepsilon \rightarrow 0$. For simplicity, let $\mu(i_4, s, \varepsilon) = \mu(i_4, s)/\varepsilon$.

The system is said to be failed iff the number of failed elements is $m+1$, $1 < m < n_1 + n_2 + n_3$.

Let $Y_\varepsilon(t)$ denote the number of failed elements at time t and let

$$\Omega_\varepsilon(m) = \inf(t : Y_\varepsilon(t) = m+1 / Y_\varepsilon(0) \leq m)$$

that is, the instant at which the system breaks down for the first time. Hence, our goal is to determine the distribution of $\Omega_\varepsilon(m)$. We have

THEOREM 1. For the system in question, under the above assumptions, independently of the initial state, the distribution of the normalized random variable $\varepsilon^m \Omega_\varepsilon(m)$ converges weakly to an exponentially distributed random variable with parameter

$$\Lambda = \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \sum_{i_3=1}^{r_3} \sum_{i_4=1}^{r_4} \prod_{i_1}^{(1)} \prod_{i_2}^{(2)} \prod_{i_3}^{(3)} \prod_{i_4}^{(4)} \frac{\prod_{s=0}^m \gamma(i_1, i_2, i_3, i_4, s)}{\prod_{s=1}^m \min(s, r) \mu(i_4, s)}$$

where $\gamma(i_1, i_2, i_3, i_4, s)$ is defined later.

Proof. The method of investigation is based on Anisimov [1, 2]. Construct the following multi-dimensional Markov chain

$$Z_\varepsilon(t) = (X_1(t), X_2(t), X_3(t), X_4(t), Y_\varepsilon(t))$$

with state space

$$((i_1, i_2, i_3, i_4 : s), i_1 = \overline{1, r_1}, i_2 = \overline{1, r_2}, i_3 = \overline{1, r_3}, i_4 = \overline{1, r_4}, s = \overline{0, n_1 + n_2 + n_3})$$

where

$X_1(t), X_2(t), X_3(t), X_4(t)$: governing Markov chains,

$Y_\varepsilon(t)$: the number of failed elements at time t .

Let us single out the subset of states

$$\langle \alpha_m \rangle = ((i_1, i_2, i_3, i_4 : q), i_1 = \overline{1, r_1}, i_2 = \overline{1, r_2}, i_3 = \overline{1, r_3}, i_4 = \overline{1, r_4}, q = \overline{0, m}).$$

Let

$$\gamma(i_1, i_2, i_3, i_4, s) = \begin{cases} n_1 \lambda(i_1, s) + n_2 \beta(i_2, s) + (n_3 - s) \nu(i_3, s), & 0 \leq s \leq n_3, \\ n_1 \lambda(i_1, s) + (n_2 + n_3 - s) \beta(i_2, s), & n_3 < s \leq n_2 + n_3, \\ (n_1 + n_2 + n_3 - s) \lambda(i_1, s), & n_2 + n_3 < s \leq n_1 + n_2 + n_3, \end{cases}$$

and

$$a_{i_1 i_1}^{(1)} + a_{i_2 i_2}^{(2)} + a_{i_3 i_3}^{(3)} + a_{i_4 i_4}^{(4)} + \gamma(i_1, i_2, i_3, i_4, s) + \min(s, r) \mu(i_4, s) / \varepsilon = R(i_1, i_2, i_3, i_4, s).$$

Hence, the problem is to determine the distribution of the first exit of $Z_\varepsilon(t)$ from $\langle \alpha_m \rangle$. It is easy to see that the sojourn time $\tau_\varepsilon(i_1, i_2, i_3, i_4, s)$ of $Z_\varepsilon(t)$ in state (i_1, i_2, i_3, i_4, s) is exponentially distributed with parameter $R(i_1, i_2, i_3, i_4, s)$. Furthermore, it can readily be verified that the transition probabilities for the embedded Markov chain, as $\varepsilon \rightarrow 0$, are

$$\begin{aligned} p_\varepsilon[(i_1, i_2, i_3, i_4, s), (j_1, i_2, i_3, i_4, s)] &= o(1), \quad s \geq 1, \\ p_\varepsilon[(i_1, i_2, i_3, i_4, s), (i_1, j_2, i_3, i_4, s)] &= o(1), \quad s \geq 1, \\ p_\varepsilon[(i_1, i_2, i_3, i_4, s), (i_1, i_2, j_3, i_4, s)] &= o(1), \quad s \geq 1, \\ p_\varepsilon[(i_1, i_2, i_3, i_4, s), (i_1, i_2, i_3, j_4, s)] &= o(1), \quad s \geq 1, \\ p_\varepsilon[(i_1, i_2, i_3, i_4, s), (i_1, i_2, i_3, i_4, s + 1)] &= \\ = \gamma[(i_1, i_2, i_3, i_4, s) \varepsilon / \min(s, r) \mu(i_4, s)] (1 + o(1)), & \quad 1 \leq s \leq n_1 + n_2 + n_3, \\ p_\varepsilon[(i_1, i_2, i_3, i_4, s), (i_1, i_2, i_3, i_4, s - 1)] &\rightarrow 1, \quad 1 \leq s \leq n_1 + n_2 + n_3, \end{aligned}$$

$$\begin{aligned}
 p_\varepsilon[(i_1, i_2, i_3, i_4, 0), (j_1, i_2, i_3, i_4, 0)] &= a_{i_1 j_1}^{(1)} / R(i_1, i_2, i_3, i_4, 0), \\
 p_\varepsilon[(i_1, i_2, i_3, i_4, 0), (i_1, j_2, i_3, i_4, 0)] &= a_{i_2 j_2}^{(2)} / R(i_1, i_2, i_3, i_4, 0), \\
 p_\varepsilon[(i_1, i_2, i_3, i_4, 0), (i_1, i_2, j_3, i_4, 0)] &= a_{i_3 j_3}^{(3)} / R(i_1, i_2, i_3, i_4, 0), \\
 p_\varepsilon[(i_1, i_2, i_3, i_4, 0), (i_1, i_2, i_3, j_4, 0)] &= a_{i_4 j_4}^{(4)} / R(i_1, i_2, i_3, i_4, 0), \\
 p_\varepsilon[(i_1, i_2, i_3, i_4, 0), (i_1, i_2, i_3, i_4, 1)] &= \gamma(i_1, i_2, i_3, i_4, 0) / R(i_1, i_2, i_3, i_4, 0).
 \end{aligned}$$

This agrees with conditions (1)–(4) in Anisimov [1] p. 151, but here the zero level is the set

$$((i_1, i_2, i_3, i_4 : s), i_1 = \overline{1, r_1}, i_2 = \overline{1, r_2}, i_3 = \overline{1, r_3}, i_4 = \overline{1, r_4}, s = \overline{0, 1})$$

while the q -th level is the set

$$((i_1, i_2, i_3, i_4 : q + 1), i_1 = \overline{1, r_1}, i_2 = \overline{1, r_2}, i_3 = \overline{1, r_3}, i_4 = \overline{1, r_4}).$$

Denote by $\Pi_\varepsilon(i_1, i_2, i_3, i_4, s)$ the stationary distribution of the Markov chain with transition matrix

$$\left\| \frac{p_\varepsilon[(i_1, i_2, i_3, i_4, s), (j_1, j_2, j_3, j_4, z)]}{1 - \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \sum_{k_3=1}^{r_3} \sum_{k_4=1}^{r_4} p_\varepsilon[(i_1, i_2, i_3, i_4, s), (k_1, k_2, k_3, k_4, m + 1)]} \right\|$$

$$i_1, j_1 = \overline{1, r_1}, i_2, j_2 = \overline{1, r_2}, i_3, j_3 = \overline{1, r_3}, i_4, j_4 = \overline{1, r_4}, s, z \leq m,$$

and let

$$\Pi_0(i_1, i_2, i_3, i_4, s) = \lim_{\varepsilon \rightarrow 0} \Pi_\varepsilon(i_1, i_2, i_3, i_4, s), \quad s = \overline{0, 1}.$$

Furthermore, denote by $(\Pi_{i_k}^{(k)}, i_k = \overline{1, r_k}, k = \overline{1, 4})$ the steady-state distribution of the governing Markov chain $(X_k(t), t \geq 0), k = \overline{1, 4}$, respectively. Clearly,

$$\Pi_{i_k}^{(k)} a_{i_k i_k} = \sum_{j \neq i_k} \Pi_j^{(k)} a_{j i_k}, \quad k = \overline{1, 4}. \tag{1}$$

Since the level 0 is in the limit and forms an essential class, the probabilities $\Pi_0(i_1, i_2, i_3, i_4, 0)$ and $\Pi_0(i_1, i_2, i_3, i_4, 1)$ satisfy the following system of equations

$$\begin{aligned}
 \Pi_0(i_1, i_2, i_3, i_4, 0) &= \sum_{j \neq i_1} \Pi_0(j, i_2, i_3, i_4, 0) a_{j i_1}^{(1)} / R(j, i_2, i_3, i_4, 0) + \\
 &+ \sum_{j \neq i_2} \Pi_0(i_1, j, i_3, i_4, 0) a_{j i_2}^{(2)} / R(i_1, j, i_3, i_4, 0) + \\
 &+ \sum_{j \neq i_3} \Pi_0(i_1, i_2, j, i_4, 0) a_{j i_3}^{(3)} / R(i_1, i_2, j, i_4, 0) + \\
 &+ \sum_{j \neq i_4} \Pi_0(i_1, i_2, i_3, j, 0) a_{j i_4}^{(4)} / R(i_1, i_2, i_3, j, 0) + \Pi_0(i_1, i_2, i_3, i_4, 1),
 \end{aligned} \tag{2}$$

$$\Pi_0(i_1, i_2, i_3, i_4, 1) = \Pi_0(i_1, i_2, i_3, i_4, 0)\gamma(i_1, i_2, i_3, i_4, 0)/R(i_1, i_2, i_3, i_4, 0). \quad (3)$$

It is not difficult to verify that the solution of (2), (3) subject to (1) is

$$\Pi_0(i_1, i_2, i_3, i_4, 0) = B\Pi_{i_1}^{(1)}\Pi_{i_2}^{(2)}\Pi_{i_3}^{(3)}\Pi_{i_4}^{(4)}R(i_1, i_2, i_3, i_4, 0),$$

$$\Pi_0(i_1, i_2, i_3, i_4, 1) = B\Pi_{i_1}^{(1)}\Pi_{i_2}^{(2)}\Pi_{i_3}^{(3)}\Pi_{i_4}^{(4)}\gamma(i_1, i_2, i_3, i_4, 0),$$

where

$$B = \left[\sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \sum_{i_3=1}^{r_3} \sum_{i_4=1}^{r_4} \Pi_{i_1}^{(1)}\Pi_{i_2}^{(2)}\Pi_{i_3}^{(3)}\Pi_{i_4}^{(4)}(\gamma(i_1, i_2, i_3, i_4, 0) + R(i_1, i_2, i_3, i_4, 0)) \right]^{-1}.$$

By the help of formulas 5.48, 5.49 in Anisimov [1] we get

$$\begin{aligned} \Pi_\varepsilon(i_1, i_2, i_3, i_4, q) &= \\ &= \varepsilon^{q-1} B \Pi_{i_1}^{(1)} \Pi_{i_2}^{(2)} \Pi_{i_3}^{(3)} \Pi_{i_4}^{(4)} \frac{\prod_{s=0}^{q-1} \gamma(i_1, i_2, i_3, i_4, s)}{\prod_{s=1}^{q-1} \min(s, r) \mu(i_4, s)} \times (1 + o(1)), \quad q > 1, \end{aligned}$$

and the probability of exit from $\langle \alpha_m \rangle$ is

$$\begin{aligned} g_\varepsilon(\langle \alpha_m \rangle) &= \varepsilon^m B \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \sum_{i_3=1}^{r_3} \sum_{i_4=1}^{r_4} \Pi_{i_1}^{(1)} \Pi_{i_2}^{(2)} \Pi_{i_3}^{(3)} \Pi_{i_4}^{(4)} \times \\ &\times \frac{\prod_{s=0}^m \gamma(i_1, i_2, i_3, i_4, s)}{\prod_{s=1}^m \min(s, r) \mu(i_4, s)} \times (1 + o(1)). \end{aligned}$$

Taking into account the exponentiality of $\tau_\varepsilon(i_1, i_2, i_3, i_4, s)$ for fixed θ we have

$$E \exp\{i\varepsilon^m \theta(\tau_\varepsilon(i_1, i_2, i_3, i_4, 0))\} = 1 + (\varepsilon^m \theta i / R(i_1, i_2, i_3, i_4, 0))(1 + o(1)),$$

$$E \exp(i\varepsilon^m \theta \tau_\varepsilon(i_1, i_2, i_3, i_4, s))\} = 1 + o(\varepsilon^m), \quad s > 0.$$

By using Corollary 5.6 in Anisimov [1] we obtain the statement, q.e.d.

Thus, for the time to the first system failure we have

$$P(\Omega_\varepsilon(m) > t) \simeq \exp(-\varepsilon^m \Lambda t).$$

In particular, if

$$n_2 = n_3 = 0, \quad \lambda(i_1, s) = \lambda(s), \quad \mu(i_4, s) = \mu(s)$$

we get the result of Sztrik [15].

Furthermore, if

$$\lambda(i_1, s) = \lambda, \quad \beta(i_2, s) = \beta, \quad \nu(i_3, s) = \nu, \quad \mu(i_4, s) = \mu$$

then the problem coincides with the model treated in Gnedenko et al. [8] or Ushakov [16].

Acknowledgement

We are very grateful to Prof. V. V. Anisimov for his helpful discussions.

References

1. Anisimov, V. V., Zakusilo, O. K., Donchenko, V. S., Elements of queueing theory and asymptotic analysis of systems, Visa Skola, Kiev, 1978 (in Russian).
2. Anisimov, V. V., Stochastic processes with discrete components, Visa Skola, Kiev, 1988 (in Russian).
3. Birolini, A., On the use of Stochastic Processes in Modelling Reliability Problems, Springer-Verlag, Berlin, 1985.
4. Burtin, Yu. D., Pittel, B. G., Asymptotic estimates of the reliability of complex systems, Engineering Cybernetics 10 (1972) pp. 445-451.
5. Franken, P., Kirstein, B.-M., Streller, A., Reliability Analysis of Complex Systems with Repair, EIK 20 (1984) pp. 407-422.
6. Gertsbakh, I. B., Asymptotic methods in reliability theory, A review, Adv. Appl. Prob. 16 (1984) pp. 157-175.
7. Gnedenko, D. B., Solovyev, A. D., Estimates of the reliability of complex renewable systems, Engineering Cybernetics 10 (1972) pp. 89-96.
8. Gnedenko, B. V., Belyayev, Yu. K., Solovyev, A. D., Mathematical methods of reliability theory, Academic Press, New York, 1969.
9. Keilson, J., Markov chain models — Rarity and exponentiality, Springer-Verlag, Berlin, 1979.
10. Korolyuk, V. S., Turbin, A. F., Semi-Markov Processes and their applications, Naukova Dumka, Kiev, 1976 (in Russian).
11. Kozlov, B. A., Ushakov, I. A., Reliability Handbook, Holt, Rinehart and Winston, Inc., New York, 1970.
12. Kovalenko, I. N., Investigations and Analysis of Reliability of Complex Systems, Naukova Dumka, Kiev, 1975 (in Russian).
13. Rukhin, A. L., Hsieh, H. K., Survey of Soviet Works in Reliability, Statistical Science 2 (1987) pp. 484-503.

14. Solovyev, A. D., Calculation and estimation of reliability characteristics, Znanie, Moscow, 1978 (in Russian).
15. Sztrik, J., Asymptotic behavior of a controlled renewable system $\vec{M}/M/r$, Theory of Probability and Mathematical Statistics **41** (1989) pp. 116–120 (in Russian).
16. Ushakov, I. A. (ed.), Reliability of Complex Systems, Handbook, Radio i Svyaz, Moscow, 1985 (in Russian).

Асимптотическое поведение сложной резервированной системы с быстрым восстановлением

А. Н. ЧЕРНЯК, Я. СТРИК

(Киев, Дебрецен)

Статья посвящена асимптотическому анализу надежности некоторой восстанавливаемой системы, функционирующей в случайных средах. При предположении «быстрого» обслуживания доказано, что распределение момента первого отказа системы при соответствующем нормировании слабо сходится к показательному закону.

A. I. Chernyak
Department of Applied Statistics
Kiev State University
252127 Kiev-127
USSR

J. Sztrik
Department of Mathematics
University of Debrecen
4010 Debrecen, Pf. 12.
Hungary

OPTIMIZATION AND STOCHASTIC DYNAMICS IN THE STATE SPACE

A. A. KRASOVSKII

(*Moscow*)

(Received December 27, 1989)

A comparison is made between the Fokker-Planck-Kolmogorov equation and Bellman equation applied to a controlled system optimal in terms of some special non-classical goal functionals to establish simple relationships linking the state space probability density with the Bellman function. A general form of the probability density is found for a linear plant and a linear observation function, with the minimal conditional mathematical expectation of the quadratic functional. An approximate general solution is obtained for a non-linear plant, under the condition of minimization of conditional mathematical expectation of the generalized work functional. The results obtained may be applied to efficiency studies in the design of optimal control systems, and to choosing the goal functionals.

1. Introduction

A certain degree of similarity between the Bellman equation [1, 6] in the theory of optimal controlled dynamic systems, and the Fokker-Planck-Kolmogorov equation [2, 3] (FPK equation) has been repeatedly noted in the literature, and used for the design of heuristic control algorithm [3-5]. It is quite probable, however, that under certain conditions this similarity is considerably deeper than previously assumed. This similarity may be used to obtain a good number of both theoretical and practical results.

Consider a Markovian continuous time process described by the following vector-valued stochastic differential equation (in the Langeven form):

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, t) + \boldsymbol{\xi}(t), \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^n$; \mathbf{F} is the differentiable vector-valued function of the above arguments, and $\boldsymbol{\xi}(t)$ is the vector-valued Gaussian white noise with the intensity matrix Q .

The FPK equation written for the logarithmic probability density $\ln p(\mathbf{x}, t)$ where $p(\mathbf{x}, t)$ is the unconditional probability density in the state space is expressed as [4, 5]

$$\begin{aligned} \frac{\partial \ln p}{\partial t} + \frac{\partial \ln p}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) = \operatorname{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right) + 0.5 \frac{\partial \ln p}{\partial \mathbf{x}} Q \left(\frac{\partial \ln p}{\partial \mathbf{x}} \right)^T + \\ + 0.5 \operatorname{tr} \left(Q \frac{\partial^2 \ln p}{\partial \mathbf{x} \partial \mathbf{x}^T} \right), \end{aligned} \quad (2)$$

with superscript T meaning transposition.

The mathematical expectation of terms in the right-hand part of equation (2)

$$\int_{-\infty}^{\infty} \operatorname{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right) p(\mathbf{x}, t) d\mathbf{x} - 0.5 \int_{-\infty}^{\infty} \frac{\partial \ln p}{\partial \mathbf{x}} Q \left(\frac{\partial \ln p}{\partial \mathbf{x}} \right)^T p(\mathbf{x}, t) d\mathbf{x}$$

has been considered in [4, 5] as the entropy stability index. Therefore, the value

$$\chi(\mathbf{x}, t) = \operatorname{tr} \left(-\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right) + 0.5 \frac{\partial \ln p}{\partial \mathbf{x}} Q \left(\frac{\partial \ln p}{\partial \mathbf{x}} \right)^T \quad (3)$$

may quite naturally be referred to as the differential entropy stability index.

If we take a linear system $\mathbf{F}(\mathbf{x}, t) = A(t)\mathbf{x}$ and the normal central probability distribution

$$\ln p(\mathbf{x}, t) = -0.5 \mathbf{x}^T P^{-1}(t)\mathbf{x} - 0.5 \ln(2^n \pi^n |P(t)|),$$

where $P(t) = M[\mathbf{x}(t)\mathbf{x}^T(t)]$ is the covariance matrix, and $|P(t)|$ is the principal determinant of this matrix, then the index of differential entropy stability (3) is expressed as

$$\chi(\mathbf{x}, t) = -\operatorname{tr} A(t) + 0.5 \mathbf{x}^T P^{-1}(t) Q P^{-1}(t)\mathbf{x}. \quad (4)$$

Thus, in this case $\chi + \operatorname{tr} A$ is a quadratic form relative to \mathbf{x} .

2. Controlled stochastic systems optimal in terms of a special non-classical functional

The equation describing a plant with linear control action $\mathbf{u} \in \mathbb{R}^r$ is written in the form

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t) + \varphi(\mathbf{x}, t)\mathbf{u} + \xi(t), \quad (5)$$

where $\mathbf{f}(\mathbf{x}, t)$ and $\varphi(\mathbf{x}, t)$ are the $(n \times 1)$ vector-valued function and $(n \times r)$ functional matrix, respectively.

As a minimizing functional, let us consider the functional of the form

$$I = M \left[V_g(\mathbf{x}(t_2)) + \int_{t_1}^{t_2} Q_g[\mathbf{x}(\theta), \theta] d\theta + \right. \\ \left. + 0.5 \int_{t_1}^{t_2} \mathbf{u}^T(\theta) K^{-1} \mathbf{u}(\theta) d\theta - 0.5 \int_{t_1}^{t_2} \mathbf{u}_{op}^T(\theta) K^{-1} \mathbf{u}_{op}(\theta) d\theta \right]. \quad (6)$$

Here V_g and Q_g are the given scalar functions of the above vector-valued arguments, and K is a non-singular symmetric matrix of the given coefficients.

Functional (6) refers to the class of the so-called non-classical functional [3, 8] since, along with the synthesized control \mathbf{u} , it contains optimal control \mathbf{u}_{op} which is unknown prior to solving the synthesis problem. However, this non-classical functional can not be regarded as the well-known generalized work functional (GWF) described in [3, 7, 8] since the last two terms in the right-hand part of (6) are subtracted from each other, rather than added. As a consequence, in this case the cost of synthesized control exceeds that of the optimal control (instead of the sum of these costs in GWF).

For the problem $\min_{\mathbf{u}} I$ in (5) and (6), the functional Bellman equation is represented as

$$\frac{\partial V}{\partial t} + \min \left\{ Q_g(\mathbf{x}, t) + 0.5 \mathbf{u}^T K^{-1} \mathbf{u} - 0.5 \mathbf{u}_{op}^T K^{-1} \mathbf{u}_{op} + \right. \\ \left. + \frac{\partial V}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x}, t) + \varphi(\mathbf{x}, t) \mathbf{u}] + 0.5 \operatorname{tr} \left(Q \frac{\partial^2 V}{\partial \mathbf{x} \partial \mathbf{x}^T} \right) \right\} = 0$$

Its solution is sought in the form

$$\mathbf{u} = \mathbf{u}_{op}^T = -K \varphi^T(\mathbf{x}, t) \left(\frac{\partial V}{\partial \mathbf{x}} \right)^T, \quad (7)$$

where

$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}, t) - \frac{\partial V}{\partial \mathbf{x}} \varphi(\mathbf{x}, t) K \varphi^T(\mathbf{x}, t) \left(\frac{\partial V}{\partial \mathbf{x}} \right)^T + \\ + 0.5 \operatorname{tr} \left(Q \frac{\partial^2 V}{\partial \mathbf{x} \partial \mathbf{x}^T} \right) = -Q_g(\mathbf{x}, t), \quad (8)$$

$$V(\mathbf{x}, t_2) = V_g(\mathbf{x}). \quad (9)$$

Let us take, as the integrand function $Q_g(\mathbf{x}, t)$ of the minimized functional (6) the differential entropy stability index (3) for closed-loop system, for which $\mathbf{F}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, t) + \varphi(\mathbf{x}, t)\mathbf{u}$, summed up with quantity $\text{tr} \left(Q \frac{\partial^2 \ln p}{\partial \mathbf{x} \partial \mathbf{x}^T} \right)$. Then equation (8) will take up the form

$$\begin{aligned} \frac{\partial V}{\partial t} + \frac{\partial V}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x}, t) + \varphi(\mathbf{x}, t)\mathbf{u}_{\text{op}}] = \text{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right) - 0.5 \frac{\partial \ln p}{\partial \mathbf{x}} Q \left(\frac{\partial \ln p}{\partial \mathbf{x}} \right)^T - \\ - 0.5 \text{tr} \left(Q \frac{\partial^2 V}{\partial \mathbf{x} \partial \mathbf{x}^T} \right) - \text{tr} \left(Q \frac{\partial^2 \ln p}{\partial \mathbf{x} \partial \mathbf{x}^T} \right). \end{aligned} \quad (10)$$

Comparison of equation (2) and (10) shows that for a controlled stochastic system (5), (7) and (8), optimal in the sense of minimization of functional (6), with $Q_g(\mathbf{x}, t) = \chi(\mathbf{x}, t) + \text{tr} \left(Q \frac{\partial^2 \ln p}{\partial \mathbf{x} \partial \mathbf{x}^T} \right)$ the following solution for the probability density in the state space is true:

$$\ln p(\mathbf{x}, t) = -V(\mathbf{x}, t), \quad (11)$$

$$\ln p(\mathbf{x}, t_2) = -V_g(\mathbf{x}). \quad (12)$$

This result may, in principle, be used for two purposes. First, after the above optimization problem is solved, i.e. the Bellman function $V(\mathbf{x}, t)$ is found, formula (11) immediately gives us the expression $\exp(-V(\mathbf{x}, t))$ that describes the behaviour of the probability density in the synthesized optimal control under final condition (12) and initial condition $p(\mathbf{x}, t) = \exp(-V(\mathbf{x}, t))$. Second, relationship (12) yields the "terminal" part of the minimized functional obtained directly from the desirable probability distribution at the finite time instant $t = t_2$. The latter consideration is quite valuable for us, like any other regulation that facilitates the selection of the minimized functional.

However, the above approach is also characterized by certain drawbacks. The solution of the nonlinear equation with partial derivatives (8) is a problem no less difficult than that of the analogous nonlinear Bellman equation in the traditional optimization statement, where the minimized functional does not contain any optimal control \mathbf{u}_{op} (see (6)).

Furthermore, because functional (6) includes the difference in control costs, this optimization problem is likely to feature some undesirable solutions.

3. Controllable semi-stochastic systems

Let us call a dynamic system with random initial conditions and no noise a semi-stochastic system. For such a system, equation (5) will be replaced by

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t) + \varphi(\mathbf{x}, t)\mathbf{u}. \quad (14)$$

For a closed-loop system, the right-hand part of equation (14), like before, will be denoted as $\mathbf{F}(\mathbf{x}, t)$. The differential entropy stability index of the closed-loop system, (3), in this case takes up the form $\chi(\mathbf{x}, t) = -\text{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right)$ while the FPK equation written with respect to the regular probability density $p(\mathbf{x}, t)$ takes up the form

$$\frac{\partial p}{\partial t} + \frac{\partial p}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) = -p \text{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right). \quad (15)$$

Let us specify the minimized functional as

$$I = M \left[V_g(\mathbf{x}(t_2)) - \int_{t_1}^{t_2} p(\mathbf{x}, \theta) \text{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, \theta) \right) d\theta + \right. \\ \left. + 0.5 \int_{t_1}^{t_2} \mathbf{u}^T(\theta) K^{-1} \mathbf{u}(\theta) d\theta - 0.5 \int_{t_1}^{t_2} \mathbf{u}_{\text{op}}^T(\theta) K^{-1} \mathbf{u}_{\text{op}}(\theta) d\theta \right]. \quad (16)$$

This is a non-classical functional that differs from (6) only in the form of its integrand function

$$Q_g = -p(\mathbf{x}, t) \text{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right).$$

The Bellman equation for problem (14), (16) is as follows:

$$\frac{\partial V}{\partial t} + \min_{\mathbf{u}} \left\{ -p(\mathbf{x}, t) \text{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right) + 0.5 \mathbf{u}^T K^{-1} \mathbf{u} - \right. \\ \left. - 0.5 \mathbf{u}_{\text{op}}^T K^{-1} \mathbf{u}_{\text{op}} + \frac{\partial V}{\partial \mathbf{x}} [\mathbf{f}(\mathbf{x}, t) + \varphi(\mathbf{x}, t) \mathbf{u}] \right\} = 0$$

and its solution is

$$\mathbf{u} = \mathbf{u}_{\text{op}} = -K \varphi^T(\mathbf{x}, t) \left(\frac{\partial V}{\partial \mathbf{x}} \right)^T, \quad (17)$$

where $V = V(\mathbf{x}, t)$ satisfies the equation

$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}, t) - \frac{\partial V}{\partial \mathbf{x}} \varphi(\mathbf{x}, t) K \varphi^T(\mathbf{x}, t) \left(\frac{\partial V}{\partial \mathbf{x}} \right)^T = \\ = p(\mathbf{x}, t) \text{tr} \left(\frac{\partial}{\partial \mathbf{x}} \mathbf{F}(\mathbf{x}, t) \right) \quad (18)$$

under the boundary condition

$$V(\mathbf{x}, t_2) = V_g(\mathbf{x}).$$

The comparison between (15) and (18) shows that in controlled semi-stochastic system (14), (17), (18) optimal in the sense of functional (16), where $p(\mathbf{x}, t_2) = -V_g(\mathbf{x})$, and $p(\mathbf{x}, t) = -V(\mathbf{x}, t)$ the current probability density in the state space is

$$p(\mathbf{x}, t) = -V(\mathbf{x}, t). \quad (19)$$

It will be interesting to note the following. The terminal term of the goal functional here is the density of the final probability distribution with the opposite sign, whereas in the previous case it was equal to the logarithm of $p(\mathbf{x}, t_2)$ taken with the opposite sign (12). For the one-dimensional distribution, this is illustrated in Fig. 1. Curve 1 here corresponds to case (12) and curve 2, to the case under consideration. Function $V_g(\mathbf{x})$ that corresponds to the first case infinitely grows with the increase of norm \mathbf{x} . This may have a negative effect on the nature of transients in the closed-loop system, for at the initial stage of the transient processes, when norm \mathbf{x} is large enough, the terminal term of the functional may "suppress" all the other terms. From this viewpoint, the second case has a certain advantage.

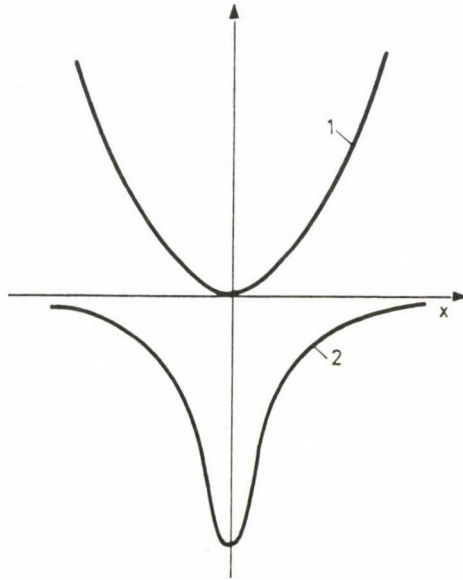


Fig. 1

The efficiency of a system is generally best expressed in probabilistic measures over probabilistic spaces. Therefore, besides selecting of a certain part of the goal functional, the above relationship may be employed for pre-estimation of the efficiency of the types of optimal systems considered. However, the potential of

such usage is rather limited. This results not only from the specific form of function Q_g and control costs in functional (6) and (16) under consideration, but also from the unconditional mathematical expectation in these functionals. The use of the unconditional mathematical expectation means that averaging is carried out throughout the entire probabilistic state space, whereas an actual control realized under concrete observation conditions is characterized by the conditional mathematical expectation [3]. Consider the solution to the general problem in the linear quadratic case.

4. General solution in linear quadratic case

It is common knowledge, that for process

$$\dot{\mathbf{x}} = A(t)\mathbf{x} + B(t)\mathbf{u} + \boldsymbol{\xi}(t) \quad (20)$$

and for the observation equation

$$\mathbf{z} = H(t)\mathbf{x} + \boldsymbol{\eta}(t), \quad (21)$$

where $\boldsymbol{\xi}(t)$ and $\boldsymbol{\eta}(t)$ are independent vector-valued white Gaussian noises with intensity matrices Q and R , respectively, and A , B , and H are specified matrices, depending, in the general case, on time, the control optimal in the sense of minimizing the functional

$$I_c = M_c \left[0.5 \mathbf{x}^T(t_2) S_g \mathbf{x}(t_2) + 0.5 \int_{t_1}^{t_2} \mathbf{x}^T(\theta) \beta \mathbf{x}(\theta) d\theta + 0.5 \int_{t_1}^{t_2} \mathbf{u}^T(\theta) K^{-1} \mathbf{u}(\theta) d\theta \right], \quad (22)$$

where M_c stands for the conditional (in observing (21)) mathematical expectation, and S_g , β , and K are specified symmetric coefficients matrices – such a control is

$$\mathbf{u} = -KB^T S \hat{\mathbf{x}}, \quad (23)$$

$$\dot{S} + SA + A^T S - SBK B^T S = -\beta, \quad S(t_2) = S_g. \quad (24)$$

Variable $\hat{\mathbf{x}}$ refers to the conditional mathematical expectation of the state vector. It serves as the output value of the Kalman-Beaucy filter (KBF):

$$\dot{\hat{\mathbf{x}}} = A\hat{\mathbf{x}} + B\mathbf{u} + PH^T R^{-1}(z - H\hat{\mathbf{x}}), \quad (25)$$

$$\dot{P} = AP + PA^T - PH^T R^{-1}HP + Q. \quad (26)$$

Expressions (23) through (26) correspond to the separation principle [1, 3, 6]. Expressions (23), (25) and (26) remain true in minimizing the non-classical functional of generalized work

$$M_c \left[0.5 \mathbf{x}^T(t_2) S_g \mathbf{x}(t_2) + 0.5 \int_{t_1}^{t_2} \mathbf{x}^T(\theta) \beta \mathbf{x}(\theta) d\theta + \right. \\ \left. + 0.5 \int_{t_1}^{t_2} \mathbf{u}^T(\theta) K^{-1} \mathbf{u}(\theta) d\theta + 0.5 \int_{t_1}^{t_2} \mathbf{u}_{op}^T(\theta) K^{-1} \mathbf{u}_{op}(\theta) d\theta \right]. \quad (27)$$

However, matrix Riccati equation (24) is in this case replaced by the matrix Lyapunov equation

$$S + SA + A^T S = -\beta, \quad S(t_2) = S_g. \quad (28)$$

Let us denote the error of estimating the state vector in the KBF as $\Delta \mathbf{x} = \hat{\mathbf{x}} - \mathbf{x}$. Substituting (23) into (25) and subtracting it from (20), we obtain

$$\Delta \dot{\mathbf{x}} = (A - PH^T R^{-1} H) \Delta \mathbf{x} + PH^T R^{-1} \boldsymbol{\eta} - \boldsymbol{\xi}, \quad (29)$$

$$\dot{\mathbf{x}} = -BK B^T S \Delta \mathbf{x} + (A - BK B^T S) \mathbf{x} + \boldsymbol{\xi}. \quad (30)$$

Present the covariance matrix of vector $(\Delta \mathbf{x}, \mathbf{x})$ in the block form:

$$M_c \left\langle \begin{bmatrix} \Delta \mathbf{x} \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} \Delta \mathbf{x} \\ \mathbf{x} \end{bmatrix}^T \right\rangle = \begin{bmatrix} M_{(11)} & M_{(12)} \\ M_{(21)} & M_{(22)} \end{bmatrix}, \quad M_{(12)} = M_{(21)}^T.$$

Using (29) and (30), we obtain the equations for the blocks of this covariance matrix:

$$\dot{M}_{(11)} = (A - PH^T R^{-1} H) M_{(11)} + M_{(11)} (A^T - H^T R^{-1} H P) + PH^T R^{-1} H P + Q, \quad (31)$$

$$\dot{M}_{(12)} = (A - PH^T R^{-1} H) M_{(12)} + M_{(12)} (A^T - SBK B^T) - M_{(11)} SBK B^T - Q, \quad (32)$$

$$\dot{M}_{(22)} = (A - BK B^T S) M_{(22)} + M_{(22)} (A^T - SBK B^T) - M_{(21)} SBK B^T - BK B^T S M_{(12)} + Q. \quad (33)$$

Comparison of (31) and (32) with (25) shows that these matrix equations have the following solutions: $M_{(11)} = P$, $M_{(12)} = -P$.

Presenting $M_{(22)}$ in the form

$$M_{(22)} = M_c[\mathbf{x}\mathbf{x}^T] = P + D, \quad (34)$$

we may find from (33) that

$$\begin{aligned} \dot{D} &= (A - BK B^T S)D + D(A^T - SBK B^T) + PH^T R^{-1}HP, \\ D(t_1) &= D_0 = P(t_1) - M_{22}(t_1). \end{aligned} \quad (35)$$

Obviously, all the processes taking place in this system are Gaussian, and with a central initial distribution the current probability density in the state space of an optimal closed-loop system is expressed by the formula

$$p(\mathbf{x}, t) = (2^n \pi^n |P(t) + D(t)|)^{-\frac{1}{2}} \exp \{ -0.5 \mathbf{x}^T [P(t) + D(t)]^{-1} \mathbf{x} \}.$$

Expressions (23), (25), (26), (34), (35), and (24) (for the case of minimizing the classical functional (22) or (28) (for the case of non-classical generalized work functional (27)), along with the initial conditions

$$P(t_1) = P_0, \quad D(t_1) = D_0 \quad (36)$$

describe the general complete solution to the problem under study.

This solution, of course, may be obtained in a different way through solving the corresponding Bellman equations and the FPK-equation in quadratic forms. However, because expressions (23), (24) or (28), (25), and (26) are commonly known, this way seems to be more convenient.

Probability density distribution may serve to estimate the efficiency of the designed optimal control system, and to choose the coefficients of the minimized functional. The respective technology demands a special description. Here, it is apt to note only the fact that, as compared with the statistical trials method, the suggested approach results in considerable savings of computer time.

5. General approximate solution for nonlinear case

The principle of separation of nonlinear systems remains true in an approximate case, which appears to be more accurate with the increase of estimation accuracy under the observation conditions considered [3].

Let us write down the corresponding expressions applying to the case of the generalized work goal functional with quadratic costs of the control and the generalized Kalman-Beaucy filter of linear approximation as a system of suboptimal estimation.

For the system

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, \mathbf{u}, t) + \boldsymbol{\xi}(t) = \mathbf{f}(\mathbf{x}, t) + \boldsymbol{\varphi}(\mathbf{x}, t)\mathbf{u} + \boldsymbol{\xi}(t) \quad (37)$$

and for the observation condition

$$\mathbf{z} = \mathbf{h}(\mathbf{x}, t) + \boldsymbol{\eta}(t) \quad (38)$$

suboptimal, in terms of minimizing the functional

$$I_c = M_c \left\{ V_g[\mathbf{x}(t_2)] + \int_{t_1}^{t_2} Q_g[\mathbf{x}(\theta), \theta] d\theta + 0.5 \int_{t_1}^{t_2} \mathbf{u}^T(\theta) K^{-1} \mathbf{u}(\theta) d\theta + \right. \\ \left. + 0.5 \int_{t_1}^{t_2} \mathbf{u}_{op}^T(\theta) K^{-1} \mathbf{u}_{op}(\theta) d\theta \right\} \quad (39)$$

is the control

$$\mathbf{u} = \mathbf{u}_{op} = -K \boldsymbol{\varphi}^T(\hat{\mathbf{x}}, t) \left(\frac{\partial V(\hat{\mathbf{x}}, t)}{\partial \hat{\mathbf{x}}} \right)^T, \quad (40)$$

where $\hat{\mathbf{x}}$ is the output value of the generalized KBF:

$$\dot{\hat{\mathbf{x}}} = \mathbf{F}(\hat{\mathbf{x}}, \mathbf{u}, t) + \hat{P} h_{\hat{\mathbf{x}}}^T(\hat{\mathbf{x}}, t) R^{-1} [\mathbf{z} - \mathbf{h}(\hat{\mathbf{x}}, t)], \quad (41)$$

$$\dot{\hat{P}} = F_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}, t) \hat{P} + \hat{P} F_{\hat{\mathbf{x}}}^T(\hat{\mathbf{x}}, t) - \hat{P} h_{\hat{\mathbf{x}}}^T(\hat{\mathbf{x}}, t) R^{-1} h_{\hat{\mathbf{x}}}(\hat{\mathbf{x}}, t) \hat{P} + Q, \quad (42)$$

$$\hat{P}(t_1) = P_0, \quad (43)$$

while $V(\mathbf{x}, t)$ is the solution to the linear equation in partial derivatives

$$\frac{\partial V}{\partial t} + \frac{\partial V}{\partial \mathbf{x}} f(\mathbf{x}, t) = -Q_g(\mathbf{x}, t) \quad (44)$$

under boundary condition $V(\mathbf{x}, t_2) = V_g(\mathbf{x})$.

Resting upon the use of expressions (40) and (44), the algorithm of optimal (suboptimal) control with prediction model is designed [3, 7, 8]. For this particular case, the analytical form of this algorithm is as follows:

$$\mathbf{u} = \mathbf{u}_{op} = -K \boldsymbol{\varphi}^T(\hat{\mathbf{x}}, t) \left[\frac{\partial}{\partial \hat{\mathbf{x}}} \left\{ V_g[X(\hat{\mathbf{x}}, t, t_2)] - \int_{t_1}^{t_2} Q_g[X(\hat{\mathbf{x}}, t, \theta)] d\theta \right\} \right]^T, \quad (45)$$

where $\mathbf{x} = X(\mathbf{x}_0, t_0, t)$ is the general solution to the equation of the free motion of the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$$

under initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$.

The algorithm with a prediction model utilizing the generalized work principle is, in the opinion of the author, the most efficient algorithm among the present-day algorithms of optimization "in the large" of complex multi-dimensional nonlinear processes. The capabilities of this class of algorithms are further extended by the latest findings in the area of fast piecewise-linear approximation of multi-argument functions on a rarefied net [9] and in the area of fast two-channel numerical integration of differential equations [10].

In the suboptimal control and estimation problem treated here, subjected to simultaneous numerical integration are process equations (37) in which \mathbf{u} is expressed by formula (45), and estimation equations (42) and (43) in which the measurement vector is described by formula (38).

The numerical integration is carried out in real time, or in some other time basis, with the use of definite difference schemes and specified initial conditions of the form $\mathbf{x}(t_1) = \mathbf{x}_0$, $\hat{\mathbf{x}}(t_1) = \hat{\mathbf{x}}_0$. Reproduced are realizations of practically white noise $\xi(t)$, $\eta(t)$. As a result of a single-step numerical modeling (integration) we obtain realizations of vector-valued functions $\mathbf{x}(t)$, $\hat{\mathbf{x}}(t)$ and $\mathbf{u}(t)$ which are taken as basic (reference) functions and designated as $\mathbf{x}_r(t)$, $\hat{\mathbf{x}}_r(t)$ and $\mathbf{u}_r(t)$. The problem of stochastic dynamics is further stated as the problem of finding the statistical characteristics of deviations

$$\Delta \mathbf{x}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t), \quad \Delta \mathbf{x}_\delta(t) = \mathbf{x}(t) - \mathbf{x}_\delta(t)$$

in a linear Gaussian approximation (owing to the negligibility of $M[|\Delta \mathbf{x}|]$, $M[|\Delta \mathbf{x}_\delta|]$). In this case, one can immediately use expressions of the type (34) and (35)

$$M_{(22)} = M(\Delta \mathbf{x}_\delta \Delta \mathbf{x}_\delta^T) = \dot{P} + D, \quad (46)$$

$$\dot{P} = F_{\hat{\mathbf{x}}}[\hat{\mathbf{x}}(t), t]\dot{P} + \dot{P}F_{\hat{\mathbf{x}}}^T[\hat{\mathbf{x}}(t), t] - \dot{P}h_{\hat{\mathbf{x}}}^T[\hat{\mathbf{x}}(t), t]R^{-1}h_{\hat{\mathbf{x}}}[\hat{\mathbf{x}}(t), t]\dot{P} + Q, \quad (47)$$

where

$$\begin{aligned} \mathbf{F}[\hat{\mathbf{x}}(t), t] &= \mathbf{f}[\hat{\mathbf{x}}(t), t] + \varphi[\hat{\mathbf{x}}(t), t]\mathbf{u}(t), \\ \dot{D} &= \frac{\partial}{\partial \hat{\mathbf{x}}} \{ \mathbf{f}[\hat{\mathbf{x}}(t), t] + \varphi[\hat{\mathbf{x}}(t), t]\mathbf{u}[\hat{\mathbf{x}}(t), t] \} D + \\ &+ D \left\langle \frac{\partial}{\partial \hat{\mathbf{x}}} \{ \mathbf{f}[\hat{\mathbf{x}}(t), t] + \varphi[\hat{\mathbf{x}}(t), t]\mathbf{u}[\hat{\mathbf{x}}(t), t] \} \right\rangle^T + \\ &+ \dot{P}(t)h_{\hat{\mathbf{x}}}^T[\hat{\mathbf{x}}(t), t]R^{-1}h_{\hat{\mathbf{x}}}[\hat{\mathbf{x}}(t), t]\dot{P}(t). \end{aligned} \quad (48)$$

The control \mathbf{u} found here is calculated by formula (45) at the previous step (in finding the realizations $\hat{\mathbf{x}}(t)$ and $\mathbf{u}(t)$).

Expression (47) coincides with equation (42), which is numerically integrated also at the first step. Thus, in order to approximately find the probability distribution that characterizes scattering of trajectories in the state space relative to

the basic trajectory, it suffices to numerically integrate only the Lyapunov matrix equation (48).

Having fulfilled this operation of numerical integration, we obtain the following probability distribution:

$$p(\Delta \mathbf{x}, t) = (2^n \pi^n |\hat{P}(t) + D(t)|)^{-\frac{1}{2}} \times \exp \left\{ -0.5 \Delta \mathbf{x}^T \left[\hat{P}(t) + D(t) \right]^{-1} \Delta \mathbf{x} \right\}. \quad (49)$$

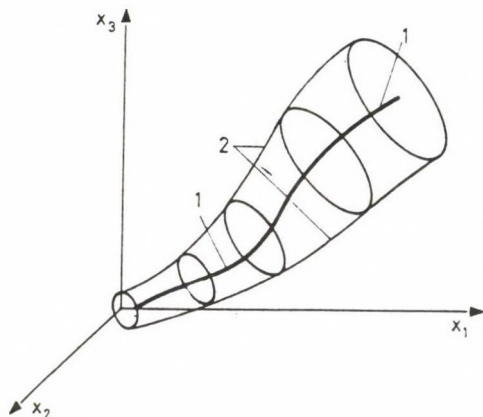


Fig. 2

This distribution may be used to both estimate the efficiency of the suboptimal control system, and choose (update) the minimized goal functional.

The simplest illustration of a reference trajectory and of trajectory scattering in the three-dimensional state space is depicted in Fig. 2. Basic trajectory 1 is obtained by numerical integration of equations (37), (41), (42), and (45) under certain initial conditions, while the tube of trajectories 2 is found by numerical integration of equation (48), also under given initial (or final) condition.

References

1. Roitenberg, Ya. N., Automatic Control, Nauka Publ., Moscow, 1978.
2. Pugachyov, V. S., Sinitsyn, I. N., Stochastic Differential Systems, Nauka Publ., Moscow, 1985.
3. Handbook on Theory of Automatic Control, Edited by A. A. Krasovskii, Nauka Publ., Moscow, 1987.

4. Krasovskii, A. A., *Statistical Theory of Transient Processes in Control Systems*, Moscow, 1968.
5. Krasovskii, A. A., *Phase Space and Statistical Theory of Dynamic Systems*, Moscow, 1974.
6. Phillis, V. A., Entropy Stability of Continuous Dynamic Systems, *Int. J. of Control*, 1982, Vol. 35, No. 2, pp. 329-340.
7. Krasovskii, A. A., Development of Generalized Work Minimum Principle, *Avtomatika i Telemekhanika*, 1987, No. 1, pp. 13-23.
8. Krasovskii, A. A., Generalization of Solution to Optimization Problem with a Non-classical Functional, *Doklady AN SSSR*, 1985, Vol. 284, No. 4, pp. 808-811.
9. Krasovskii, A. A., Approximation of Multi-argument Functions in Systems of Numerical Modeling, *Izv. AN SSSR, Technical Cybernetics*, 1989, No. 3, pp. 3-11.
10. Krasovskii, A. A., A Method of Fast Numerical Integration for One Class of Dynamic System, *Izv. AN SSSR, Technical Cybernetics*, 1989, No. 1, pp. 3-14.

Оптимизация и стохастическая динамика в пространстве состояний

А. А. КРАСОВСКИЙ

(Москва)

Сопоставляются уравнение Фоккера-Планка-Колмогорова и уравнение Беллмана для управляемой стохастической системы, оптимальной в отношении некоторых специально подобранных неклассических целевых функционалов. Для таких функционалов получены простые соотношения между плотностью вероятности в пространстве состояний и функцией Беллмана. Для линейного объекта и линейной функции наблюдения при минимизации условного математического ожидания квадратичного функционала получено в общем виде решение для плотности вероятности в пространстве состояний.

Приближенное общее решение найдено для случая нелинейного объекта (с линейно входящим управлением) при минимизации условного математического ожидания функционала обобщенной работы. Полученные результаты могут быть использованы при исследовании эффективности синтезированных систем оптимального управления и выборе целевых функционалов.

А. А. Красовский
СССР, Москва, 125083,
Петровско-Разумовская аллея, 16

INFINITE-DIMENSIONAL SYSTEMS: APPROXIMATE CONTROLLABILITY AND OBSERVABILITY. PART I

S. K. KOROVIN, M. G. NIKITINA AND S. V. NIKITIN

(*Moscow*)

(Received December 27, 1989)

In most applications it is sufficient to know that the system is approximately controllable in the sense that it can be made to change any state for any desired one with some (generally, as low as desired) error. This paper proposes a new practicable way to analyse approximate controllability and observability.

Introduction

The mathematical theory of the infinite-dimensional system finds numerous applications. Some industrial and physical processes are described as distributed parameter systems. The mathematical theory of infinite-dimensional systems is concerned with various open-loop systems that are subjected to exogenous signals (or having an input) and whose state is observable (or having an output). The state-of-the-art in this field has been described in surveys [27, 36].

Control of distributed parameter systems has been the subject of numerous books [1-4, 7, 8, 20]. The subject of this article is the approximate controllability, observability, and stabilization of infinite-dimensional systems whose input and output are assumed finite-dimensional. Papers on these systems [19, 21-23, 32, 33, 37-39] concentrate on controllability and observability which have been shown to be dual [36]. Infinite-dimensional analogs of ranking observability and global controllability criteria have been formulated that are important mainly for the theory, since their direct application assumes a family of functions which makes a basic and complex problem in itself in every specific case [6].

In most applications it is sufficient to know that the system is approximately controllable in the sense that it can be made to change any state for any desired one with some (generally, as low as desired) error. Approximate controllability has recently become the subject of research [41, 42]. Criteria for approximate and accurate controllability through pulsed signals have been obtained [9]. Approximate reachability of zero has been studied [11] through reduction to analysis of dense

solvability of some equation. This paper proposes a practical way to analyse approximate controllability with the use of the conventional Fourier procedure. The method is, in a sense, an extension of the procedure proposed in [41]. Criteria of the simultaneous controllability and observability of a family of finite-dimensional systems are proved in Section 2. Application of these criteria to controllability (observability) analysis of finite-dimensional Fourier approximations leads to practicable sufficient conditions of approximate controllability (observability) of infinite-dimensional systems in Section 3. A new modification of the well-known Kalman criterion is proposed in Section 3, which is the core of the theory described in Sections 3 and 4.

1. Problem statement

A linear stationary control system is analysed in the form

$$\begin{aligned}\Sigma(C, A, B) : \dot{x} &= Ax + Bu, \\ y &= Cx,\end{aligned}$$

where A is a linear endomorphism of the Hilbert space H over a field of complex numbers, in other words, $A \in \text{End}(H)$ or

$$A : \mathcal{D}(A) \subset H \rightarrow I(A) \subset H,$$

$\mathcal{D}(A)$ is the domain of the operator A , and $I(A)$ its range; B and C are linear operators, $B : \mathcal{D}(B) \subset U \rightarrow I(B) \subset H$ and $C : \mathcal{D}(C) \subset H \rightarrow I(C) \subset Y$ where U and Y are complex finite-dimensional Hilbert spaces, or $Bu = \sum_{i=1}^m b_i u_i$ where $b_i \in H$, $i = 1, 2, \dots, m$, and C is a finite number of linear functionals $c_i \in H^*$.

The operator A is assumed to meet the following conditions:

a1) A is an infinitesimal generator of the C_0 -semi-group e^{At} in H ;

a2) the operator spectrum $\sigma(A)$ is a discrete set from \mathbb{C} and every $\lambda \in \sigma(A)$ has a finite multiplicity;

a3) $\overline{\mathcal{D}(A)} = H$ (the domain $\mathcal{D}(A)$ is dense everywhere in H) and there exists a basis $\{\xi_\lambda\}$ in H which consists of eigen- and adjoint vectors of the operator A , or $\text{span}_{\mathbb{C}} \{\xi_\lambda\}_{\lambda \in \sigma(A)}$ is dense everywhere in H ($\text{span}_K Z$ is the set of all possible linear combinations from Z with factors from K).

This paper will concentrate on the approximate control of the state, and observability of an infinite-dimensional process $\Sigma(C, A, B)$ with a finite-dimensional input. Special attention will be given to the analysis of controllability. The proposed spectral form of the controllability criterion for finite-dimensional systems is practicable in applications.

In applications the system need not necessarily be accurately controllable or observable, e.g. in the sense of the definitions from [1, 20, 36] and approximate controllability and observability are sufficient. A weaker variety of controllability and observability is introduced in

DEFINITION 1. If for any two points $z_1, z_2 \in H$ and $\varepsilon > 0$ there are $T > 0$, $\bar{z}_1 \in O_\varepsilon(z_1)$, $\bar{z}_2 \in O_\varepsilon(z_2)$ (here $O_\varepsilon(z) = \{x \in H : \|x - z\|_H < \varepsilon\}$) and a control $u(t) : [0, T] \rightarrow U$ such that

$$\exp \left(\int_0^T (Ax(\tau) + Bu(\tau)) d\tau \right) \bar{z}_1 = \bar{z}_2,$$

with $\exp \left(\int_0^T (Ax(\tau) + Bu(\tau)) d\tau \right)$ denoting the flow in H generated by the system $\dot{x} = Ax + Bu(t)$, then $\Sigma(C, A, B)$ is approximately controllable.

This fact is denoted as an inclusion

$$\Sigma(C, A, B) \in \mathcal{AV}(H)$$

DEFINITION 2. The system $\Sigma(C, A, B)$ is approximately observable if the conjugate system

$$\begin{aligned} \Sigma(B^*, A^*, C^*) : \dot{\xi} &= A^*\xi + C^*u, \\ y &= B^*\xi \end{aligned}$$

is approximately controllable. If $\Sigma(B^*, A^*, C^*) \in \mathcal{AV}(H)$, then $\Sigma(C, A, B) \in \mathcal{AH}(H)$. Practicable methods of analysing approximate controllability and observability are developed in Section 3.

2. Structural features of controllable and observable systems

The objective of this Section is to obtain a form of the rank Kalman-Krasovsky controllability condition which would be better suited for computation. What is important is that for an infinite-dimensional system these rank conditions can not be checked. On the other hand, in numerous problems the operator has been thoroughly studied and its eigenvalues and eigenvectors are computable. This is so, in particular in some mathematical-physics problems. Consequently, it would be useful to express the controllability and observability conditions as characteristics of the process. In this Section the controllability criterion will be formulated in terms of the spectrum and invariant subspaces of A .

For this purpose we will need the following

DEFINITION 3. The Jordan index $IJ(\lambda, A)$ of the number $\lambda \in \sigma(A)$ is the number of Jordan cells that are associated with the eigenvalue λ , or

$$IJ(\lambda, A) = \begin{cases} 0 & \text{for } \lambda \notin \sigma(A), \\ \text{the number of Jordan cells associated with } \lambda \in \sigma(A). \end{cases}$$

We will need the following notation. Let A be a linear operator mapping \mathbb{C}^n into \mathbb{C}^n with $\sigma(A) = \{\lambda_1, \dots, \lambda_\nu\}$, $\lambda_i \neq \lambda_j$, $1 \leq i < j \leq \nu$. Then $\xi_{ij}(\lambda_\alpha)$ is understood as an adjoint vector of power j which is associated with the eigenvalue $\lambda_\alpha \in \sigma(A)$ i.e. $(\lambda_\alpha E - A)^j \xi_{ij}(\lambda_\alpha) = 0$ while $(\lambda_\alpha E - A)^{j-1} \xi_{ij}(\lambda_\alpha) \neq 0$. $\{\mu_i^\alpha; i = 1, \dots, IJ(\lambda_\alpha, A)\}$ are dimensions of the Jordan cells with eigenvalue λ_α in their diagonals.

By virtue of the above, $\{\xi_{1i}(\lambda_\alpha), \dots, \xi_{\mu_{\alpha,i}}^\alpha(\lambda_\alpha)\}_{i=1}^{IJ(\lambda_\alpha, A)}$ are bases of eigenvalue subspaces of A . Let us consider a finite-dimensional system $\Sigma(C, A, B)$. Associate it with a totality of numbers

$$\begin{aligned} \varphi_{ij}^\alpha = \det & \left[\{\xi_{1,1}(\lambda_1), \xi_{2,1}(\lambda_1), \dots, \xi_{\mu_{1,1}}^1(\lambda_1), \xi_{1,2}(\lambda_1), \dots, \xi_{\mu_{2,2}}^1(\lambda_1), \dots, \right. \\ & \left. \xi_{\mu_{\kappa_1, \kappa_1}}^1(\lambda_1)\}, \dots, \right. \\ & \left. \{\xi_{1,1}(\lambda_\alpha), \xi_{2,1}(\lambda_\alpha), \dots, \xi_{\mu_{1,1}}^1(\lambda_\alpha), \dots, \xi_{1,i}(\lambda_\alpha), \dots, \xi_{\mu_{i-1,1}}^\alpha(\lambda_\alpha), b_j, \right. \\ & \left. \xi_{1,i+1}(\lambda_\alpha), \dots, \right. \\ & \left. \{\xi_{1,1}(\lambda_\nu), \dots, \xi_{\mu_{1,1}}^\nu(\lambda_\nu), \dots, \xi_{1,\kappa_\nu}(\lambda_\nu), \dots, \xi_{\mu_{\kappa_\nu, \kappa_\nu}}^\nu(\lambda_\nu)\} \right], \\ & 1 \leq \alpha \leq \nu, \quad 1 \leq i \leq IJ(\lambda_\alpha, A), \quad 1 \leq j \leq m, \end{aligned}$$

where $\kappa_i = IJ(\lambda_\alpha, A)$ and b_j is the j -th column of the matrix $B = \{b_1, \dots, b_j, \dots, b_m\}$. In the above notation the rank Kalman-Krasovsky criterion can be made to take the form of

THEOREM 1. A finite-dimensional system $\Sigma(C, A, B)$ is controllable iff

$$\text{rank} \{ \varphi_{ij}^\alpha; 1 \leq j \leq m, 1 \leq i \leq IJ(\lambda_\alpha, A) \} \geq IJ(\lambda_\alpha, A) \quad (1)$$

for every $1 \leq \alpha \leq \nu$.

Proof. The necessity of the conditions (1) follows from the obvious fact that the controllability of the system entails that of every subsystem. Assume that at least one condition of (1) is not met. For instance $\text{rank} \{ \varphi_{ij}^1 \}_{i,j} < IJ(\lambda_1, A)$. Therefore, if $IJ(\lambda_1, A) = 1$, then $\varphi_{ij}^1 = 0$ and there exists an uncontrollable subsystem which is associated with the above Jordan cell of A . If, however, $IJ(\lambda_1, A) \geq 2$, then there exist at least two Jordan cells, $J_1(\lambda_1)$ and $J_2(\lambda_1)$ (that are associated with one λ_1) which must be handled by one control, which is impossible by the Kalman-Krasovsky criterion. This proves the necessity of the condition (1).

Let us prove the inverse implication for the case when the input is one variable or $B = b \in \mathbb{C}^n$ (for a multi-dimensional input the proof is similar). The control system is

$$\dot{x} = Ax + bu, \quad x \in \mathbb{C}^n.$$

Then the system

$$\begin{aligned} \dot{z} &= \lambda z + \rho u, \\ \dot{x} &= Ax + bu, \end{aligned} \quad (2)$$

where $\lambda \in \mathbb{C}$, $z \in \mathbb{C}$ and $\rho \in \mathbb{C} \setminus \{0\}$ is controllable iff $\lambda \notin \sigma(A)$. Indeed, the Kalman matrix for this system has the form

$$K = \begin{pmatrix} \lambda^n \rho & \lambda^{n-1} \rho & \dots & \lambda \rho & \rho \\ A^n b & A^{n-1} b & \dots & Ab & b \end{pmatrix}.$$

Multiply the second column of the matrix K by λ and subtract the product from the first one, then multiply the third column by λ and subtract it from the second one, etc. Finally, multiply the last column by λ and subtract it from the last but one. As a result we have

$$\begin{aligned} |\det K| &= |\rho| \cdot |\det(A^n b - \lambda A^{n-1} b, A^{n-1} b - \lambda A^{n-2} b, \dots, Ab - \lambda b)| = \\ &= |\rho| \cdot |\det(A^{n-1} b, A^{n-2} b, \dots, Ab, b)| \cdot |P_A(\lambda)|, \end{aligned}$$

$P_A(\lambda)$ is the characteristic polynomial of the matrix A .

Consequently, the system (20) is controllable iff $\lambda \notin \sigma(A)$.

Let us show that when a controllable Jordan cell $J(\lambda)$ is added to the system, controllability is preserved iff $\lambda \notin \sigma(A)$. Take up the case where $J(\lambda)$ is a Jordan cell of order 2×2 . In a general situation the reasoning is similar. For the system

$$\begin{aligned} \dot{z}_1 &= \lambda z_1 + z_2 + \rho_1 u, \\ \dot{z}_2 &= \lambda z_2 + \rho_2 u, \quad \rho_2 \neq 0, \\ \dot{x}_1 &= Ax + bu \end{aligned} \quad (3)$$

the Kalman matrix is

$$K = \begin{pmatrix} \lambda^{n+1} \rho_1 + (n+1) \lambda^n \rho_2 & \lambda^n \rho_1 + n \lambda^{n-1} \rho_2 & \dots & \lambda \rho_1 + \rho_2 & \rho_1 \\ \lambda^{n+1} \rho_2 & \lambda^n \rho_2 & \dots & \lambda \rho_2 & \rho_2 \\ A^{n+1} b & A^n b & \dots & Ab & b \end{pmatrix}.$$

Multiply the second row by ρ_1/ρ_2 and subtract the product from the first one. Following this, multiply the second column by λ and subtract it from the first one, the third column by λ and subtract it from the second one, etc. Finally, multiply the last one by λ and subtract it from the last but one. As a result we have

$$\det K = \det \begin{pmatrix} \lambda^n \rho_2 & \lambda^{n-1} \rho_2 & \dots & \rho_2 & 0 \\ 0 & 0 & \dots & 0 & \rho_2 \\ A^{n+1} b - \lambda A^n b & A^n b - \lambda A^{n-1} b & \dots & Ab - \lambda b & b \end{pmatrix}.$$

Using the reasoning of the proof of controllability for the system (2) we have

$$|\det K| = |\rho_2|^2 \cdot |\det(A^{n-1}b, A^{n-2}b, \dots, Ab, b)| \cdot |P_A(\lambda)|^2,$$

where $P_A(\lambda)$ is the characteristic polynomial of the matrix A . Consequently, the system (30) is controllable iff $\lambda \notin \sigma(A)$ and $\rho_2 \neq 0$. Proof that the system

$$\begin{aligned} \dot{z} &= J(\lambda)z + \rho u, & z &\in \mathbb{C}^l, \\ \dot{x} &= Ax + bu, & x &\in \mathbb{C}^l, \end{aligned}$$

is controllable iff $\rho_l \neq 0$ and $\lambda \notin \sigma(A)$ is the same. In this case condition (1) transforms into the requirement that $\rho_l \neq 0$ and so $|\det K| = |\rho_l|^l \cdot |P_A(\lambda)|^l \times |\det(A^{n-1}b, \dots, Ab, b)|$.

Because any linear system can be obtained by adding subsystems with Jordan cells, sufficiency of condition (1) is proved in the one-dimensional case. In the general case the proof is analogous. This proves the Theorem.

The Theorem makes possible a relatively simple analysis of a high-dimensional linear system which is decomposed into subsystems that are associated with different Jordan cells of the operator A .

Then elements φ_{ij}^α (condition (1)) are determinants of matrices that are made of eigen- and adjoint vectors that are associated with given λ_α , with the adjoint vector of the maximal power replaced by the j -th column of the matrix B . Consequently, a certain independence of condition a1) of the dimension of the system $\Sigma(C, A, B)$ offers Theorem 1 certain advantages over the Kalman-Krasovsky criterion. True, with $n \leq 4$ (n being the dimension of the system $\Sigma(C, A, B)$), condition (1) requires a larger computer load than computation of the Kalman matrix rank.

Corollary 1. If $\Sigma(C, A, B)$ has one input, the system is controllable iff $\lambda \in \sigma(A)$ is associated with a unique eigenvector and conditions (1) hold which in this case take the form $\varphi(\lambda_\alpha) \neq 0$ for any $\lambda_\alpha \in \sigma(A) = \{\lambda_1, \lambda_2, \dots, \lambda_\nu\}$ where

$$\begin{aligned} \varphi(\lambda_\alpha) &= \det \{ \xi_1(\lambda_1), \dots, \xi_{\mu_1}(\lambda_1), \xi_1(\lambda_2), \dots, \xi_{\mu_2}(\lambda_2), \dots, \xi_1(\lambda_\alpha), \dots, \xi_{\mu_\alpha-1}(\lambda_\alpha), \\ &\quad b, \xi_1(\lambda_{\alpha+1}), \dots, \xi_{\mu_{\alpha+1}}(\lambda_{\alpha+1}), \dots, \xi_{\mu_\nu}(\lambda_\nu) \}. \end{aligned}$$

Consider simple examples which illustrate the application of the criterion and Corollary 1.

Example 1. Because any matrix over a field of the complex numbers is reducible to the Jordan form, take up controllability of the system $\dot{x} = Ax + bu$ whose matrix A consists of two Jordan cells which are associated with different eigenvalues

$$A = \begin{pmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix}.$$

The eigenvector which is associated with λ_1 is $(1, 0, 0, 0, 0)^T$. The adjoint vectors are $(0, 1, 0, 0, 0)^T$ and $(0, 0, 1, 0, 0)^T$. The eigenvector which is associated with λ_2 is $(0, 0, 0, 1, 0)^T$ and the adjoint vector is $(0, 0, 0, 0, 1)^T$. By virtue of Corollary 1 it is necessary to check when the determinants made of eigen- and adjoint vectors for λ_1 and λ_2 are nonzero

$$\varphi(\lambda_1) = \begin{vmatrix} \lambda_1 & 0 & b_1 & 0 & 0 \\ 0 & 1 & b_2 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & b_4 & 1 & 0 \\ 0 & 0 & b_5 & 0 & 1 \end{vmatrix} = b_3, \quad \varphi(\lambda_2) = \begin{vmatrix} 1 & 0 & 0 & 0 & b_1 \\ 0 & 1 & 0 & 0 & b_2 \\ 0 & 0 & 1 & 0 & b_3 \\ 0 & 0 & 0 & 1 & b_4 \\ 0 & 0 & 0 & 0 & b_5 \end{vmatrix} = b_5$$

Consequently, the system is controllable iff $b_3 \neq 0$ and $b_5 \neq 0$.

Example 2. The matrix A consists of two Jordan cells that are associated with one eigenvalue λ :

$$A = \begin{pmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 \\ 0 & 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & 0 & \lambda \end{pmatrix}, \quad B = \begin{pmatrix} b_1 & d_1 \\ b_2 & d_2 \\ b_3 & d_3 \\ b_4 & d_4 \\ b_5 & d_5 \end{pmatrix}.$$

The eigenvectors are in this case $(1, 0, 0, 0, 0)^T$, $(0, 0, 0, 1, 0)^T$ and the adjoint vectors $(0, 1, 0, 0, 0)^T$, $(0, 0, 1, 0, 0)^T$ and $(0, 0, 0, 0, 1)^T$.

Then

$$\varphi_{11}(\lambda) = \begin{vmatrix} 1 & 0 & b_1 & 0 & 0 \\ 0 & 1 & b_2 & 0 & 0 \\ 0 & 0 & b_3 & 0 & 0 \\ 0 & 0 & b_4 & 1 & 0 \\ 0 & 0 & b_5 & 0 & 1 \end{vmatrix} = b_3, \quad \varphi_{12}(\lambda) = \begin{vmatrix} 1 & 0 & d_1 & 0 & 0 \\ 0 & 1 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 & 0 \\ 0 & 0 & d_4 & 1 & 0 \\ 0 & 0 & d_5 & 0 & 1 \end{vmatrix} = d_3,$$

$$\varphi_{21}(\lambda) = \begin{vmatrix} 1 & 0 & 0 & 0 & b_1 \\ 0 & 1 & 0 & 0 & b_2 \\ 0 & 0 & 1 & 0 & b_3 \\ 0 & 0 & 0 & 1 & b_4 \\ 0 & 0 & 0 & 0 & b_5 \end{vmatrix} = b_5, \quad \varphi_{22}(\lambda) = d_5.$$

Consequently, the system specified by the matrix A and the input matrix B is controllable iff

$$\text{rank} \begin{pmatrix} b_3 & d_3 \\ b_5 & d_5 \end{pmatrix} \geq 2$$

or at least equal to the number of Jordan cells that are associated with a given λ .

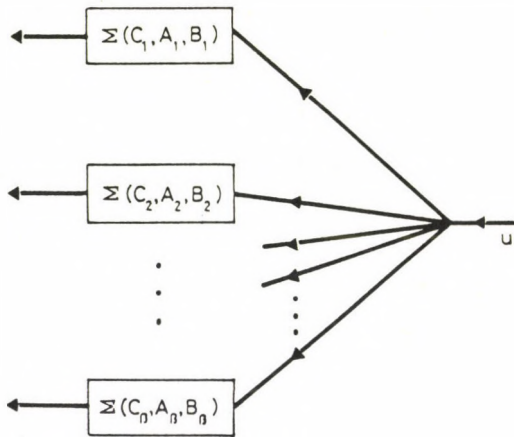


Fig. 1

Remark. In this Example the matrix B can not, by virtue of Corollary 5 below, consist of one column, since otherwise the system is certainly uncontrollable.

Corollary 2. If the system $\Sigma(C, A, B)$ is controllable, then, for any $\lambda \in C$, $IJ(\lambda, A) \leq \text{rank } B$.

This proposition immediately follows from conditions (1).

The above results make it possible to analyse the simultaneous controllability of a finite totality of independent processes, Fig. 1.

The state of the family $\{\Sigma(C_i, A_i, B_i)\}_{i=1}^{\beta}$ is obviously controllable if so is the component process

$$\begin{aligned} \dot{x}_i &= A_i x_i + B_i u, & i &= 1, 2, \dots, \beta. \\ y_i &= C_i x_i. \end{aligned}$$

For convenience, $\Sigma(C, A, B)$ will be said to have no internal resonance if all the eigenvalues of the operator are different. For the totality $\{\Sigma(C_i, A_i, B_i)\}_{i=1}^{\beta}$ the notion of external resonance is important. The totality of the systems $\{\Sigma(C_i, A_i, B_i)\}_{i=1}^{\beta}$ has no external resonance if $\sigma(A_i) \cap \sigma(A_j) = \emptyset$ for all $1 \leq i < j \leq \beta$, $IJ(\lambda, A_i)$ will be referred to as multiplicity of the internal resonance of frequency for the i -th system. $IJ(\lambda, A)$ is the multiplicity of the resonance of frequency λ for the system of Fig. 1 where A is a block diagonal matrix associated with the system

$$\dot{x}_i = A_i x_i + B_i u, \quad i = 1, 2, \dots, \beta.$$

Using this terminology, let us formulate

Corollary 3. If every process $\Sigma(C_i, A_i, B_i)$ of the family $\mathfrak{S} = \{\Sigma(C_i, A_i, B_i)\}_{i=1}^{\beta}$ is controllable and no external resonance is present, the family \mathfrak{S} is controllable.

The necessary condition for the controllability of the family \mathfrak{S} is proved by

Corollary 4. If the family $\mathfrak{S} = \{\Sigma(C_i, A_i, B_i)\}_{i=1}^\beta$ is controllable, then the multiplicity of the resonance of any frequency $\lambda \in \bigcup_i \sigma(A_i)$ does not exceed $\max_i \text{rank } B_i$.

In the application necessary conditions are very important, which can only be checked if the properties of the operator A are known. This includes Corollary 2 re-formulated into

Corollary 5. If $\text{rank } B < \max_{\lambda \in \sigma(A)} \text{IJ}(\lambda, A)$, the system $\Sigma(C, A, B)$ is not controllable.

Theorem 1 and its Corollaries are very helpful in analysing finite-dimensional systems because checking the rank Kalman-Krasovsky criterion reduces to analysing the spectra of subsystems.

Propositions, dual of Theorem 1 and its Corollaries, define the spectral observability criterion for linear systems. The system $\Sigma(C, A, B)$ is well-known [1, 36] to be observable iff the system $\Sigma(B^*, A^*, C^*)$ is controllable (A^* being an operator conjugate with A). This proposition and Theorem 1 yield the form of the observability criterion for finite-dimensional linear stationary systems. Let us limit ourselves here to formulating the observability criterion for the family $\{\Sigma(C_i, A_i, B_i)\}_{i=1}^\beta$. This criterion is dual of controllability of a family of systems. The structure of a system made of subsystems is shown in Fig. 2.

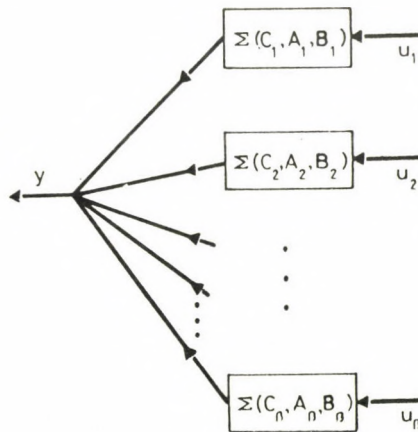


Fig. 2

It would be natural to assume that the family $\{\Sigma(C_i, A_i, B_i)\}_{i=1}^\beta$ is observable if so is the system

$$\dot{x}_i = A_i x_i + B_i u_i, \quad i = 1, 2, \dots, \beta,$$

$$y = \sum_{i=1}^{\beta} \bar{y}_i$$

where the first components of the vector \bar{y}_i are those of the output of the i -th subsystem and the remaining ones are equal to zero. The dimension of every \bar{y}_i ($1 \leq i \leq \beta$) is equal to the highest dimension of the subsystem outputs.

Using propositions dual of Theorem 1 and Corollaries 1 and 2 we have the following results on simultaneous observability.

THEOREM 2. With $\mathfrak{S} = \{\Sigma(C_i, A_i, B_i)\}_{i=1}^{\beta}$ being a family of finite-dimensional systems the following propositions are true:

b1) if every system $\Sigma(C_i, A_i, B_i)$ of the family $\mathfrak{S} = \{\Sigma(C_i, A_i, B_i)\}_{i=1}^{\beta}$ is observable and no external resonance occurs, then the family \mathfrak{S} is observable;

b2) if the family $\mathfrak{S} = \{\Sigma(C_i, A_i, B_i)\}_{i=1}^{\beta}$ is observable, then the multiplicity of the resonance of any frequency $\lambda \in \cup_i \sigma(A_i)$ does not exceed $\max_i \text{rank } C_i$.

These results make it possible to analyse the approximate observability and controllability of infinite-dimensional systems. However, let us first take up cases of observable and controllable finite-dimensional systems.

Example 3. A system of N oscillators has the form

$$\begin{aligned} \frac{d^2}{dt^2} x_i(t) + w_i^2 x_i(t) &= b_i u(t), \quad i = 1, 2, \dots, N, \\ y(t) &= \sum_{i=1}^N c_i x_i(t), \end{aligned}$$

where y is the system output. By virtue of Corollary 3 of Theorem 1 and of Theorem 2 this system is controllable and observable iff the natural frequencies of the oscillators are different ($w_i \neq w_j$ for $i \neq j$, or there is no resonance) and $b_i \neq 0$, $c_i \neq 0$ for any $i = 1, 2, \dots, N$.

Example 4. Consider a family of differential equations of the form

$$\begin{aligned} a_{n,i} \frac{d^{n_i} x_i(t)}{dt^{n_i}} + a_{n_i-1,i} \frac{d^{n_i-1} x_i(t)}{dt^{n_i-1}} + \dots + a_{0,i} x_i(t) &= b_i u_i(t), \\ y(t) &= \sum_{i=1}^N c_i x_i(t), \quad i = 1, 1, \dots, N \end{aligned} \quad (4)$$

where

$$c_i \neq 0, \quad b_i \neq 0, \quad a_{ij} \in \mathbb{C}, \quad a_{n_i,i} \neq 0 \quad (0 \leq j \leq n_i, \quad i = 1, 2, \dots, N).$$

Let us see if the system is controllable. Every i -th subsystem is obviously controllable in an n -dimensional space of variables $(x_i(t), \frac{d}{dt} x_i(t), \dots, \frac{d^{n_i-1}}{dt^{n_i-1}} x_i(t))$. By virtue of Corollaries 3 and 4, for controllability of the family it is necessary and sufficient that for any $1 \leq i < j \leq N$ the polynomials

$$\begin{aligned} g_i(\lambda) &= a_{n_i,i} \lambda^{n_i} + a_{n_i-1,i} \lambda^{n_i-1} + \dots + a_{0,i} = 0, \\ g_j(\lambda) &= a_{n_j,j} \lambda^{n_j} + a_{n_j-1,j} \lambda^{n_j-1} + \dots + a_{0,j} = 0 \end{aligned}$$

have no common roots. The latter is not true unless the resultant $\text{Res}(g_i, g_j)$ of the polynomials g_i and g_j is nonzero. Consequently, N differential equations such as (4) are not controllable iff for any $1 \leq i < j \leq N$

$$\begin{aligned} \text{Res}(g_i, g_j) = & \left| \begin{array}{cccccccccccc} a_{n,i} & a_{n-1,i} & \cdot & \cdot & \cdot & \cdot & \cdot & a_{1i} & a_{0i} & 0 & \cdot & \cdot \\ 0 & a_{n,i} & \cdot & \cdot & \cdot & \cdot & \cdot & a_{2i} & a_{1i} & a_{0i} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & a_{n,i} & a_{n-1,i} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & a_{0i} \\ a_{n,j} & a_{n-1,j} & \cdot & a_{1j} & a_{0j} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & a_{n,j} & \cdot & a_{2j} & a_{1j} & a_{0j} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & a_{n,j} & a_{n-1,j} & \cdot & \cdot & a_{0j} \end{array} \right| \begin{array}{l} \left. \vphantom{\begin{array}{c} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{array}} \right\} n_j \\ \left. \vphantom{\begin{array}{c} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{array}} \right\} n_i \end{array} \neq 0. \end{aligned}$$

If $\text{Res}(g_i, g_j) \neq 0$ with any $1 \leq i < j \leq N$, then by Theorem 2, this family is observable. What is more, this is the necessary condition for observability.

3. Approximate controllability and observability

Now, let us proceed to the approximate controllability of infinite-dimensional systems.

By Condition a2) the eigenvalues of the operator can be enumerated. Let us assume that the inequality $\text{Re } \lambda > 0$ holds only for the finite part of the sets $\sigma(A)$. Enumerate the set $\sigma(A)$ as follows

$$\text{Re } \lambda_1 \geq \text{Re } \lambda_2 \geq \text{Re } \lambda_3 \geq \dots \geq \text{Re } \lambda_n \geq 0 \geq \text{Re } \lambda_{n+1} \geq \dots \tag{5}$$

Every eigenvalue $\lambda_i \in \sigma(A)$ repeats the number of times equal to its multiplicity. Let $\xi(\lambda_i)$ be adjoint or eigenvector which is associated with λ_i and in the basis $\{\xi(\lambda_i)\}$ the first eigenvectors and then adjoint vectors are written in an increasing order of their power. This is always possible by virtue of Assumption a3). The sufficient condition of approximate controllability is formulated as

THEOREM 3. Let $Bu = \sum_{i=1}^m b_i u_i$ where $b_i \in H$ ($i = 1, 2, \dots, m$), A satisfy a1, a2 and a3, there exists only a finite set of $\lambda \in \sigma(A)$ such that $\text{Re } \lambda > 0$ over $\sigma(A)$, the enumeration (5) be introduced, and $\sigma(A) \subset \Pi_{\alpha, \beta} = \{z \in \mathbb{C}; -\alpha \leq \text{Re } z \leq \beta\}$ for some real $\alpha, \beta > 0$. If for a natural N the system

$$\dot{x}_N = A_N x_N + B_N u \tag{6a}$$

(where $H = H_N \oplus \tilde{H}_N$ is the spectral decomposition associated with the basis $\{\xi(\lambda_i)\}_{i=1}^N, \{\xi(\lambda_i)\}_{i=N+1}^\infty$; consequently, $A = A_N \oplus \tilde{A}_N$ and $B_N = P_N B$ where P_N is a natural mapping on H_N) is globally controllable, then the system

$$\dot{x} = Ax + Bu \tag{6b}$$

is approximately controllable.

Proof. Take any two points $z^1, z^2 \in H$ and design a control which moves the system (6a) from z_N^1 into z_N^2 , where $z_N^i = \sum_{j=1}^N z_{N_j}^i \xi(\lambda_j)$, $i = 1, 2$ are elements with the first N coordinates of the point z^i . By the condition of the Theorem the system $\dot{x}_N = A_N x_N + B_N u$ is controllable; therefore, at any $T \neq 0$ the matrix

$$U_N(T) = \int_0^T e^{-A_N \tau} B_N B_N^* e^{-A_N^* \tau} d\tau$$

is nonsingular. Consequently, the control

$$U_N(t) = B_N^* e^{-A_N^* t} U_N^{-1}(T) (e^{-A_N T} z_N^2 - z_N^1)$$

moves the system $\Sigma(A_N, B_N)$ from the state z_N^1 into z_N^2 within time T . Denote

$$\tilde{z}_N^i = \sum_{j=N+1}^{\infty} z_{N_j}^i \xi(\lambda_j).$$

Then to show that

$$\hat{z}_N^2 = e^{\tilde{A}_N T} \tilde{z}_N^1 + \int_0^T e^{\tilde{A}_N(T-\tau)} \tilde{B}_N B_N^* e^{-A_N^* \tau} d\tau U_N^{-1}(T) (e^{-A_N T} z_N^2 - z_N^1) + o(1) \quad (7)$$

as $N \rightarrow \infty$ (here $\tilde{B}_N = (I - P_N)B$, I is an identity mapping and $o(1) \rightarrow 0$ as $N \rightarrow \infty$) is to prove the Theorem. Let us first show that at some T

$$\|U_N^{-1}(T)\| \leq \gamma(T)$$

with any natural N . It is well-known that

$$\|U_N^{-1}(T)\|^{-1} \geq \min_{|z|_{\mathbb{C}^N}=1} \langle z, U_N(T)z \rangle_{\mathbb{C}^N},$$

where $\langle \cdot, \cdot \rangle_{\mathbb{C}^N}$ is a standard scalar product in \mathbb{C}^N . At the same time

$$\begin{aligned} \langle z, U_N(T)z \rangle_{\mathbb{C}^N} &= \int_0^T \sum_{i=1}^m |\langle e^{-A_N \tau} b_N^i, z \rangle|^2 d\tau \geq \\ &\geq \frac{1}{T} \sum_{i=1}^m \left(\int_0^T |\langle e^{-A_N \tau} b_N^i, z \rangle| d\tau \right)^2 \geq \\ &\geq \frac{1}{T} \sum_{i=1}^m \left(\max_{0 \leq t \leq T} |\langle A_N^{-1} e^{-A_N t} b_N^i, z \rangle - \langle A_N^{-1} b_N^i, z \rangle| \right)^2 \end{aligned}$$

(the latter inequality follows from the fact that

$$\int_0^T |\langle e^{-AN\tau} b_N^i, z \rangle| d\tau \geq \max_{0 \leq t \leq T} \left| \int_0^t \langle e^{-AN\tau} b_N^i, z \rangle d\tau \right|,$$

where b_N^i is the i -th column of the matrix

$$B_N = \{b_N^1, \dots, b_N^m\}.$$

Denote

$$C_N(T) = \sum_{i=1}^m \left(\min_{|z|=1} \max_{0 \leq t \leq T} |\langle A_N^{-1} e^{-ANt} b_N^i, z \rangle - \langle A_N^{-1} b_N^i, z \rangle| \right)^2 / T.$$

Then

$$\|U_N^{-1}(T)\| \leq \frac{1}{C_N(T)}.$$

It is easily seen that

$$0 \leq C_{N+1}(T) \leq C_N(T), \quad C_N(T + \Delta T) \geq C_N(T)$$

at any $\Delta T \geq 0$.

Consequently, there is a limit

$$\lim_{N \rightarrow \infty} C_N(T) = C_\infty(T).$$

Let us show that at some T it is true that $C_\infty(T) > 0$. Assume that the opposite is true, i.e. $C_\infty(T) = 0$ at any T . From the weak compactness of the sphere in H follows the existence of $|\hat{z}(T)| = 1$ such that

$$\langle A^{-1} e^{-At} b_i, \hat{z}(T) \rangle - \langle A^{-1} b_i, \hat{z}(T) \rangle = 0$$

with any $0 \leq t \leq T$, $1 \leq i \leq m$. Again, from the weak compactness of the sphere in H we have the existence of a point $|\hat{z}(\infty)| = 1$ such that

$$\langle A^{-1} e^{-At} b_i, \hat{z}(\infty) \rangle = \langle A^{-1} b_i, \hat{z}(\infty) \rangle = 0, \quad i = 1, 2, \dots, m$$

at any $0 \leq t \leq \infty$. The latter equality, together with the finiteness of the eigenvalues for which $\text{Re } \lambda > 0$ entail $\hat{z}(\infty) = 0$, is in conflict with the equality $|\hat{z}(\infty)| = 1$. Consequently, there exists $T > 0$ such that $C_\infty(T) > 0$. This implies a limited norm of the operator $U_N^{-1}(T)$ over N , i.e.

$$\|U_N^{-1}(T)\| < C_\infty^{-1}(T)$$

for any natural number N . From the latter inequality, the formula for control, the condition $\sigma(A) \subset \Pi_{\alpha, \beta}$, and the fact that $b_i \in H$ it follows that

$$\lim_{N \rightarrow \infty} \left\{ \int_0^T \|e^{\tilde{A}_N(T-\tau)} \tilde{B}_N B_N^* e^{-A_N^* \tau}\| d\tau C_\infty^{-1}(T) \|e^{-A_N T} z_N^2 - z_N^1\| \right\} = 0.$$

This proves the truth of the relation (7) and the Theorem.

This Theorem reduces the analysis of the approximate controllability of an infinite-dimensional system to that of the controllability of a totality of finite-dimensional systems. Numerous examples show that the controllability of the resultant finite-dimensional systems is most easily analysed by the method of Section 2. From Theorem 3 immediately follows the sufficient condition for the approximate controllability.

Corollary 1. If all the condition of Theorem 3 hold and for any N the system

$$\begin{aligned} \dot{x}_N &= A_N x_N + B_N u, \\ y &= C_N x_N \end{aligned}$$

is observable, where $C_N = CP_N$ and P_N is a natural mapping of H on H_N , then the system $\Sigma(C, A, B)$ is approximately observable.

Theorem 3 and its Corollaries are applicable to various oscillatory systems. For parabolic systems and diffusion systems a weaker version of Theorem 3 holds. To formulate this we will need

DEFINITION 4. The point $z_0 \in H$ is approximately reachable within $T > 0$ if, for any point $z \in H$ and any $\varepsilon > 0$, there are a control $u : [0, T] \rightarrow U$ and a point $\bar{z} \in O_\varepsilon(z)$ such that

$$\exp \left(\int_0^T (Ax(\tau) + Bu(\tau)) d\tau \right) \bar{z} \in O_\varepsilon(z_0).$$

A weaker analogue of Theorem 3 for parabolic systems is

THEOREM 4. If a1, a2, a3 hold, only a finite number of eigenvalues $\lambda \in \sigma(A)$ satisfy the inequality $\operatorname{Re} \lambda > 0$, the elements of $\sigma(A)$ are numerated as above, and there is $\delta > 0$ such that from $\lambda_j \neq \lambda_i$ it follows that $|\lambda_j - \lambda_i| > \delta > 0$ where δ is independent of i and j .

If for any natural N the system

$$\dot{x}_N = A_N x_N + B_N u$$

is controllable (A_N and B_N are as in the condition of Theorem 3), then for the system (6b) the zero element of the Hilbert space is approximately reachable.

Example which illustrate the application of the above methods are reported in Part II of the present article. In [11] sufficient conditions for approximate reachability of zero within finite time have also been obtained but with weaker constraints imposed on the operator.

4. Acknowledgement

The authors are grateful to academician S. V. Emelyanov for his important comments and constant active support of the research.

References

1. *Balakrishnan, A. V.*, Applied functional analysis, New York, Heidelberg, Berlin, Springer-Verlag, 1976.
2. *Bensoussan, A., Lions, J. L.*, Contrôle impulsionnel et inequations quasivariationnelles. Méthodes Mathématiques de l'Informatique, 11, Dunod, 1982.
3. *Butkovskiy, A. G.*, Metodi upravleniya sistemami s raspredelennymi parametrami, Moscow, Nauka, 1975, 568 p. (in Russian).
4. *Butkovskiy, A. G.*, Struktural'naya teoriya raspredelennykh sistem, Moscow, Nauka, 1977.
5. *Dyachenko, S. N.*, Spectral'noe razlozheniye beskonечnomernykh sistem s konechnomernym vkhodom, in "Slozh. syst. upr.", Kiev, 1987, pp. 52-57.
6. *Ilyin, V. A.*, Diff. uravn., Vol. 22, 1986.
7. *Lions, T. L.*, Contrôle des systemes distribués singuliers, Méthodes Mathématiques de l'Informatique, 13, Gauthier Villars, 1983.
8. *Lions, T. L.*, Some methods in the mathematical analysis of systems and their control, Beijing, Science Press, 1981.
9. *Lyashko, S. I., Maniakovsky, A. A.*, Upravlyaemost parabolicheskikh sistem s impulsnym upravleniyem, Dokl. Nauk USSR, 1989, Vol. 306, No. 2, pp. 276-279.
10. *Nefedov, S. A., Sholokhovich, F. A.*, Kriteriy stabiliziruемости dinamicheskikh sistem s konechnomernym vykhodom, Diff. Uravn., Vol. 22, No. 2, 1986, pp. 223-228.
11. *Shkliar, B. Sh.*, K upravlyaemosti lineinykh sistem s raspredelennymi parametrami, Dokl. Akad. Nauk USSR, 1989, Vol. 3-7, No. 3, pp. 560-563.
12. *Ahmed, N. V., Li, P.*, Stabilizability of perturbed linear systems on Hilbert spaces, Proc. 27th IEEE Conf. Decis. and Contr., Austin, Texas, Dec. 7-9, 1988, Vol. 3, New York (N. Y.), pp. 1972-1976.
13. *Bomer, A.* On control problems, Periodica Mathematica Hungarica, Vol. 20, No. 1, 1988, pp. 13-25.
14. *Balakrishnan, A. V.*, Boundary control of parabolic equations: L-Q-h-theory, Proc. Conf. on Theory of Nonlinear Equations, Sept. 1977, Berlin, Akademie Verlag, 1978.
15. *Cabrera, J. B. D., Furuta, K.*, Improving the robustness of Nussbaum-type regulator by the use of ϵ -modification — Local results, System & Control Letters, Vol. 12 (1989), pp. 421-429.
16. *Chen, G.*, Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain, Math. Pures Appl., 58, 1979, pp. 249-273.

17. *Chen, G.*, A note on the boundary stabilization of the wave equation, *SIAM J. Control Optim.*, Vol. **19**, 1981, pp. 106-113.
18. *Cirina, M. A.*, Boundary Controllability of Nonlinear Hyperbolic Systems, *SIAM J. Control*, Vol. **7**, 1969.
19. *Curtain, R. B.*, Finite dimensional compensators for some parabolic systems with unbounded control and observations, *SIAM J. Control and Optim.*, Vol. **22**, 1984, pp. 255-276.
20. *Curtain, R. F., Pritchard, A. J.*, Infinite-dimensional linear systems Theory, Lecture Notes in Control and Information Sci. **8**, Berlin, New York, Springer-Verlag, 1978.
21. *Curtain, R. F., Salamon, D.*, Finite-dimensional compensators for infinite-dimensional systems with unbounded input operators, *SIAM J. Control Optim.*, Vol. **24**, 1986, pp. 797-816.
22. *Kovayashi, T.*, Finite-dimensional adaptive control for infinite-dimensional systems, *Int. J. Control*, 1988-48, No. **1**, pp. 289-302.
23. *Kovayashi, T.*, Remarks on discrete-time servomechanism design for parabolic distributed parameter systems, *Int. J. Syst. Sci.*, 1988-19, No. **7**, pp. 1323-1333.
24. *Kunimatsu, N., Ito, K.*, Stabilization of a nonlinear distributed parameter vibratory system, *Int. J. Control*, 1988, Vo. **48**, No. **6**, pp. 2389-2415.
25. *Lasiecka, I., Triggiani, R.*, Finite rank, relatively bounded perturbations of semi-group generators, Part 1, *Ann. Scuola Norm. Sup. Pisa, Cl. Sci.*, Vol. **4**, **12**, 1985, pp. 641-668.
26. *Lagnese, J.*, Exact boundary value controllability of a class of hyperbolic equations, *SIAM J. Control and Optimiz.*, Vol. **16**, 1978, pp. 1000-1017.
27. *Lions, J. L.*, Exact controllability, stabilization and perturbations for distributed systems, *SIAM Review*, Vol. **30**, No. **1**, pp. 1-67.
28. *Nakagiri, S. I.* Identifiability of Linear Systems in Hilbert Spaces, *SIAM J. Control and Optim.*, **21**, 1983, pp. 501-530.
29. *Namou, T.*, Feedback Stabilization for Distributed Parameter Systems of Parabolic Type, *J. Diff. Eq.*, **33**, 1979, pp. 167-188.
30. *Ostalczyk, P.*, Controllability criteria based on matrix pencil description of a linear system, *Adv. Modell. and Simul.*, 1989, **16**, No. **2**, pp. 43-52.
31. *Pedersen, M.*, Boundary feedback stabilization of distributed parameter systems: An application of pseudo-differential boundary operators, *Proc. 27th IEEE Conf. Decis. and Contr.*, Austin, Texas, Dec. 7-9, 1988, Vol. **1**, , New York, pp. 366-368.
32. *Pohjolainen, S. A.*, Robust multivariable PI-controller for infinite-dimensional systems, *IEEE Trans. Aut. Control*, 1982, Vol. **AC-27**, No. **1**, pp. 17-30.
33. *Polak, E., Harn, Y-P.*, On the design of finite-dimensional stabilizing compensators for infinite-dimensional feedback systems via semi-infinite optimization, *Proc. 27th IEEE Conf. Decis. and Contr.*, Austin, Texas, Dec. 7-9, 1988, Vol. **3**, New York (N.Y.), 1988, pp. 2453-2458.
34. *Prätzel Wolters, D., Ilchmann, A., Owens, D. H.*, High-gain robust adaptive controllers for multivariable systems, *Systems Control Letters*, Vol. **8**, 1987, pp. 397-404, North-Holland.
35. *Rebarber, R.*, Conditions for stability of distributed parameter systems, *Proc. 27th IEEE Conf. Decis. and Contr.*, Austin, Texas, Dec. 7-9, 1988, Vol. **1**, New York (N.Y.), 1988, pp. 369-372.
36. *Russel, D. L.*, Controllability and stabilizability theory for partial differential equations. Recent progress and open questions, *SIAM Review*, Vol. **20**, 1978, pp. 639-739.
37. *Sakawa, Y.*, Feedback control of second-order evolutions with damping. *SIAM J. Contr. Optim.*, Vol. **22**, (3), 1984.
38. *Sakawa, Y.*, Feedback control of second-order evolutions equations with unbounded observation, *Int. J. Control*, Vol. **41**, (3), 1985, pp. 717-733.

39. Sakawa, Y., Matsuno, F., Fukushima, S., Modeling and feedback control of a flexible arm, *J. Robotic Syst.*, Vol. 2, No. 4, 1985.
40. Triggiani, R., On the stability problem in Banach space, *J. of Mathematical Analysis and Applications*, Vol. 52, 1975, pp. 383-403.
41. You, Y., Controllability and stabilizability of vibrating simply supported plate with pointwise control, *Advance in applied mathematics*, Vol. 10, 1989, pp. 324-343.
42. Jreugering, G., Schmidt Georg, E. J. P., Boundary control of a vibrating plate with internal damping, *Math. Meth. App. Sci.*, 1989, Vol. 11, No. 5, pp. 573-586.
43. Operator methods for optimal control problems, Ed. Zee Sung T., New York, Basel: Marcel Dekker, 1987, X, 316 p. (Pure and App. Math., Vol. 108).

Бесконечномерные системы: аппроксимативная управляемость и наблюдаемость. Часть I

С. К. КОРОВИН, М. Г. НИКИТИНА, С. В. НИКИТИН

(Москва)

Работа посвящена аппроксимативной управляемости и наблюдаемости бесконечномерных систем с конечномерным входом и выходом. Для анализа управляемости (наблюдаемости) систем большой размерности вида $\dot{x} = Ax + Bu$ предложена модификация критерия Калмана, позволяющая по спектральным свойствам оператора A судить об управляемости (наблюдаемости) систем. Приведены критерии управляемости (наблюдаемости) семейства конечномерных систем. Предложен метод анализа аппроксимативной управляемости (наблюдаемости) бесконечномерных систем, основанный на процедуре Фурье. Метод прост и удобен в применении. Результаты работы иллюстрированы примерами.

С. К. Коровин
ВНИИ системных исследований
СССР, 117312, Москва, В-312,
пр. 60летия Октября, 9

Typesetting by TYPOT_EX Kft, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

or to L. GYÖRFI
Technical University of Budapest
H-1111 Budapest, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

The papers, preferably in English or Russian, should be typed double spaced on one side of good-quality paper with wide margins (4–5 cm). The first page of the paper should carry the title, the author(s) names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as $\text{diag}(d_1, d_2, \dots, d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском языке в трех экземплярах, на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылаются верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ


<i>Vaněček A.</i> : Strongly nonlinear and other control systems (<i>Ванечек А.</i> Строго нелинейные и другие системы управления)	3
<i>Kurzanski A. B., Pschenichnyi B. N., Pokotilo V. G.</i> : Optimal inputs for guaranteed identification (<i>Куржанский А. Б., Пиеничный Б. Н., Покотило В. Г.</i> Оптимальные входы в задаче гарантированной идентификации)	13
<i>Taras'ev A. M.</i> : The function of an optimal guaranteed result of control problems with a vector criterion (<i>Тарасев А. М.</i> Функция оптимального гарантированного результата в задачах управления с векторным критерием)	25
<i>Chernyak A. I., Sztrik J.</i> : Asymptotic behaviour of a complex renewable standby system with fast repair (<i>Черняк А. И., Стрик Я.</i> Асимптотическое поведение сложной резервированной системы с быстрым восстановлением)	37
<i>Krasovskii A. A.</i> : Optimization and stochastic dynamics in the state space (<i>Красовский А. А.</i> Оптимизация и стохастическая динамика в пространстве состояний)	45
<i>Korovin S. K., Nikitina M. G. and Nikitin S. V.</i> : Infinite-dimensional systems: Approximate controllability and observability. Part I (<i>Коровин С. К., Никитина М. Г., Никитин С. В.</i> Бесконечномерные системы: аппроксимативная управляемость и наблюдаемость. Часть I)	59

316920

9

VOL. 20 • NUMBER 2
TOM HOMEP

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



PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
УПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1991

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

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

- theory of control processes
- theory of adaptive systems
- theory of estimation and identification
- theory of controlling robot-technologic and flexible manufacturing systems
- information theory
- information-theoretic aspects of multiple access networks.

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

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

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

- теория процессов управления;
- теория адаптивных систем;
- теория оценивания и идентификации;
- теория управления робототехническими и гибкими производственными системами;
- теория информации;
- теория информации в области сетей с множественным доступом.

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

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czech and Slovak Federal Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia:

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

For all other countries:

PERGAMON PRESS PLC Headington Hill Hall, Oxford OX3 0BW, England

or

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

1991 Subscription Rate DM 627,— per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITOR

N. N. KRASOVSKII (USSR)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

K. V. FROLOV

A. B. KURZHANSKI

I. A. OVSEEVICH

E. D. TERYAEV

R. Z. KHASHMINSKII

HUNGARY

T. VÁMOS

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

L. GYÖRFI

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJC

I. VAJDA

РЕДАКТОР ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

К. В. ФРОЛОВ

А. Б. КУРЖАНСКИЙ

И. А. ОВСЕВИЧ

Е. Д. ТЕРЯЕВ

Р. З. ХАСЬМИНСКИЙ

ВНР

Т. ВАМОШ

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

Л. КЕВИЦКИ

Л. ДЪЕРФИ

Я. КОЧИШ

ЧССР

Й. БЕНЕШ

В. СТРЕЙЦ

И. ВАЙДА



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

 MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

DEFINITION AND RECOGNITION OF CLASSICAL SETS BY THE ROUGH ONES

I. KRAMOSIL

(Prague)

(Received November 28, 1990)

Rough sets are a rather new branch of modern mathematics with interesting applications. Here we show how rough sets can be used to describe and solve the problem of statistical induction in a way as general as possible, covering a number of important particular cases.

1. A model of statistical induction based on rough sets

From the viewpoint of its origins the notion of rough sets naturally follows from that of indiscernibility relation. Let X be a nonempty set, let \mathcal{A} be a nonempty collection of predicates applicable to all elements of X . Hence for each $x \in X$, $A \in \mathcal{A}$, $A(x)$ is a well-formed formula of a formalized language which is either true or false; the truth-value of $A(x)$ is denoted by $Tv(A(x))$. Define for each $x, y \in X$,

$$x \approx y \stackrel{\text{def}}{\iff} (\forall A \in \mathcal{A})(Tv(A(x)) = Tv(A(y))). \quad (1)$$

Evidently, \approx is an equivalence relation on X . It is called *indiscernibility relation on X with respect to \mathcal{A}* , in order to pick up the fact that if $x \approx y$, then there is no possibility to discern between x and y using the predicates from \mathcal{A} . The definition can be extended to empty set \mathcal{A} of predicates setting, in this case $x \approx y$, for each $x, y \in X$. In general, any equivalence relation on X can be taken as an indiscernibility relation.

Let \approx be an indiscernibility relation on X , let $V \subset X$. Set

$$\underline{V} = \{x : x \in V, (\forall y \in X)(y \approx x \Rightarrow y \in V)\} \quad (2)$$

$$\overline{V} = \{x : x \in X, (\exists y \in V)(y \approx x)\}. \quad (3)$$

Denoting, for all $x \in X$, by $[x]$ the corresponding equivalence class, i.e.

$$[x] = \{y : y \in X, y \approx x\}, \quad (4)$$

we may write

$$\underline{V} = U\{[x] : [x] \subset V\}, \quad \overline{V} = U\{[x] : [x] \cap V \neq \emptyset\}. \tag{5}$$

Informally, \underline{V} is the set of all elements which can be surely stated to be in V on the ground of the truth-values of the predicates generating the indiscernibility relation in question. $X - \overline{V}$ is the set of elements which are certainly not in V , and for elements from $\overline{V} - \underline{V}$ the membership relation with respect to V cannot be evaluated within the given framework. The pair $\langle \underline{V}, \overline{V} \rangle$ is called *rough set* generated by V and \approx , cf., e.g. [4]. Evidently, each pair $\langle \underline{W}, \overline{W} \rangle$, $\underline{W} \subset \overline{W}$ of subsets of X can be taken as rough set generated by a set W , $\underline{W} \subset W \subset \overline{W}$, and by the equivalence (indiscernibility) relation \approx such that $x \approx y$ iff either $x, y \in \underline{W}$, or $x, y \in \overline{W} - \underline{W}$, or, finally, $x, y \in X - \overline{W}$.

In order to avoid the technical difficulties connected with the measurability of the corresponding mappings let us limit ourselves, in the sequel, only to finite or countably infinite spaces. Let $A = \{a_1, a_2, \dots\}$ be a nonempty countable set, let $\mathcal{V} = \{ \langle \underline{V}(i), \overline{V}(i) \rangle \}_{i=1}^{\infty}$ be a sequence of rough sets in A , i.e. $\underline{V}(i) \subset \overline{V}(i) \subset A$ for all $i = 1, 2, \dots$. Suppose that $\bigcup_{i=1}^{\infty} \underline{V}(i) = \bigcap_{i=1}^{\infty} \overline{V}(i)$ and denote by V this subset of A ; we say that \mathcal{V} defines V . Perhaps a weaker way of definition of V by $\{ \langle \underline{V}(i), \overline{V}(i) \rangle \}$ could be also considered but here we prefer the most simple, even if also a rather restrictive one.

Denoting by χ_V the characteristic function (identifier) of V as a subset of A and supposing that \mathcal{V} defines V we immediately have

$$\chi_V(x) = \sup_{1 \leq i < \infty} \chi_{\underline{V}(i)}(x) = \inf_{1 \leq i < \infty} \chi_{\overline{V}(i)}(x) \tag{6}$$

for each $x \in A$. For the sake of unambiguity let us recall explicitly, that $\chi_{V(i)}(x) = 1$, if $x \in \underline{V}(i)$, $\chi_{V(i)}(x) = 0$, if $x \in A - \underline{V}(i)$.

Take an $x \in A$, if \mathcal{V} is finite, or, what is the same, if $\underline{V}(i) = \emptyset$ and $\overline{V}(i) = A$ for all but a finite number of indices, then $\chi_V(x)$ can be effectively computed by (6) using $\chi_{\overline{V}(i)}$ or $\chi_{\underline{V}(i)}$. If \mathcal{V} is infinite, this method is theoretically ineffective; in practice, it is ineffective also for finite but very large \mathcal{V} 's. However, an immediately and intuitive idea yields that if we "sample at random" a finite number i_1, i_2, \dots, i_n of indices, then

$$\sup_{1 \leq j \leq n} \chi_{\underline{V}(i_j)}(x), \tag{7}$$

or

$$\inf_{1 \leq j \leq n} \chi_{\overline{V}(i_j)}(x) \tag{8}$$

approximate, under certain regularity conditions and in a reasonable sense, the desired value $\chi_V(x)$. In fact,

$$\sup_{1 \leq j \leq n} \chi_{\underline{V}(i_j)}(x) \leq \chi_V(x) \leq \inf_{1 \leq j \leq n} \chi_{\overline{V}(i_j)}(x). \tag{9}$$

If the set A is finite, such an approximative decision can be taken, sequentially, for all $x \in A$, in order to obtain an approximation for the set V as a whole. In the opposite case, i.e. for an infinite A , we may, again, apply the same elementary idea of statistical approximation, sampling at random some x_1, x_2, \dots, x_m from A and approximating V either by

$$\left\{ x_k : k \leq m, \sup_{1 \leq j \leq n} \chi_{\underline{V}(i_j)}(x_k) = 1 \right\} \tag{10}$$

or

$$\left\{ x_k : k \leq m, \inf_{1 \leq j \leq n} \chi_{\overline{V}(i_j)}(x_k) = 1 \right\}. \tag{11}$$

These informal reasonings can be formalized as follows.

Let $\langle \Omega, \mathcal{S}, P \rangle$ be an abstract probability space, hence, Ω is a nonempty set, \mathcal{S} is a σ -field of subsets of Ω , and P is a probability measure on \mathcal{S} . Let $\mathbb{N} = \{0, 1, 2, \dots\}$ be the set of non-negative integers, let $\mathbb{N}^+ = \mathbb{N} - \{0\}$. Let $\mathcal{X} = \{X_i\}_{i=1}^\infty$ be a sequence of random variables (i.e. measurable mappings) defined on $\langle \Omega, \mathcal{S}, P \rangle$ and taking their values in \mathbb{N}^+ , let $\mathcal{Y} = \{Y_i\}_{i=1}^\infty$ be another such sequence of random variables. Let $\mathcal{V} = \{\{\underline{V}(i), \overline{V}(i)\}\}_{i=1}^\infty$ be a sequence of rough sets which defines the set $V \subset A = \{a_1, a_2, \dots\}$. Set, by induction,

$$\mathcal{V}_{*0} = \emptyset, \quad \mathcal{V}_0^* = A, \tag{12}$$

$$\begin{aligned} \mathcal{V}_{*,i} &= \mathcal{V}_{*,i}(\omega) = \mathcal{V}_{*,i-1}(\omega) \cup \{a(X_i(\omega))\}, \\ &\quad \text{if } a(X_i(\omega)) \in \underline{V}(Y_i(\omega)), \end{aligned} \tag{13}$$

$$\mathcal{V}_{*,i} = \mathcal{V}_{*,i-1} \text{ otherwise.}$$

For the sake of notational simplicity we write $a(i)$ instead of a_i for the elements of the basic space A .

$$\begin{aligned} \mathcal{V}_i^* &= \mathcal{V}_i^*(\omega) = \mathcal{V}_{i-1}^*(\omega) - \{a(X_i(\omega))\}, \\ &\quad \text{if } a(X_i(\omega)) \in A - \overline{V}(Y_i(\omega)), \\ \mathcal{V}_i^* &= \mathcal{V}_{i-1}^* \text{ otherwise.} \end{aligned} \tag{14}$$

Evidently, $\mathcal{V}_{*,i}(\omega) \subset V \subset \overline{\mathcal{V}}_i^*(\omega)$ for each $\omega \in \Omega$ and $i = 1, 2, \dots$, so that the sets $\mathcal{V}_{*,i}$ and \mathcal{V}_i^* may serve as first and very rough approximations of the unknown set V . Given a random variable Z which takes $\langle \Omega, \mathcal{S}, P \rangle$ into A , the quality of the approximation $\langle \mathcal{V}_{*,i}, \mathcal{V}_i^* \rangle$ of V can be quantitatively measured by

$$P(\{\omega : \omega \in \Omega, Z(\omega) \in \mathcal{V}_i^* - \mathcal{V}_{*,i}\}); \tag{15}$$

the closer to zero this value may be, the better the approximation. In what follows, we shall investigate the conditions under which the value of (15) can be as close to

zero as desired taking i large enough, and with respect to a sufficiently large class of random variables Z .

It is perhaps worth of explicit introducing that the *stochastic* nature of this approximation process is principal, its *deterministic* alternative being, evidently, of almost no worth and sense.

2. Basic theorem for the proposed statistical induction model

The pair $\langle \mathcal{X}, \mathcal{Y} \rangle$ of sequences of random variables defined above, with each X_i and Y_i taking (Ω, \mathcal{S}, P) into \mathbb{N}^+ , is called *regular*, if the vector random variables $\langle X_i, Y_i \rangle$, $i = 1, 2, \dots$, are mutually statistically independent; in symbols, if for all $m \in \mathbb{N}^+$, $\langle i_1, i_2, \dots, i_m \rangle \in (\mathbb{N}^+)^m$, and $\langle z_1, z_2, \dots, z_m \rangle \in (\mathbb{N}^+ \times \mathbb{N}^+)^m$,

$$\begin{aligned} & P\left(\bigcap_{j=1}^m \{\omega : \omega \in \Omega, \langle X_{i_j}(\omega), Y_{i_j}(\omega) \rangle = z_j\}\right) \\ &= \prod_{j=1}^m P(\{\omega : \omega \in \Omega, \langle X_{i_j}(\omega), Y_{i_j}(\omega) \rangle = z_j\}), \end{aligned} \quad (16)$$

and if, moreover, for all $z \in \mathbb{N}^+ \times \mathbb{N}^+$,

$$\sum_{j=1}^{\infty} P(\{\omega : \omega \in \Omega, \langle X_{i_j}(\omega), Y_{i_j}(\omega) \rangle = z\}) = \infty. \quad (17)$$

In what follows, we shall omit the symbols $\dots \omega : \omega \in \Omega \dots$ in expressions like (16) or (17), supposing that no misunderstanding menaces. Sequence \mathcal{X} is called *independent identically and non-trivially distributed* (i.i.n.d.-sequence, in short), if the random variables X_i are statistically independent and if, moreover, for each $i, j \in \mathbb{N}^+$

$$P(\{X_i(\omega) = j\}) = P(\{X_1(\omega) = j\}) > 0. \quad (18)$$

A very simple example of regular pairs of sequences can be obtained as follows. Two i.i.n.d.-sequences \mathcal{X} and \mathcal{Y} are called *independent*, if $\{X_1, Y_1, X_2, Y_2, \dots\}$ is a sequence of mutually statistically independent random variables.

Lemma 1. Let \mathcal{X}, \mathcal{Y} be independent i.i.n.d.-sequences, then $\langle \mathcal{X}, \mathcal{Y} \rangle$ is a regular pair of sequences. \square

Proof. Let $m \in \mathbb{N}^+$, $\langle i_1, \dots, i_m \rangle \in \mathbb{N}^m$, $\langle z_1, \dots, z_m \rangle \in (\mathbb{N}^+ \times \mathbb{N}^+)^m$, let $z_i = \langle n_i^1, n_i^2 \rangle$ for each $i \leq m$, then

$$\begin{aligned}
 & P\left(\bigcap_{j=1}^m \{ \langle X_{i_j}(\omega), Y_{i_j}(\omega) \rangle = \langle n_j^1, n_j^2 \rangle \}\right) = \\
 & = P\left(\bigcap_{j=1}^m (\{X_{i_j}(\omega) = n_j^1\} \cap \{Y_{i_j}(\omega) = n_j^2\})\right) = \\
 & = \left(\prod_{j=1}^m P(\{X_{i_j}(\omega) = n_j^1\}) \cdot \prod_{j=1}^m P(\{Y_{i_j}(\omega) = n_j^2\})\right) = \tag{19} \\
 & = \prod_{j=1}^m (P(\{X_{i_j}(\omega) = n_j^1\}) \cdot P(\{Y_{i_j}(\omega) = n_j^2\})) = \\
 & = \prod_{j=1}^m P(\{X_{i_j}(\omega) = n_j^1, Y_{i_j}(\omega) = n_j^2\})
 \end{aligned}$$

so that (16) holds. Moreover, for each $\langle n^{(1)}, n^{(2)} \rangle \in \mathbb{N}^+ \times \mathbb{N}^+$,

$$\begin{aligned}
 & \sum_{j=1}^{\infty} P(\{X_j(\omega) = n^{(1)}, Y_j(\omega) = n^{(2)}\}) = \\
 & = \sum_{j=1}^{\infty} P(\{X_j(\omega) = n^{(1)}\}) \cdot P(\{Y_j(\omega) = n^{(2)}\}) = \tag{20} \\
 & = \sum_{j=1}^{\infty} P(\{X_1(\omega) = n^{(1)}\}) \cdot P(\{Y_1(\omega) = n^{(2)}\}) = \infty
 \end{aligned}$$

due to (18), so that (17) holds. The lemma is proved. □

THEOREM 1 (SOUNDNESS THEOREM). Let $\mathcal{V} = \{\langle \underline{V}(i), \overline{V}(i) \rangle\}_{i=1}^{\infty}$ be a sequence of rough sets which defines the set $V \subset A = \{a_1, a_2, \dots\}$, let $\langle \mathcal{X}, \mathcal{Y} \rangle$, $\mathcal{X} = \{X_i\}_{i=1}^{\infty}$, $\mathcal{Y} = \{Y_i\}_{i=1}^{\infty}$, be a regular pair of systems of random variables with each X_i and Y_i taking $\langle \Omega, \mathcal{S}, P \rangle$ into \mathbb{N}^+ , then

$$P\left(\left\{ \lim_{n \rightarrow \infty} \mathcal{V}_{*,n}(\omega) = \lim_{n \rightarrow \infty} \mathcal{V}_n^*(\omega) = V \right\}\right) = 1. \tag{21}$$

□

Proof. Due to the definition of $\mathcal{V}_{*,n}$ and \mathcal{V}_n^* , a necessary (but not sufficient) condition for $a \in \mathcal{V}_n^*$ reads that $a \in \underline{V}(i)$ for some $i \in \mathbb{N}^+$, hence, $a \in \bigcup_{i=1}^{\infty} \underline{V}(i) = \underline{V}$, so that $\mathcal{V}_{*,n} \subset V$. Dually, a necessary (but not sufficient) condition for $a \in \mathcal{V}_{*,n}$ reads that $a \in A - \overline{V}(i)$ for some $i \in \mathbb{N}^+$, hence $a \in \bigcup_{i=1}^{\infty} (A - \overline{V}(i)) =$

$= A - \bigcap_{i=1}^{\infty} \overline{V}(i) = A - V$, so that $A - \mathcal{V}_n^* \subset A - V$ and $V \subset \mathcal{V}_n^*$. Consequently, for each $\omega \in \Omega$, $n \in \mathbb{N}^+$,

$$\mathcal{V}_{*,n}(\omega) \subset \mathcal{V}_{*,n+1}(\omega) \subset \lim_{n \rightarrow \infty} \mathcal{V}_{*,n} \subset \lim_{n \rightarrow \infty} \mathcal{V}_n^*(\omega) \subset \mathcal{V}_{n+1}^*(\omega) \subset \mathcal{V}_n^*(\omega). \quad (22)$$

Set, for each $j \in \mathbb{N}^+$,

$$\lambda(j) = \{i : i \in \mathbb{N}^+, a_j \in \underline{V}(i) \cup (A - \overline{V}(i))\}. \quad (23)$$

If $a_j \in V$, then $\lambda(j) = \{i : a_j \in \underline{V}(i)\} \neq \emptyset$, as $\bigcup_{i=1}^{\infty} \underline{V}(i) = V$ implies that there exists $i \in \mathbb{N}^+$ such that $a_j \in \underline{V}(i)$. A necessary and sufficient condition for $a_j \in \bigcup_{n=1}^{\infty} \mathcal{V}_{*,n}(\omega) (= \lim_{n \rightarrow \infty} \mathcal{V}_{*,n}(\omega))$ reads that there exists $i \in \mathbb{N}^+$ such that $X_i(\omega) = j$ and $Y_i(\omega) \in \lambda(j)$. The probability of this random event reads:

$$\begin{aligned} & P\left(\bigcup_{i=1}^{\infty} \{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\}\right) = \\ & = P\left(\Omega - \bigcap_{i=1}^{\infty} (\Omega - \{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})\right) = \\ & = 1 - P\left(\bigcap_{i=1}^{\infty} (\Omega - \{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})\right) = \\ & = 1 - \prod_{i=1}^{\infty} (1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})). \end{aligned} \quad (24)$$

This value is one iff $\sum_{i=1}^{\infty} P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\}) = \infty$, which is true because of (17) and of the fact that $\lambda(j) \neq \emptyset$. Hence,

$$P\left(\left\{\lim_{n \rightarrow \infty} \mathcal{V}_{*,n}(\omega) = V\right\}\right) = 1. \quad (25)$$

If $a_j \in A - V$, then the necessary and sufficient condition for $a_j \in A - \lim_{n \rightarrow \infty} \mathcal{V}_n^*(\omega)$ formally is the same as above, but now $\lambda(j) = \{i : a_j \in A - \overline{V}(i)\}$. Hence,

$$P\left(\left\{V = \lim_{n \rightarrow \infty} \mathcal{V}_n^*(\omega)\right\}\right) = 1, \quad (26)$$

and the theorem is proved. \square

Common sense and mathematician's everyday experience yield that limit results like (21) are important when the theoretical or philosophical correctness and soundness of an algorithm are investigated, but from the practical point of view the non-limit properties are ultimately decisive. Therefore, in the sequel, an appropriately defined "speed of convergence" of $\mathcal{V}_{*,n}$ and \mathcal{V}_n^* to V will be investigated.

3. Some results concerning the speed of convergence

Considering the same model and notations as above, an element $a_j \in A$ is called *undecided* in the n -th step, if $a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)$. Set, for each $\varepsilon > 0$,

$$L(\varepsilon, j) = \min\{n : P(\{a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) \leq \varepsilon\}, \tag{27}$$

hence, $L(\varepsilon, j)$ denotes the (minimal) number of steps necessary to decide about a_j with the probability at least $1 - \varepsilon$. If Z is a random variable taking (Ω, \mathcal{S}, P) into \mathbb{N}^+ , then $LE_Z(\varepsilon)$ will denote the expected value of $L(\varepsilon, j)$ with respect to Z , i.e.,

$$LE_Z(\varepsilon) = \sum_{j=1}^{\infty} [L(\varepsilon, j)P(\{Z(\omega) = j\})]. \tag{28}$$

THEOREM 2. Under the conditions of Theorem 1, $L(\varepsilon, j)$ is finite for all $\varepsilon > 0$ and for all $j \in \mathbb{N}^+$. If, moreover, for all $i \in \mathbb{N}^+$ and for $\lambda(j)$ defined by (23),

$$P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\}) \geq Q_j \tag{29}$$

for some $Q_j > 0$, then $L(\varepsilon, j) \leq Q_j^{-1} \ln(1/\varepsilon)$. □

Proof. Fix $j \in \mathbb{N}^+$, then $a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)$ iff for no $i \leq n$ the random events $X_i(\omega) = j, Y_i(\omega) \in \lambda(j)$ simultaneously occur. Due to the supposed statistical independence of vector random variables $\langle X_i, Y_i \rangle$,

$$P(\{a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) = \prod_{i=1}^n \left(1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})\right). \tag{30}$$

Hence, this probability is majorized by an $\varepsilon > 0$, if

$$\sum_{i=1}^n \ln\left(1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})\right) < \ln \varepsilon. \tag{31}$$

As $\ln(1 - x) \leq -x$ for each $0 \leq x < 1$, a sufficient condition for (31) reads

$$\sum_{i=1}^n P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\}) \geq \ln(1/\varepsilon). \tag{32}$$

Relation (17) together with the fact that $\lambda(j) \neq \emptyset$ for each $j \in \mathbb{N}^+$ imply that there exists a finite n_0 satisfying (32), evidently $L(\varepsilon, j) \leq n_0$. Under the supplementary condition (29) a sufficient condition for (32) reads $n_0 Q_j \geq \ln(1/\varepsilon)$ from which the assertion immediately follows.

As can be easily seen, the conditions of Theorem 2 do not permit to state that $LE_Z(\varepsilon) < \infty$ for all random variables Z , neither when (29) holds. Or, take \mathcal{X} and \mathcal{Y} in such a way that

$$P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\}) = 2^{-k} \quad (33)$$

for all $i = 2^k$, $k \in \mathbb{N}^+$, this probability being equal to a $\delta_z > 0$ independently of i for all other pairs $z = \langle n^1, n^2 \rangle \in \mathbb{N}^+ \times \mathbb{N}^+$ of indices. Suppose that (16) holds, then, evidently, (17) holds as well and $\langle \mathcal{X}, \mathcal{Y} \rangle$ satisfies the conditions of Theorem 1. Let Z take $\langle \Omega, \mathcal{S}, P \rangle$ into \mathbb{N}^+ in such a way that for all $k \in \mathbb{N}^+$

$$P(\{Z(\omega) = 2^k\}) = 2^{-k}; \quad (34)$$

for other indices their probability of sampling by Z is trivially zero. Then $L(\varepsilon, 2^k) = 2^k \ln(1/\varepsilon)$, so that $LE_Z(\varepsilon) = \infty$ by (28).

Corollary. If the conditions of Theorem 2 and (29) hold, then for each $a \in A$ the probability that a will not be decided yet in the n -th step tends exponentially to zero with n increasing (an immediate consequence of (30)).

Let us consider the case with finite sets A and \mathcal{V} and with uniform probability distribution over these sets. This situation can be formally embedded within the presented formalization as follows.

THEOREM 3. Let $\mathcal{X} = \{X_i\}_{i=1}^\infty$, $\mathcal{Y} = \{Y_i\}_{i=1}^\infty$ be two sequences of statistically independent random variables taking $\langle \Omega, \mathcal{S}, P \rangle$ into \mathbb{N}^+ and such that, for each $i \in \mathbb{N}^+$,

$$\begin{aligned} P(\{X_i(\omega) = j\}) &= 1/N, & \text{if } 1 \leq j \leq N, \\ P(\{Y_i(\omega) = k\}) &= 1/m, & \text{if } 1 \leq k \leq m, \\ P(\{X_i(\omega) = j\}) &= P(\{Y_i(\omega) = k\}) = 0 & \text{if } j > N, k > m. \end{aligned} \quad (35)$$

Mutual statistical independence of all X_i 's and Y_i 's also supposed. Let Z be a random variable taking $\langle \Omega, \mathcal{S}, P \rangle$ into A , statistically independent of each X_i and Y_i and with the uniform probability distribution over $\{1, 2, \dots, m\}$. Let

$$(\forall j \leq N)(\exists l, k \leq m) \left[a_j \in \left(\underline{V}(l) \cap (A - \overline{V}(k)) \right) \right]. \quad (36)$$

Then, for each $j \leq N$,

$$\begin{aligned} P(\{a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) &= P(\{Z(\omega) \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) \leq \\ &\leq (1 - (mN)^{-1})^n, \end{aligned} \quad (37)$$

$$L(\varepsilon, j) \leq mN \ln(1/\varepsilon). \quad (38)$$

□

Proof. An easy calculation yields

$$\begin{aligned}
 & P(\{a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) \leq \\
 & \leq \prod_{i=1}^n \left(1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})\right) = \\
 & = \prod_{i=1}^n \left(1 - P(\{X_i(\omega) = j\})P(\{Y_i(\omega) \in \lambda(j)\})\right) = \tag{39} \\
 & = \prod_{i=1}^n \left(1 - N^{-1}m^{-1} \text{card}(\lambda(j) \cap \{1, 2, \dots, m\})\right) \leq \\
 & \leq (1 - (mN)^{-1})^n
 \end{aligned}$$

due to (36).

$$\begin{aligned}
 & P(\{Z(\omega) \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) = \\
 & = \sum_{j=1}^N P(\{Z(\omega) = a_j, a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) = \\
 & = \sum_{j=1}^N P(\{Z(\omega) = a_j\})P(\{a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)\}) = \\
 & = \sum_{i=1}^N N^{-1}(1 - (mN)^{-1})^n,
 \end{aligned}$$

as the supposed statistical independence of random variables Z and X_i, Y_i implies that the random events $Z(\omega) = a_j$ and $a_j \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)$ are also statistically independent. (38) immediately follows from Theorem 2, setting $Q_j = (mN)^{-1}$. \square

4. Statistical decision functions defined by rough sets

The pair $\langle \mathcal{V}_{*,n}(\omega), \mathcal{V}_n^*(\omega) \rangle$ defined above by $\mathcal{V} = \left\{ \langle \underline{V}(i), \overline{V}(i) \rangle \right\}_{i=1}^\infty$, $\mathcal{X} = \{X_i\}_{i=1}^\infty$, and $\mathcal{Y} = \{Y_i\}_{i=1}^\infty$, can be taken as a rough set which approximates, in the reasonable sense explained and proved above, the subset $V \subset A$, completely defined by $\cup_{i=1}^\infty \underline{V}(i)$ or by $\cap_{i=1}^\infty \overline{V}(i)$. From another point view, $\langle \mathcal{V}_{*,n}(\omega), \mathcal{V}_n^*(\omega) \rangle$ may be seen as a definition of a three-valued failure-proof decision function for the membership predicate for V , if $a \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)$ we are not able to decide. However, admitting the possibility of a probabilistically quantifiable error connected with the decisions, we may use $\langle \mathcal{V}_{*,n}(\omega), \mathcal{V}_n^*(\omega) \rangle$ in order to define a classical two-valued statistical decision function for the membership predicate for V as follows.

Let Z be a random variable taking $\langle \Omega, \mathcal{S}, P \rangle$ into \mathbb{N} set

$$p = \frac{P(\{a_{Z(\omega)} \in \mathcal{V}_{*,n}(\omega)\})}{P(\{a_{Z(\omega)} \in (\mathcal{V}_{*,n}(\omega) \cup (A - \mathcal{V}_n^*(\omega)))\})}. \quad (40)$$

Let U_p be a random variable taking $\langle \Omega, \mathcal{S}, P \rangle$ into the binary set $\{0, 1\}$ in such a way that $P(\{U_p(\omega) = 1\}) = p$. Both the random variables U, Z are mutually statistically independent and are also statistically independent of each X_i and Y_i . Let $D = \{d_0, d_1\}$ be a binary set of decisions, d_0 means "the tested element of A is in $A - V$ ", d_1 means "... is in V ". For $a \in A, m \in \mathbb{N}^+$, decision function $\delta_Z(a, n, \cdot)$ reads

$$\delta_Z(a, n, \omega) = d_1, \quad \text{or} \quad \delta_Z(a, n, \omega) = 1, \quad (41)$$

if either $a \in \mathcal{V}_{*,n}(\omega)$ or $a \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)$ and $U_p(\omega) = 1$,

$$\delta_Z(a, n, \omega) = d_0 \quad \text{or} \quad \delta_Z(a, n, \omega) = 0 \quad (42)$$

otherwise, i.e. if either $a \in A - \mathcal{V}_n^*(\omega)$ or $a \in \mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)$ and $U_p(\omega) = 0$.

Intuitively, δ_Z decides in a very simple way. If an element was already decided when forming $\langle \mathcal{V}_{*,n}(\omega), \mathcal{V}_n^*(\omega) \rangle$, δ_Z repeats this decision, in the other case it flips a coin. The probabilities of the two results are in proportion to the sizes of the two corresponding sets of already decided elements, the size being quantified by the probability of sampling generated by Z . For example, supposing that A is finite and Z samples each element with the same probability, then the coin yields both the answers with probabilities defined by the ratios of the two results among the already obtained ones.

Evidently, δ_Z admits both kinds of wrong decision. An element $a \in A - V$, being in $\mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega)$, may be wrongly proclaimed to be in V and vice versa; $a \in V \cap (\mathcal{V}_n^*(\omega) - \mathcal{V}_{*,n}(\omega))$ may be, again wrongly, proclaimed to be in $A - V$. In order to be able to define rigorously the probabilities of both kinds of error as corresponding conditional probabilities we suppose that the decision function δ_Z is applied to the element of A sampled by X_{n+1} , i.e., to the first element in the sequence $\langle a_{X_1(\omega)}, a_{X_2(\omega)}, \dots \rangle$ not more tested by the three-valued failure-proof decision function mentioned above. Set

$$PE_1 = PE_1(Z, n) = P(\{\delta_Z(a_{X_{n+1}(\omega)}, n, \omega) = 1\} / \{a_{X_{n+1}(\omega)} \in A - V\}), \quad (43)$$

$$PE_2 = PE_2(Z, n) = P(\{\delta_Z(a_{X_{n+1}(\omega)}, n, \omega) = 0\} / \{a_{X_{n+1}(\omega)} \in V\}), \quad (44)$$

the other parameters not being explicitly introduced, PE_1 , (PE_2 , resp.) is called the *probability of the first (second, resp.) kind of error* for the decision function $\delta_Z(\cdot, n, \cdot)$ and with respect to X_{n+1} . It is quite reasonable to expect that both PE_1 and PE_2 tend to zero with n increasing; in the next chapter we shall see under which conditions this is true.

5. Basic theorem for the proposed statistical decision function

THEOREM 4. Let $\mathcal{V}, \mathcal{X}, \mathcal{Y}$ as in Theorem 1, but with regularity condition replaced by a weaker one

$$\sum_{j=1}^{\infty} P(\{X_j(\omega) = i, Y_j(\omega) \in \lambda(i)\}) = \infty \tag{45}$$

for each $i \in \mathbb{N}^+$. Let there exist a sequence c_1, c_2, \dots of positive reals and two other positive reals Q_1, Q_2 such that

$$Q_1 \leq P(\{X_n(\omega) = i\})/c_i \leq Q_2 \tag{46}$$

for all $i, n \in \mathbb{N}^+$, then

$$\lim_{n \rightarrow \infty} PE_1(X_{n+1}, n) = \lim_{n \rightarrow \infty} PE_2(X_{n+1}, n) = 0. \tag{47}$$

□

Proof. For $\tilde{X}_n = a_{X_n}$, (43) implies

$$\begin{aligned} PE_1(X_{n+1}, n) &= P(\{\delta_{\tilde{X}_{n+1}}(\tilde{X}_{n+1}(\omega), n, \omega) = 1\} / \{\tilde{X}_{n+1}(\omega) \in A - V\}) = \\ &= \frac{P(\{\delta_{\tilde{X}_{n+1}}(\tilde{X}_{n+1}(\omega), n, \omega) = 1, \tilde{X}_{n+1}(\omega) \in A - V\})}{P(\{\tilde{X}_{n+1}(\omega) \in A - V\})}. \end{aligned} \tag{48}$$

Due to (46), for $\tilde{V} = \{i : i \in \mathbb{N}^+, a_i \in V\}$,

$$\begin{aligned} P(\{\tilde{X}_{n+1}(\omega) \in A - V\}) &= \sum_{a \in A - V} P(\{\tilde{X}_{n+1}(\omega) = a\}) = \\ &= \sum_{i \in \mathbb{N}^+ - \tilde{V}} P(\{X_{n+1}(\omega) = i\}) \leq Q_1 \left(\sum_{i \in \mathbb{N}^+ - \tilde{V}} c_i \right) > 0, \end{aligned} \tag{49}$$

moreover,

$$\begin{aligned} &P(\{\delta_{x_{n+1}}(\tilde{X}_{n+1}, n, \omega) = 1, \tilde{X}_{n+1}(\omega) \in A - V\}) = \\ &= P(\{\delta_{x_{n+1}}(\tilde{X}_{n+1}, n, \omega) = 1, \tilde{X}_{n+1}(\omega) \in A - \mathcal{V}_n^*(\omega)\}) + \\ &+ P(\{\delta_{x_{n+1}}(\tilde{X}_{n+1}, n, \omega) = 1, \tilde{X}_{n+1}(\omega) \in A - \mathcal{V}_n^*(\omega) - V\}). \end{aligned} \tag{50}$$

If $a \in A - \mathcal{V}_n^*(\omega)$, then $\delta_{X_{n+1}}(a, n, \omega) = 0$ so that the first summand vanishes. Hence,

$$\begin{aligned}
 PE_1(X_{n+1}, n) &= \\
 &= \left(P(\{\tilde{X}_{n+1}(\omega) \in A - V\}) \right)^{-1} \times \\
 &\quad \times P\left(\left\{ \delta_{X_{n+1}}(\tilde{X}_{n+1}(\omega), n, \omega) = 1, \tilde{X}_{n+1}(\omega) \in A \in \mathcal{V}_n^*(\omega) - V \right\} \right) = \\
 &= \left(P(\{\tilde{X}_{n+1}(\omega) \in A - V\}) \right)^{-1} \times \tag{51} \\
 &\quad \times \sum_{a \in A - V} P\left(\left\{ \tilde{X}_{n+1}(\omega) = a, \delta_{X_{n+1}}(a, n, \omega) = 1, a \in \mathcal{V}_n^*(\omega) - V \right\} \right) = \\
 &= \left(P(\{\tilde{X}_{n+1}(\omega) \in A - V\}) \right)^{-1} \times \\
 &\quad \times \sum_{a \in A - V} P\left(\left\{ \tilde{X}_{n+1}(\omega) = a, U_p(\omega) = 1, a \in \mathcal{V}_n^*(\omega) - V \right\} \right),
 \end{aligned}$$

where

$$p = p_n = \frac{P(\{\tilde{X}_{n+1}(\omega) \in \mathcal{V}_{*,n}(\omega)\})}{P(\{\tilde{X}_{n+1}(\omega) \in (\mathcal{V}_{*,n}(\omega) \cup (A - \mathcal{V}_n^*(\omega)))\})}. \tag{52}$$

A short reviewing of the proof of Theorem 1 yields that even under the weakened conditions (45), $\mathcal{V}_{*,n}(\omega)$ and $\mathcal{V}_n^*(\omega)$ tend to V almost surely with n increasing, so that p_n tends to $P(\{\tilde{X}_{n+1}(\omega) \in V\})$. The supposed statistical independence of the random variable U_p yields that

$$PE_1(X_{n+1}, n) = p \left(P(\{\tilde{X}_{n+1}(\omega) \in A - V\}) \right)^{-1} P(\{\tilde{X}_{n+1}(\omega) \in \mathcal{V}_n^*(\omega) - V\}). \tag{53}$$

Due to (49) and (52), $PE_1(X_{n+1}, n)$ tends to zero iff

$$\lim_{n \rightarrow \infty} P(\{\tilde{X}_{n+1}(\omega) \in \mathcal{V}_n^*(\omega) - V\}) = 0. \tag{54}$$

However,

$$\begin{aligned}
 &\sum_{a \in A - V} P(\{\tilde{X}_{n+1}(\omega) = a, a \in \mathcal{V}_n^*(\omega) - V\}) = \\
 &= \sum_{j \in \mathbf{N}^+ - \tilde{V}} P\left(\left\{ X_{n+1}(\omega) = j \right\} \cap \bigcap_{i=1}^n \left(\Omega - \{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\} \right) \right) = \tag{55} \\
 &= \sum_{j \in \mathbf{N}^+ - \tilde{V}} \left[P(\{X_{n+1}(\omega) = j\}) \prod_{i=1}^n \left(1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\}) \right) \right].
 \end{aligned}$$

Given $\varepsilon > 0$, take a finite set $B \subset \mathbf{N}^+ - \tilde{V}$ such that

$$\sum_{i \in B} c_i \geq (1 - (\varepsilon'/2)) \sum_{i \in \mathbf{N}^+ - \tilde{V}} c_i, \tag{56}$$

hence,

$$\sum_{i \in (\mathbf{N}^+ - \tilde{V}) - B} c_i \leq (\varepsilon'/2) \sum_{i \in \mathbf{N}^+ - \tilde{V}} c_i, \tag{57}$$

where

$$\varepsilon' = \varepsilon \left(Q_2 \sum_{i \in \mathbf{N}^+ - \tilde{V}} c_i \right)^{-1}. \tag{58}$$

Now, (45) implies that

$$\lim_{n \rightarrow \infty} \prod_{i=1}^n P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\}) = 0 \tag{59}$$

for each j , take $n_0 \in \mathbf{N}^+$ such that

$$\prod_{i=1}^n (1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})) < \varepsilon'/2 \tag{60}$$

for all $j \in B$ and $n \geq n_0$. Now, the last expression in (55) can be written as

$$\begin{aligned} & \sum_{j \in (\mathbf{N}^+ - \tilde{V}) \cap B} \left[P(\{X_{n+1}(\omega) = j\}) \prod_{i=1}^n (1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})) \right] + \\ & + \sum_{j \in (\mathbf{N}^+ - \tilde{V}) - B} \left[P(\{X_{n+1}(\omega) = j\}) \prod_{i=1}^n (1 - P(\{X_i(\omega) = j, Y_i(\omega) \in \lambda(j)\})) \right] \leq \\ & \leq (\varepsilon'/2) \sum_{j \in (\mathbf{N}^+ - \tilde{V}) \cap B} P(\{X_{n+1}(\omega) = j\}) + \sum_{j \in (\mathbf{N}^+ - \tilde{V}) - B} P(\{X_{n+1}(\omega) = j\}) \leq (61) \\ & \leq (\varepsilon'/2) \sum_{j \in (\mathbf{N}^+ - \tilde{V}) \cap B} (Q_2 c_j) + \sum_{j \in (\mathbf{N}^+ - \tilde{V}) - B} (Q_2 c_j) \leq \\ & \leq (\varepsilon'/2) \sum_{j \in \mathbf{N}^+ - \tilde{V}} (Q_2 c_j) + (\varepsilon'/2) \sum_{j \in \mathbf{N}^+ - \tilde{V}} (Q_2 c_j) = \varepsilon, \end{aligned}$$

due to the definition of ε' . Hence, (54) holds and, consequently, $PE_1(X_{n+1}, n) \rightarrow 0$ for $n \rightarrow \infty$. For $PE_2(X_{n+1}, n)$, the proof is quite analogous with $A - V$ and V , and $\mathcal{V}_n^*(\omega) - V$ and $V - \mathcal{V}_{*,n}(\omega)$, mutually replaced. \square

As an illustration of this general result, consider the finite case with $A = \{a_1, a_2, \dots, a_N\}$, $\mathcal{V} = \{\langle \underline{V}(1), \overline{V}(1) \rangle, \langle \underline{V}(2), \overline{V}(2) \rangle, \dots, \langle \underline{V}(m), \overline{V}(m) \rangle\}$ such that $\bigcup_{i=1}^m \underline{V}(i) = \bigcap_{i=1}^m \overline{V}(i) = V$. Let \mathcal{X} and \mathcal{Y} generate two sequences of statistically independent, uniformly and identically distributed random samples from the corresponding sample spaces, as formally described by the conditions (35) of Theorem 3. In this case, using the same way of reasoning as in the proof of Theorem 4 and setting $v = \text{card } V$, $\lambda(a_i) = \lambda(i)$,

$$\begin{aligned}
 & PE_2(X_{n+1}, n) = \\
 &= \frac{P(\{\tilde{X}_{n+1}(\omega) \in \mathcal{V}_{*,n}(\omega)\})}{P(\{\tilde{X}_{n+1}(\omega) \in (\mathcal{V}_{*,n}(\omega) \cup (A - \mathcal{V}_n^*(\omega)))\})} \cdot \frac{P(\{\tilde{X}_{n+1}(\omega) \in \mathcal{V}_{*,n}(\omega) - V\})}{P(\{\tilde{X}_{n+1}(\omega) \in A - V\})} \approx_n \\
 &\approx_n \left(\frac{v}{N}\right) \frac{\sum_{a \in A - V} P(\{\tilde{X}_{n+1}(\omega) = a, a \in \mathcal{V}_n^*(\omega) - V\})}{1 - (v/N)} = \tag{62} \\
 &= \left(\frac{v}{N}\right) \left(1 - \frac{v}{N}\right)^{-1} \times \\
 &\times \sum_{a \in A - V} \left[P(\{\tilde{X}_{n+1}(\omega) = a\}) \prod_{i=1}^n \left(1 - (P(\{\tilde{X}_i(\omega) = a\})P(\{Y_i(\omega) \in \tilde{\lambda}(a)\}))\right) \right] \leq \\
 &\leq \left(\frac{v}{n}\right) \left(1 - \frac{v}{N}\right)^{-1} \sum_{a \in A - V} \frac{1}{N} \left(1 - \frac{v}{Nm}\right)^n = \left(\frac{v}{N}\right) \left(1 - \frac{1}{Nm}\right)^n.
 \end{aligned}$$

An analogous calculation yields

$$PE_1(X_{n+1}, n) \approx_n \left(1 - \frac{v}{N}\right) \left(1 - \frac{1}{Nm}\right)^n, \tag{63}$$

where $a_n \approx b_n$ means that $\lim_{n \rightarrow \infty} (a_n/b_n) = 1$.

6. Frequential approximation of the statistical decision function

When applying the decision function δ defined above, the weak point is how to determine the value p to construct the random variable U_p . Or, this value is defined as the ratio of two abstract probability values, in general, not immediately obtainable from the observations and data being at our disposal. A strong temptation would bring us to the idea to replace the value p by

$$\tilde{p} = \frac{\text{card } \mathcal{V}_{*,n}(\omega)}{\text{card } \mathcal{V}_{*,n}(\omega) + \text{card}(A - \mathcal{V}_n^*(\omega))}, \tag{64}$$

but a short re-consideration yields that such a simplification may imply a serious and undesirable modification of the statistical decision function δ . For example, let $\{a_{i_0}\} = V \subset A = \{a_1, a_2, \dots\}$, let $P(\{\tilde{X}_i(\omega) = a_{i_0}\}) = 0.9$ for each $i \in \mathbb{N}^+$, let $P(\{\tilde{X}_i(\omega) = a\}) > 0$ and independent of i for each $a \in A$, let there exist, for infinitely many $a \in A$, an index $i(a)$ such that $a \in \bar{V}(i(a))$. Then $p \rightarrow 0.9$ for $n \rightarrow \infty$, but $\tilde{p} \rightarrow 0$, as $\text{card } \mathcal{V}_{*,n}(\omega) \leq 1$ for each n and $\text{card } \mathcal{V}_n^*(\omega) \rightarrow \infty$ with probability 1, if $n \rightarrow \infty$.

This difficulty can be solved in two ways. Either, (64) is applied only in the case of uniformly distributed and statistically independent random variables X_i , when, due to the strong law of large numbers, p and \tilde{p} tend, with probability one (almost surely), to the same value. Or, we may slightly modify the definition of the induction processes by computing, together with $\mathcal{V}_{*,n}(\omega)$ and $\mathcal{V}_n^*(\omega)$, also the values

$$v(n, \omega) = \sum_{i=1}^n \chi_{V(Y_i(\omega))}(\tilde{X}_i(\omega)), \tag{65}$$

and

$$w(n, \omega) = \sum_{i=1}^n \chi_{A-\bar{V}(Y_i(\omega))}(\tilde{X}_i(\omega)). \tag{66}$$

So, the values $v(n, \omega)$ and $w(n, \omega)$ express the numbers of cases, when at random sampled element $\tilde{X}_i(\omega) \in A$ was decided (positively, in the case of $v(n, \omega)$, or negatively, in the case of $w(n, \omega)$), using at random sampled rough set $\langle \underline{V}(Y_i(\omega)), \bar{V}(Y_i(\omega)) \rangle$. Evidently, repeated decisions concerning the elements already decided are repeatedly registered by $v(n, \omega)$ and $w(n, \omega)$. Now, set

$$p^* = p_n^* = \frac{v(n, \omega)}{v(n, \omega) + w(n, \omega)}, \tag{67}$$

and define the decision function δ^* in the same way as δ , just with p replaced by p^* . Let $PE_i(X_{n+1}, n, \delta)$ and $PE_i(X_{n+1}, n, \delta^*)$, $i = 1, 2$, denote the corresponding probabilities of errors of both kinds.

Let $E_{V,n}(P, \lambda)$ be the expected value of the probability of decidability of an at random sampled element $X_n(\omega) \in V$, i.e.,

$$E_{V,n}(P, \lambda) = \sum_{i \in \tilde{V}} \frac{P(\{X_n(\omega) = i\})}{P(\{X_n(\omega) \in \tilde{V}\})} P(\{Y_n(\omega) \in \lambda(i)\}), \tag{68}$$

analogously,

$$E_{A-V,n}(P, \lambda) = \sum_{i \in \mathbb{N}^+ - \tilde{V}} \frac{P(\{X_n(\omega) = i\})}{P(\{X_n(\omega) \in \mathbb{N}^+ - \tilde{V}\})} P(\{Y_n(\omega) \in \lambda(i)\}). \tag{69}$$

THEOREM 5. Let \mathcal{X} and \mathcal{Y} be independent i.i.n.d. sequences (cf. (18)), let $V \neq \emptyset$, let

$$E_{V,1}(p, \lambda) = E_{A-V,1}(p, \lambda) > 0, \tag{70}$$

then, for both $i = 1, 2$ and for each $k \in \mathbb{N}^+$,

$$PE_i(X_k, n, \delta^*) \approx_n PE_i(X_k, n, \delta). \tag{71}.$$

□

Remark. Condition (70) means that the ability of system \mathcal{X} , \mathcal{Y} , \mathcal{V} to decide the membership relation for V is “in average” the same for elements as well as for non-elements of V . Assertion (71) immediately implies that $\lim_{n \rightarrow \infty} PE_i(X_k, n, \delta^*) = \lim_{n \rightarrow \infty} PE_i(X_k, n, \delta)$ ($= 0$ as Theorem 4 yields), but it is stronger, as it claims the speed of convergence to be qualitatively the same for both the decision functions δ and δ^* and for both kinds of probabilities of errors.

Proof. Due to (65), $n^{-1}v(n, \omega)$ denotes the relative frequency of indices $i \leq n$ for which $\tilde{X}_i(\omega) \in \underline{V}(Y_i(\omega))$. The fact that \mathcal{X} and \mathcal{Y} are independent i.i.n.d. sequences and strong law of large numbers (cf. [1], e.g.) imply that

$$P\left(\left\{\lim_{n \rightarrow \infty} n^{-1}v(n, \omega) = P(\{\tilde{X}_1(\omega) \in \underline{V}(Y_1(\omega))\})\right\}\right) = 1. \tag{72}$$

However,

$$\begin{aligned} P\left(\{\tilde{X}_1(\omega) \in \underline{V}(Y_1(\omega))\}\right) &= \sum_{j \in \tilde{V}} P\left(\{X_1(\omega) = j, a_j \in \underline{V}(Y_1(\omega))\}\right) = \\ &= \sum_{j \in \tilde{V}} P\left(\{X_1(\omega) = j, Y_1(\omega) \in \lambda(j)\}\right) = \\ &= \sum_{j \in \tilde{V}} [P(\{X_1(\omega) = j\})P(\{Y_1(\omega) \in \lambda(j)\})] = (E_{V,1}(P, \lambda))P(\{\tilde{X}_1(\omega) \in V\}) > 0, \end{aligned} \tag{73}$$

due to (68). Quite similarly, (69) yields

$$P\left(\left\{\lim_{n \rightarrow \infty} n^{-1}w(n, \omega) = P(\{\tilde{X}_1(\omega) \in A - \bar{V}(Y_1(\omega))\})\right\}\right) = 1 \tag{74}$$

and

$$P\left(\{\tilde{X}_1(\omega) \in A - \bar{V}(Y_1(\omega))\}\right) = (E_{A-V,1}(P, \lambda))P(\{\tilde{X}_1(\omega) \in A - V\}) > 0 \tag{75}$$

Hence, with probability one,

$$\begin{aligned} \lim_{n \rightarrow \infty} p_n^*(\omega) &= \lim_{n \rightarrow \infty} \frac{v(n, \omega)}{v(n, \omega) + w(n, \omega)} = \\ &= \lim_{n \rightarrow \infty} \frac{n^{-1}v(n, \omega)}{n^{-1}v(n, \omega) + n^{-1}w(n, \omega)} = \\ &= \lim_{n \rightarrow \infty} \frac{(E_{V,1}(P, \lambda))P(\{\tilde{X}_1(\omega) \in V\})}{(E_{V,1}(P, \lambda))P(\{\tilde{X}_1(\omega) \in V\}) + (E_{A-V,1}(P, \lambda))P(\{\tilde{X}_1(\omega) \in A - V\})} = \\ &= P(\{\tilde{X}_1(\omega) \in V\}) \end{aligned} \tag{76}$$

due to (70). But, $\lim_{n \rightarrow \infty} p_n = P(\{\tilde{X}_1(\omega) \in V\})$ as well (cf. the proof of Theorem 4), so that

$$P(\{\lim_{n \rightarrow \infty} p_n^*(\omega) = \lim_{n \rightarrow \infty} p_n\}) = 1. \tag{77}$$

Referring, again, to the prof of Theorem 4, we can easily obtain that for each $k, n \in \mathbb{N}^+$,

$$PE_1(X_k, n, \delta) = (1 - p_n)K_n, \tag{78}$$

$$PE_2(X_k, n, \delta) = p_n L_n, \tag{79}$$

for appropriate K_n, L_n independently of p_n and with positive limit values for $n \rightarrow \infty$. For δ^* , the situation is more difficult, as p_n^* is a random variable. Instead of (51) we obtain

$$PE_1(X_k, n, \delta^*) = \int \left\{ \left[P(\{\tilde{X}_k(\omega) \in A - V\}) \right]^{-1} \cdot \left[\sum_{a \in A - V} P(\{\tilde{X}_k(\omega) = a, U_{p_n^*(\omega)}(\omega) = 1, a \in \mathcal{V}_n^*(\omega) - \underline{V}\}) \right] \right\} dF_{p,n}(\omega), \tag{80}$$

where $F_{p,n}$ is the distribution function of the random variable p_n^* . Following the pattern of the proof of Theorem 4 we obtain that

$$PE_1(X_k, n, \delta^*) = \int (1 - p_n^*(\omega)) K_n dF_{p,n}(\omega) = (1 - EP_n^*(\cdot)) K_n, \tag{81}$$

with the same K_n as in (78), where

$$EP_n^*(\cdot) = \int p_n^*(\omega) dF_{p,n}(\omega) \tag{82}$$

is the expected value of the random variable $p_n^*(\cdot)$. Quite analogously,

$$PE_2(X_k, n, \delta^*) = (EP_n^*(\cdot)) L_n. \tag{83}$$

Set $p_0 = \lim_{n \rightarrow \infty} p_n$, take an $\varepsilon > 0$. Due to the Jęgorov theorem ([1], e.g.) and (77) there exists measurable $Q_\varepsilon \subset \Omega$ such that $P(Q_\varepsilon) > 1 - (\varepsilon/2)$ and $p_n^*(\omega) \rightarrow p_0$ uniformly on Q_ε . So, take n_0 such that $|p_n^*(\omega) - p_0| < \varepsilon/2$ for each $n \geq n_0$ and each $\omega \in Q_\varepsilon$. Then $E(|p_n^*(\cdot) - p_0|) \leq (\varepsilon/2)P(Q_\varepsilon) + \varepsilon/2 \leq \varepsilon$, hence, $|EP_n^*(\cdot) - p_0| \leq \varepsilon$, so that $EP_n^* \rightarrow p_0$ for $n \rightarrow \infty$. This fact and (78), (79), (81), (83) imply the assertion (71). □

7. Comments, remarks, and conclusions

The results presented above are very elementary and perhaps even trivial, and they could be developed in more details or replaced by more sophisticated ones.

Nevertheless, some basic ideas of statistical induction through rough sets seem to be illustrated, by these results, in a degree sufficient enough to subject them to a brief discussion.

First of all, let us pick up the fact, that our presentation of statistical induction through rough sets as a part of artificial intelligence is quite legitimate. Rough sets can be taken as the most general tools, at least within the framework of classical set-theoretic language, to describe and to deal with incomplete knowledge expressible, otherwise, in the three-valued Lukasiewicz logic with the third value interpreted as "it is not known, whether. . ." There are numerous particular cases in which such a situation may occur, remember, e.g., missing values of some observations in the GUHA method, undecidability of certain assertions in formalized theories because of theoretical or time-space limitations, etc. The case when a system of rough sets defines a (classical) set can be seen as a process of appropriate combination of pieces of partial or uncertain knowledge to obtain the complete knowledge, and processes like this are fully covered by the domain of artificial intelligence. If such a process yields this complete knowledge only asymptotically, as it is the case even A or \mathcal{V} are infinite, our approach offers, at a level as general as possible, an approximative statistical solution reasonable from the point of view of simple statistical qualitative criteria. The same approach can be applied in the case of practical intractability, when the sets A and \mathcal{V} are finite but too large to be checked systematically. The notions and apparatus of rough sets enable us to pick out what is common for many processes of statistical approximative combination of partial (incomplete) knowledge and what is, in particular problems, often hidden behind the specific features of the problem in question.

A further development may proceed at least in the two following directions. Either, a supplementary structure may be imposed on the sets A and V , and the properties of elements of A and V , or some relations among these elements, involved by this structure, can be used to propose more sophisticated statistical induction procedures than the most simple one described above. These auxiliary structures on A and V may be more or less closely inspired by and connected with the intended practical applications, but they may be also rather general. For example, the set A may be equipped by a topological structure according to which the elements of A which are "close" to elements of V are also in V with a greater probability than those which are rather isolated from elements of V . An appropriate union of neighbourhoods of elements of $\mathcal{V}_{*,n}(\omega)$ will then serve as a reasonable approximation of V . Or, going in the opposite direction, we may still weaken the conditions of the model investigated above. For example, we may consider the case when, given $a \in A$ and $(\underline{V}, \overline{V}) \in \mathcal{V}$ the answers whether $a \in \underline{V}$ and $a \in A - \overline{V}$ are charged with positive probabilities of error. This error cumulates with that one connected with the decision function D and deteriorates the statistical qualities of the induction procedure in question. Both these modifications will be investigated in the next future.

Being rather trivial, the contribution is almost self-explanatory. All the references, if any, concerning the most simple combinatorial probability can be consulted with any textbook of elementary probability theory; let us introduce [1] as a very good one. [4] is a foundatory paper on rough sets and [3] is an example of latest contributions to this theory. Finally, [2] is mentioned as it contains a more detailed of references accessible in our country.

References

1. *Feller, W.*, An Introduction to Probability Theory and its Applications, vol. I and II, second edition. John Wiley and Sons, New York, 1962, 1965 (Russian translation: Mir, Moscow, 1964, 1967).
2. *Kubát, M.*, More attention to rough set theory. In: Aplikace umělé inteligence AI89, Praha, 1989, pp. 361–368.
3. *Orłowska, E.*, Semantics knowledge operators. Bull. Polish Acad. Sci., ser. Mathematics, vol. 35 (1987), no. 5–6, pp. 255–263.
4. *Pawlak, Z.*, Rough sets. International Journal on Computer and Information Sciences 11 (1982), no. 3, pp. 341–356.

Определение и распознавание классических множеств посредством грубых множеств

ИВАН КРАМОСИЛ

(Прага)

Грубые множества являются относительно новой областью современной математики с интересными применениями. В работе показано, каким образом возможно воспользоваться грубыми множествами, чтобы описать и решить проблему статистической индукции самым общим образом, покрывающим большое число важных частных случаев.

Ivan Kramosil
 Institute of Information Theory and Automation
 Czechoslovak Academy of Sciences
 Pod Vodárenskou věží 4
 18208 Praha 8
 Czechoslovakia

INFINITE-DIMENSIONAL SYSTEMS:
DESIGN OF SAKAWA CONTROLLERS.
PART II

S. K. KOROVIN, M. G. NIKITINA AND S. V. NIKITIN

(*Moscow*)

(Received December 27, 1989)

In this paper the design of a finite-dimensional Sakawa controller is described. A lower estimate is proposed for the dimension of this controller in a closed-loop infinite-dimensional system. Case studies are reported of designing such finite-dimensional controllers, and a relation is established between the controller parameters and the spectral properties of the associated elliptic operator.

1. Introduction

The stabilizability of distributed parameter systems has been investigated in a number of papers [2-5, 7, 8, 11-14, 17-23]. Stabilization criteria have been obtained [2, 3, 8] with a finite-dimensional input, and extended to a wider range of processes with some conditions on input operator compactness [17]. For finite-dimensional input and output systems adaptive controllers [9] and PI-controllers [15] have been obtained. In [10] the stabilizability of parabolic distributed systems, with certain conditions imposed on the operator spectrum, is studied through controllability and observability analysis of some linear finite-dimensional systems. Existence has been proved [11] of stabilizing feedback for a nonlinear oscillatory distributed parameter system with a specified indicator of exponential stability. Existence has been proved of and an explicit form proposed [19, 20] for finite-dimensional stabilizing feedback for an infinite-dimensional system. For these findings to be applied, however, the dimension of the controller has to be known which would provide the desired exponential stability. This dimension is determined in [21] by selection with subsequent computer modeling of the process. In this paper a lower estimate is proposed for the dimension of this controller in a closed-loop infinite-dimensional system, case studies are reported of designing such finite-dimensional controllers, and the relation is established between the controller parameters and spectral properties of the associated elliptic operator.

2. Design of Sakawa controllers

This Section will describe the design of finite-dimensional Sakawa controllers [19–21].

A controller has been proposed [19, 20] for the stabilization of distributed parameter system. In [21] these methods are used to offset bending oscillations of a flexible element in a manipulator tracking the desired path. These papers do not, however, provide any estimates of the controller state space dimension which depends on feedback gains, the observer parameters as well as on the process parameters. The lower estimate of the controller dimension must be known for the design.

In this Section this estimate is computed for the Sakawa controller which provides μ -exponential stability of a closed-loop system. By assumption the input of the system $\Sigma(C, A, B)$ is finite-dimensional and, therefore, this system is assumed to be stabilizable by finite-dimensional feedback.

This stabilizability has been shown [2, 3, 8, 17] possible iff

e1) for some $\gamma \in \mathbf{R}$, $\gamma < \mu$ the subset $\sigma_\gamma = \{\lambda \in \sigma(A); \operatorname{Re} \lambda \geq \gamma\}$ of spectrum $\sigma(A)$ is finite and, moreover, the multiplicity of every eigenvalue is also finite and can be separated by a simple closed-loop from the remainder of the spectrum $\sigma(A)$ or has a spectral decomposition [3], $A = A_N \oplus \tilde{A}_N$ and $H = H_N \oplus \tilde{H}_N$, where N is the dimension of the invariant subspace H_N associated with the set σ_γ ;

e2) (A_N, B_N) is a controllable pair with $B_N = P_N B$, where $P_N: H \rightarrow H_N$ is a natural mapping of H on H_N ;

e3) $\tilde{A}_N = (I - P_N)A$ is an infinitesimal generator of a γ -exponentially stable semi-group $e^{\tilde{A}_N T}$ or a negative γ -type of a semi-group.

In the design of finite-dimensional linear stabilizing feedback let us start, as in the preceding Section, with a finite-dimensional system $\Sigma(C, A, B)$. For this system $\Sigma(C, A, B)$ linear dynamic and static feedback are designed in various ways. One of them is very efficient in stabilizing finite-dimensional approximations of infinite-dimensional systems with a discrete spectrum.

Linear stabilizing output $y = Cx$ feedback has the form

$$\begin{aligned} u &= Kz, \\ \dot{z} &= Az + BKz + L(Cz - y), \end{aligned}$$

where the operators $K: \mathbb{C}^N \rightarrow \mathbb{C}^m$ and $L: \mathbb{C}^l \rightarrow \mathbb{C}^N$ are chosen so as to make systems

$$\begin{aligned} \dot{x} &= (A + BK)x, \\ \dot{z} &= (A + LC)z \end{aligned}$$

μ -exponential.

Let us show the parameters K and L are related with the spectral properties of the operator A , in particular its resolvent $R(\lambda, A) = (\lambda I - A)^{-1}$.

THEOREM 1. Any number $\lambda \in \sigma(A + BK) \setminus \sigma(A)$ satisfies the relation

$$\det(I - KR(\lambda, A)B) = 0$$

and the eigenvectors $\xi(\lambda)$ associated with this λ are computed by the formula

$$\xi(\lambda) = R(\lambda, A)B\zeta, \quad (1)$$

where ζ is some nontrivial solution of the system of equations

$$\zeta = KR(\lambda, A)B\zeta. \quad (2)$$

If $\lambda \in \sigma(A + BK) \cap \sigma(A)$, then the associated eigenvector $\xi(\lambda)$ satisfies the relations

$$\langle \hat{\xi}(\bar{\lambda}), BK\xi(\lambda) \rangle = 0,$$

where $\hat{\xi}(\bar{\lambda})$ is any solution of the equation $(\bar{\lambda}I - A^*)\hat{\xi}(\bar{\lambda}) = 0$.

Proof. If $\lambda \in \sigma(A + BK) \setminus \sigma(A)$ and $(A + BK)\xi(\lambda) = \lambda\xi(\lambda)$, then

$$\xi(\lambda) = R(\lambda, A)BK\xi(\lambda), \quad K\xi(\lambda) \neq 0, \quad (3)$$

because in the opposite case $\lambda \in \sigma(A)$. Left-multiplying (3) by K we can see that the homogeneous system of linear equations

$$(I - KR(\lambda, A)B)\zeta = 0$$

has a nontrivial solution $\zeta = K\xi(\lambda)$. Consequently, with $\det(I - KR(\lambda, A)B) = 0$ with $\lambda \in \sigma(A + BK) \setminus \sigma(A)$.

Let us check that with $\lambda \in \sigma(A + BK) \setminus \sigma(A)$ the eigenvector $\xi(\lambda)$ is computed by formulas (1) and (2). It is easily seen that $(\lambda I - (A + BK))R(\lambda, A)\zeta - BK R(\lambda, A)B\zeta = 0$. This proves the truth of formula (1).

If $\lambda \in \sigma(A + BK) \cap \sigma(A)$ then

$$BK\xi(\lambda) = (\lambda I - A)\xi(\lambda). \quad (4)$$

Because $\det(\lambda I - A) = 0$ the system (4) is solvable iff for any solution $\hat{\xi}(\bar{\lambda})$ of the adjoint equation $(\bar{\lambda}I - A^*)\hat{\xi}(\bar{\lambda}) = 0$. It is true that $\langle \hat{\xi}(\bar{\lambda}), BK\xi(\lambda) \rangle = 0$, which proves the Theorem.

Theorem 1 is formulated in a most straightforward way when the input is one-dimensional, or $B = b \in \mathbb{C}^n$. More specifically, the following proposition is true.

Corollary 1. If the input is one-dimensional, then

- 1) any number $\lambda \in \sigma(A + bk) \setminus \sigma(A)$ satisfies the relation $\langle k, R(\lambda, A)b \rangle_{\mathbb{R}} = 1$, and the eigenvector $\xi(\lambda)$ associated with this λ is computed by the formula $\xi(\lambda) = R(\lambda, A)b$, with $\langle p, q \rangle_{\mathbb{R}} = \sum_{i=1}^N p_i q_i$ hereafter;
- 2) if $\lambda \in \sigma(A + bk) \cap \sigma(A)$ then the associated eigenvector satisfies the relation

$$\langle \hat{\xi}(\bar{\lambda}), b \rangle_{\mathbb{R}} \cdot \langle k, \xi(\lambda) \rangle_{\mathbb{R}} = 0,$$

where $\hat{\xi}(\bar{\lambda})$ is any solution of the equation $(\bar{\lambda}I - A^*)\hat{\xi}(\bar{\lambda}) = 0$.

Propositions following from the Theorem 1 and Corollary 1 are naturally true which lead to the same results for the system $\dot{z} = Az + LCz$. Let us use these results to compute the dimension of the Sakawa controllers for μ -exponential stabilizability of the infinite-dimensional system $\Sigma(C, A, b)$, where $C \in H$, $b \in H$. Recall the main stages of the design of Sakawa controllers.

Operator A is assumed to have no multiple eigenvalues and satisfy a1)–a3) [23], e1)–e3) and $\operatorname{Re} \lambda_n \rightarrow -\infty$ as $n \rightarrow \infty$. Then elements of $\sigma(A)$ are enumerated

$$\operatorname{Re} \lambda_1 > \operatorname{Re} \lambda_2 > \operatorname{Re} \lambda_3 > \dots > \operatorname{Re} \lambda_n > 0 > \operatorname{Re} \lambda_{n+1} > \dots \quad (5)$$

A Sakawa controller is designed for a fixed $\mu \in \mathbb{R}$. The system with this controller in the closed-loop is required to be μ -exponentially stable. The design proceeds in the following stages.

Stage 1. Under the conditions of (5) and with $\operatorname{Re} \lambda_n \rightarrow -\infty$ as $n \rightarrow \infty$, choose a natural N such that with $n \geq N$, $\operatorname{Re} \lambda_n < \mu$.

Stage 2. The spectral decomposition of A is $A = A_N \oplus \tilde{A}_N$, $H = H_N \oplus \tilde{H}_N$, where N is chosen at Stage 1. For a finite-dimensional system $\Sigma(C_N, A_N, b_N)$ (here $C_N = P_N C$, $b_N = P_N b$, and $P_N: H \rightarrow H_N$ is a natural mapping) a stabilizing dynamic feedback is designed

$$\begin{aligned} u &= \sum_{i=1}^N k_i z_i, \\ \dot{z} &= A_N z + b_N u + \tilde{l}_N \left(\sum_{i=1}^N c_i z_i - y \right), \end{aligned} \quad (6)$$

where the parameters (k_1, \dots, k_N) and $(l_1, \dots, l_N)^T$ are chosen by conventional methods [1] so as to make the closed-loop N -dimensional system μ -exponentially stable and, as before, $\langle c, x \rangle_{\mathbb{R}}$ is understood as the sum $\sum_{i=1}^{\infty} c_i x_i$ or $y = \langle c, x \rangle_{\mathbb{R}}$.

Stage 3. The feedback (6) is modified

$$\begin{aligned}
 u &= \sum_{i=1}^N k_i z_i, \\
 \dot{z} &= A_{N+m} z + b_{N+m} u + \bar{l}_N \left(\sum_{i=1}^{N+m} c_i z_i - y \right),
 \end{aligned}
 \tag{7}$$

where \bar{l}_N denotes an $(N + m)$ -dimensional vector whose first N coordinates are $(l_1, \dots, l_N)^T$ and the latter m ones are zero, $z \in \mathbb{C}^{N+m}$, and $A = A_{N+m} \oplus \tilde{A}_{N+m}$ is the associated spectral decomposition of A . This dynamic feedback is referred to as a Sakawa controller [19]. It has been shown in [19, 20] that the feedback system $\Sigma(C, A, B)$ is μ -exponentially stable for m being fairly large.

Let us determine the lower estimate of m with which μ -exponential stability is ensured for the closed-loop system.

THEOREM 2. For the system $\Sigma(C, A, b)$, $(C, b \in H)$ conditions a1)–a3) [23] hold, in (5) the elements of $\sigma(A)$ are enumerated and $\text{Re } \lambda_n \rightarrow -\infty$ as $n \rightarrow \infty$ or, for any fixed $\mu \in \mathbb{R}$, there exists λ_N such that $\text{Re } \lambda_N < \mu$ and if the system $\Sigma(C_N, A_N, b_N)$ is controllable and observable, then there exists a Sakawa controller (7) which μ -exponentially stabilizes the system $\Sigma(C, A, b)$ when the natural number m satisfies the conditions

$$\begin{aligned}
 &\sum_{j=1}^N \left(1 / \left| \sum_{i=1}^N \frac{c_i l_i}{(\tilde{\lambda}_j - \lambda_i)^2} \right|^2 \right) + \sum_{\rho=1}^N \left| \sum_{\mu=1}^N l_\mu^2 \sum_{\nu=1}^N c_\nu k_\nu \times \right. \\
 &\times \sum_{j=1}^N 1 / \left\{ (\tilde{\lambda}_j - \lambda_\mu)(\tilde{\lambda}_j - \lambda_\nu)(\omega_\rho - \tilde{\lambda}_j) \cdot \sum_{i=1}^N \frac{c_i l_i}{(\tilde{\lambda}_j - \lambda_i)^2} \right\}^2 \times \\
 &\times \frac{1}{\left| \sum_{i=1}^N \frac{b_i k_i}{(\omega_\rho - \lambda_i)^2} \right|^2} < \frac{(\mu - \text{Re } \lambda_{N+m+1})^2}{\sum_{\nu=N+m+1}^{\infty} |c_\nu|^2},
 \end{aligned}
 \tag{8a}$$

$$\begin{aligned}
 &\sum_{i=1}^N \left| \sum_{\mu=1}^N \frac{l_\mu k_\mu}{(\tilde{\lambda}_i - \lambda_\mu)} \right|^2 \cdot \left| 1 + \sum_{\nu=1}^N 1 / (\tilde{\lambda}_i - \omega_\nu) \left(\sum_{\mu=1}^N \frac{b_\mu k_\mu}{(\omega_\nu - \lambda_\mu)^2} \right) \right|^2 + N < \\
 &< \frac{(\mu - \max\{\max \text{Re } \omega_i, \max \text{Re } \lambda_i\})^2}{\sum_{\nu=N+m+1}^{\infty} |b_\nu|^2}
 \end{aligned}
 \tag{8b}$$

with $b_i = \langle b, \xi(\lambda_i) \rangle$ and $c_i = \langle c, \xi(\lambda_i) \rangle$, where $\langle \cdot, \cdot \rangle$ is a scalar product in H , $\xi(\lambda_i)$ is the eigenvector of the operator A associated with $\lambda_i \in \sigma(A)$, $\bar{k}_N = (k_1, \dots, k_N)$ and

$\bar{l}_N = (l_1, \dots, l_N)^T$ are parameters of the feedback and the controller, respectively. N is chosen so that $\text{Re } \lambda_N < \mu \{ \tilde{\lambda}_1, \dots, \tilde{\lambda}_N \}$ are eigenvalues of the operator $A_N + \bar{l}_N c_N$, $\{ \omega_1, \dots, \omega_N \}$ are the eigenvalues of the operator $A_N + b_N \bar{k}_N$; it is assumed that

$$\{ \{ \tilde{\lambda}_1, \dots, \tilde{\lambda}_N \} \cup \{ \omega_1, \dots, \omega_N \} \} \cap \sigma(A) = \emptyset, \quad \tilde{\lambda}_i \neq \tilde{\lambda}_j, \quad \omega_i \neq \omega_j$$

with $i \neq j$ and $\{ \tilde{\lambda}_1, \dots, \tilde{\lambda}_N \} \cap \{ \omega_1, \dots, \omega_N \} = \emptyset$.

Proof. From a3) it follows that a basis $\{ \xi(\lambda_i) \}_{i=1}^\infty$ can be chosen in H where $\{ \lambda_i \}_{i=1}^\infty$ is the spectrum of the operator A with the enumeration (5). Let the parameters \bar{k}_N and \bar{l}_N and natural numbers N and m be such that are formulated in the Theorem. The system $\Sigma(C, A, b)$ in the basis $\{ \xi(\lambda_i) \}_{i=1}^\infty$ is represented in the form:

$$\begin{aligned} \dot{z}_i &= \lambda_i z_i + b_i \langle \bar{k}_N, z \rangle_{\mathbb{R}} + l_N \left(\sum_{i=1}^{N+m} c_i z_i - y \right), & i = 1, 2, \dots, N, \\ \dot{z}_j &= \lambda_j z_j + b_j \langle \bar{k}_N, z \rangle_{\mathbb{R}}, & j = N+1, \dots, N+m, \\ \dot{x}_\nu &= \lambda_\nu x_\nu + b_\nu \langle \bar{k}_N, z \rangle_{\mathbb{R}}, & \nu = 1, \dots, N+m, \\ \dot{\tilde{x}}_\mu &= \lambda_\mu \tilde{x}_\mu + b_\mu \langle \bar{k}_N, z \rangle_{\mathbb{R}}, & \mu = N+m+1, \dots, \infty, \\ y &= \sum_{\nu=1}^\infty c_\nu x_\nu = \langle c_{N+m}, x_{N+m} \rangle_{\mathbb{R}} + \langle \tilde{c}_{N+m}, \tilde{x}_{N+m} \rangle_{\mathbb{R}}, \end{aligned}$$

where $\tilde{x}_{N+m} = (x^{N+m+1}, \dots)$ and $x_{N+m} = (x^1, \dots, x^{N+m})$. Following the substitution $\alpha = z_{N+m} - x_{N+m}$, $\zeta = x_{N+m}$, and $\tilde{x}_{N+m} = x_{N+m}$, the equations of the system take the form

$$\dot{\alpha}_i = \lambda_i \alpha_i + l_i \left(\sum_{i=1}^N c_i \alpha_i + \sum_{i=N+1}^{N+m} c_i \alpha_i \right) + l_i \langle \tilde{x}_{N+m}, \tilde{c}_{N+m} \rangle_{\mathbb{R}}, \quad i = 1, \dots, N, \quad (9a)$$

$$\dot{\alpha}_j = \lambda_j \alpha_j, \quad j = N+1, \dots, N+m, \quad (9b)$$

$$\dot{\zeta}_j = \lambda_j \zeta_j + b_j (\langle \bar{k}_N, \alpha \rangle_{\mathbb{R}} + \langle \bar{k}_N, \zeta \rangle_{\mathbb{R}}), \quad j = 1, \dots, N+m, \quad (9c)$$

$$\dot{\tilde{x}}_j = \lambda_j \tilde{x}_j + b_j (\langle \bar{k}_N, \alpha \rangle_{\mathbb{R}} + \langle \bar{k}_N, \zeta \rangle_{\mathbb{R}}), \quad j = N+m+1, \dots, \infty. \quad (9d)$$

Because $\text{Re } \lambda_{N+1} < \mu$ the system (9) is exponentially stable iff so is the system

$$\begin{aligned} \dot{\xi} &= S_N \xi + l_N \langle \tilde{c}_{N+m}, \tilde{x}_{N+m} \rangle_{\mathbb{R}}, \\ \dot{\tilde{x}}_j &= \lambda_j \tilde{x}_j + b_j \langle \bar{k}_N, \xi \rangle_{\mathbb{R}}, \quad j = N+m+1, \dots, \infty, \end{aligned}$$

where $\xi = (\alpha, \zeta)^T$, $k_N = (k_1, \dots, k_N, k_1, \dots, k_N)$ and the matrix S_N is block-wise

$$S_N = \left(\begin{array}{c|c} -\mathcal{E}_0(\lambda) + \bar{l}_N c_N & 0 \\ \hline b_N \bar{k}_N & -\mathcal{E}_0(\lambda) + b_N \bar{k}_N \end{array} \right),$$

where

$$\mathcal{E}_\omega(\lambda) = \begin{pmatrix} \omega - \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \omega - \lambda_N \end{pmatrix}, \quad \lambda = \{\lambda_1, \dots, \lambda_N\}.$$

The eigenvalues of the operator $S_N: \mathbb{C}^{2N} \rightarrow \mathbb{C}^{2N}$ are $\{\tilde{\lambda}_i\}_{i=1}^N$, or a spectrum associated with the vector \bar{l}_N are $\{\lambda_i\}_{i=1}^N \cap \sigma(A) = \emptyset$ and it follows from Corollary 1 of Theorem 1 that $\{\tilde{\lambda}_i\}_{i=1}^N$ are roots of the equation

$$\langle c_N, \mathcal{E}_{\tilde{\lambda}}^{-1}(\lambda) \bar{l}_N \rangle_{\mathbb{R}} = 1,$$

where $\{\omega_i\}_{i=1}^N$ are eigenvalues associated with the vector \bar{k}_N , or roots of the equation

$$\langle \bar{k}_N, \mathcal{E}^{-1}(\lambda) b_N \rangle_{\mathbb{R}} = 1.$$

The eigenvalues of the operator S_N are assumed to be enumerated in the order in which they are listed. Let Ξ_N be a matrix whose columns are eigenvectors of the operator S_N , or $\Xi_N = \{\xi_1, \dots, \xi_{2N}\}$ or

$$S_N \xi_i = \begin{cases} \tilde{\lambda}_i \xi_i, & \text{with } 1 \leq i \leq N \\ \omega_i \xi_i, & \text{with } N + 1 \leq i \leq 2N. \end{cases}$$

Then the rows for the matrix Ξ_N^{-1} are transposed eigenvectors of the operator S_N^T (where T denotes transposition, $(S_{ij})^T = S_{ji}$). Indeed, if $S_N^T \gamma_j = \mu_j \gamma_j$, then

$$\langle \gamma_j, \xi \rangle_{\mathbb{R}} = \sum_{i=1}^{2N} \gamma_{ji} \xi_{\nu i} = \frac{1}{\mu_j} \langle S_N^T \gamma_j, \xi_\nu \rangle_{\mathbb{R}} = \frac{1}{\mu_j} \langle \gamma_j, S_N \xi_\nu \rangle_{\mathbb{R}} = \frac{\alpha_\nu}{\mu_j} \langle \gamma_j, \xi_\nu \rangle, \quad \alpha_\nu \in \sigma(S_N).$$

Consequently, if $\alpha_\nu \neq \mu_j$, then $\langle \gamma_j, \xi_\nu \rangle_{\mathbb{R}} = 0$ and $\langle \gamma_j, \xi_\nu \rangle_{\mathbb{R}} \neq 0$ with $\alpha_\nu = \mu_j$. The latter follows from the fact that the vectors $\{\xi_i\}_{i=1}^{2N}$ are linearly independent. In the light of the above, $\{\gamma_i\}_{i=1}^{2N}$ can be taken such that $\langle \gamma_i, \xi_j \rangle_{\mathbb{R}} = 0$ with $i \neq j$ and $\langle \gamma_i, \xi_i \rangle_{\mathbb{R}} = 1$. Then

$$\Xi_N^{-1} = \begin{pmatrix} \gamma_1^T \\ \vdots \\ \gamma_{2N}^T \end{pmatrix}.$$

The matrices Ξ_N and Ξ_N^{-1} are computed by using Corollary 1 of Theorem 1.

$$\Xi_N = \left\{ \left(\begin{array}{c} \mathcal{E}_{\tilde{\lambda}_i}^{-1}(\lambda) \bar{l}_N \\ \mathcal{B} \mathcal{E}_{\tilde{\lambda}_i}^{-1}(\omega) \mathcal{B}^{-1} b_n \langle \bar{k}_n, \mathcal{E}_{\tilde{\lambda}_i}^{-1} \bar{l}_n \rangle \end{array} \right)_{i=\overline{1, N}} \left(\begin{array}{c} 0 \\ b_1 \\ \frac{\omega_k - \lambda_1}{\omega_k - \lambda_1} \\ \vdots \\ b_N \\ \frac{\omega_k - \lambda_N}{\omega_k - \lambda_N} \end{array} \right)_{k=\overline{1, N}} \right\},$$

$$\Xi_N^{-1} = \left\{ \begin{array}{l} \left[\left(\frac{c_1}{\lambda_k - \lambda_1} \quad \cdots \quad \frac{c_N}{\lambda_k - \lambda_1} \quad 0 \right) \left(\sum_{i=1}^N \frac{c_i l_i}{(\lambda_k - \lambda_i)^2} \right)^{-1} \right]_{k=\overline{1, N}} \\ \left[\left(\mathcal{A} \mathcal{E}_{\omega_i}^{-1}(\tilde{\lambda}) \mathcal{A}^{-1} \bar{k}_n \langle b_N, \mathcal{E}_{\omega_i}^{-1}(\lambda) \bar{k}_n \rangle \right)^T \left(\mathcal{E}_{\omega_i}^{-1}(\lambda) \bar{k}_n \right)^T \sum_{j=1}^N \frac{b_j k_j}{(\omega_i - \lambda_j)^2} \right]_{i=\overline{1, N}} \end{array} \right\},$$

where the matrices \mathcal{A} , \mathcal{B} , and \mathcal{A}^{-1} , \mathcal{B}^{-1} have the form

$$\mathcal{A} = \begin{pmatrix} \frac{l_1}{\tilde{\lambda}_1 - \lambda_1} & \cdots & \frac{l_1}{\tilde{\lambda}_N - \lambda_1} \\ \vdots & & \vdots \\ \frac{l_N}{\tilde{\lambda}_1 - \lambda_N} & \cdots & \frac{l_N}{\tilde{\lambda}_N - \lambda_N} \end{pmatrix}, \quad \mathcal{B} = \begin{pmatrix} \frac{b_1}{\omega_1 - \lambda_1} & \cdots & \frac{b_1}{\omega_N - \lambda_1} \\ \vdots & & \vdots \\ \frac{b_N}{\omega_1 - \lambda_N} & \cdots & \frac{b_N}{\omega_N - \lambda_N} \end{pmatrix},$$

$$(\mathcal{B}^{-1})^T = \left\{ \begin{array}{l} \left(\frac{k_1}{\omega_1 - \lambda_1} \right) \left(\sum_{i=1}^N \frac{b_i k_i}{(\omega_1 - \lambda_i)^2} \right)^{-1} \quad \cdots \quad \left(\frac{k_1}{\omega_N - \lambda_1} \right) \left(\sum_{i=1}^N \frac{b_i k_i}{(\omega_N - \lambda_i)^2} \right)^{-1} \\ \left(\frac{k_N}{\omega_1 - \lambda_N} \right) \left(\sum_{i=1}^N \frac{b_i k_i}{(\omega_1 - \lambda_i)^2} \right)^{-1} \quad \cdots \quad \left(\frac{k_N}{\omega_N - \lambda_N} \right) \left(\sum_{i=1}^N \frac{b_i k_i}{(\omega_N - \lambda_i)^2} \right)^{-1} \end{array} \right\},$$

$$(\mathcal{A}^{-1})^T = \left\{ \begin{array}{l} \left(\frac{c_1}{\tilde{\lambda}_1 - \lambda_1} \right) \left(\sum_{i=1}^N \frac{c_i l_i}{(\tilde{\lambda}_1 - \lambda_i)^2} \right)^{-1} \quad \cdots \quad \left(\frac{c_1}{\tilde{\lambda}_N - \lambda_1} \right) \left(\sum_{i=1}^N \frac{c_i l_i}{(\tilde{\lambda}_N - \lambda_i)^2} \right)^{-1} \\ \left(\frac{c_N}{\tilde{\lambda}_1 - \lambda_N} \right) \left(\sum_{i=1}^N \frac{c_i l_i}{(\tilde{\lambda}_1 - \lambda_i)^2} \right)^{-1} \quad \cdots \quad \left(\frac{c_N}{\tilde{\lambda}_N - \lambda_N} \right) \left(\sum_{i=1}^N \frac{c_i l_i}{(\tilde{\lambda}_N - \lambda_i)^2} \right)^{-1} \end{array} \right\}.$$

Once the coordinates $p = \Xi^{-1} \xi$ are changed, the matrix of the operator S_N becomes diagonal and the system equations take the form

$$\dot{p} = \begin{pmatrix} \tilde{\lambda}_1 & & & & 0 \\ & \ddots & & & \\ & & \tilde{\lambda}_N & & \\ & & & \omega_1 & \\ 0 & & & & \ddots \\ & & & & & \omega_N \end{pmatrix} p + \Xi^{-1} \bar{l}_N \langle \bar{c}_{N+m}, \tilde{x}_{N+m} \rangle_{\mathbb{R}},$$

$$\dot{\tilde{x}}_j = \lambda_j \tilde{x}_j + b_j \langle \Xi^T k_N, p \rangle_{\mathbb{R}}, \quad j = N + m + 1, \dots, \infty; \quad \bar{l}_n = (l_1, \dots, l_N, l_1, \dots, l_N)^T.$$

Take the Lyapunov function in the form $V = (\|p\| + \|\tilde{x}_{N+m}\|)$, where $\|p\|^2 = \langle p, p \rangle_{\mathbb{R}}$ and $\|\tilde{x}_{N+m}\|^2 = \langle \tilde{x}_{N+m}, \tilde{x}_{N+m} \rangle_{\mathbb{R}}$. Having differentiated it by virtue of the system we have

$$\frac{dV}{dt} \leq \left(\max_i \left\{ \max_i \operatorname{Re} \tilde{\lambda}_i, \max_i \operatorname{Re} \omega_i \right\} + \|\bar{b}_{N+m}\| \cdot \|\Xi_N^T k_N\| \right) \cdot \|p\| + \\ + (\operatorname{Re} \lambda_{N+m+1} + \|\bar{c}_{N+m}\| \cdot \|\Xi_N^{-1} l_N\|) \cdot \|\tilde{x}_{N+m}\|.$$

Consequently, if

$$\begin{aligned} & \operatorname{Re} \lambda_{N+m+1} + \|\tilde{c}_{N+m}\| \cdot \|\Xi_N^{-1} \bar{l}_N\| < \mu, \\ & \max \left\{ \max_i \operatorname{Re} \tilde{\lambda}_i, \max_i \operatorname{Re} \omega_i \right\} + \|\tilde{b}_{N+m}\| \cdot \|\Xi_N^T k_N\| < \mu, \end{aligned} \tag{10}$$

then $\frac{dV}{dt} \leq \mu V$, or the system is μ -exponentially stable. From the latter two inequalities a natural number m is chosen. For this purpose squares of the norms $\|\Xi_N^T k_N\|^2$ and $\|\Xi_N^{-1} \bar{l}_N\|^2$ are computed. We have

$$\begin{aligned} \|\Xi_N^T k_N\|^2 &= \sum_{i=1}^N \left| \sum_{\mu=1}^N l_\mu k_\mu / (\lambda_i - \lambda_\mu) \right|^2 \times \\ &\quad \times \left| 1 + \sum_{\nu=1}^N \left[(\tilde{\lambda}_i - \omega_\nu) \sum_{j=1}^N b_j k_j / (\omega_\nu - \lambda_j)^2 \right] \right|^2 + N, \\ \|\Xi_N^{-1} \bar{l}_N\|^2 &= \sum_{j=1}^N 1 / \sum_{i=1}^N c_i l_i / (\tilde{\lambda}_j - \lambda_i)^2 + \sum_{\rho=1}^N \left(\sum_{\mu=1}^N l_\mu^2 \cdot \sum_{\nu=1}^N c_\nu k_\nu \times \right. \\ &\quad \times \sum_{j=1}^N 1 / \left. \left\{ (\tilde{\lambda}_j - \lambda_\mu)(\tilde{\lambda}_j - \lambda_\nu)(\omega_\rho - \tilde{\lambda}_j) \cdot \sum_{i=1}^N c_i l_i / (\tilde{\lambda}_j - \lambda_i)^2 \right\} \right)^2 \times \\ &\quad \times \left| \sum_{i=1}^N b_i k_i / (\omega_\rho - \lambda_i)^2 \right|^{-2}. \end{aligned}$$

From these relations and inequalities (10) follows the truth of the Theorem.

This theorem makes it possible to analyse the stabilizability of and design stabilization algorithms for a wide range of systems. In the following Sections examples are provided illustrating the application of the results. For computation inequalities (8) have to be made cruder and the computation simpler. It is easily shown that for $\|\Xi_N^T k_N\|^2$, $\|\Xi_N^{-1} \bar{l}_N\|^2$ the following estimates hold

$$\begin{aligned} \|\Xi_N^T \bar{l}_N\|^2 &\leq N \frac{\|\bar{l}_N\|^2 \cdot \|k_N\|^2}{\rho^2(\tilde{\lambda}, \lambda)} \cdot \left(1 + \sum_{\nu=1}^N (\rho(\tilde{\lambda}, \omega_\nu) \cdot \left| \sum_{\mu=1}^N b_\mu k_\mu / (\omega_\nu - \lambda_\mu)^2 \right|) \right)^{-1} + N, \\ \|\Xi_N^{-1} \bar{l}_N\|^2 &\leq \sum_{j=1}^N \left| \sum_{i=1}^N c_i l_i / (\tilde{\lambda}_j - \lambda_i)^2 \right|^{-2} + \sum_{\rho=1}^N \sum_{j=1}^N \|\bar{l}_N\|^4 \cdot \|c_N\|^2 \cdot \|k_N\|^2 \times \\ &\quad \times \left(\rho^4(\tilde{\lambda}, \lambda) \rho^2(\omega, \tilde{\lambda}) \left| \sum_{i=1}^N c_i l_i / (\tilde{\lambda}_j - \lambda_i)^2 \right|^2 \cdot \left| \sum_{i=1}^N b_i k_i / (\omega_\rho - \lambda_i)^2 \right|^2 \right)^{-1}, \end{aligned} \tag{11}$$

where $\rho(\omega, \lambda) = \min_{1 \leq i, j \leq N} |\omega_i - \lambda_j|$.

Conditions (8) may be simplified by using these equations.

3. Examples

Take up diffusional equations

$$\begin{aligned} \frac{\partial v}{\partial t} &= \frac{\partial^2 v}{\partial x^2} + f(v) + g(x)u, \\ v(0, t) &= v(\pi, t) = 0 \quad \text{at any } t \geq 0, \\ v(x, 0) &= \varphi(x) \quad \text{with any } x \in (0, \pi), \end{aligned} \tag{12}$$

$y = \frac{2}{\pi} \int_0^\pi p(x)v \, dx$ is the system output. Functions $g(x)$ and $p(x)$ reflect the properties of the input and output, respectively, or the resolution of the measuring equipment and the nature of the control action. Function $f(v)$ is assumed to be uniformly Lipschitz, or $|f(v)| < F|v|$ with any $v \in \mathbb{R}$ where $F \in \mathbb{R}_+$.

It is assumed that $g(x)$ and $p(x)$ belong to, at least, $L_2(0, \pi)$ and $\varphi(x) \in H_2(0, \pi)$. Then solution of the problem (12) is understood in a generalized sense. Design a Sakawa controller for the system (12) in the case of

$$\begin{aligned} p(x) &= 10 \sin x + \sum_{i=2}^{\infty} \left(\frac{1}{2}\right)^{\frac{i-1}{2}} \sin ix, \\ g(x) &= 5 \sin x + \sum_{i=2}^{\infty} \left(\frac{1}{3}\right)^{\frac{i-1}{2}} \sin ix, \\ F &= 1. \end{aligned}$$

It is required to stabilize the system (12) μ -exponentially with $\mu = -2$. To do this it is sufficient to design μ -exponentially stabilizing feedback for the linear part of the system (12) assuming that $\mu = -3$. The Hilbert space H will be $L_2(0, \pi)$ with a scalar product $\langle v, w \rangle = \frac{2}{\pi} \int_0^\pi v\bar{w} \, dx$. The operator $A = \frac{\partial^2}{\partial x^2}$, $D(A) = \{v(0) = v(\pi) = 0; \frac{\partial^2 v}{\partial x^2} \in L_2(0, \pi)\}$ is dense everywhere, its spectrum $\sigma(A)$ is discrete, $\sigma(A) = \{-n^2; n = 1, 2, \dots\}$ and the eigenfunctions

$$\sin x, \sin 2x, \dots, \sin nx, \dots$$

add up to a basis $L_2(0, \pi)$. In this basis the linear part of the system (12) has the form

$$\begin{cases} \frac{d}{dt}T_1 = -T_1 + 5u, \\ \frac{d}{dt}T_n = -n^2T_n + \left(\frac{1}{3}\right)^{\frac{n-1}{2}}u, & n = 2, \dots, \infty \\ y = 10T_1 + \sum_{i=2}^{\infty} \left(\frac{1}{2}\right)^{\frac{i-1}{2}}T_i. \end{cases} \quad (13)$$

Because every subsystem is controllable and external resonance does not occur, any N -dimensional subsystem of the system (13) is controllable and observable. Consequently, all conditions of Theorem 2 are met and a Sakawa controller can be designed.

Stage 1. For $\bar{\mu} = -3$ with $n = 2$ we have $\lambda_2 = -4 < -3$. Consequently, $N = 2$.

Stage 2. The system $\Sigma(C_N, A_N, b_N)$ ($N = 2$) has the form

$$\begin{cases} \frac{d}{dt}T_1 = -T_1 + 5u, \\ \frac{d}{dt}T_2 = -4T_2 + \frac{1}{\sqrt{3}}u, \\ y = 10T_1 + \frac{1}{\sqrt{2}}T_2. \end{cases} \quad (14)$$

Using Corollary 3 of Theorem 3 [23] we see that the system (14) is controllable and observable. By conventional techniques the feedback is designed

$$\begin{cases} u = -2z_1 + 2\sqrt{3}z_2, \\ \dot{z}_1 = -z_1 + 5u - \frac{2}{5}(10z_1 + \frac{1}{\sqrt{2}}z_2 - y), \\ \dot{z}_2 = -4z_2 + \frac{1}{\sqrt{3}}u. \end{cases}$$

The system (14) with this feedback is easily shown to be 3-exponentially stable and with the notation of Theorem 2 we have $\omega_1 = -6$, $\omega_2 = -7$, $\tilde{\lambda}_1 = -4$, $\tilde{\lambda}_2 = -5$, $\lambda_1 = -1$, and $\lambda_2 = -4$.

Stage 3. The natural number m in (17) must be chosen so that conditions (8) of Theorem 2 hold. Using inequalities (11) we have

$$\begin{aligned} \|\Xi_N^T k_n\| &\leq 130, \\ \|\Xi_N^{-1} l_n\| &< 201\,604. \end{aligned}$$

Now, the number m is found from the inequalities

$$130 \leq \frac{1}{\sum_{\nu=2+m+1}^{\infty} \left(\frac{1}{3}\right)^{\nu-1}}$$

$$201\,604 \leq ((2+m+1)^2 - 3)^2 / \sum_{\nu=2+m+1}^{\infty} \left(\frac{1}{2}\right)^{\nu-1}.$$

From these inequalities it follows that $m \geq 6$. Consequently, the final form of the 3-exponentially stabilizing feedback is

$$u = -2z_1 + 2\sqrt{3}z_2,$$

$$\dot{z}_1 = -z_1 + 5u - \frac{2}{5} \left(10z_1 + \frac{1}{\sqrt{2}}z_2 + \frac{1}{2}z_3 + \frac{1}{2\sqrt{2}}z_4 + \frac{1}{4}z_5 + \right. \\ \left. + \frac{1}{4\sqrt{2}}z_6 + \frac{1}{8}z_7 + \frac{1}{8\sqrt{2}}z_8 - y \right),$$

$$\dot{z}_2 = -4z_2 + \frac{1}{\sqrt{3}}u, \quad \dot{z}_5 = -25z_5 + \left(\frac{1}{3}\right)^2 u,$$

$$\dot{z}_3 = -9z_3 + \frac{1}{3}u, \quad \dot{z}_6 = -36z_6 + \frac{1}{9\sqrt{3}}u,$$

$$\dot{z}_4 = -16z_4 + \frac{1}{3\sqrt{3}}u, \quad \dot{z}_7 = -49z_7 + \left(\frac{1}{3}\right)^3 u,$$

$$\dot{z}_8 = -64z_8 + \frac{1}{27\sqrt{3}}u.$$

Sakawa controllers are designed in a similar way for systems of the form

$$\frac{\partial v}{\partial t} = Lv + g(x)u,$$

$$Bv|_{\partial\Omega} = 0,$$

$$y = \int_{\Omega} p(x)v(t, x) dx,$$

where L and B are certain linear differential operators and the stationary problem

$$Lv = 0,$$

$$Bv|_{\partial\Omega} = 0$$

is elliptical.

Let us now consider the approximate controllability and observability of the system

$$\begin{aligned} \frac{\partial^2 v}{\partial t^2} &= \operatorname{div} k(x) \operatorname{grad} v + g(x)u, \\ v|_{\partial\Omega} &= 0, \\ y &= \int_{\Omega} p(x)v(t, x) dx, \end{aligned} \quad (15)$$

where $\Omega \subset \mathbb{R}^n$ is a limited region with a smooth boundary $\partial\Omega$. The Hilbert space H will be the space $L_2(\Omega)$ with a scalar product $\langle v, w \rangle = \int_{\Omega} v(x)\bar{w}(x) dx$. In this case a1)–a3) [23] are easily shown to hold, where $A v = \operatorname{div} k(x) \operatorname{grad} v$ and $D(A) = \{v|_{\partial\Omega} = 0, Av \in L_2(\Omega)\}$. It is well known that $\sigma(A) = \{\lambda_1, \lambda_2, \dots\}$ is a discrete set $0 > \lambda_1 > \lambda_2 > \dots > \lambda_n \rightarrow -\infty$ as $n \rightarrow \infty$ and the set of eigenfunctions $\{\xi(\lambda_i)\}_{i=1}^{\infty}$ form a base in $L_2(\Omega)$. In this base the problem (15) re-arranges into

$$\begin{aligned} \ddot{T}_i &= -\lambda_i T_i + g_i u, \\ y &= \sum_{i=1}^{\infty} p_i T_i, \quad i = 1, 2, \dots, \infty, \end{aligned}$$

where

$$\begin{aligned} v &= \sum_{i=1}^{\infty} T_i \xi(\lambda_i), \\ \operatorname{div} k(x) \operatorname{grad} \xi(\lambda_i) &= \lambda_i \xi(\lambda_i), \\ \xi(\lambda_i)|_{\partial\Omega} &= 0, \\ p_i &= \int_{\Omega} p(x) \bar{\xi}(\lambda_i) dx, \quad g_i = \int_{\Omega} g(x) \bar{\xi}(\lambda_i) dx. \end{aligned} \quad (16)$$

For this system all conditions of Theorem 3 [23] are met and so the system is approximately controllable and observable iff for any N the system

$$\begin{aligned} \ddot{T}_i &= -\lambda_i T_i + g_i u, \quad i = 1, 2, \dots, N, \\ y &= \sum_{i=1}^N p_i T_i \end{aligned}$$

is controllable and observable. The latter is true iff $p_i \neq 0$ and $g_i \neq 0$ for any $1, 2, \dots$. Consequently, the system (20) is approximately controllable and observable iff for any natural i

$$\int_{\Omega} p(x) \bar{\xi}(\lambda_i) dx \neq 0, \quad \int_{\Omega} g(x) \bar{\xi}(\lambda_i) dx \neq 0,$$

where $\xi(\lambda_i)$ is a solution of the elliptical boundary-value problem (14).

4. Conclusions

The findings of this paper are applicable to the analysis of a broad class of distributed parameter systems. Thus, if the stationary part of the equation system is elliptical, then the methods devised above make it possible to investigate its controllability and to design, in numerous cases, a finite-dimensional stabilizing feedback. Sakawa's basic ideas not only make it possible to design linear dynamic feedbacks such as (5) but also open up vistas for using various well-tried finite-dimensional controllers in infinite-dimensional cases. Thus, infinite-dimensional systems under compact uncertainty can be stabilized by using Sakawa modification of the adaptive Nussbaum controllers [6, 16] or binary stabilization algorithms. Especially promising is the adaptive tuning of the natural parameter m from (5) in the case of a nonlinear uniformly Lipschitz disturbance but with the Lipschitz constant unknown.

References

1. *Balakrishnan, A. V.*, Applied functional analysis. Springer-Verlag, New York, Heidelberg, Berlin, 1976.
2. *D'achenko, S. N.*, Spektral'noye razlozheniye beskonechnomernykh sistem s konechnomernym vkhodom. In "Slozh. sist. upr.", Kiev, 1987, pp. 52-57 (in Russian).
3. *Nefedov, S. A., Sholohovich, F. A.*, Kriterii stabiliziruyemosti dinamicheskikh sistem s konechnomernym vkhodom. Diff. Uranv., vol. 22, no. 2, 1986, pp. 223-228.
4. *Shkliar, B. Sh.*, K upravlyayemosti lineinykh sistem s raspredelennymi parametrami. Dokl. Nauk. USSR-1989-307, no. 3, pp. 560-563.
5. *Balakrishnan, A. V.*, Boundary control of parabolic equations: L-Q-h-Theory. Proc. Conf. on Theory of nonlinear equations, Sept. 1977. Akademie Verlag, Berlin, 1978.
6. *Cabrera, J. B. D., Furuta, K.*, Improving the robustness of Nussbaum-type regulator by the use of ϵ -modification - Local results. System & Control Letters 12 (1989), pp. 421-429.
7. *Chen, G.*, A note on the boundary stabilization of the wave equation. SIAM J. Control Optim., 19 (1981), pp. 106-113.
8. *Curtain, R. F., Pritchard, A. J.*, Infinite-dimensional linear systems theory. Lecture Notes in Control and Information Sci. B. Springer-Verlag, Berlin, New York, 1978.
9. *Kobayashi, T.* Finite-dimensional adaptive control for infinite-dimensional systems. Int. J. Contr.-1988-48, no. 1, pp. 289-302.
10. *Kobayashi, T.*, Remarks on discrete-time servomechanism design for parabolic distributed parameter system. Int. J. Syst. Sci.-1988-19, no. 7, pp. 1323-1333.
11. *Kunimatsu, N., Ito, K.*, Stabilization of a nonlinear distributed parameter vibratory system. Int. J. Control, 1988, vol. 48, no. 6, pp. 2389-2415.

12. *Lasiecka, I., Triggiant, R.*, Finite Rank, Relatively bounded perturbations of semigroup generators, Part I. Ann. Scuola Norm. Sup. Pisa, CL. Sci., (4), 12 (1985), pp. 641-668.
13. *Lions, J. L.*, Exact controllability, stabilization and perturbations for distributed systems. SIAM Review, vol. 30, no. 1 (1988), pp. 1-67.
14. *Pedersen, M.*, Boundary feedback stabilization of distributed parameter systems: an application of pseudo-differential boundary operators. Proc. 27th IEEE Conf. Decis. and Contr., Austin, Tex., Dec. 7-9, 1988. vol. 1, New York, 1988, pp. 366-368.
15. *Pohjolainen, S. A.*, Robust multivariable PI-controller for infinite-dimensional systems. IEEE Trans. Aut. Control (1982), vol. AC-27, no. 1, pp. 17-30.
16. *Pratzel-Wolters, D., Ilchmann, A., Owers, D. H.*, High-gain robust adaptive controllers for multivariable systems. Systems Control Letters 8 (1987), North Holland, pp. 397-404.
17. *Rebarber, R.*, Conditions for stability of distributed parameter systems. Proc. 27th IEEE Conf. Decis. and Contr., Austin, Tex., Dec. 7-9, 1988. vol. 1, New York, 1988, pp. 369-372.
18. *Russel, D. L.*, Controllability and stabilizability theory for partial differential equations. Recent Progress and Open Questions. SIAM Review, vol. 20, (1978), pp. 639-739.
19. *Sakawa, Y.*, Feedback control of second order evolution with damping. SIAM J. Contr. Optim., vol. 22 (3), 1984.
20. *Sakawa, Y.*, Feedback control of second order evolutions equations with unbounded observation. Int. J. Contr., vol. 41 (3), 1985, pp. 717-733.
21. *Sakawa, Y., Matsuno, F., Fukushima, S.*, Modeling and feedback control of a flexible arm. J. Robotic Syst., vol. 2, no. 4, 1985.
22. *Triggiant, R.*, On the stability problem in Banach space. J. of Mathematical Analysis and Applications, vol. 52, (1975), pp. 383-403.
23. *Korovin, S. K., Nikitina, M. G., Nikitin, S. V.*, Infinite-dimensional systems: approximate controllability and observability, Part I. Problems of Control and Information Theory, vol. 20, no. 1, Akadémiai Kiadó, Budapest, 1991, pp. 91-107.

Бесконечномерные системы. Синтез регуляторов Сакавы. Часть II

С. К. КОРОВИН, М. Г. НИКИТИНА, С. В. НИКИТИН

(Москва)

В работе рассматривается стабилизация бесконечномерных систем с конечномерными входом и выходом. Дана нижняя оценка размерности регулятора, предложенного Сакавой, стабилизирующего бесконечномерную систему с заданным показателем экспоненциальной устойчивости. Приведены примеры синтеза стабилизирующей обратной связи.

С. К. Коровин
ВНИИ системных исследований
СССР, 117312, Москва, В-312,
пр. 60-летия Октября, 9

LUMPED INPUT AND DISTRIBUTED OUTPUT SYSTEMS AT THE CONTROL OF DISTRIBUTED PARAMETER SYSTEMS

G. HULKÓ

(*Bratislava*)

(Received January 15, 1990)

The paper deals with fundamental problems of control of some classes of Distributed Parameter Systems by means of Lumped Input and Distributed Output Systems.

Keywords: Lumped Parameter Systems, Distributed Parameter Systems, Distributed Input and Distributed Output System, Lumped Input and Distributed Output System, Distributed Parameter System of Control.

1. Introduction

Distributed Parameter Systems (DPS) are mostly interpreted as Distributed input and Distributed output Systems (DDS) by means of Partial Differential Equations (PDE).

Many times it is advantageous to interpret some classes of DPS as Lumped input and Distributed output Systems (LDS), Hulkó, (1979-1990).

LDS is a new concept in Systems and Control theory. The fundamental problems of control of some classes of DPS are formulated and solved by means of LDS in the paper.

For illustration some results will be indicated from engineering practice: Self-tuning control of a temperature field at fluidized combustion in energetics.

2. Distributed input and distributed output systems

Systems the state quantities of which are given by quantity fields, by infinite-dimensional quantities or by spatial distributed parameters are DPS.

In the present theory of DPS these systems are mostly interpreted by means of PDE. It is well demonstrated, for example by the definition in Systems and Control Encyclopedia: "Distributed systems are systems which can be described or modelled by partial differential equations."; Lions (1987).

If we consider state quantities $X(x, t)$ identical with output quantities $Y(x, t)$ then partial differential relations

$$P[Y(x, t)] = Q[U(x, t)] \quad (2.1)$$

—at the specified conditions—give in the input/output relation Distributed input and Distributed output System (DDS). (Figure 1. x —space axis co-ordinate, t —time axis co-ordinate.) The present DPS theory is in fact the theory of DDS.

At attempts of applying DDS theory results principal difficulties arise in current engineering conditions. For example: How to generate in the direct way the infinite-dimensional input quantity $U(x, t)$?!

3. Lumped input and distributed output systems

The study of the operation of various classes of DPS in technology, the live and lifeless nature shows that $U(x, t)$ is very often generated indirectly by means of some generators $\{G_i\}_i = G$ of distributed input quantities $\{U_i(x, t)\}_i$. Then $U(x, t) = \sum U_i(x, t)$. Figure 2a, Figure 9, Hulkó, Mikulecký (1984), Hulkó, Kocsis (1985), Hulkó *et al.* (1985–1988), Hulkó *et al.* (1983–1989).

The system between the vector $\mathbf{U}(t) = \{U_i(t)\}_i$ and $Y(x, t)$ is an LDS, Hulkó (1987–1990), Figure 2.

Real LDS occurs frequently in practice, when

- * the controlled quantity is given as the quantity field $Y(x, t)$ and
- * practically manipulable input quantities are at disposal only as lumped quantities $\{U_i(t)\}_i$, Figure 9.

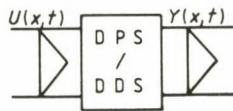


Fig. 1. Distributed Parameter System (DPS) as Distributed input and Distributed output System (DDS), $U(x, t)/Y(x, t)$ — Distributed Input/Output Quantities

Generally, control problems of LDS can not be solved on the basis of PDE theory results.

Let us model the LDS dynamics by Multi Input and Distributed Output (MIDO) system:

$$Y(x, k) = \sum_{i=1}^n gH(x, i, k) \odot U_i(k), \quad (3.1)$$

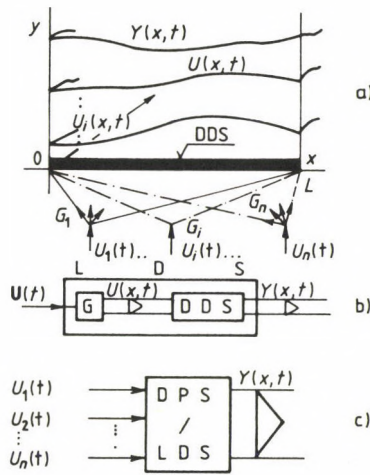


Fig. 2. Distributed Parameter System (DPS) as Lumped input and Distributed output System (LDS). $\{U_i(t)\}_i = \mathbf{U}(t)$ — Lumped input quantities; $\{G_i\}_i = G$ — Generators of distributed input quantities: $\{U_i(x, t)\}_i = \sum_{i=1}^n U_i(x, t) = U(x, t)$ — Distributed input quantity; $Y(x, t)$ — Distributed output quantity

Multi Input and Multi Distributed Output (MIMDO) system:

$$\begin{bmatrix} Y_1(x, k) \\ \vdots \\ Y_n(x, k) \end{bmatrix} = \begin{bmatrix} gH(x, 1, k) & & \\ & \ddots & \\ & & gH(x, n, k) \end{bmatrix} \circledast \begin{bmatrix} U_1(k) \\ \vdots \\ U_n(k) \end{bmatrix}, \quad (3.2)$$

Multi Input and Multi Output (MIMO) system:

$$\begin{bmatrix} Y_1(x_1, k) \\ \vdots \\ Y_n(x_n, k) \end{bmatrix} = \begin{bmatrix} gH(x_1, 1, k) & & \\ & \ddots & \\ & & gH(x_n, n, k) \end{bmatrix} \circledast \begin{bmatrix} U_1(k) \\ \vdots \\ U_n(k) \end{bmatrix} \quad (3.3)$$

where $\{gH(x, i, k)\}_{i,k}$ are discrete pulse (weighting) characteristics of LDS system and of shapers $\{H_i\}_i = H$: HLDS. “ \circledast ” is the convolution sum sign, Figures 3, 4.

Let us introduce the reduced characteristics

$$\{gHR(x, i, k) = gH(x, i, k)/gH(x_i, i, k)\}_{i,k}, \quad (3.4)$$

Figure 4. Let us transcribe models (3.1–3) by means of them:

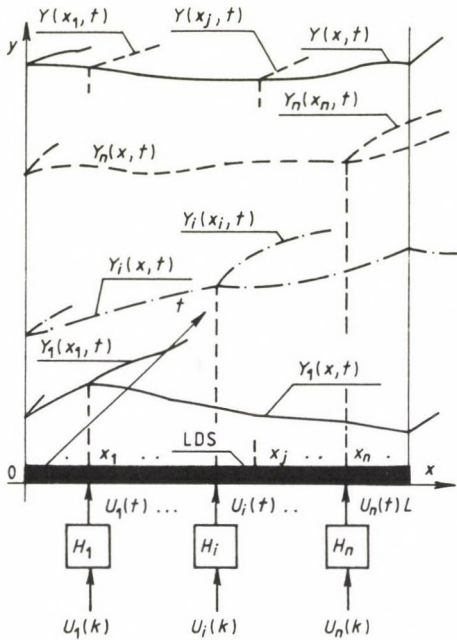


Fig. 3. Input/output quantities of LDS and pulse shapers $\{H_i\}; i = H$: HLDS.

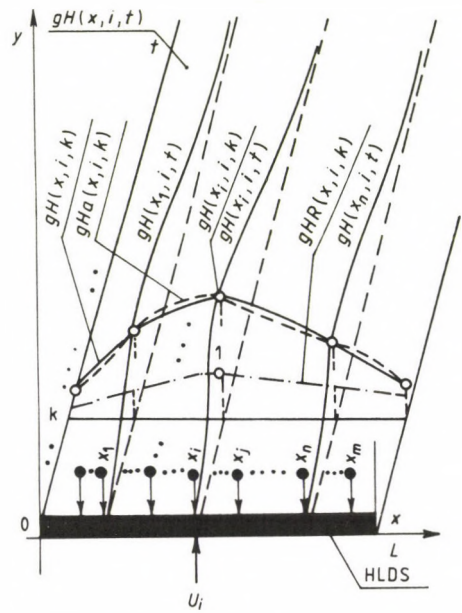


Fig. 4. i -th distributed pulse characteristics: $gH(x, i, t)$

$$Y(x, k) = \sum_{i=1}^n gHR(x, i, k)gH(x_i, i, k) \otimes U_i(k), \quad (3.5)$$

$$\begin{bmatrix} Y_1(x, k) \\ \vdots \\ Y_n(x, k) \end{bmatrix} = \begin{bmatrix} gHR(x, 1, k)gH(x_1, 1, k) \\ \ddots \\ gHR(x, n, k)gH(x_n, n, k) \end{bmatrix} \otimes \begin{bmatrix} U_1(k) \\ \vdots \\ U_n(k) \end{bmatrix} \quad (3.6)$$

$$\begin{bmatrix} Y_1(x_1, k) \\ \vdots \\ Y_n(x_n, k) \end{bmatrix} = \begin{bmatrix} gH(x_1, 1, k) \\ \ddots \\ gH(x_n, n, k) \end{bmatrix} \otimes \begin{bmatrix} U_1(k) \\ \vdots \\ U_n(k) \end{bmatrix} \quad (3.7).$$

So, the HLDS dynamics is decomposed into

- * the time, finite-dimensional components (TC):

$$\{gH(x_i, i, k)\}_{i=1, n; k}, \quad (3.8)$$

- * the space, infinite-dimensional components (SC):

$$\{gHR(x, i, k)\}_{i=1, n; k}. \quad (3.9)$$

The time component of the output quantity, trajectories $\{Y_i(x_i, k)\}_i$, (3.7) are given by the time, finite-dimensional components of the HLDS dynamics, (3.8).

Linear combinations of elements of space components of HLDS dynamics, (3.9) are shifted on trajectories (3.7) and they give distributed quantities $\{Y_i(x, k)\}_i$.

Their sum gives the whole distributed output quantity: $Y(x, k) = \sum_{i=1}^n Y_i(x, k)$.

Let us consider the actions of step functions $\{U_i(k) = 1(k)\}_{i=1, n}$ on the MIMDO model inputs at zero steady state. We obtain on the outputs distributed transient characteristics $\{hH(x, i, k)\}_{i, k}$. Let us define reduced courses of these characteristics at $t = \infty$ (Figure 5a).

$$\{hHR(x, i, \infty) = hH(x, i, \infty)/hH(x_i, i, \infty)\}_i. \tag{3.10}$$

At actions of constant input quantities $\{U_i(\infty)\}_i$ the output quantity can be expressed on MIMO model output level by means of $\{hHR(x, i, \infty)\}_i$:

$$\begin{aligned} Y(x, \infty) &= \sum_{i=1}^n hHR(x, i, \infty)hH(x_i, i, \infty)U_i(\infty) = \\ &= \sum_{i=1}^n hHR(x, i, \infty)Y_i(x_i, \infty), \end{aligned} \tag{3.11}$$

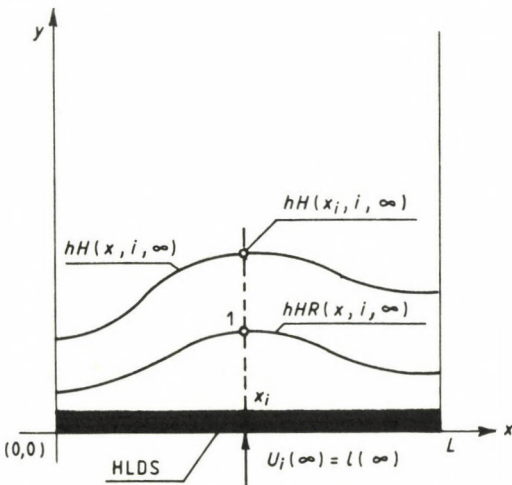


Fig. 5a.

$$hHR(x, i, \infty) = hH(x, i, \infty)/hH(x_i, i, \infty)$$

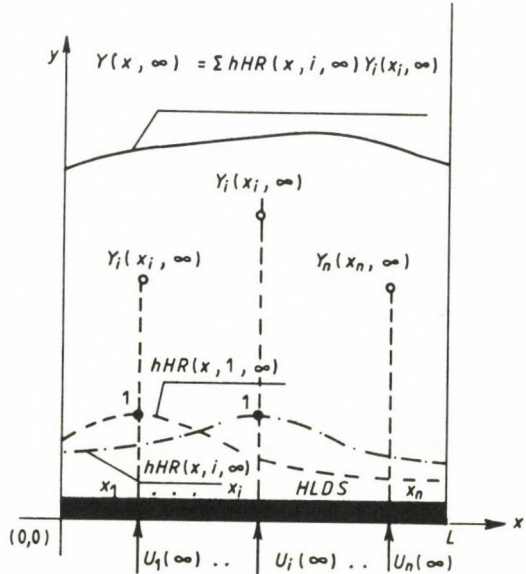


Fig. 5b. Steady values of input/output quantities

where $\{Y_i(x_i, \infty) = hH(x_i, i, \infty)U_i(\infty)\}_i$ are output quantities on the MIMO model output level (Figure 5b).

4. LDS control

Let us start from the following Distributed Parameter Discrete Control System (DPDCS), Figure 6 in the analysis of fundamental problems of LDS control.

The dynamics of controlled system HLDS is described by discrete MIDO, MIMDO, and MIMO models. The output quantity $Y(x, k)$ is modelled by means of a MIDO system in block MHLDS: $YM(x, k)$. At the same time, the relation

$$Y(x, k) = YM(x, k) \quad (4.1)$$

is assumed. The sampling “ K ” is considered only in time discretion.

Let us start from the following formulation of control task according to limited extent of this paper:

FR: The aim of the LDS control is to secure that at disturbances and desired quantity changes

α : In “ x ” direction the controlled system output quantity — on MIDO model output level; in time $t = \infty$: $\check{Y}(x, \infty)$ will in the δ -neighbourhood of the steady desired quantity $W(x, \infty)$

$$\|W(x, \infty) - \check{Y}(x, \infty)\| = \|\check{E}(x, \infty)\| \leq \delta. \quad (4.2)$$

where δ is a given real positive number and

β : in time direction the control process is of prescribed quality. For example $\{\check{E}_i(x_i, \infty) = 0\}_{i=1, n}$ on the MIMO model output values level.

First let us solve this elementary task:

FI: The aim of the LDS control is to transfer of the distributed output quantity from the infinite-dimensional steady-state value: $Y(x, \infty) = 0$

α : in “ x ” direction into δ -neighbourhood of the desired quantity $W(x, \infty)$; on the MIDO model output level

$$\|W(x, \infty) - \check{Y}(x, \infty)\| = \|\check{E}(x, \infty)\| \leq \delta \quad (4.3)$$

when δ is a given real positive number and

β : in the time direction so that $\{\check{E}_i(x_i, \infty)\}_{i=1, n}$ on the MIMO model output values level.

THEOREM TFI. Let the controlled HLDS dynamics be represented by discrete MIDO, MIMDO, and MIMO models. The control problem FI has the solution in DPDCS, (Figure 6) if

$$\alpha : \|W(x, \infty) - \sum \check{E}_i(l(hHR)x, i, \infty)\| \leq \delta \quad (4.4)$$

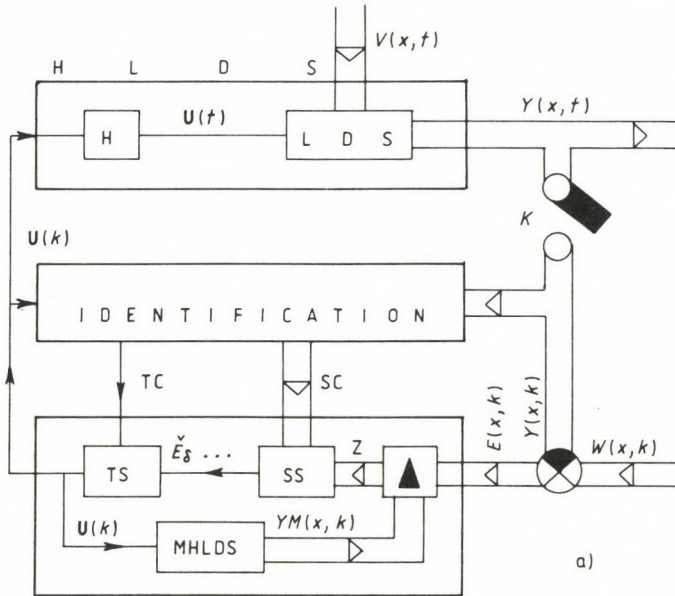


Fig. 6a. Discrete Distributed Parameter System of Control: TC/SC — Time/Space Components of HLDS dynamics; TS/SS — Time/Space Components of Control Synthesis

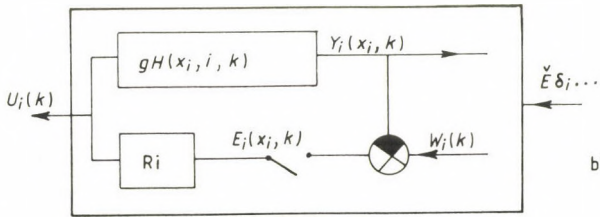


Fig. 6b. i -th lumped control system from the set $\{gH(x_i, i, k); R_i\}_i$

and β : regulators $\{R_i\}_{i=1,n}$ are of PI (proportional and integral) type.

In the followings the constructive proof of the assertions of the theorem is given from the assumptions. With regard to the limited extent of this paper it is assumed that the used mathematical objects have the necessary properties for the considered operations. At the same time these operations and formulated problems indeed have solutions. For example, the δ -controllability is assumed at the given control task. The problems of controllability is not studied here. Changes of the desired quantity and disturbances are assumed as distributed step functions. Further, the suitable relation between the sampling interval T and sampling time τ is assumed.

Proof. DPDCS is in zero steady-state in point "0":

$$\{\check{E}\delta_i(0)\}_i = \mathbf{U}(0) = \{U_i(0)\}_i = \mathbf{0}, \quad (4.5)$$

$$W(x, 0) = E(x, 0) = Z(x, 0) = V(x, 0) = Y(x, 0) = YM(x, 0) = 0. \quad (4.6)$$

A distributed step function of desired quantity $W(x, 1) = W(x, t) = W(x, \infty)$ operates on the DPDCS in the first sampling interval $\mathbf{1} = (0, 1)$, Figure 7. Then in point 1 we obtain:

$$W(x, 1) = W(x, 1) = W(x, \infty); \quad (4.7)$$

further,

$$E(x, 1) = W(x, 1) = W(x, \infty). \quad (4.8)$$

Let the following quantities be defined:

$$\Delta E(x, k) = E(x, k) - E(x, k - 1), \quad (4.9)$$

$$\Delta YM(x, k) = YM(x, k) - YM(x, k - 1). \quad (4.10)$$

These quantities are compared in block \blacktriangle :

$$Z(x, k) = \Delta E(x, k) + \Delta YM(x, k). \quad (4.11)$$

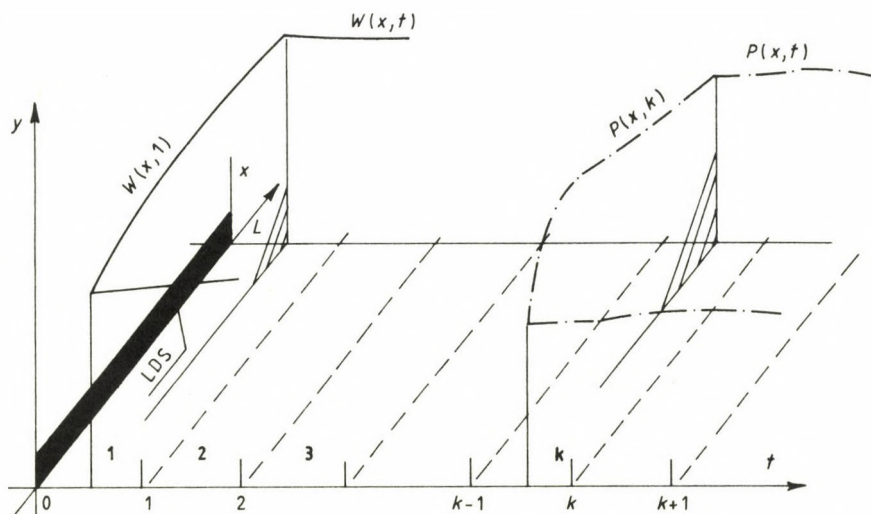


Fig. 7. Changes of distributed disturbances and desired quantities

Then we obtain on the output of this block:

$$Z(\mathbf{x}, 1) = E(\mathbf{x}, 1) = W(\mathbf{x}, \infty). \quad (4.12)$$

Let us solve this approximation task in block SS

$$\|Z(\mathbf{x}, 1) - \sum_{i=1}^n E\delta_i(1)hHR(\mathbf{x}, i, \infty)\| \leq \delta. \quad (4.13)$$

This task has the solution according to relations (4.4), (4.12) and it gives the vector

$$\check{E}\delta_i(1) = \{\check{E}\delta_1(1), \dots, \check{E}\delta_n(1)\}. \quad (4.14)$$

At the same time the approximation deviation is $R(\mathbf{x}, 1)$, Figure 8a. Components of this vector give lumped desired quantities of single discrete control loops $\{gH(\mathbf{x}_i, i, k); R_i\}_i$ in block TS, Figure 6b:

$$\{W_i(1) = \check{E}\delta_i(1)\}_i. \quad (4.15)$$

The vector of the lumped input quantities:

$$\mathbf{U}(2) = \{U_1(2), \dots, U_n(2)\} \quad (4.16)$$

is obtained from these discrete loops for interval $\mathbf{2} = (1, 2)$. The MIMO model is diagonal, therefore, the single components $\{U_i(2)\}_i$ are obtained from single one-dimensional discrete control loops $\{gH(\mathbf{x}_i, i, k); R_i\}_i$, Figure 6b. We obtain, after applying vector $\mathbf{U}(2)$ in interval $\mathbf{2}$, in point 2:

$$E(\mathbf{x}, 2) = W(\mathbf{x}, \infty) - Y(\mathbf{x}, 2). \quad (4.17)$$

Further, by relations (4.9), (4.10), (4.12):

$$\begin{aligned} \Delta E(\mathbf{x}, 2) &= E(\mathbf{x}, 2) - E(\mathbf{x}, 1) = W(\mathbf{x}, \infty) - Y(\mathbf{x}, 2) - W(\mathbf{x}, \infty) = \\ &= -Y(\mathbf{x}, 2) \end{aligned} \quad (4.18)$$

$$\Delta YM(\mathbf{x}, 2) = YM(\mathbf{x}, 2) - YM(\mathbf{x}, 1). \quad (4.19)$$

Since $\mathbf{U}(1) = 0$ and $YM(\mathbf{x}, 1) = 0$ then

$$\Delta YM(\mathbf{x}, 2) = YM(\mathbf{x}, 2). \quad (4.20)$$

Comparing $\Delta E(\mathbf{x}, 2)$ and $\Delta YM(\mathbf{x}, 2)$ we obtain:

$$\Delta E(\mathbf{x}, 2) + \Delta YM(\mathbf{x}, 2) = -Y(\mathbf{x}, 2) + YM(\mathbf{x}, 2) = Z(\mathbf{x}, 2). \quad (4.21)$$

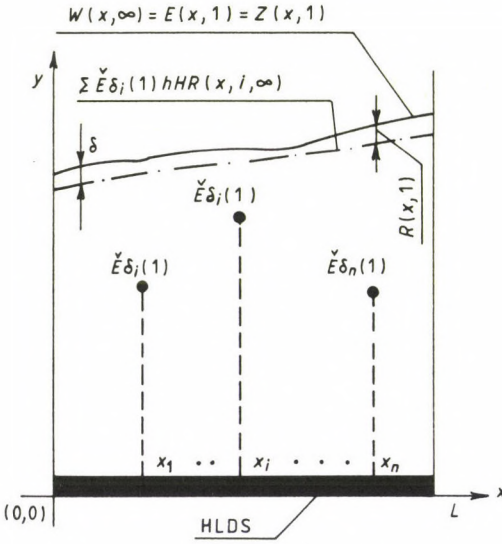


Fig. 8a. Approximation task in block SS

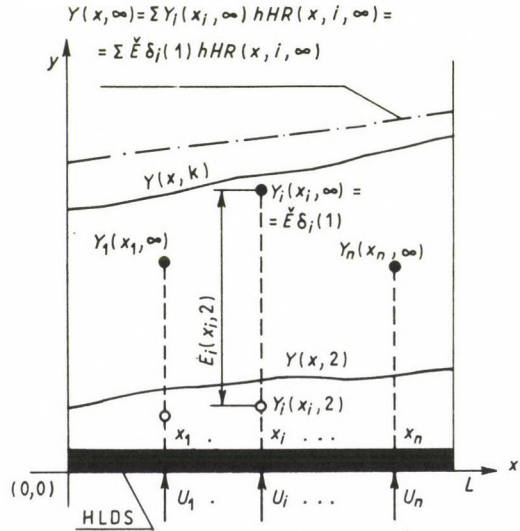


Fig. 8b. Changes of HLDS output quantities

With regard to relation (4.1)

$$Z(x, 2) = 0. \tag{4.22}$$

The SS block input quantity is zero, therefore, the output quantity is zero, too. The the lumped desired quantities $\{W_i(k)\}_{i,k}$ remain unchanged in block TS:

$$\{W_i(2) = W_i(1) = \check{E}\delta_i(1)\}_i. \tag{4.23}$$

At this procedure it holds for the further steps $k = 3, 4, \dots$

$$\{Z(x, k) = 0\}; \quad k = 3, 4, \dots \tag{4.24}$$

$$\{W_i(k) = W_i(1) = \dots = \check{E}\delta_i(1)\}; \quad k = 3, 4, \dots \tag{4.25}$$

On the basis of PI type of regulator $\{R_i\}_i$ we obtain in $t \rightarrow \infty$ on the MIMO level, Figure 6b:

$$\{\check{Y}_i(x_i, \infty) = W_i(\infty) = \dots = W_i(1) = \check{E}\delta_i(1)\}_i. \tag{4.26}$$

On the MIDO model output level this means:

$$\check{Y}(x, \infty) = \sum_{i=1}^n \check{Y}_i(x_i, \infty) hHR(x, i, \infty) = \sum_{i=1}^n \check{E}\delta_i(1) hHR(x, i, \infty); \tag{4.27}$$

(Figure 8b). So, with regard to relation (4.4) FI: α holds:

$$\|W(x, \infty) - \check{Y}(x, \infty)\| \leq \delta.$$

With regard to relation (4.26) the following relations hold in control loops $\{gH(x_i, i, k); R_i\}_i$:

$$\check{E}_i(x_i, \infty) = W_i - \check{Y}_i(x_i, \infty) = \check{E}\delta_i(1) - \check{Y}_i(x_i, \infty) = 0\}_i. \quad (4.28)$$

This means that FI: β also hold, Figure 8b. ■

Let us show that DPDCS gives the solution of the more general problem FR:, too. Starting out from the solution of FI: Up to interval "k" let the DPDCS operate so that at the proof of Theorem TFI a disturbance $P(x, k)$ appears on the HLDS output in this interval, Figure 7. Then we obtain in point "k":

$$E(x, k) = W(x, \infty) - Y(x, k) - P(x, k). \quad (4.29)$$

Since

$$E(x, k - 1) = W(x, \infty) - Y(x, k - 1), \quad (4.30)$$

$$\Delta YM(x, k) = YM(x, k) - YM(x, k - 1) \quad (4.31)$$

for

$$\Delta E(x, k) + \Delta YM(x, k) = -P(x, k). \quad (4.32)$$

In point "k" $Z(x, k)$ is given by

$$Z(x, k) = -P(x, k) + R(x, 1). \quad (4.33)$$

In point "k" $Z(x, k)$ is approximated in block SS and a vector $\{\check{E}\delta_i(k)\}_i$ is obtained. Lumped desired quantities are modified in control loops of block TS: $\{gH(x_i, i, k); R_i\}_i$, Figure 6b.

$$\{W_i(k) = \check{E}\delta_i(1) + \check{E}\delta_i(k)\}_i \quad (4.34)$$

and $U(k + 1)$ is generated, etc.

The deviation between $Z(x, k)$ and $\sum_{i=1}^n \check{E}\delta_i(k)hHR(i, k\infty)$ let be marked by $R(x, k)$. Then this deviation is added to the further distributed disturbance, which acts e.g. in the interval h ($h > k$): $Q(x, h)$

$$Z(x, h) = -Q(x, h) + R(x, h) \quad (4.35)$$

etc.

Further, changes of the desired quantity: $W(x, p), \dots, W(x, q), p > h$, are evaluated by the above scheme, too. These distributed step functions act in further time intervals on DPDCS. Then the steady distributed desired quantity is

$$W(x, \infty) = \sum_{r=1}^q W(x, r), \quad (4.36)$$

where q is a finite integer.

This means that the relation

$$\|W(x, \infty) - \check{Y}(x, \infty)\| = \|\check{E}(x, \infty)\| \leq \delta \quad (4.37)$$

is fulfilled in time $t \rightarrow \infty$ at the action of further disturbances of desired quantities. This finally, means the FR: α fulfilment.

At the control, on the MIMO level, given conditions are held, which refer to PI type of regulators $\{R_i\}_i$:

$$\check{E}_i(x_i, \infty) = W_i(\infty) - \check{Y}_i(x_i, \infty) = 0 \quad (4.38)$$

in time direction. It means that FR: β hold.

The approximated values $Ya(x, k), YMa(x, k), gHRa(x, i, k), hHRa(x, i, \infty), \dots$ etc. are considered, when "K" is for space/time sampling. The given accuracy of the approximation ε is secured by an appropriate choice of sampling in space direction $\{x_j\}_j$, Figure 4:

$$\|Ya(x, k) - Y(x, k)\| = \|YMa(x, k) - YM(x, k)\| \leq \varepsilon. \quad (4.39)$$

Then tasks of further type are solved, instead of type (4.13);

$$\left\| W(x, \infty) - \sum_{i=1}^n E\delta_i(k)hHRa(x, i, \infty) \right\| \leq \delta - \varepsilon. \quad (4.40)$$

(The control deviation $E(x, k)$ is continuously evaluated in practice in block at \blacktriangle the control process. When e.g. $\|E(x, \infty)\| \geq \delta$ at some steady-state then the output quantity of \blacktriangle is considered:

$$Z(x, \infty) = E(x, \infty),$$

etc.)

5. Self-tuning control of temperature fields at fluidized combustion in energetics

The concise example of self-tuning control of fluidized fireplace as LDS is indicated in this Section, Hulkó *et al.* (1983–89). This problem can not be solved on the basis of PDE theory. Here we show the possibility of PDE utilization to a priori information expression.

Large quantity of sulphur is often produced at the low-grade fuel combustion in the fluidized layer, Figure 9. It considerably deteriorates the quantity of environment by emissions into the atmosphere and devastates the nature, for example,

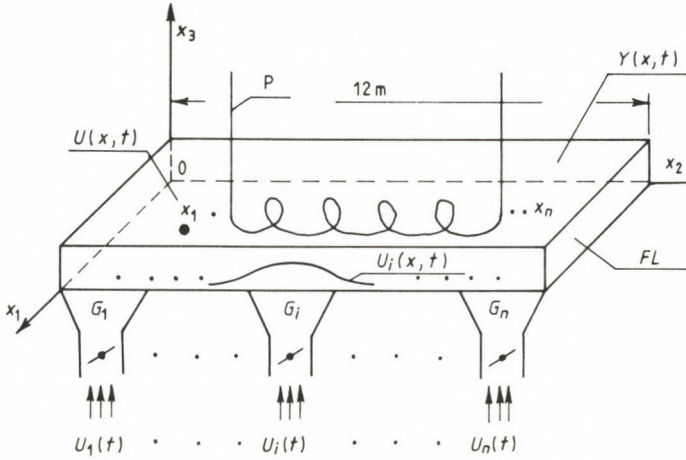


Fig. 9. Stationary fluidized layer in fireplace with pneumatic haulage: $U_i(t)$ — supplied quantity of fuel and additives, i -th lumped input quantity; G_i — i -th subsystem of pneumatic haulage, i -th generator of the distributed input quantity: $U_i(x, t)$; $U(x, t) = \sum_{i=1}^n U_i(x, t)$ — distributed input quantity; FL — fluidized layer; P — high-pressure parts of steam generator; ● — measuring points: x_1, \dots, x_m ; $Y(x, t)$ — temperature field of the fluidized layer, distributed output quantity

in the form of acid rains. The effective desulfurization is reached by suitable calcareous additives at the optimum desulphurization temperature T_{opt} . The control of temperature field will be solved by a system of self-tuning control because of the considerable fluctuation of the low-grade fuel calorific value.

Let us interpret the fluidized fireplace as an LDS, Figures 2, 9. Let us represent the FL as DDS. Its dynamics is approximated by a parabolic PDE of second order. Let us relate the Green function $G(x, \xi, t)$ to this PDE.

The dynamics of generators $\{G_i\}_i$ let be given, for example, by relation:

$$\left\{ U_i(x, t) = \int_0^t G_i(x, i, \tau) U_i(\tau) d\tau \right\}_i \quad (5.1)$$

Then distributed weighting characteristics of LDS are

$$\left\{ g(x, i, t) = \int_0^L G(x, \xi, t) G_i(\xi, i, t) d\xi \right\}_i \quad (5.2)$$

which represent the controlled system's a priori information.

The actual courses of $\{gH(x_i, i, k)\}_i$ and $\{hHR(x, i, \infty)\}_i$ are determined by on-line identification procedures.

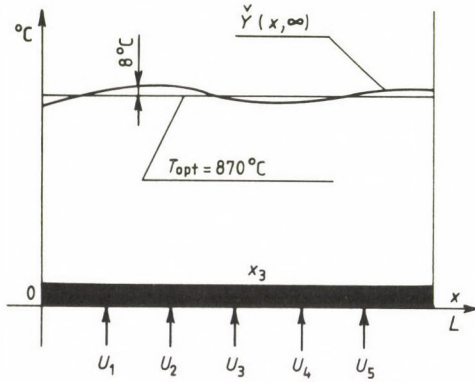


Fig. 10a. Steady-state of temperature field at self-tuning control

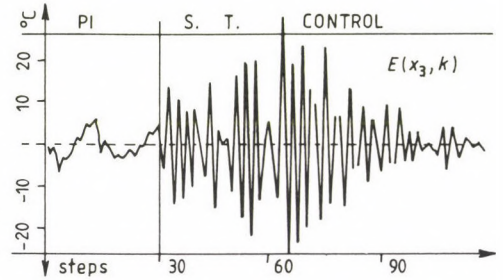


Fig. 10b. Derivation of control at x_3 : T_{opt} — $Y(x_3, k) = E(x_3, k)$. PI — preliminary identification; S.T. CONTROL — self-tuning control

Tasks of self-tuning control synthesis are solved in Section 4.

Typical courses of controlled quantities are shown in Figure 10 at self-tuning control of fluidized fireplace.

6. Conclusion

Fundamental problems of control of some classes of DPS are analysed by means of LDS in this paper. LDS is a new concept of system and control theory. DPDCS was designed for quantity fields control i.e. control of DPS.

Further results and research on these fundamental structures lead to

- * MIDO, MIMDO and MIMO systems/models identification;
- * LDS deterministic, stochastic, and adaptive control, control of nonlinear LDS;
- * optimal design of machines and machineries for quantity fields interactions as distributed objects;
- * algorithmization of tasks of optimal control and optimal design problems, Hulkó (1979–90).

These are only the basic structures of LDS theory but already these results give for the control practice of various classes of DPS the same possibilities which are at the disposal for the control of lumped parameter systems.

References

1. *Butkovskij, A. G.*, Characteristics of Distributed Parameter Systems, Nauka, Moscow, 1979 (in Russian).
2. *Lions, J. L.*, In: Systems and Control Encyclopedia — Comprehensive user's guide, Pergamon Press, Oxford, 1987.
3. *Hulkó, G.*, On Nonparametric Representation of Distributed Parameter Systems, Proc. 5th IFAC Symposium on Identification and Systems Parameter Estimation, Darmstadt, 1979.
4. *Hulkó, G., Sapák, Ľ.*, On Multilevel Decomposition of Distributed Parameter Systems, Proc. 2nd IFAC Symposium on Large Scale Systems Theory and Applications, Toulouse, 1980.
5. *Hulkó, G., Rohal-Ľikiv, B., Sapák, Ľ.*, On Adaptive Control of Distributed Parameter Systems, Preprints 8th World Congress of IFAC, Kyoto, 1981.
6. *Hulkó, G., Sapák, Ľ., Rohal-Ľikiv, B.*, Towards New Adaptive Control Algorithms for Distributed Parameter Systems, Preprints IFAC Workshop on Adaptive Control and Signal Processing, San Francisco, 1983.
7. *Hulkó, G., Mikulecký, M.*, Distributed and Nonparametric Models of Blood Circulation and Hepatobiliary Transport, In: Preprints "International Symposium on Mathematical Modelling of Liver Dye Excretion", Smolenice, Czechoslovakia, 1984.
8. *Hulkó, G.*, Distributed Parameter System — Hierarchical Control, In: Preprints of European Workshop on the Real Time Control of Large Scale Systems, Patras, 1984.
9. *Hulkó, G., Kocsis, M.*, Quantity fields control in nuclear energy, In: Preprints of "Automatized systems of control in nuclear energy symposium", Tále, Czechoslovakia, 1985.
10. *Hulkó, G.*, Control of Distributed Parameter Systems by Means Multi-Input and Multi Distributed Output Systems, Preprints 10th World Congress of IFAC, München, 1987.
11. *Hulkó, G.*, Distributed Parameter Systems Control by Means of Multi-Input and Multi Distributed Output Systems I., II., Proc. IMACS/IFAC Symposium "Distributed Parameter Systems '87", Hiroshima, 1987.
12. *Hulkó, G.*, Control of Lumped Input and Distributed Output Systems, Preprints 5th IFAC/IMACS/IFIP Symposium on Control of Distributed Parameter Systems, Perpignan, 1989.
13. *Hulkó, G. et al.*, Identification of Lumped Input and Distributed Output Systems, Preprints 5th IFAC/IMACS/IFIP Symposium on Control of Distributed Parameter Systems, Perpignan, 1989.
14. *Hulkó, G. et al.*, New Engineering Methods for Quantities Fields — Distributed Parameter Systems Control, Preprints of IFAC Symposium on Low Cost Automation, Milano, 1989.
15. *Hulkó, G. et al.*, Control of rolling-mill heat furnaces as distributed parameter systems, Research program reports, Bratislava, 1985–88.
16. *Hulkó, G. et al.*, Temperature fields control of fluidized fireplaces, Research program reports for Slovak Energetic Engineering Works, Bratislava, 1983–88.
17. *Hulkó, G. et al.*, Computer-Aided Design of Distributed Parameter Systems of Control, 11th World Congress of IFAC, Tallinn, 1990 (accepted for presentation).

**Системы со сосредоточенным входом и распределенным выходом
при управлении системами с распределенными параметрами**

Г. Хулко

(Братислава)

В статье изложена новая концепция управления системами с распределенными параметрами на базе систем со сосредоточенным входом и с распределенным выходом.

G. Hulkó

Department of Automatic Control and Measurements

Slovak Technical University

Nám. Slobody, č. 17

812 31 Bratislava, Czechoslovakia

ON THE CONSTRUCTION OF SOLUTION TO NONREGULAR PROBLEMS OF OPTIMAL CONTROL

A. G. CHENTSOV

(*Sverdlovsk*)

(Received January 20, 1990)

In this paper a problem of asymptotical optimization under perturbed constraints is investigated. Conditions of partial stability and non-sensitivity with respect to certain kinds of perturbations are obtained. A correct extension in a special class of vector-valued finitely-additive measures is suggested; natural relations between exact, generalized and approximate solutions are stated; the relations are expressed in terms of closures of admissible sets and the sets of optimal and "almost" optimal solutions. Applications to certain problems of optimal control are considered.

1. Introduction

The need to investigate extremal problems with approximate methods, to realize algorithms under conditions deviating from "nominal" ones, to seek a priori estimations for the result and its dependence on the parameters of a problem, as well as a number of other questions of practical importance, motivate a special consideration of perturbations of an initial optimization problem within a certain class corresponding to a concrete situation arising in practice. An entire investigation implies certain stability, or non-sensitivity with respect to small perturbations of data. If this does not take place, a special regularization of the problem is needed, otherwise a solution found without taking perturbations into account can lose practical importance.

In general, various components of an extremal problem can be perturbed; in this paper, however, we consider perturbations of the entire system of constraints determining the admissible set (the set of all admissible elements) of the problem. To illustrate the sense of the questions treated below, let us consider the following example relating to optimal control. (More general statements concerned with perturbations of control problems will be considered after the formulation of the basic extremal problem.) Let a system described by a vector differential equation

$$\dot{\mathbf{x}}(t) = A^0(t)\mathbf{x}(t) + f(t)\mathbf{b}^0(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0,$$

be given. The matrix $A^0(t)$ is assumed (for simplicity) to be continuous on a given interval $T_0 = [t_0, \vartheta_0]$, a control $f(t)$ is scalar and non-negative, $\mathbf{b}^0(t)$ being, in general, non-continuous, is such that there exists a solution $\mathbf{x}_f = (\mathbf{x}_f(t), t \leq t \leq \vartheta_0)$ generated by a control f which is piece-wise constant and continuous from the right on $[t, \vartheta_0[$. Let f satisfy the constraint

$$\int_{t_0}^{\vartheta_0} f(t) dt \leq c.$$

Besides, let the following constraint on possible laws of resource expenditure be imposed. Assume that a partition of $[t_0, \vartheta_0[$ with points $t_0 < t_1 < \dots < t_m = \vartheta_0$ is fixed, and a control at each interval $[t_{i-1}, t_i[$ may take one of two forms: 1) pause implying the integral $(f(t), t_{i-1} \leq t < t_i)$ to be no larger than a given a , $a > 0$; 2) impulse implying the above integral to be no smaller than b , $b > a$. In addition, we require each two impulses to be separated in time by at least one pause. The physical sense of this condition is obvious: the intervals of intensive work must be separated by the intervals of "reduction" needed for restoration of the capacity for work. Within the framework of the possibilities provided by the imposed constraints, there is, naturally, a certain space for choice. If integrals over the intervals $[t_{i-1}, t_i[$ are considered as m -dimensional vectors, the additional constraint on control functions f is reduced to the condition

$$\left(\int_{t_0}^{t_1} f(t) dt, \dots, \int_{t_{m-1}}^{t_m} f(t) dt \right) \in Y.$$

Here Y is the set of all vectors \mathbf{y} from the m -dimensional space having non-negative components and satisfying the following conditions: 1) each component (coordinate) y_i of \mathbf{y} either no larger than a or satisfies the inequality $b \leq y_i$, 2) for any y_k and y_l , $k < l$, such that

$$(b \leq y_k) \& (b \leq y_l),$$

there exists an integer r , $k < r < l$ such that $y_r \leq a$. The set Y is, in general, not convex, but it is closed. Indeed, if $\mathbf{y} = (y_1, \dots, y_m)$ is the limit of a sequence $(\mathbf{y}^{(j)} \in Y; j = 1; 2; \dots)$ then, due to the convergence $y_i^{(j)} \rightarrow y_i$ ($j \rightarrow \infty$) taking place for each integer i , $1 \leq i \leq m$, and according to 1) and the inequality $b - a > 0$ we have $(y_i \leq a) \vee (b \leq y_i)$ for the components of \mathbf{y} , since otherwise there exists a δ , $\delta > 0$ such that an interval $[y_i - \delta, y_i + \delta[$ does not contain a single point among $y_i^{(1)}, y_i^{(2)}, \dots$; the fact y_1, \dots, y_m are non-negative is obvious, too. If two components y_k and y_l , where $k < l$, lie in $[b, \infty[$, then there exists an integer r between k and l such that $y_r \leq a$. Indeed, assume to the contrary, that $y_r > a$ for any r , $k < r < l$ ($l = k + 1$ is not excluded). Then (as it was stated already) $b \leq y_i$, $k \leq i \leq l$. On

the other hand, for all sufficiently large j , the deviations $|y_k^{(j)} - y_k|, \dots, |y_l^{(j)} - y_l|$ are smaller than $b - a$, which implies (so far as $y^{(j)} \in Y$) that $b \leq y_k^{(j)}, \dots, b \leq y_l^{(j)}$; here condition 1) used in the definition of Y is taken into account. This contradicts condition 2). So, the assumption is wrong, and $y_r \leq a$ for a certain r , $k < r < l$. The proof of the closedness of Y was given in detail, for it is important for further investigation.

Having constraints upon (open-loop) controls specified, introduce for any control f , a quality functional identifying it with a value of a certain continuous function g_0 of a phase state at the terminal instant; in other words, our goal in the considered problem is the minimization of a value $g_0(\mathbf{x}_f(\vartheta_0))$ by a rational choice of f . We obtained a dynamical optimization problem of a certain practical interest. However, the natural question arises: to what extent is the problem sensitive to the perturbations, from the point of view of its optimal result (value) and the solutions f close to the value with respect to the criterion $g_0(\mathbf{x}_f(\vartheta_0))$? How does the result depend on the "energetic" parameter c and the set Y ? Stability with respect to a corresponding class of possible perturbations is of special importance. Several properties weaker than stability may be of interest, too. Thus, the above example provides a non-linear infinite dimensional problem of dynamical optimization; this problem is, in general, not convex which complicates its investigation to a considerable extent. We see that the statement of the problem is simple enough; nevertheless, the question of its stability is actual even for the following "incomplete" class of perturbations: c is replaced by $c + \varepsilon$ where $\varepsilon > 0$, and Y is replaced by its ε -neighbourhood. Indeed, the considered question transforms in this case to the following one: how close are the results of the problem of minimization of $g_0(\mathbf{x}_f(\vartheta_0))$, if all the constraints are kept, except the latter two which are disturbed slightly. The question can be formulated, however, in a more general form, for the problems analogous to the considered one may appear in other, more complicated situations. The investigation carried out below and involving the above example is qualitative, it does not have a purpose to provide algorithms. However, the investigation, in its general form, needs using rather abstract mathematical tools. Besides, the initial problem can be considered not necessarily as that of optimal control. Having in mind various applications, we formulate it in an abstract form typical for the general theory of extremal problems. Namely, in this paper we consider a non-linear infinite-dimensional extremal problem

$$\begin{aligned} W(\mathbf{f}) &\rightarrow \inf, \quad \mathbf{f} = (f_1, \dots, f_r), \\ f_1 &\in B_0^+(E, \mathcal{L}), \dots, f_r \in B_0^+(E, \mathcal{L}), \\ \sum_{i=1}^r \int_E f_i(\mathbf{x}) \eta(d\mathbf{x}) &\leq c, \quad \int_E S(\mathbf{x}) \mathbf{f}(\mathbf{x}) \eta(d\mathbf{x}) \in Y. \end{aligned} \quad (1.1)$$

Here (E, \mathcal{L}) is a measurable space with a semi-algebra [1, p. 40] of sets ($E \neq \emptyset$), $B_0^+(E, \mathcal{L})$ is the positive (in the sense of the point-wise order) cone in the lattice

$B_0(E, \mathcal{L})$ of all \mathcal{L} -step functionals on E , η is a positive (real-valued) finitely-additive measure [2, ch. III, IV] on \mathcal{L} , $c \in [0, \infty[$, S is a matrix-valued mapping on E , whose components admit uniform approximation (on E) by elements of $B_0(E, \mathcal{L})$, Y is a non-empty closed set in a corresponding finite-dimensional space. In the sequel conditions on W will be formulated, which will imply the representation $W(\mathbf{f}) = w(\mathbf{f} * \eta)$, where $\mathbf{f} * \eta$ is the indefinite (component-wise) integral of \mathbf{f} , and w is continuous in an appropriate ($*$ -weak) sense. Note that the functional W can be, in particular, be of the form

$$W(\mathbf{f}) = g_0 \left(\int_E \tilde{S}(\mathbf{x}) f(\mathbf{x}) \eta(d\mathbf{x}) \right), \quad (1.2)$$

where \mathbf{f} corresponds to (1.1), \tilde{S} is a matrix-valued mapping on E (satisfying conditions analogous to those imposed on S in (1.1)), and g_0 is a continuous function on a corresponding finite-dimensional space. This is, in particular, the case considered in the above example of a control problem.

Formulations (1.1), (1.2) can, in particular, be provided by a more general optimal control problem for a system

$$\dot{\mathbf{x}}(t) = A^0(t)\mathbf{x}(t) + B^0(t)\mathbf{f}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.3)$$

Here $A^0(\cdot)$ and $B^0(\cdot)$ are matrix-valued functions, the first one is continuous, and the second one is Borel measurable; $\mathbf{x}(t)$ is realized in an appropriate finite-dimensional space; \mathbf{x}_0 is a given initial state; \mathbf{f} corresponds to (1.1), provided E coincides with an interval $T = [t_0, \vartheta_0[$ ($t_0 < \vartheta_0$); η is the trace of the Lebesgue measure on the corresponding semi-algebra \mathcal{L} . The latter is determined in accordance with the sense of the considered problem; in some cases \mathcal{L} may coincide with the σ -algebra of all Borel subsets of T . At the same time, if only piece-wise constant and continuous from the right vector-functions \mathbf{f} are admissible, such a broad measurable space is not needed, and, provided the components of $B^0(\cdot)$ admit uniform approximation by step-functions of the above type, one can identify \mathcal{L} with the family of all half-open intervals $[a, b[$, $t_0 \leq a \leq b \leq \vartheta_0$, and η with the function of length. The condition of non-negativeness of f_i corresponds to the assumption that these functions are resources for a control device; in this case the c -constraint in (1.1) provides a condition on the entire resource, and the Y -constraint provides (as in the example) the admissible set of resource expenditure laws. We have a finite system of "shields" which can not be used in an arbitrary way (for instance, as it was mentioned, a prolonged "extremal" resource expenditure may not be admissible). As for (1.2), not that for the control system (1.3), this condition is satisfied, if the quality of a process is estimated by a continuous function of a terminal state. If the dependence on a trajectory is more complicated, a representation of W in terms of a special functional w should be used instead of (1.2).

The general extremal problem (1.1) is not stable, if only relaxation of the constraints is admissible (this is a typical class of perturbations considered in the theory of extension, see [3, ch. III]). This gives reason to investigate various methods of asymptotical optimization under the "weak" perturbations of the constraints. Two kinds of such perturbations will be considered: 1) "entire" relaxation implying perturbations of both c - and Y -constraints, and 2) "partial" relaxation keeping c -constraints unperturbed. The rate of perturbation will be characterized by a scalar parameter ε , $\varepsilon > 0$. However, in a formal statement of the problem ε is normally not fixed. In this case the "limit" situation corresponding to $\varepsilon \downarrow 0$ is of practical interest. Every ε -perturbed problem is characterized by a finite or infinite value monotonically depending on ε , $\varepsilon > 0$; the smaller is an ε , the more "hard" is an ε -perturbed problem and, consequently, the worse is the quality provided by a value. The limits of such ε -values (corresponding to the cases 1) and 2)) as $\varepsilon \downarrow 0$, characterize a special optimality of the problem (1.1) which may differ from the "usual" optimality. Further, we investigate conditions and possible variants of partial stability and non-sensitivity to the perturbations of certain kinds; besides, we consider the optimization in a class of approximate solutions, having a sense of special regularization. Here we find natural analogies with the theory of ill-posed problems [4, 5]. However, the most important analogies concern numerous investigations on extensions of extremal problem [1, 6-10] (see, in particular, [1, ch. III]). The basic element here is the compactification of solutions of a problem, it is close to the extensions or compactifications of topological spaces (see, for instance, [11, 12]). In standard procedures of extension of an extremal problem a considerable role plays "convexivization" realized usually with the help of measures (as a rule, Borel, regular, non-negative and normed) [13-16, 3, 6, 8, 1]. Sometimes it is necessary to apply finitely-additive measures; this is connected with using non-continuous functionals in the statement of a problem. Such a situation takes place in the problem (1.1); usage of finitely-additive measures is advisable here also in the cases, where the measure η is countably-additive, as in the control problem for the system (1.3). Extensions (of extremal problems) within the class of finitely-additive measures were considered in [17-20] and other papers. Here we follow the approach of [21-23] embedding "ordinary" solutions \mathbf{f} into a corresponding space of finitely-additive measures through the indefinite integrals. As the final result for the problem (1.1) we obtain conditions of "partial" stability and regularizability with respect to the perturbations of certain classes; all conditions are given in terms of the initial problem and can be verified directly. The next Section contains a list of general mathematical notions and, by first reading, can be omitted without prejudice to understanding of the main results.

2. Finitely-additive measures

In what follows we use quantors, connectives, special symbols \triangleq , def, etc. For an arbitrary set H , 2^H stands for the set of all non-empty subsets of H , and $\text{Fin}(H)$ stands for the set of all finite sets from 2^H . If A and B are non-empty sets, denote by B^A the set of all mappings from A to B ; if $g \in B^A$ and $G \in 2^A$, then $(g | G) \in B^G$ is, as usual, [2, p. 13] the trace or the contraction of g to the set G . If (T, τ) , $T \neq \emptyset$, is a topological space, then $\mathcal{C}(T, \tau)$ is def the set of all τ -continuous functionals on T , and $\text{cl}(\cdot, \tau)$ is the operator of closure in (T, τ) . In the sequel, \mathbf{R} is the real line, $\mathcal{N} \triangleq \{1; 2; \dots\}$. We set $\forall k \in \mathcal{N} : \overline{1, k} \triangleq \{i : i \in \mathcal{N}, i \leq k\}$. $R_k \triangleq \mathbf{R}^{\overline{1, k}}$ (k -dimensional arithmetical space). Fix positive integers $n \in \mathcal{N}$ and $r \in \mathcal{N}$; we let for brevity $\mathfrak{N} \triangleq R_n$, $\mathfrak{R} \triangleq R_r$, fixing the n -dimensional and r -dimensional spaces, respectively, which will be used for the description of the problem (1.1). Besides, in (1.1) S is a mapping with values in $\mathfrak{M} \triangleq \mathbf{R}^{\overline{1, n} \times \overline{1, r}}$, i.e. $S \in \mathfrak{M}^E$. Other assumptions and notations of Section 1 we keep without additional explanations (note only that $Y \in 2^{\mathfrak{N}}$ is closed in \mathfrak{N} , which is equipped for definiteness with the sup-norm $\|\cdot\|_n$). The mentioned notations are general: we fix certain notations connected with the finitely-additive measures [2, ch. III, IV] (for details see also [24, 25]). Let $(\text{add})_+[\mathcal{L}]$ be the set of all non-negative finitely-additive measures on \mathcal{L} , and $\mathbf{A}(\mathcal{L})$ be the set of all measures

$$\mu - \nu, (\mu, \nu) \in (\text{add})_+[\mathcal{L}] \times (\text{add})_+[\mathcal{L}],$$

it is a linear subspace $\mathbf{R}^{\mathcal{L}}$ equipped with the traditional norm (variation). Besides, note that $\eta \in (\text{add})_+[\mathcal{L}]$ (see ch. 1); denote $(\text{add})^+[\mathcal{L}; \eta]$ the set of all measures $\mu \in (\text{add})_+[\mathcal{L}]$ such that $\forall L \in \mathcal{L}$ from $\eta(L) = 0$, it follows that $\mu(L) = 0$. Assume also that $\forall b \in [0, \infty[$:

$$\Pi[b] \triangleq \{\mu \in (\text{add})_+[\mathcal{L}] \mid \mu(E) \leq b\}, \quad (2.1)$$

$$\begin{aligned} \Xi[b] &\triangleq (\text{add})^+[\mathcal{L}; \eta] \bigcap \Pi[b] = \\ &= \{\mu \in \Pi[b] \mid \forall L \in \mathcal{L} : (\eta(L) = 0) \Rightarrow (\mu(L) = 0)\}. \end{aligned} \quad (2.2)$$

Further, we shall introduce “vector” analogue of the sets (2.1), (2.2), necessary for an extension of the problem (1.1).

Denote by $\mathbf{B}(E)$ the set of all bounded functionals on E equipped with the natural sup-norm $\|\cdot\|$ [2, p. 261], and denote by $B(E, \mathcal{L})$ the closure of $B_0(E, \mathcal{L})$ (see Section 1) in $(\mathbf{B}(E), \|\cdot\|)$; if \mathcal{L} is a σ -algebra of sets, then $B(E, \mathcal{L})$ coincides with the set of all \mathcal{L} -measurable functionals from $\mathbf{B}(E)$. In this connection $B^*(E, \mathcal{L})$ (topologically conjugate to $B(E, \mathcal{L})$) and $\mathbf{A}(\mathcal{L})$ with the norm defined as variation are isometrically isomorphic, that allows us to consider the pair $(B(E, \mathcal{L}), \mathbf{A}(\mathcal{L}))$ as a duality and the simplest integral [24, p. 75] (used below) as a bilinear form, respectively. We equip $\mathbf{A}(\mathcal{L})$ with the natural $*$ -weak topology $\tau_*(\mathcal{L})$, and interpret

$B(E, \mathcal{L})$ as a pre-conjugate space. The conditions of compactness in $(A(\mathcal{L}), \tau_*(\mathcal{L}))$ are determined by the well-known Alaoglu's theorem [2, p. 459]. In what follows $\forall \mathbf{f} \in B(E, \mathcal{L})$ the measure $\mathbf{f} * \eta \in A(\mathcal{L})$ is the indefinite η -integral of \mathbf{f} [24, p. 76]; we shall need (as in Section 1) a vector analogue of this notion implying component-wise integration. Now, consider subspaces $(A(\mathcal{L}), \tau_*(\mathcal{L}))$, important in further investigation. Namely, denote by $\tau_\eta^*(\mathcal{L})$ the trace of $\tau_*(\mathcal{L})$ at the (non-empty) set $(\text{add})^+[\mathcal{L}; \eta]$. Besides, $\forall b \in [0, \infty[$ denote by $\tau_b^*(\mathcal{L})$ the trace of $\tau_*(\mathcal{L})$ at $\Pi[b]$. At last, $\forall b \in [0, \infty[$ consider (following [21, 22]) the topology $\tau_b^0(\mathcal{L})$ of the set $\Pi[b]$, generated by the base of the sets

$$\{\nu \in \Pi[b] \mid \forall X \in \mathcal{X} : \mu(X) = \nu(X)\},$$

$$(\mu, \mathcal{X}) \in \Pi[b] \times \text{Fin}(\mathcal{L});$$

note that $\tau_b^*(\mathcal{L}) \subset \tau_b^0(\mathcal{L})$. Thus, in the last case we actually consider $\Pi[b]$ as a subspace of the Tikhonov's product of \mathcal{L} samples of the real line with the discrete topology. The indicated structure was introduced in [21, 22]. It is convenient for us to pass $\forall b \in [0, \infty[$ to the relative topology $\Xi[b]$ setting $\tilde{\tau}_b^0[\mathcal{L}]$ to be the trace of $\tau_b^0(\mathcal{L})$ at $\Xi[b]$ so that $(\Xi[b], \tilde{\tau}_b^0[\mathcal{L}])$ is a subspace of $(\Pi[b], \tau_b^0(\mathcal{L}))$. Now, consider vector analogue of the notions from [21, 22] (for details, see [23]). In this connection we need to use vector functions on E and vector measures. Further, we denote components of a vector function and a vector measure by a letter and initial object equipping it with a subindex. The same agreement we set for components of matrix-valued functions. By $(\mathcal{R}\text{-add})^+[\mathcal{L}; \eta]$ we denote the set of all functions $\mu \in \mathcal{R}^{\mathcal{L}}$ such that $\mu_i \in (\text{add})^+[\mathcal{L}; \eta]$ ($i \in \overline{1, r}$); the needed set of vector measures is defined. Let $B_0^+[E; \mathcal{L}; \mathcal{R}]$ be the set of all functions from \mathcal{R}^E , whose components belong to $B_0^+(E, \mathcal{L})$ and $B[E; \mathcal{L}; \mathfrak{N}]$ be the set of all functions from \mathfrak{N}^E with components from $B(E, \mathcal{L})$. All components of S are assumed to lie in $B(E, \mathcal{L}) : S_{i,j} \in B(E, \mathcal{L})$ for $i \in \overline{1, n}, j \in \overline{1, r}$. Integration of vector functions with respect to the measure η is component-wise, we keep all the above notations for definite and indefinite η -integrals, this will not lead to ambiguity: $\forall \mathbf{f} \in B_0^+[E; \mathcal{L}; \mathcal{R}]$ we have $\mathbf{f} * \eta \in (\mathcal{R}\text{-add})^+[\mathcal{L}; \eta]$. The topologization of $(\mathcal{R}\text{-add})^+[\mathcal{L}; \eta]$ is realized through the r -multiple product of $\tau_\eta^*(\mathcal{L})$, it is identically characterized by the class of convergence: a directedness [11, p. 96] $\mu^{(\alpha)}$ in $(\mathcal{R}\text{-add})^+[\mathcal{L}; \eta]$ converges to a measure μ from this set in the sense of the topology $\Theta_\eta^*(\mathcal{L})$ in question, iff $\forall i \in \overline{1, r}$ the directedness $(\mu_i^{(\alpha)})$ converges to μ_i in the sense of $\tau_\eta^*(\mathcal{L})$. Note that $\forall b \in [0, \infty[$:

$$\hat{\Xi}_{\mathcal{R}}[b] \triangleq \left\{ \mu \in (\mathcal{R}\text{-add})^+[\mathcal{L}; \eta] \mid \sum_{i=1}^r \mu_i(E) \leq b \right\} \tag{2.3}$$

is a non-empty and compact set in the sense of $\Theta_\eta^*(\mathcal{L})$. Consider also an other topological lattice, introducing $\forall b \in [0, \infty[$ the set $\hat{\Xi}_{\mathcal{R}}^\times[b]$ of all measures $\mu \in$

$(\mathcal{R}\text{-add})^+[\mathcal{L}; \eta]$ such that $\forall i \in \overline{1, r} : \mu_i \in \Xi[b]$; we topologize this set through the product of r samples of the topologies $\tilde{\tau}_b^0[\mathcal{L}]$. Namely, we equip $\forall b \in [0, \infty[$ the set $\Xi_{\mathcal{R}}^x[b]$ with the topology $\Theta_b^0[\mathcal{L}]$ determined identically by the following convergence; a directedness $(\mu^{(\alpha)})$ in $\Xi_{\mathcal{R}}^x[b]$ converges (in the sense of $\Theta_b^0[\mathcal{L}]$) to a $\mu \in \Xi_{\mathcal{R}}^x[b]$, iff $\forall i \in \overline{1, r}$ the directedness $(\mu_i^{(\alpha)})$ converges to μ_i in the sense of $\tilde{\tau}_b^0[\mathcal{L}]$. Besides, assume that $\forall b \in [0, \infty[$ $\hat{\nu}_b^*(\mathcal{L})$ is def the trace of $\Theta_{\eta}^*(\mathcal{L})$ at $\hat{\Xi}_{\mathcal{R}}[b]$ and $\hat{\nu}_b^0[\mathcal{L}]$ is the trace of $\Theta_b^0[\mathcal{L}]$ at $\hat{\Xi}_{\mathcal{R}}[b]$; as the result we obtain the compactum $(\hat{\Xi}_{\mathcal{R}}[b], \hat{\nu}_b^*(\mathcal{L}))$ and the auxiliary topological space $(\hat{\Xi}_{\mathcal{R}}[b], \hat{\nu}_b^0[\mathcal{L}])$ with a "degenerate" type of convergence (see [24]). The indicated constructions concern an extension of the problem (1.1). In order to give a more brief characterization of the problem, assume in addition that

$$M_b^* \triangleq \left\{ \mathbf{f} \in B_0^+[E; \mathcal{L}; \mathcal{R}] \mid \sum_{i=1}^r \int_E f_i(\mathbf{x}) \eta(d\mathbf{x}) \leq b \right\},$$

$$\tilde{M}_b^* \triangleq \{ \mathbf{f} * \eta : \mathbf{f} \in M_b^* \};$$

the latter is a subset of (2.3); moreover, we have the following important property of density [23]:

$$\hat{\Xi}_{\mathcal{R}}[b] = \text{cl}(\tilde{M}_b^*, \hat{\nu}_b^0[\mathcal{L}]) = \text{cl}(\tilde{M}_b^*, \hat{\nu}_b^*(\mathcal{L})). \quad (2.4)$$

Now, we end the list of notions concerned with measures and pass to the exact statement of the extremal problems.

3. Asymptotical optimization

Fix a functional

$$w \in C((\mathcal{R}\text{-add})^+[\mathcal{L}; \eta], \Theta_{\eta}^*(\mathcal{L})), \quad (3.1)$$

$$W \triangleq (w(\mathbf{f} * \eta))_{\mathbf{f} \in B_0^+[E; \mathcal{L}; \mathcal{R}]}. \quad (3.2)$$

This will be used as the criterion for the problem (1.1).

Let us consider $\forall \varepsilon \in [0, \infty[$ the problem

$$W(\mathbf{f}) \longrightarrow \inf, \quad \mathbf{f} \in M_{c+\varepsilon}^*,$$

$$\int_E S(\mathbf{x}) f(\mathbf{x}) \eta(d\mathbf{x}) \in Y_{\varepsilon} \quad (3.3)$$

where

$$Y_{\varepsilon} \triangleq \{ \mathbf{z} : \mathbf{z} \in \mathfrak{N}, \inf_{\mathbf{y} \in Y} \|\mathbf{y} - \mathbf{z}\|_n \leq \varepsilon \}.$$

Then (1.1) is (3.3) with $\varepsilon = 0$. Let us also introduce $\forall \varepsilon \in [0, \infty[$ the problem

$$W(\mathbf{f}) \longrightarrow \inf, \quad \mathbf{f} \in M_c^*,$$

$$\int_E S(\mathbf{x})f(\mathbf{x})\eta(d\mathbf{x}) \in Y_\varepsilon. \tag{3.4}$$

Further, (3.3) and (3.4) will define two families of relaxed perturbed problems. If $\varepsilon \in [0, \infty[$, then denote by $\Lambda[\varepsilon]$ (by $\Lambda_*[\varepsilon]$) the set of all admissible elements for the problem (3.3) (for the problem (3.4)). Finally, put $F \triangleq \Lambda[0]$ getting the set of all admissible elements for the non-perturbed problem.

Now, let us introduce a generalized construction putting

$$\forall \mu \in (\mathcal{R}\text{-add})^+[\mathcal{L}; \eta]$$

$$\int_E S(\mathbf{x})\mu(d\mathbf{x}) \triangleq \left(\sum_{j=1}^r \int_E S_{i,j}(\mathbf{x})\mu_j(d\mathbf{x}) \right)_{i \in \overline{1, n}}$$

and obtaining, naturally, a vector from \mathfrak{N} . The problems

$$w(\mu) \longrightarrow \min, \quad \mu \in \hat{\Xi}_{\mathcal{R}}[c + \varepsilon], \quad \int_E S(\mathbf{x})\mu(d\mathbf{x}) \in Y_\varepsilon \tag{3.5}$$

where $\varepsilon \in [0, \infty[$ will be called generalized ones (however, the case $\varepsilon = 0$ will more frequently be used). Denote $\forall \varepsilon \in [0, \infty[$ by $\Xi_{\mathcal{R}}^0(\varepsilon)$ the set of all admissible elements for the problem (3.5).

THEOREM 3.1. The set of admissible elements for the generalized problem and those for the perturbed problems are connected by the following limit relation

$$\begin{aligned} \Xi_{\mathcal{R}}^0(0) &= \bigcap_{\varepsilon > 0} \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \Lambda[\varepsilon]\}, \Theta_\eta^*(\mathcal{L})) = \\ &= \bigcap_{\varepsilon > 0} \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \Lambda_*[\varepsilon]\}, \Theta_\eta^*(\mathcal{L})) = \bigcap_{\varepsilon > 0} \text{cl}(\{\mathbf{f} * \eta : \\ &\mathbf{f} \in \Lambda_*[\varepsilon]\}, \hat{\vartheta}_c^*(\mathcal{L})) = \bigcap_{0 < \varepsilon \leq 1} \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \Lambda[\varepsilon]\}, \hat{\vartheta}_{c+1}^*(\mathcal{L})). \end{aligned} \tag{3.6}$$

Only the first two equalities in (3.6) are actually to be proved. Denote the second and the third intersections in (3.6) by A_1 and A_2 , respectively. The inclusion $A_1 \subset \Xi_{\mathcal{R}}^0(0)$ follows from the definition of the $*$ -weak topology, here the property of continuous dependence of an integral on a corresponding measure and the fact that Y is closed are used. The inclusion $\Xi_{\mathcal{R}}^0(0) \subset A_2$ follows from (2.4). At last

$\Lambda_*[\varepsilon] \subset \Lambda[\varepsilon]$ for $\varepsilon > 0$ passing to the closures of the corresponding sets, we get $A_2 \subset A_1$.

THEOREM 3.2. Let $\forall i \in \overline{1, n}, j \in \overline{1, r} : S_{i,j} \in B_0(E, \mathcal{L})$.

Then

$$\begin{aligned} \Theta_{\mathcal{R}}^0(0) &= \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in F\}, \hat{\vartheta}_c^*(\mathcal{L})) = \\ &= \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in F\}, \vartheta_c^0[\mathcal{L}]). \end{aligned}$$

The proof exploits (2.4) and the embedding $\hat{\vartheta}_c^*(\mathcal{L}) \subset \vartheta_c^0[\mathcal{L}]$ is deduced easily from the definitions

THEOREM 3.3. The following three conditions are equivalent:

- 1) $\Xi_{\mathcal{R}}^0(0) \neq \emptyset$;
- 2) $\Lambda[\varepsilon] \neq \emptyset$ ($\varepsilon > 0$);
- 3) $\Lambda_*[\varepsilon] \neq \emptyset$ ($\varepsilon > 0$).

THEOREM 3.4. If $\forall i \in \overline{1, n}, j \in \overline{1, r}$ the inclusion $S_{i,j} \in B_0(E, \mathcal{L})$ is true, then the conditions $F \neq \emptyset$ and $\Xi_{\mathcal{R}}^0(0) \neq \emptyset$ are equivalent.

Theorem 3.3 is a trivial corollary of Theorem 3.1, and Theorem 3.4 follows from Theorem 3.2. We assume in the sequel the following condition to be fulfilled:

Condition 3.1. $\Xi_{\mathcal{R}}^0(0) \neq \emptyset$.

Now, $\forall \varepsilon \in]0, \infty[$ we have: $\Lambda[\varepsilon] \in 2^{M_{c+\varepsilon}^*}$, $\Lambda_*[\varepsilon] \in 2^{M_c^*}$. Taking into account (3.1) and (3.2) and the fact that a continuous functional on a compact space is bounded [11, 12], we get that the "ordinary" ε -values

$$v_\varepsilon \triangleq \inf_{\mathbf{f} \in \Lambda[\varepsilon]} W(\mathbf{f}), \quad v_\varepsilon^* \triangleq \inf_{\mathbf{f} \in \Lambda_*[\varepsilon]} W(\mathbf{f})$$

are finite for $\varepsilon > 0$. More than that, since $v_\delta \leq v_\delta^*$ for $\delta > 0$, the set $\{v_\varepsilon^* : \varepsilon \in]0, \infty[\}$ (and, consequently, the set $\{v_\varepsilon : \varepsilon \in]0, \infty[\}$) is bounded above, so

$$V \triangleq \sup_{\varepsilon > 0} v_\varepsilon \in \mathbf{R}, \quad V^* \triangleq \sup_{\varepsilon > 0} v_\varepsilon^* \in \mathbf{R} \quad (3.7)$$

are asymptotical values from Section 1. Besides

$$\tilde{v}[\varepsilon] \triangleq \min_{\mu \in \Xi_{\mathcal{R}}^0(\varepsilon)} w(\mu) \in \mathbf{R} \quad (3.8)$$

is obviously the value of the problem (3.5), where $\varepsilon \geq 0$. A connection between (3.7) and (3.8) will be considered in the next Section, devoted to the application of compactifications for the description of asymptotical values of practical interest.

4. Asymptotics and extensions

Expressing of various asymptotical values (3.7) in terms of generalized (compactified) problems traditionally played a considerable role in investigations on

extremal problems (see [3]). This investigation has, in addition, the purpose to compare various variants of perturbed problems; from this point of view $\tilde{v}[0]$ (3.8) plays the role of a “mediator” that is seen from the following assertion on an asymptotical non-sensitivity to the perturbations of the energetic parameter “ c ”.

THEOREM 4.1. The values (3.7) coincide and are determined by (3.8) for $\varepsilon = 0$; i.e. $V = V^* = \tilde{v}[0]$.

The proof exploits Theorem 3.1 and concretizes [27]. Note that the problem (1.1) in “rigid” statement is unstable provided the parameter c , $c \geq 0$, allowed to grow, and therefore, the “relaxation” of the Y -restriction has a sense of a special regularization [4]. As shows [18], the opposite combination, i.e. a regularization of a non-continuous dependence from Y for a perturbation of a form $Y \rightarrow Y_\varepsilon$ ($\varepsilon > 0$) through an additional perturbation of the parameter c , is, in general, impossible.

THEOREM 4.2. Let $\mu_0 \in \Xi_{\mathcal{R}}^0(0)$ be a solution of the problem (3.5) for $\varepsilon = 0$: $w(\mu_0) = \tilde{v}[0]$. Let, in addition, (\mathbf{f}_α) be a directedness [11, p. 96] in M_c^* such that $(\mathbf{f}_\alpha * \eta)$ converges to μ_0 in $\hat{\vartheta}_c^*(\mathcal{L})$. Then (\mathbf{f}_α) is an optimal approximate solution [27], i.e.

1) $\forall \varepsilon \in [0, \infty[$ inclusions $\mathbf{f}_\alpha \in \Lambda_*[\varepsilon]$ take place starting from a certain instant [11, p. 96];

2) $(W(\mathbf{f}_\alpha))$ converges to V^* .

The proof follows rather evidently from the definition of the $*$ -weak topology (3.1) and (3.2). Note that, as it is seen from (2.4) and (3.5), a directedness (\mathbf{f}_α) can be built up constructively (the “component-wise” variant of an approximate directedness [21] for the “scalar” modification of (2.4) can be applied), provided a solution μ_0 of (3.5) is fixed; the existence of μ_0 follows from the well-known properties of continuous functionals on compact spaces [11, p. 217].

5. Some questions, connected with stability

For each problem (3.5) (by Condition 3.1) we have $\Xi_{\mathcal{R}}^0(\varepsilon) \neq \emptyset$ for $\varepsilon \geq 0$. Introduce $\forall \varepsilon \in [0, \infty[$ the (non-empty) set $\tilde{V}[\varepsilon]$ of all solutions of (3.5); $\tilde{V}[\varepsilon]$ is the set of all $\mu_0 \in \Xi_{\mathcal{R}}^0(\varepsilon)$ such that $w(\mu_0) = \tilde{v}[\varepsilon]$. As $\varepsilon \downarrow 0$ the directedness $(\tilde{v}[\varepsilon], \varepsilon > 0)$ converges to $\tilde{v}[0]$ (the ordering of the semi-axis $[0, \infty[$ is dual to the natural one), so $\tilde{v}[0] = \sup(\{\tilde{v}[\varepsilon] : \varepsilon \in]0, \infty[\})$. Thus, the generalized problem is always stable with respect to the value.

THEOREM 5.1. Let H_* be an arbitrary $\Theta_\eta^*(\mathcal{L})$ -neighbourhood [28, p. 19] of the set $\tilde{V}[0]$. Then $\exists \delta \in]0, \infty[: \forall \varepsilon \in]0, \delta[$ it holds $\tilde{V}[\varepsilon] \subset H_*$.

The proof utilizes standard representations of the compactness in terms of converging subdirectedness [12, p. 203] (the opposite assumption leads easily to a contradiction). Note that, as it is shown in [18], the initial problem (1.1) is not stable in the above sense (even with respect to the value). In the sequel we suppose the following condition (together with Condition 3.1) to be fulfilled:

Condition 5.1. The matrix-valued function S is a step-function: $S_{i,j} \in B_0(E, \mathcal{L})$ for $i \in \overline{1, n}$ and $j \in \overline{1, r}$.

According to Theorem 3.4 $F \neq \emptyset$, this allows us to introduce the ordinary value $V^0 \triangleq \inf(\{W(\mathbf{f}) : \mathbf{f} \in F\}) \in \mathbb{R}$ and to extend the dependence $\varepsilon \mapsto v_\varepsilon :]0, \infty[\rightarrow \mathbb{R}$ to the point $\varepsilon = 0$; namely, we put $v_0 = v_\varepsilon |_{\varepsilon=0} \triangleq V^0$.

THEOREM 5.2. Ordinary, generalized and asymptotical values coincide: $V^0 = \tilde{v}[0] = V = V^*$.

The proof follows evidently from Theorems 3.2 and 4.1 (see [23]).

THEOREM 5.3. Let $\mu^0 \in \tilde{V}[0]$; (\mathbf{f}_α) be a directedness in M_c^* such that $(\mathbf{f}_\alpha * \eta)$ converges to μ^0 in $\vartheta_c^0[\mathcal{L}]$. Then (\mathbf{f}_α) is asymptotically optimal as a "precision" solution:

- 1) $\mathbf{f}_\alpha \in F$, starting from a certain instant [11, p. 96];
- 2) $(W(\mathbf{f}_\alpha))$ converges to V^0 .

Remark. The direction (\mathbf{f}_α) from Theorem 5.3 can be constructed for any μ^0 from the (non-empty) set $\tilde{V}[0]$ if one uses the corresponding construction from [21] applied there for proving an equality similar to (2.4).

Let $\forall \varepsilon \in [0, \infty[$, $\delta \in]0, \infty[$: $\mathcal{V}_\varepsilon(\delta) \triangleq \{\mathbf{f} \in \Lambda[\varepsilon] \mid W(\mathbf{f}) \leq v_\varepsilon + \delta\}$ and besides $\forall \varkappa \in]0, \infty[$: $\mathcal{V}[\varkappa] \triangleq \mathcal{V}_0(\varkappa) = \mathcal{V}_\varepsilon(\varkappa) |_{\varepsilon=0}$ the last set is that of all \varkappa -optimal solutions of (1.1); the sense of $\mathcal{V}_\varepsilon(\delta)$ is analogous and concerns (3.3).

THEOREM 5.4. Let $\varkappa \in]0, \infty[$, $\gamma \in]\varkappa, \infty[$. Then

$$\bigcap_{\varepsilon > 0} \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \bigcup_{0 < \theta \leq \varepsilon} \mathcal{V}_\theta(\varkappa)\}, \Theta_\eta^*(\mathcal{L})) \subset \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \mathcal{V}[\gamma]\}, \vartheta_c^0[\mathcal{L}]). \quad (5.1)$$

The proof uses essentially (2.4) (for details see [23]). Due to the relations of the topologies we mentioned while discussing Theorem 3.2, the right-hand side of (5.1) can be made "more rough", i.e. replaced by its closure in $\Theta_\eta^*(\mathcal{L})$ with keeping the enclosure. In general, (5.1) is an indirect characteristic of stability in "bad" functional spaces of step-mappings. A regularizing increment $\gamma - \varkappa > 0$ has no essential influence on the sense of the statement, important only for small \varkappa and γ .

Using Theorem 3.2 and the obvious (see Theorem 5.2) equality $v_\varepsilon = \tilde{v}[\varepsilon]$ ($\varepsilon \geq 0$) we obtain the equality

$$\begin{aligned} \tilde{V}[0] &= \bigcup_{\delta > 0} \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \mathcal{V}[\delta]\}, \\ \Theta_\eta^*(\mathcal{L}) &= \bigcup_{\delta > 0} \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \mathcal{V}[\delta]\}, \\ \hat{\vartheta}_c^*(\mathcal{L}) &= \bigcup_{\delta > 0} \text{cl}(\{\mathbf{f} * \eta : \mathbf{f} \in \mathcal{V}[\delta]\}, \vartheta_c^0(\mathcal{L})) \end{aligned} \quad (5.2)$$

that characterizes the generalized solution for the most important variant of the problem (3.5) (case $\varepsilon = 0$), as the limits of "almost" optimal ordinary solutions.

The representation of the form (5.2) is frequently not true for a general extremal problem with constraints (it can also be violated for the considered problem, if Condition 5.1 is not fulfilled; see an example from [18]).

Acknowledgement

The author thanks N. N. Krasovskii for attention to the work and valuable advises.

References

1. Neveu, J., *Mathematical basis of probability theory*. Moscow, Mir, 1969. 309 p.
2. Dunford, N., Schwarz, J. T., *Linear operators. General theory*. Moscow, 1962. 892 p.
3. Varga, J., *Optimal control of differential and functional equations*. Moscow, 1977. 624 p.
4. Tikhonov, A. N., Arsenin, W. J., *Methods of solving ill-posed problems*. Moscow, 1975. 224 p.
5. Ivanov, V. K., *Ill-posed problems in topological spaces*. *Sibir. Mat. Zhurnal*, 1969. vol. 10, No. 5, pp. 1065–1074.
6. Young, L. C., *Lectures on the calculus of variations and optimal control theory*. Moscow, Mir, 1974. 488 p.
7. Ioffe, A. D., Tikhomirov, V. M., *Theory of extremal problems*. Moscow, 1974. 286 p.
8. Gamkrelidze, R. V., *Foundations of optimal control theory*. Tbilisi: Izdat. Tbil. Univ., 1977. 253 p.
9. Duffin, R. Y., *Infinite programs*. In: *Linear inequalities and related systems*. Moscow, 1959, pp. 263–276.
10. Golstein, E. G., *Theory of duality in mathematical programming*. Moscow, 1971. 351 p.
11. Kelley, J. L., *General topology*. Moscow, 1981. 431 p.
12. Engelking, R., *General topology*. Moscow, 1986. 751 p.
13. Ekeland, I., Temam, R., *Convex analysis and variational problems*. Moscow, Mir, 1979. 399 p.
14. Krasovskii, N. N., Subbotin, A. I., *Positional differential games*. Moscow, 1974. 456 p.
15. Krasovskii, N. N., *Controlling of a dynamical system*. Moscow, 1985. 520 p.
16. Subbotin, A. I., Chentsov, A. G., *Optimization of a guarantee in control problems*. Moscow, 1981. 286 p.
17. Pashaev, A. B., Chentsov, A. G., *Generalized control problem in a class of finitely-additive measures*. *Kibernetika*, 1986, No. 2, pp. 110–112.
18. Seseikin, A. N., Chentsov, A. G., *On a finite-dimensional problem of mathematical programming*. *Kibernetika*, 1988, No. 2, pp. 115–116.
19. Chentsov, A. G., *Finitely-additive measures and problems of minimum*. *Kibernetika*, 1988, No. 3, pp. 67–70.
20. Chentsov, A. G., *Two-valued measures on a semi-algebra of sets and certain applications to infinite-dimensional problems of mathematical programming*. *Kibernetika*, 1988, No. 6, pp. 72–76.

21. Chentsov, A. G., On some representations of finitely-additive measures approximated by indefinite integrals. Sverdlovsk, 1987. 36 p. (Dep. in VINITI, No. 8511-B87).
22. Chentsov, A. G., Finitely-additive measures in extensions of extremal problems. Sverdlovsk, 1988, 74 p. (Dep. in VINITI, No. 5690-B88).
23. Chentsov, A. G., Vector finitely-additive measures and extensions in a certain class of nonlinear extremal problems. Sverdlovsk, 1988, 38 p. (Dep. in VINITI, No. 8191-B88).
24. Chentsov, A. G., To the question of universal integrability of bounded functions. Mat. sbornik, 1986, vol. 131, No. 1, pp. 73–93.
25. Belov, E. G., Chentsov, A. G., Some properties of two-valued measures and conditions of universal integrability. Mat. zametki, 1987, vol. 42, No. 2, pp. 228–297.
26. Chentsov, A. G., Applications of measure theory to control problems. Sverdlovsk: Sredne-Ural. Izdat., 1985. 127 p.
27. Chentsov, A. G., Optimization under non-accurate constraints. Sverdlovsk, 1986, 54 p. (Preprint) (Inst. Math. and Mech. Ural Sci. Center).
28. Bourbaki, N., General topology. Moscow, 1968, 272 p.

О конструкции решений нерегулярных задач оптимального управления

А. Г. ЧЕНЦОВ

(Свердловск)

Рассматривается асимптотическая постановка нелинейной экстремальной задачи, а также ее расширение до стандартной задачи на минимум функционала, непрерывного на компакте. Исследуется вопрос об устойчивости по результату в условиях возмущений полной системы ограничений структуры асимптотически оптимальных приближенных решений. Основным инструментом исследования являются процедуры компактификации пространства решений $\mathbf{f} = (f_1, \dots, f_r)$ исходной задачи в классе векторных конечно-аддитивных мер $\mu = (\mu_1, \dots, \mu_r)$. При этом интегральные ограничения на сумму полных импульсов $f_i \geq 0, i = 1, \dots, r$, естественным образом переходят в соответствующее условие на сумму величин $\mu_i(E)$, где E — область определения f_i ; кроме того, компоненты μ_1, \dots, μ_r векторной меры должны удовлетворять условию «зануления»: $\mu_i(L) = 0$, если $\eta(L) = 0$, где η — неотрицательная скалярная конечно-аддитивная мера, участвующая в интегральном ограничении на выбор \mathbf{f} . Дополнительные ограничения на интегрант \mathbf{f} , имеющие смысл включения, на уровне исходной постановки возмущаются до ε -окрестностей; в обобщенной задаче аналогичное ограничение переходит в «обычное» включение для интеграла матричнозначной функции по векторной мере μ . При некоторых специальных условиях, гарантирующих устойчивость по результату и сводящихся к требованию ступенчатости упомянутой матричнозначной функции, получены утверждения, имеющие смысл «слабо» регуляризированной топологической устойчивости почти оптимальных решений исходной задачи при их погружении в компактификатор. Кроме того, в последнем случае используется и несколько более сильная нульмерная топология, отвечающая произведению подпространств соответствующей (измеримой структуре пространства решений)

тихоновской степени прямой с дискретной топологией. При тех же условиях ступенчатости установлено, что экстремальные точки компактификатора и обобщенные пределы почти оптимальных решений невозмущенной исходной задачи — суть одно и то же. Упомянутое достаточное условие (ступенчатость) существенно и не может быть ослаблено даже до требования равномерной непрерывности.

А. Г. Ченцов

Институт математики и механики УрО АН СССР

620219, Свердловск,

ГСП-384,

ул. С. Ковалевской, 16.

LONG-RANGE ADAPTIVE CONTROL OF ARMAX PLANTS WITH ACCESSIBLE DISTURBANCES

J. M. LEMOS

(Lisbon)

(Received April 11, 1990)

Feedforward adaptive control of ARMAX plants is considered in order to reduce the influence of disturbances that can be measured. The algorithm used is an extension of the long-range, multipredictive adaptive control algorithm named MUSMAR. Three main results are presented: first, a parametrization of multipredictive models, identifiable by standard RLS, is developed for ARMAX plants with ARMA accessible disturbances, working in closed-loop. Second, the algorithm, resulting from coupling this implicit plant representation with a multistep quadratic cost, is shown to present a robust tuning property of the controller gains. Third, a simulation example is presented.

Keywords: Adaptive control. Predictive control. Self-tuning. Feedforward control. Convergence analysis. Linear Quadratic Stochastic control.

1. Introduction

There are many practical situations in which the control performance can be greatly improved by exploiting the knowledge of the accessible disturbances acting on the plant. Examples with industrial relevance include drum boiler level and pressure control in power plants [1], control of gas-cooled reactors in nuclear power stations [2], frequency control of hydro power stations [3] and the bottom temperature control of glass furnaces [4]. More exotic applications are the exploitation of auxiliary signals for time-series forecasting [5] and active sound [6].

All the above examples include feedforward terms of one form or another, to cancel out in anticipation the effects of the accessible disturbances. As opposite to feedback, however, feedforward action requires greater precision in the models used for control purposes. Thus, in performing feedforward control of uncertain and/or time-varying plants, one is naturally lead to the use of adaptive techniques.

Several kinds of adaptive feedforward controllers can be envisaged, by extending in a natural way adaptive feedback controllers. Clarke-Gawthrop type

controllers [4], LQG feedforward controllers [3, 2], explicit criterion minimization [3, 6] and multistep predictive control [7], are all possibilities.

The work reported in this paper is concerned with the latest class of feedforward controllers. The algorithm to be discussed and analysed is an extension of the long-range multipredictive adaptive controller named MUSMAR [8].

2. Problem formulation

With relation to long-range multipredictive feedforward controllers, three problems are solved in this paper.

The first is the *modelling problem*. For the least-squares plus minimum variance self-tuner, it is a classic result that ARMAX plants controlled by this algorithm admit in closed-loop an ARX model correctly describing its output [9]. In [10], this result is extended to multipredictive models (i.e. models describing the output over a certain horizon) by applying Implicit Modelling Theory [11]. The first problem considered in this paper is the development of multipredictive models for ARMAX plants with ARMA accessible disturbances, working in closed-loop. The models to obtain are such that its residuo is orthogonal to the subspace generated by the available data, thus being amenable of identification by *standard* Recursive Least Squares (RLS).

The second problem is to develop an adaptive control algorithm, by coupling the models above with a control law obtained by the minimization of a multistep quadratic cost function of the type

$$J_T = E \left\{ \sum_{k=1}^T [\tilde{y}^2(t+k) + \rho u^2(t+k-1)] \mid I^t \right\} \quad (1)$$

in which $\tilde{y} \triangleq y - r$ is the tracking error of the output of the plant y with respect to the reference r to be tracked, u is the plant input, ρ is a non-negative control weight and I^t is the information pattern available at time t . I^t contains observations of the past values of u , y , the accessible disturbance v and the reference r , this last taken in this paper as zero.

Finally, the third problem is the assessment of the resulting controller performance *in the presence of unmodelled plant dynamics*.

3. Modelling issues

Consider the SISO plant described by the ARMAX model with accessible disturbance

$$A(q)y(t) = B(q)u(t) + D(q)v(t) + C(q)e(t). \quad (2)$$

in which A and B are polynomials in the forward shift operator q , such that $\partial A - \partial B = d \geq 1$, ∂A denoting the degree of A , A is monic and all the common factors between A and B are stable. The innovations signal $\{e\}$ is a sequence of independent identically distributed random variables with zero mean and variance σ_e^2 . Further, $\partial D = \partial C = \partial A = n$. Polynomial C is Hurwitz and v is the accessible disturbance modelled as the AR process

$$A_v(q)v(t) = q^{n_v}e_v(t) \quad (3)$$

with A_v an Hurwitz polynomial of degree n_v and $\{e_v\}$ is a white noise sequence with variance σ_v^2 , independent of $\{e\}$.

The input is given by the *stabilizing* control law [12]

$$R(q)u(t) = -S(q)y(t) + M(q)v(t) + C(q)\eta(t) \quad (4)$$

where R and S are coprime polynomials, such that R is monic and

$$\partial R = n_R, \quad \partial S = n_R - 1, \quad \partial M = p$$

where n_R and p are the orders of the optimal controllers [12].

The sequence $\{\eta\}$ is a zero mean white dither noise, uncorrelated with $\{e\}$ and $\{e_v\}$ such that $\sigma_\eta^2 \triangleq E[\eta_i^2] \ll \min(\sigma_e^2, \sigma_v^2)$.

Remark 1. The above formulation encompasses the situation in which the output of the ARMA plant

$$A(q)y(t) = B(q)u(t) + C(q)e(t) \quad (5)$$

is to follow a reference v_t given by the output of (3). Multiplying (5) by A_v , (3) by A and subtracting both equations, produces

$$A(q)A_v(q)\tilde{y}(t) = B(q)A_v(q)u(t) - A(q)A_v(q)v(t) + C(q)A_v(q)e(t) \quad (6)$$

in which $\tilde{y}(t) \triangleq y(t) - v(t)$ is the tracking error. Equation (6) is of the form (2) and the results obtained for one case may be specialized for the other.

3.1. Implicit models

The first step in solving the modelling problem is to study under what conditions the plant (2, 3) coupled with the controller defined by (4) admits an ARX representation. This ARX model is only valid for the controlled system, and thus is called an *implicit* model [11]. The following theorem, which extends to plants with accessible disturbances similar results given in [10], provides an answer to this problem. In order to improve clarity, the results in this section are self-contained.

Proposition 1. The controlled system obtained by coupling (2, 3) with the controller defined by (4) admits, in stochastic steady state (s.s.s.) an ARX representation

$$\mathcal{A}(q)y(t) = \mathcal{B}(q)u(t) + \mathcal{D}(q)v(t) = \eta(t) + e(t) \quad (7)$$

with $\mathcal{A}(q)$ monic, iff the characteristic polynomial of the closed-loop system satisfies:

$$\text{i) } \exists Q(q) : A(q)R(q) + B(q)S(q) = C(q)Q(q) \quad (8)$$

and

$$\text{gcd}(R, S) = \text{gcd}(R, S, Q) \quad (9)$$

$$\text{ii) } \exists Q_v(q) : D(q)R(q) + B(q)M(q) = C(q)Q_v(q) \quad (10)$$

and

$$\text{gcd}(R, M) = \text{gcd}(R, M, Q_v). \quad (11)$$

Further, under these conditions, the following identities hold:

$$\mathcal{A}(q)R(q) + \mathcal{B}(q)S(q) = Q(q) \quad (12)$$

$$\mathcal{D}(q)R(q) + \mathcal{B}(q)M(q) = Q_v(q) \quad (13)$$

$$R(q) = B(q) - C(q)\mathcal{B}(q) \quad (14)$$

$$S(q) = -A(q) + C(q)\mathcal{A}(q) \quad (15)$$

$$M(q) = -D(q) + C(q)\mathcal{D}(q). \quad (16)$$

□

Proof. The proof follows similar ones found in [10, 11] and is given in the Appendix.

3.2. The τ -UCPP property

Implicit models are not unique [10]. Among the possible models are of special interest here the ones that correctly describe the closed-loop system, no matter what the inputs over a certain time interval are. In order to formalize this idea, the following concept is used:

DEFINITION 1 [10, 11]. An ARX implicit model is said to enjoy the *unconstrained control prediction property* of order τ (or to be τ -UCPP) if it correctly gives the output up to time $t + 1$ for an arbitrary input sequence between the instants $t - r + 1$ and t .

Proposition 1. Let the inputs of the ARMAX plant (2) be given up to time $t - \tau$ by a control law of the form (4) satisfying:

$$\text{i) } \exists Q(q) : A(q)R(q) + B(q)S(q) = Q(q)C(q) \quad (17)$$

$$\text{ii) } \exists Q_v(q) : D(q)R(q) + B(q)M(q) = Q_v(q)C(q). \quad (18)$$

Then, there exist an implicit τ -UCPP ARX model giving the correct output up to time $t + 1$ no matter what the inputs from $t - r + 1$ up to t are used.

A model $(\mathcal{A}, \mathcal{B})$ enjoys the τ -UCPP property if

$$B^*(q^{-1}) = Q_B^*(q^{-1}) + q^{-(\tau+1)}\mathcal{G}_B^*(q^{-1}) \quad (19)$$

where $\partial\mathcal{G}_B^*(q^{-1}) = 0$ and

$$Q_B^*(q^{-1}) = \beta_1q^{-1} + \dots + \beta_\tau q^{-1} \quad (20)$$

satisfies

$$B^*(q^{-1}) = Q_B^*(q^{-1})C^*(q^{-1}) + q^{-(\tau+1)}G_B^*(q^{-1}) \quad (21)$$

with $G_B^*(q^{-1}) = 0$.

In addition, (19) is satisfied if and only if

$$\mathcal{A}^*(q^{-1}) = Q_A^*(q^{-1}) + q^{-(\tau+1)}\mathcal{G}_A^*(q^{-1}) \quad (22)$$

and

$$\mathcal{D}^*(q^{-1}) = Q_D^*(q^{-1}) + q^{(\tau+1)}\mathcal{G}_D^*(q^{-1}) \quad (23)$$

with $\partial\mathcal{G}_A^*(q^{-1}) = 0$, $\partial\mathcal{G}_D^*(q^{-1}) = 0$, where $\partial X^*(q^{-1})$ denotes the value of the smallest exponent in q^{-1} with nonzero coefficient in the polynomial $X^*(q^{-1})$, and

$$Q_A^*(q^{-1}) = 1 + \alpha_1q^{-1} + \dots + \alpha_\tau q^{-\tau} \quad (24)$$

$$Q_D^*(q^{-1}) = 1 + \delta_1q^{-1} + \dots + \delta_\tau q^{-\tau} \quad (25)$$

satisfy

$$A^*(q^{-1}) = Q_A^*(q^{-1})C^*(q^{-1}) + q^{-(\tau+1)}G_A^*(q^{-1}) \quad (26)$$

$$D^*(q^{-1}) = Q_D^*(q^{-1})C^*(q^{-1}) + q^{-(\tau+1)}G_D^*(q^{-1}) \quad (27)$$

$$\partial G_A^*(q^{-1}) = 0, \quad \partial G_D^*(q^{-1}) = 0.$$

Further, under (19) or, equivalently (22, 23):

$$(G_A - \mathcal{G}_A C)R = (G_B - \mathcal{G}_B C)S \quad (28)$$

and

$$(G_D - \mathcal{G}_D C)R = (G_D - \mathcal{G}_B C)M. \quad (29)$$

□

Proof. See the Appendix.

Proposition 3. The plant (2) working in closed-loop under the control law (4) satisfying (8–11) admits the τ -UCPP implicit model given by:

$$\begin{aligned} y_{t+\tau} + \alpha_1 y_{t+\tau-1} + \dots + \alpha_{\tau-1} y_{t+1} &= \\ &= \beta_1 u_{t+\tau-1} + \dots + \beta_{\tau-1} u_{t+1} + \beta_{\tau} u_t + \\ + \gamma_0 v_{t+\tau} + \gamma_1 v_{t+\tau-1} + \dots + \gamma_{\tau-1} v_{t+1} + \Sigma'_{\tau} s_t + e_{t+\tau} \end{aligned} \quad (30)$$

with

$$\frac{A}{C} = \sum_{i=0}^{\infty} \alpha_i q^{-i} \quad \frac{B}{C} = \sum_{i=1}^{\infty} \beta_i q^{-i} \quad \frac{D}{C} = \sum_{i=0}^{\infty} \gamma_i q^{-i} \quad (31)$$

Σ_{τ} is a vector whose entries are made up by the last $n_R - 1$ coefficients of \mathcal{A} , the last n_R coefficients of \mathcal{B} , the last $n_R + n_v + 1$ coefficients of \mathcal{D} , and

$$s_t = [y_t \dots y_{t-n_R+1} u_{t-1} \dots u_{t-n_R} v_t \dots v_{t-n_R-n_v}]'. \quad (32)$$

This model is denoted \mathcal{M}_{τ} . □

Proof. See the Appendix.

Consider T τ -UCPP implicit models, for $\tau = 1, \dots, T$. Using this pencil of models, y_{t+1} is eliminated from \mathcal{M}_2 using \mathcal{M}_1 ; then y_{t+1} and y_{t+2} are eliminated from \mathcal{M}_3 using \mathcal{M}_1 and \mathcal{M}_2 , and so on. Also, project “future” samples of the accessible disturbance, $v_{t+\tau}, \dots, v_{t+1}$ in the samples of v in s_t . In this way, the following multipredictor model correctly describing the output of the plant from $t + 1$ up to $t + T$, is obtained:

$$Y_t = WU_t + \Pi'_{s_t} + \mathcal{E}_t \quad (33)$$

with W a Toeplitz lower triangular matrix of parameters, Π a matrix of parameters of convenient dimensions

$$Y_t \triangleq [y_{t+1} \dots y_{t+T}]' \quad (34)$$

$$U_t \triangleq [u_t \dots u_{t+T-1}]' \quad (35)$$

and \mathcal{E} a vector of residues orthogonal to the data in each predictor.

By performing the minimization of (1) with respect to U_t using (33), it is possible to derive a control law which approximates the steady state LQS control if T is large enough. A better choice is to assume that a fixed gain is acting on the plant from $t + 1$ to $t + T - 1$, the following multipredictor model being thus obtained:

$$y_{t+i} = \theta_i u_t + \psi'_i s_t + \nu_i^y(t) \quad (36)$$

$$u_{t+i-1} = \mu_{i-1} u_t + \phi'_{i-1} s_t + \nu_{i-1}^u(t) \quad (37)$$

$$i = 1, \dots, T$$

$$\nu_i^y(t), \nu_{i-1}^u(t) \perp [u_t, s_t]. \tag{38}$$

Here, θ_i and μ_i are vectors and Ψ_i and Φ_i are matrices of parameters whose entries depend on W and Π , and the gain acting on the plant on steady-state. The exact form of this dependence is given by lengthy expressions which are omitted, since they are irrelevant hereafter.

Two points are, however, to be remarked: first, the imposition of a constant feedback allows a tighter approximation of the LQS control with respect to what is obtained by leaving all the entries in U_t free. Second, models (37, 38) will give rise to an adaptive algorithm with its only possible equilibrium points given by the local minima of the LQS steady-state cost.

4. The adaptive feedforward controller

Using the predictive models (36, 37) to minimize the multistep quadratic cost (1), the following adaptive feedforward control algorithm is derived:

MUSMAR with feedforward

At each sampling period t , recursively execute the following steps:

1. Using standard RLS, estimate the parameters in the predictive models (36, 37).
2. Calculate the vector of updated feedback gains by

$$f(t) = -\frac{1}{\alpha(t)} \sum_{i=0}^{T-1} [\theta_{i+1}(t)\psi_{i+1}(t) + \rho\mu_i(t)\phi_i(t)] \tag{39}$$

$$\alpha(t) \triangleq \sum_{i=0}^{T-1} [\theta_{i+1}^2(t) + \rho\mu_i^2(t)] \tag{40}$$

where $[\theta_i(t) \ \psi_i'(t)]$, $[\mu_i(t) \ \phi_i'(t)]'$ are RLS estimates of the homonimous parameters of (36, 37) and ρ is as in (1).

3. Apply to the plant the control given by

$$u_t = f'(t)s_t + \eta_t \tag{41}$$

with η_t a low intensity dither noise injected in order to fulfill a persistent excitation condition.

Equations (39, 40) are obtained simply by minimizing (1) with respect fo f , assuming $u(t)$ given by (41) with $\eta = 0$ and that (37, 38) hold. Since s_t contains

samples of the accessible disturbance, it is remarked that the above algorithm actually includes feedforward terms.

5. Robustness analysis

This Section is concerned with the robustness properties against unmodelled plant dynamics of MUSMAR with feedforward. Available theoretic results for the regulation problem [13] are extended to plants with accessible AR disturbances. The main conclusion is that, *in the presence of any structural mismatching* between the plant and their models, as T increases, MUSMAR equilibrium points approach the local minima of the steady-state LQ criterion. No assumption is made on the regressor complexity. The main interest is in the possible convergence points of the MUSMAR feedback gain vector f . The analysis is based on the O.D.E. approach described in [14] whose applicability rely on the following assumption:

The sequence of regulator parameters $f(t) \in D_s$ and $\|s_t\|$ is bounded for infinitely many t , w. p. one. Here, D_s is a compact set in which $f(t)$ defines a closed-loop system with poles strictly inside the unit circle.

Since the multipredictor coefficients of (36, 37) are estimated via a standard RLS algorithm, the asymptotic average evolution of their estimates is described by the following set of O.D.E.'s:

$$\begin{bmatrix} \dot{\theta}_{i+1}(\tau) \\ \dot{\psi}_{i+1} \end{bmatrix} = R^{-1}(\tau)E \{z_t(f(\tau)) \times \quad (42)$$

$$\times [y_{t+i+1}(f(\tau)) - (\theta_{i+1}(\tau)f(\tau) + \psi_{i+1}(\tau))'s_t(f(\tau) - \theta_i(\tau))]\}$$

$$\begin{bmatrix} \dot{\mu}_i(\tau) \\ \dot{\phi}_i(\tau) \end{bmatrix} = R^{-1}(\tau)E \{z_t(f(\tau)) \times \quad (43)$$

$$\times [u_{t+i}(f(\tau)) - (\mu_i(\tau)f(\tau) + \phi_i(\tau))'s_t(f(\tau)) - \mu_i(\tau)\eta_t]\}$$

$$\dot{R}(\tau) = -R(\tau) + R_z(\tau) \quad (44)$$

where

$$z_t \triangleq |u_t s_t'|' \quad (45)$$

$$R_z(\tau) \triangleq E[z_t(f(\tau))z_t'(f(\tau))] = \begin{bmatrix} f'(\tau)R_s(\tau)f(\tau) + \sigma_\eta^2 & f'(\tau)R_s(\tau) \\ R_s(\tau)f(\tau) & R_s(\tau) \end{bmatrix} \quad (46)$$

$$R_s(\tau) \triangleq E[s_t(f(\tau))s_t'(f(\tau))] \quad (47)$$

and $f(\tau)$ is, as in (39), with t replaced by τ . In the above equations, *dot* denotes derivative with respect to τ and $E[\cdot]$ expectation with respect to the probability

density function induced on $\{u\}$ and $\{y\}$ by $\{e\}$ and $\{\eta\}$, assuming the system in the s.s.s. corresponding to the constant control law

$$u_t = f'(\tau)s_t + \eta_t. \tag{48}$$

By using similar arguments as in [13] it is shown that the corresponding O.D.E. for $f(\tau)$ can be written as

$$\dot{f}(\tau) = -\alpha^{-1}(\tau)R_s^{-1}(\tau)p(\tau) + o(|\dot{f}(\tau)|) \tag{49}$$

where $\tilde{f}(\tau) \triangleq f(\tau) - f^*$; f^* denotes any equilibrium point of (40); $o(|x|)$ is such that $\lim_{x \rightarrow 0} \frac{o(|x|)}{|x|} = 0$;

$$p(\tau) \triangleq \sigma_\eta^{-2} \sum_{i=10}^{T-1} [R_{y\eta}(i+1; \tau)R_{ys}(i+1; \tau) + \rho R_{u\eta}(i; \tau)R_{us}(i; \tau)] \tag{50}$$

and

$$R_{y\eta}(i+1; \tau) \triangleq E[y_{t+i+1}(f(\tau))\eta_t] \tag{51}$$

$$R_{ys}(i+1; \tau) \triangleq E[y_{t+i+1}(f(\tau))s_t(\tau)] \tag{52}$$

with similar definitions for $R_{u\eta}(i; \tau)$ and $R_{us}(i; \tau)$.

In order to give a convenient interpretation to (50), the following lemma is introduced.

Lemma 1. Let $Q^*(q^{-1}; \tau) \triangleq Q^*(q^{-1}; f(\tau))$ be the closed-loop characteristic polynomial corresponding to $f(\tau)$. Then

$$\sigma_\eta^{-2}R_{y\eta}(i+1; \tau) = \left[\frac{q^{-1}B^*(q^{-1})}{Q^*(q^{-1}; \tau)} \right]_{i+1} \tag{53}$$

$$\sigma_\eta^{-2}R_{u\eta}(i+1; \tau) = \left[\frac{A^*(q^{-1})}{Q^*(q^{-1}; \tau)} \right]_i \tag{54}$$

where $[H^*(q^{-1})_i]$ denotes the i -th impulse response sample associated with the transfer function $H^*(q^{-1})$.

Proof. See the Appendix.

According to (53, 54), (50) is rewritten as

$$\begin{aligned} p(t) &= \sum_{i=0}^{T-1} E \left\{ \left[\frac{q^{-1}B^*(q^{-1})}{Q^*(q^{-1}; \tau)} \right]_{i+1} y_{t+i+1}(\tau)s_t(\tau) + \rho \left[\frac{A^*(q^{-1})}{Q^*(q^{-1}; \tau)} \right]_i u_{t+i}(\tau)s_t(\tau) \right\} \\ p(t) &= E \left\{ y_t(\tau) \left[\left[\frac{q^{-1}B^*(q^{-1})}{Q^*(q^{-1}; \tau)} \right] |_T s_t(\tau) \right] + \rho u_t(\tau) \left[\left[\frac{A^*(q^{-1})}{Q^*(q^{-1}; \tau)} \right] |_{T-1} s_t(\tau) \right] \right\} \end{aligned} \tag{55}$$

where $H|_T$ denotes the truncation to the power q^{-T} of the power series expansion in q^{-1} of the transfer function H .

Consider the unconditional cost

$$J(f(\tau)) = \frac{1}{2}e[y_t^2 + \rho u_t^2] \quad (56)$$

as a function of the constant gains $f(\tau)$. As shown in [3] for plants with accessible disturbances,

$$\begin{aligned} \nabla J(f(\tau)) &\triangleq \frac{\partial J(f(\tau))}{\partial f(\tau)} = \\ &= E \left\{ y_t(\tau) \left[\frac{q^{-1}B^*(q^{-1})}{Q^*(q^{-1}; \tau)} s_t(\tau) \right] + \rho u_t(\tau) \left[\frac{A^*(q^{-1})}{Q^*(q^{-1}; \tau)} s_t(\tau) \right] \right\}. \end{aligned} \quad (57)$$

Thus, comparing (55) with (57), the former is seen to be a good approximation to the latter, whenever

$$|\lambda[Q^*(q^{-1}; \tau)]|^{2(T-1)} \gg 1 \quad (58)$$

where $\lambda[Q]$ denotes any root of Q . Therefore, in a neighbourhood of any equilibrium point satisfying (58), O.D.E. (50) can be approximated by

$$\dot{f}(\tau) = -[\alpha(\tau)]^{-1} R_s^{-1}(\tau) T \nabla J(f(\tau)) + o(|\dot{f}(\tau)|). \quad (59)$$

The above results are summarized in the following

Proposition 4. Consider the MUSMAR algorithm with $\rho > 0$ for any i/o transport delay smaller or equal than T , arbitrary C innovation polynomial with no root outside the unit circle, and any regressor complexity. Then, amongst the equilibria f^* giving rise to a closed-loop system with well damped modes relatively to the control horizon T such that (55) can be replaced by (57), the only MUSMAR possible converging points approach to local minima or edge points of the unconditional cost (56).

Proof. The proof is done recalling [14] that the only possible converging points of a recursive stochastic algorithm are the locally stable equilibrium points of the associated O.D.E. Since in (59) $\alpha(\tau) > 0$ and $R_s(\tau) > 0$, the conclusion follows. \square

According to the Remark 1 of Chapter 2, substituting \tilde{y} for y , the above analysis encompasses the servo problem.

6. Simulation example

Consider the nonminimum-phase plant

$$y_{t+2} - 1.5y_{t+1} + 0.7y_t = u_{t+1} + 2u_t + v_{t+1} + e_{t+2} \quad (60)$$

where $\{e_t\}$ is a zero mean, unit variance, white noise Gaussian sequence. Figure 1 displays the results obtained by controlling this plant using MUSMAR, with and without feedforward. The improvement in the figure of merit obtained by using feedforward is of about 10. The structure of the feedforward controller is defined by

$$T = 3, \quad \rho = 10^{-5}, \quad \sigma_\eta^2 = 10^{-4},$$

$$s_t = [y_t \ y_{t-1} \ u_t \ v_t]'$$

Without feedforward, the term in v_t is removed from the pseudostate.

7. Conclusion and final remarks

A long-range adaptive controller for ARMAX plants subject to accessible disturbances, modelled as AR signals, has been developed and analysed. The final algorithm turns out to be a modification of the MUSMAR controller, obtained by the incorporation of feedforward terms in a natural way.

Although related to other feedforward algorithms, this new controller is developed from a distinct standpoint, and presents a number of important advantages. Indeed, it includes the Clarke–Gawthrop type feedforward controller of [4] as a special case, obtained by choosing the optimization horizon $T = 1$. By extending T , important properties are acquired by the algorithm, e.g. a tight approximation to the LQS feedforward control is yielded. However, this approximation is obtained by implicit methods, as opposite to [2], in which involved Diophantine equations are to be solved. A close connection with the explicit minimization methods of [3] also exist since, as shown in Section 4, the controller gains progress in the opposite of a modified gradient direction.

Besides the derivation of the algorithm, two orders of questions have been considered. The first involves modelling issues and is concerned with conditions under which an ARMAX plant with an AR accessible disturbance can be correctly described in closed-loop by ARX models. The relevance of this result stems from the fact that *standard* recursive least squares may be used for identification, thus yielding simpler and less restrictive algorithms.

The second aspect is the tuning ability of the algorithm in the presence of unmodelled plant dynamics. It is shown, that, *even in the presence of unmodelled plant dynamics*, the only possible convergence points of the gains tightly approximate the local minima of the LQS steady state cost. It is remarked that, although this result does not ensure convergence (it only characterizes the possible points of convergence), there is enough simulation evidence that MUSMAR will actually converge, whenever a minimum exists. The robustness properties of MUSMAR are not unexpected. Indeed, since MUSMAR does *not* rely on a single plant model (as it happens e.g. in both [2, 3]), but on a set of *separately* identified models, it is expected that the redundancy, thereby introduced renders, algorithm is more robust

with respect to deviations to ideal behaviour caused by nonlinearities, unmodelled plant dynamics or uncertain i/o transport delay than algorithms relying on the extrapolation of a single predictive model.

Appendix

Proof of Proposition 1

(only if)

Assume that an ARX implicit model exists, characterized by \mathcal{A} , \mathcal{B} and \mathcal{D} . Using (4) in (2) multiplied by $R(q)$ and then (3), the following closed-loop model is obtained:

$$y(t) = \frac{BM + DR}{AR + BS} \cdot \frac{q^{n_v}}{A_v} e_v(t) + \frac{BC}{AR + BS} \eta(t) + \frac{CR}{AR + BS} e(t). \quad (61)$$

On the other way, using (4) in (7) multiplied by $R(q)$, and then (3):

$$y(t) = \frac{BM + DR}{AR + BS} \cdot \frac{q^{n_v}}{A_v} e_v(t) + \frac{CB + R}{AR + BS} \eta(t) + \frac{R}{AR + BS} e(t). \quad (62)$$

Let $Q(q)$ and $Q_v(q)$ be given by (12) and (13). Due to the fact that $\{e\}$, $\{e_v\}$ and $\{\eta\}$ are uncorrelated, the transfer functions from the corresponding signals to the output must be equal.

Thus, from $\{e\}$ to $\{y\}$ and by (12):

$$\frac{CR}{AR + BS} = \frac{R}{Q}$$

which yields (8).

From $\{e_v\}$ to $\{y\}$ and by (13)

$$\frac{BM + DR}{AR + BS} = \frac{Q_v}{Q}$$

and, from (8) this yields (9).

From $\{\eta\}$ to $\{y\}$, (8) and by (12)

$$\frac{BC}{CQ} = \frac{CB + R}{Q}$$

which yields (14).

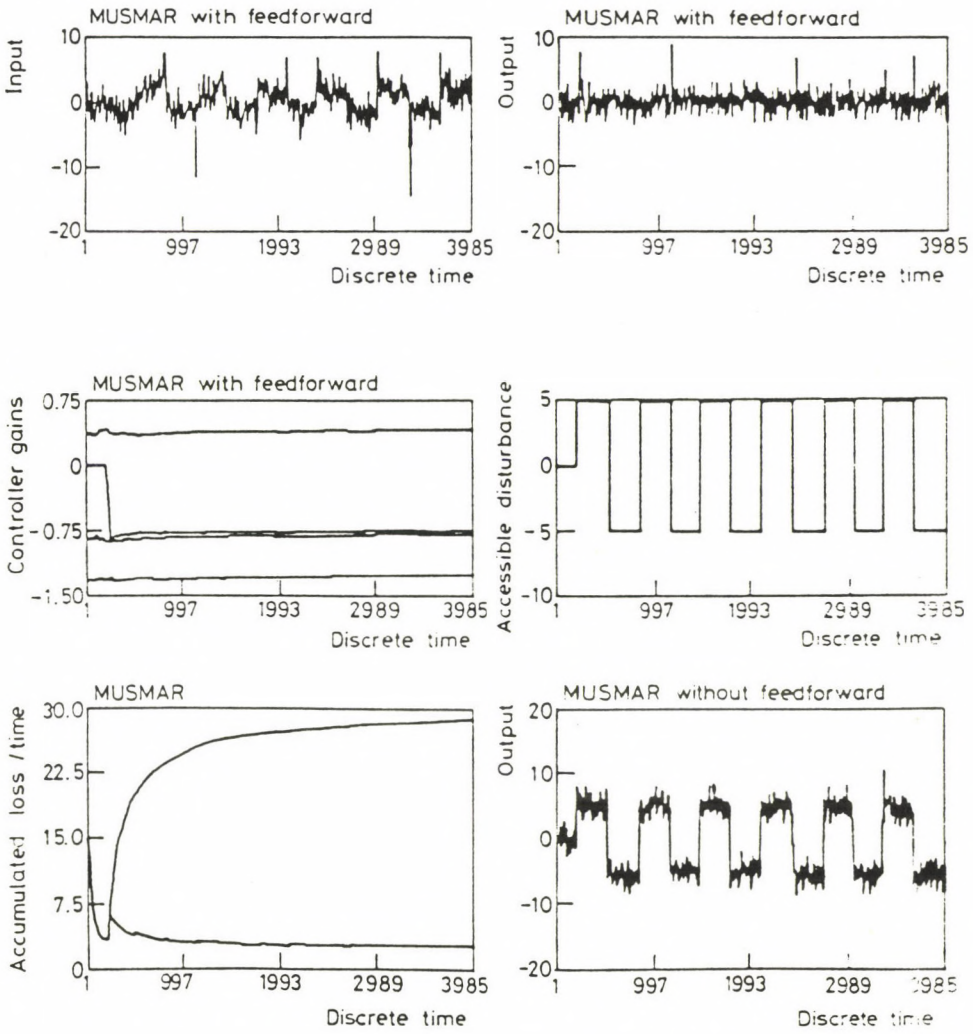


Fig. 1. Comparison of MUSMAR with and without feedforward

Multiplying (14) by S and M and using (12) and (13), respectively, (15) and (16) are obtained.

(if)

In closed-loop, the plant (2) with the controller (4) is characterized by the transfer functions from $\{e_v\}$, $\{e\}$ and $\{\eta\}$ to the output and the input. These are

$$y(t) = \frac{BM + DR}{AR + BS} \cdot \frac{q^{n_v}}{A_v} e_v(t) + \frac{BC}{AR + BS} \eta(t) + \frac{CR}{AR + BS} e(t) \quad (63)$$

$$y(t) = \frac{AM - DS}{AR + BS} \cdot \frac{q^{n_v}}{A_v} e_v(t) + \frac{AC}{AR + BS} \eta(t) - \frac{CS}{AR + BS} e(t). \quad (64)$$

Consider an implicit model given by (7) with the control law (4). For such a model, the same transfer functions are

$$y(t) = \frac{BM + DR}{AR + BS} \cdot \frac{q^{n_v}}{A_v} e_v(t) + \frac{CB + R}{AR + BS} \eta(t) + \frac{R}{AR + BS} e(t) \quad (65)$$

$$y(t) = \frac{AM - DS}{AR + BS} \cdot \frac{q^{n_v}}{A_v} e_v(t) + \frac{AC - S}{AR + B} S_\eta(t) - \frac{S}{AR + B} e(t). \quad (66)$$

For these models to be identical, the following conditions must be *simultaneously* satisfied:

$$AR + BS = C(AR + BS) \quad (67)$$

$$BM + DR = C(DR + BM) \quad (68)$$

$$AM - DS = C(AM + DS) \quad (69)$$

$$B = CB + R \quad (70)$$

$$A = AC - S. \quad (71)$$

To see that they can be simultaneously satisfied, consider the Diophantine equation

$$XA - YB = 0. \quad (72)$$

All its solutions are given by

$$X = LB \quad (73)$$

$$Y = LA \quad (74)$$

for any polynomial L .

Take (67) and write it as

$$(A - CA)R - (-B + CB) = 0. \quad (75)$$

The solution of equation (75) satisfy

$$A - CA = LS \quad (76)$$

$$-B + CB = LR \quad (77)$$

for any L .

Similarly, for (68)

$$(D - \mathcal{D}C)R - (-B + CB)M = 0 \quad (78)$$

whose solutions satisfy

$$D - \mathcal{D}C = LM \quad (79)$$

$$-B + CB = LR \quad (80)$$

again for any L .

Comparing (70) with (77) and (71) with (76) it is concluded that a choice of $L = -1$ is compatible with all the conditions.

□

Proof of Proposition 2

According to (12), all possible B have the form

$$B = X + LR \quad (81)$$

with L generic. Let $L^*(q^{-1})$ be found by long division of $Q_B^*(q^{-1}) - X^*(q^{-1})$ by $R^*(q^{-1})$ so as to satisfy

$$Q_B^*(q^{-1}) - X^*(q^{-1}) = L^*(q^{-1})R^*(q^{-1}) - q^{-(\tau+1)}\mathcal{G}_B^*(q^{-1})$$

with $\partial L = \tau$ and $\partial \mathcal{G}_B^*(q^{-1}) = 0$.

With this choice, B verifies

$$Q_B^*(q^{-1}) - B^*(q^{-1}) + L^*(q^{-1})R^*(q^{-1}) = L^*(q^{-1})R^*(q^{-1}) - q^{-(\tau+1)}\mathcal{G}_B^*(q^{-1})$$

which proves that a B can be found in the form (19).

In order to prove (22, 23), note that (8) imply:

$$AR + BS = C(AR + BS) \quad (82)$$

$$DR + BN = C(\mathcal{D}R + \mathcal{B}M) \quad (83)$$

$$(\mathcal{B}C - B)S = -(\mathcal{A}C - A)R \quad (84)$$

$$(\mathcal{B}C - B)M = -(\mathcal{D}C - D)R \quad (85)$$

$$\begin{aligned} [B^*(q^{-1})C^*(q^{-1}) - B^*(q^{-1})]S^*(q^{-1}) &= \\ &= -[A^*(q^{-1})C^*(q^{-1}) - A^*(q^{-1})]R^*(q^{-1}) \end{aligned} \quad (86)$$

$$\begin{aligned} [B^*(q^{-1})C^*(q^{-1}) - B^*(q^{-1})]M^*(q^{-1}) &= \\ &= -[D^*(q^{-1})C^*(q^{-1}) - D^*(q^{-1})]R^*(q^{-1}). \end{aligned} \quad (87)$$

If $B^*(q^{-1})$ is, as in (19), the above equations become

$$\begin{aligned} q^{-(\tau+1)}[\mathcal{G}_B^*(q^{-1})C^*(q^{-1}) - G_B^*(q^{-1})]S^*(q^{-1}) &= \\ &= -[A^*(q^{-1})C^*(q^{-1}) - A^*(q^{-1})]R^*(q^{-1}) \end{aligned} \quad (88)$$

$$\begin{aligned} q^{-(\tau+1)}[\mathcal{G}_B^*(q^{-1})C^*(q^{-1}) - G_B^*(q^{-1})]M^*(q^{-1}) &= \\ &= -[D^*(q^{-1})C^*(q^{-1}) - D^*(q^{-1})]R^*(q^{-1}). \end{aligned} \quad (89)$$

Write $A^*(q^{-1})$ and $D^*(q^{-1})$, as in (26, 27), and use it in (88, 89) to get:

$$\begin{aligned} q^{-(\tau+1)}[\mathcal{G}_B^*(q^{-1})C^*(q^{-1}) - G_B^*(q^{-1})]S^*(q^{-1}) &= \\ = -[(A^*(q^{-1}) - Q_D^*(q^{-1}))C^*(q^{-1}) - q^{-(\tau+1)}G_A^*(q^{-1})]R^*(q^{-1}) \end{aligned} \quad (90)$$

$$\begin{aligned} q^{-(\tau+1)}[\mathcal{G}_B^*(q^{-1})C^*(q^{-1}) - G_B^*(q^{-1})]M^*(q^{-1}) &= \\ = -[(D^*(q^{-1}) - Q_D^*(q^{-1}))C^*(q^{-1}) - q^{-(\tau+1)}G_D^*(q^{-1})]R^*(q^{-1}). \end{aligned} \quad (91)$$

Using the fact that R is monic, (22, 23) are obtained. □

Proof of proposition 3

Since (equation 8)

$$AR + BS = CQ$$

for some polynomial Q , the degree of Q is determined as follows:

From the structure of the control law

$$\partial R = n_R \quad (92)$$

$$\partial S = n_R - 1. \quad (93)$$

Since it is assumed that $\partial B \leq n - 1$, the term AR has a higher degree than BS . In order to equate the coefficients of the highest powers of AR and CQ , since $\partial C = n$, it must be true that

$$\partial Q = n_R. \quad (94)$$

Once ∂Q is known, equation (10) is used to relate the degrees of \mathcal{A} and \mathcal{B} :

$$\mathcal{A}R + \mathcal{B}S = Q.$$

Since

$$\begin{aligned}\partial Q &= n_R \\ \partial(\mathcal{A}R) &= \partial\mathcal{A} + n_R \\ \partial(\mathcal{B}S) &= \partial\mathcal{B} + n_R - 1\end{aligned}$$

it must be

$$\partial\mathcal{B} = \partial\mathcal{A} + 1 \quad (95)$$

in order that the coefficients of the highest powers of q^{-1} in $\mathcal{A}R$ and $\mathcal{B}S$ cancel out.

By (19) and (22)

$$\mathcal{A}^*(q^{-1}) = Q_A^*(q^{-1}) + q^{-(\tau+1)}\mathcal{G}_A^*(q^{-1}) \quad (96)$$

$$\mathcal{B}^*(q^{-1}) = Q_B^*(q^{-1}) + q^{-(\tau+1)}\mathcal{G}_B^*(q^{-1}) \quad (97)$$

and since $\partial Q_A = \partial Q_B = \tau$, relation (95) gives

$$\partial\mathcal{G}_B = \partial\mathcal{G}_A + 1. \quad (98)$$

Equations (22), (26), (21) and (19) used in (82) give

$$(G_A - \mathcal{G}_A C)R = (G_B - \mathcal{G}_B C)S \quad (99)$$

From (26), (21):

$$G_A = (A - Q_A C)q^{\tau+1} \quad (100)$$

$$G_B = (B - Q_B C)q^{\tau+1}. \quad (101)$$

Multiply (100) by R , (101) by S , subtract and use (8) to get

$$G_A R - G_B S = C[Q - Q_A R + Q_B S]q^{\tau+1}. \quad (102)$$

From (102) and (99):

$$\mathcal{G}_A R - \mathcal{G}_B S = \text{known polynomial}. \quad (103)$$

Equation (103) is used to obtain the minimum degree of \mathcal{G}_B by equating the number of equations to the number of unknowns.

The number of equations is equal to the number of coefficients, which in turn is given by the degree plus one. Thus:

$$\text{Number of equations} = (\partial\mathcal{G}_B + n_R - 1) + 1 = \mathcal{G}_B + n_R \quad (104)$$

$$\text{Number of unknowns} = \partial\mathcal{G}_A + 1 + \partial\mathcal{G}_B + 1 = 2\partial\mathcal{G}_B + 1. \quad (105)$$

Equating the number of equations to the number of unknowns:

$$\partial \mathcal{G}_B + n_R = 2\partial \mathcal{G}_B + 1$$

i.e.

$$\partial \mathcal{G}_B = n_R - 1. \quad (106)$$

From (19)

$$\partial \mathcal{B} = \tau + 1 + \partial \mathcal{G}_B = n_R + \tau. \quad (107)$$

To get the degree of \mathcal{D} , a similar technique is applied to (83). Write it as

$$(G_D - \mathcal{G}_D C)R = (G_B - \mathcal{G}_B C)M. \quad (108)$$

From (19) and (27)

$$G_D = -(D + CQ_D)q^{\tau+1} \quad (109)$$

$$G_B = (B - Q_B C)q^{\tau+1}. \quad (110)$$

From (108):

$$\mathcal{G}_D R - \mathcal{G}_B M = \text{known polynomial}. \quad (111)$$

Then:

$$\text{Number of unknowns} = \partial \mathcal{G}_D + 1 + \partial \mathcal{G}_B + 1 = \mathcal{G}_D + \partial \mathcal{G}_B + 2 \quad (112)$$

$$\text{Number of equations} = \partial \mathcal{G}_B + n_R + n_v - 1 + 1 = \partial \mathcal{G}_B + n_R + n_v. \quad (113)$$

Equating the number of unknowns to the number of equations, the minimum degree for \mathcal{G}_D is obtained:

$$\partial \mathcal{D} = \tau + n_R + n_v - 1. \quad (114)$$

Finally, by (107, 95):

$$\partial \mathcal{A} = n_R + \tau - 1. \quad (115)$$

Thus, there exist *finite* orders for A , B and D given by (107), (114) and (115) such that (7) enjoys the τ -UCPP property. Writing (7) for t replaced by $t + \tau$ yields (32). □

Proof of Lemma 1

In closed-loop

$$R^*(q^{-1}; \tau)u_t = -S^*(q^{-1}; \tau)y_t + M^*(q^{-1}; \tau)v_t + \eta_t.$$

Consequently, if

$$\begin{aligned}
 Q^*(q^{-1}; \tau) &= A^*(q^{-1})R^*(q^{-1}; \tau) + q^{-1}B^*(2^{-2})S^*(q^{-1}; \tau), \\
 y_t(\tau) &= \frac{q^{-1}B^*(q^{-1})C^*(q^{-1})}{Q^*(q^{-1}; \tau)}\eta_t + \frac{C^*(q^{-1})R^*(q^{-1}; \tau)}{Q^*(q^{-1}; \tau)}e_t + \\
 &+ \frac{q^{-1}B^*(q^{-1})M^*(q^{-1}; \tau) + D^*(q^{-1})R^*(q^{-1}; \tau)}{Q^*(q^{-1}; \tau)} \cdot \frac{q^{n_v}}{A_v^*(q^{-1})}e_v(t) \\
 u_t(\tau) &= \frac{A^*(q^{-1})C^*(q^{-1})}{Q^*(q^{-1}; \tau)}\eta_t + \frac{C^*(q^{-1})S^*(q^{-1}; \tau)}{Q^*(q^{-1}; \tau)}e_t + \\
 &+ \frac{A^*(q^{-1})M^*(q^{-1}; \tau) - D^*(q^{-1})S^*(q^{-1}; \tau)}{Q^*(q^{-1}; \tau)} \cdot \frac{q^{n_v}}{A_y^*(q^{-1})}e_v(t).
 \end{aligned}$$

Since $\{\eta\}$ is uncorrelated with both $\{e\}$ and $\{e_v\}$, the lemma follows. \square

References

1. Coito, F., Garcia, F., Lemos, J. M., Mano, A., "Modelling and Control of a drum boiler." IFAC Symp. Power Systems. Brussels, Belgium, 1988.
2. Hunt, K.J., Stochastic Optimal Control Theory with Application in Self-tuning Control. Springer-Verlag, 1989. Lecture Notes in Control and Information Sciences 117.
3. Sternad, M., Optimal and Adaptive Feedforward Regulators. Ph. D. Thesis, Uppsala Univ., Sweden, 1987.
4. Wertz, V., Demeuse, P., Application of Clarke-Gawthrop Type Controllers for the Bottom Temperature of a Glass Furnace." *Automatica*, **23**, 2, 1987. pp. 215-220.
5. Lemos, J.M., Ferreira, M., "Algorithms for adaptive filtering using control techniques." International Conference on Digital Signal Processing, Florence. Italy, 1987.
6. Elliot, S.J., Nelson, D.A., "A stochastic gradient algorithm for multichannel sound control." Prep. 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing. Lund, Sweden, 1986.
7. Lemos, J.M., Adaptive LQ Control. Ph. D. Thesis. Technical University of Lisbon, Portugal, 1989.
8. Greco, C., Menga, G., Mosca E., Zappa, G., "Performance improvements of self-tuning controllers by multistep horizons: The MUSMAR approach." *Automatica*, **20**, 1984. pp. 681-699.
9. Åstrom, K.J., Wittenmark, B., "On self-tuning regulators. *Automatica*, **9**, 1973. pp. 185-199.
10. Mosca, E., Zappa, G. "ARX modelling of controlled ARMAX plants and its application to robust multipredictor adaptive control." Proc. 24th IEEE CDC, Ft. Lauderdale, Florida, 1985. pp. 856-861. Also IEEE trans. Aut. Contr., Jan. 1989.
11. Casalino, G., Davoli, F., Minciardi, R., Zappa, G., "On implicit modelling theory: basic concepts and application to adaptive control. *Automatica*, **23**, 2. 1986. pp 189-201.

12. Mosca, E., Zappa, G., "Matrix fraction solution to the stochastic LQ disturbance rejection and servo problems." 26th IEEE CDC, 1987.
13. Mosca, E., Zappa, G., Lemos, J.M., "Robustness of multipredictor adaptive regulators: MUSMAR." 8th IFAC Symp. Ident. Syst. Param. Est., Peching, China, 1988. Also July 1989 issue of Automatica.
14. Ljung, L., "Analysis of recursive stochastic algorithms." IEEE Trans. Aut. Control, AC-22, 4. 1977. pp 551-575.

Адаптивное управление широкого диапазона для систем ARMAX с доступными помехами

И. М. ЛЕМОС

(Лиссабон)

Рассматривается адаптивное управление с прямой связью для систем ARMAX с целью уменьшить влияние измеряемых помех. Используемый алгоритм представляет собой расширенный мультипредсказывающий адаптивный алгоритм управления широкого диапазона, который обозначен MUSMAR. Показаны три основных результата: первое, параметризация мультипредсказывающих моделей, которые возможно идентифицировать при помощи стандартных наименьших квадратов для систем ARMAX с измеряемыми помехами типа ARMA, воздействующими в замкнутой цепи управления; второе, показывается алгоритм, вытекающий из соединения неявного представления системы с мультиступенчатым квадратным критерием и его робастные свойства настройки усиления регуляторов; третье, показан пример симуляции.

J. M. Lemos
INESC,
Rua Alves Redol, 9, Apartado 10105,
1017 Lisboa
Portugal

(E, Δ) -ACHIEVABLE RATES FOR MULTIPLE DESCRIPTIONS OF RANDOM VARYING SOURCE

E. A. HAROUTUNIAN, R. SH. MAROUTIAN

(Yerevan)

(Received November 28, 1989)

We study the coding by two encoders and two decoders of discrete random varying memoryless sources, when both decoders use side information about source states. We derive the inner and outer bounds for the (E, Δ) -achievable rates region, that is the rates achievable for a given pair of exponents $E = (E_1, E_2)$ of the probabilities of exceeding the distortion levels $\Delta = (\Delta_1, \Delta_2)$, respectively.

1. Introduction. Problem statement. Formulation of results

A random varying memoryless source $\{X, Y\}$ is a sequence of independent identically distributed pairs of random variables $\{(X_i, Y_i)\}_{i=1}^{\infty}$ given by the probability distribution

$$P^* \circ W^* = \{P^* \circ W^*(x, y) = P^*(x)W^*(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}.$$

The signals x of alphabet \mathcal{X} of the principal source $\{X\}$ must be transmitted to the receivers. The information about signals $y \in \mathcal{Y}$ of the additional source $\{Y\}$, considered as the states of the source $\{X\}$, can be used for the better transmission of the principal source.

Multiple description of a discrete memoryless source is a simultaneous encoding of the source by several encoders and, correspondingly, decoding by several decoders, each of which is connected to a part of the encoders [1-5].

In the present paper we study the problem of the multiple description of random varying sources by two encoders and two decoders. One of the decoders is connected only with the first encoder, and the other one with both encoders. It is supposed that both decoders have the full information about the additional source (see Fig. 1).

Let \mathcal{U}, \mathcal{V} be two finite reproduction alphabets on the first and second decoders, respectively, and $d_1: \mathcal{X} \times \mathcal{U} \rightarrow [0, \infty)$, $d_2: \mathcal{X} \times \mathcal{V} \rightarrow [0, \infty)$ be the corresponding distortion measures. For the length n sequences $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{u} \in \mathcal{U}^n$, $\mathbf{v} \in \mathcal{V}^n$ of

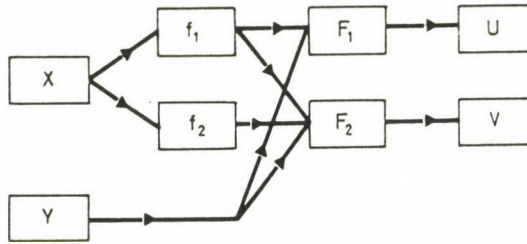


Fig. 1

distortions are defined as the average of the distortions between their corresponding elements, i.e.

$$d_1(x, u) \triangleq \frac{1}{n} \sum_{i=1}^n d_1(x_i, u_i),$$

$$d_2(x, u) \triangleq \frac{1}{n} \sum_{i=1}^n d_2(x_i, v_i).$$

An $(f, F) = (f_1, f_2, F_1, F_2)$ block code of length n is formed by two encoding functions $f_1: \mathcal{X}^n \rightarrow \{l_1, \dots, l_{L(n)}\}$, $f_2: \mathcal{X}^n \rightarrow \{k_1, \dots, k_{K(n)}\}$, and two decoding functions $F_1: \{l_1, \dots, l_{L(n)}\} \times \mathcal{Y}^n \rightarrow \mathcal{U}^n$, $F_2: \{l_1, \dots, l_{L(n)}\} \times \{k_1, \dots, k_{K(n)}\} \times \mathcal{Y}^n \rightarrow \mathcal{V}^n$. Probabilities $e_{1,n}$ and $e_{2,n}$ of exceeding the distortion levels $\Delta_1 \geq 0$ and $\Delta_2 \geq 0$, given on the first and second decoders, respectively, are defined as

$$e_{1,n} = e_1(f_1, F_1, d_1, \Delta_1, n) \triangleq \sum_x P^*(x) W^*(\{d_1(x, F_1(f_1(x), y)) > \Delta_1\} | x),$$

$$e_{2,n} = e_2(f_1, f_2, F_2, d_2, \Delta_2, n) \triangleq \sum_x P^*(x) W^*(\{d_2(x, F_2(f_1(x), f_2(x), y)) > \Delta_2\} | x).$$

Our aim is to study the characteristics of codes ensuring the exponential decrease of probabilities $e_{1,n}$ and $e_{2,n}$ with given exponents $E_1 \geq 0$ and $E_2 \geq 0$, respectively,

$$e_{j,n} \leq \exp(-nE_j), \quad j = 1, 2. \tag{1.1}$$

Two non-negative numbers R_1, R_2 are called (E, Δ) -achievable pairs of rates, if for any $\varepsilon > 0$ and for all $n \geq n(\varepsilon, R_1, R_2)$ there exists a code (f, F) satisfying (1.1) and such that

$$\frac{1}{n} \log L(n) \leq R_1 + \varepsilon, \quad \frac{1}{n} \log K(n) \leq R_2 + \varepsilon.$$

The region of all (E, Δ) -achievable rate pairs is denoted by $\mathfrak{R}(E, \Delta)$.

Let $P \triangleq \{P(x), x \in \mathcal{X}\}$ be a probability distribution on \mathcal{X} , $W \triangleq \{W(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$ be a matrix of the conditional probabilities on \mathcal{Y} for given $x \in \mathcal{X}$, and $Q \triangleq \{Q(u, v|x), x \in \mathcal{X}, u \in \mathcal{U}, v \in \mathcal{V}\}$ be a matrix of conditional probabilities of pairs $(u, v) \in \mathcal{U} \times \mathcal{V}$ for given $x \in \mathcal{X}$.

We shall use the following notations: for divergences

$$D(P||P^*) \triangleq \sum_x P(x) \log \frac{P(x)}{P^*(x)},$$

$$D(P \circ W||P^* \circ W^*) \triangleq \sum_x P(x)W(y|x) \log \frac{P(x)W(y|x)}{P^*(x)W^*(y|x)},$$

$$D(W||W^*|P^*) \triangleq \sum_x P(x)W(y|x) \log \frac{W(y|x)}{W^*(y|x)},$$

for entropies

$$H_P(X) \triangleq -\sum_x P(x) \log P(x),$$

$$H_{P,W}(Y|X) \triangleq \sum_{x,y} P(x)W(y|x) \log W(y|x),$$

for mutual information

$$I_{P,W}(X \wedge Y) \triangleq H_P(X) - H_{P,W}(X|Y).$$

Let

$$B(E) \triangleq \{P, W: D(P \circ W||P^* \circ W^*) \leq E\}$$

and

$$M_{P,Q}d(X, U) \triangleq \sum_{x,u,v} P(x)Q(u, v|x)d(x, u).$$

Define the function

$$\Phi(P) = Q_P \triangleq \{Q_P(u, v|x), x \in \mathcal{X}, u \in \mathcal{U}, v \in \mathcal{V}\}$$

determining the correspondence of a Q to each P , such that, if $D(P||P^*) \leq E_1$, then $M_{P, \Phi(P)}d_1(X, U) \leq \Delta_1$, and if $D(P||P^*) \leq E_2$, then $M_{P, \Phi(P)}d_2(X, V) \leq \Delta_2$. Let $\mathfrak{M}(E, \Delta)$ be the set of all such functions Φ for given $E = (E_1, E_2)$ and $\Delta = (\Delta_1, \Delta_2)$.

Denote by $\mathfrak{R}_r(E, \Delta, \Phi)$ the set of all pairs R_1, R_2 such that the two following inequalities hold

$$R_1 \geq \min \left\{ \max_{P, W \in B(E_1)} I_{P, \Phi(P)}(X \wedge U) - I_{P, \Phi(P), W}(Y \wedge U) + \right. \\ \left. + E_1 - D(P \circ W||P^* \circ W^*); \max_{P: D(P||P^*) \leq E_1} I_{P, \Phi(P)}(X \wedge U) \right\},$$

$$R_1 + R_2 \geq \min \left\{ \max_{P, W \in B(E_2)} I_{P, \Phi(P)}(X \wedge UV) - I_{P, \Phi(P), W}(Y \wedge UV) + \right. \\ \left. + E_2 - D(P \circ W||P^* \circ W^*); \max_{P: D(P||P^*) \leq E_2} I_{P, \Phi(P)}(X \wedge UV) \right\}.$$

Consider the “random coding” region

$$\mathfrak{R}_r(E, \Delta) \triangleq \bigcup_{\Phi \in \mathfrak{M}(E, \Delta)} \mathfrak{R}_r(E, \Delta, \Phi),$$

and the “sphere packing” region

$$\mathfrak{R}_{sp}(E, \Delta) \triangleq \bigcup_{\Phi \in \mathfrak{M}(E, \Delta)} \mathfrak{R}_{sp}(E, \Delta, \Phi),$$

where $\mathfrak{R}_{sp}(E, \Delta, \Phi)$ is the set of all pairs R_1, R_2 such that the two following inequalities hold

$$\begin{aligned} R_1 &\geq \max_{P, W \in \mathcal{B}(E_1)} [I_{P, \Phi(P)}(X \wedge U) - I_{P, \Phi(P), W}(Y \wedge U)], \\ R_1 + R_2 &\geq \max_{P, W \in \mathcal{B}(E_2)} [I_{P, \Phi(P)}(X \wedge UV) - I_{P, \Phi(P), W}(Y \wedge UV)]. \end{aligned}$$

The following theorems will be proved in Section 3 and Section 4, respectively.

THEOREM 1. For positive $E_1, E_2, \Delta_1, \Delta_2$ the following inclusion holds

$$\mathfrak{R}(E, \Delta) \supseteq \mathfrak{R}_r(E, \Delta).$$

THEOREM 2. For positive $E_1, E_2, \Delta_1, \Delta_2$ the following inclusion holds

$$\mathfrak{R}(E, \Delta) \subseteq \mathfrak{R}_{sp}(E, \Delta).$$

Remark. As in [4] it can be proved that in the Theorems it is sufficient to use the sets \mathcal{U} and \mathcal{V} with $|\mathcal{U}| \leq |\mathcal{X}| + 2$ and $|\mathcal{V}| \leq (|\mathcal{X}| + 1)^2$. Here and later we denote by $|A|$ the cardinality of the finite set A .

Now, let us carry out the comparison of our result with earlier known ones. Gray and Wyner [1] first considered the problem of the multiple descriptions of a standard source for the same diagramme of encoders and decoders as our ones. They found the region of achievable rates for a given distortion levels. If in Theorem 1 and 2 one takes $E_1 \rightarrow 0, E_2 \rightarrow 0$ and $|\mathcal{Y}| = 1$, then he obtains the result of [1].

Multiple descriptions for sources with two encoders and three decoders are considered in papers of El Gamal, Cover [2], Ahlswede [3], Gelfand, Pinsker [5]. Heegard and Berger [4] describe the set of achievable rates of random varying source coding for given distortion criterion, when the side information can be absent at one of the decoders. In [6, 7] for some models of discrete memoryless sources the problem of determination of the E -achievable rates is considered. The same problem as in this paper but in the case of absence of the additional information on the decoders is studied in [8].

The works [9–12] (see also [13, 14]) were devoted to the study of the coding problem for correlated sources.

The region of (E, Δ) -achievable rates for some models of random varying sources is determined in [15].

The result of this paper were presented in the IXth All-Union conference on coding theory and information transmission in Odessa [16].

The proof of Theorem 1 uses the Lemma of Section 2, which is a generalization of the covering lemma from [17].

2. Typical sequences and the Covering Lemma

Let us denote the number of positions with \mathbf{x} in x by $n(\mathbf{x}|x)$. The sequence $\mathbf{x} \in \mathcal{X}^n$ has the type P if $n(\mathbf{x}|x) = nP(x)$ for all $x \in \mathcal{X}$. The set of all sequences of type P is denoted by $T_P(X)$. The number of different types of sequences in \mathcal{X}^n is less than $(n + 1)^{|\mathcal{X}|}$.

We shall use the following well-known combinatorial relations [13]:

$$(n + 1)^{-|\mathcal{X}|} \exp\{nH_P(X)\} \leq |T_P(X)| \leq \exp\{nH_P(X)\},$$

for $\mathbf{x} \in T_P(X)$

$$P^*(\mathbf{x}) = \exp\{-n(D(P||P^*) + H_P(X))\},$$

$$(n + 1)^{-|\mathcal{X}|} \exp\{-nD(P||P^*)\} \leq P^*(T_P(X)) \leq \exp\{-nD(P||P^*)\}.$$

One says that a sequence $\mathbf{y} \in \mathcal{Y}^n$ has conditional type W for given $\mathbf{x} \in \mathcal{X}^n$, if $n(\mathbf{x}, \mathbf{y}|x, y) = n(\mathbf{x}|x)W(y|x)$ for every $x \in \mathcal{X}, y \in \mathcal{Y}$. Denote the set of these sequences by $T_W(Y|\mathbf{x})$. It is known [13] that if $\mathbf{x} \in T_P(X)$, then

$$(n + 1)^{-|\mathcal{X}||\mathcal{Y}|} \exp\{nH_{P,W}(Y|X)\} \leq |T_W(Y|\mathbf{x})| \leq \exp\{nH_{P,W}(Y|X)\},$$

for $\mathbf{y} \in T_W(Y|\mathbf{x})$

$$W^*(\mathbf{y}|x) = \exp\{-n(D(W||W^*|P) + H_{P,W}(Y|X))\},$$

and

$$(n + 1)^{-|\mathcal{X}||\mathcal{Y}|} \exp\{-nD(W||W^*|P)\} \leq W^*(T_W(Y|\mathbf{x})|x) \leq \exp\{-nD(W||W^*|P)\}.$$

Denote $Q_1 \triangleq \{Q_1(x|u), x \in \mathcal{X}, u \in \mathcal{U}\}$, where

$$Q_1(x|u) = \sum_v Q(u, v|x)P(x) \left(\sum_{x,v} Q(u, v|x)P(x) \right)^{-1}.$$

We say that $\mathbf{x} \in T_P(X)$ has conditional type Q_1 , for given $\mathbf{u} \in \mathcal{U}^n$, if $n(\mathbf{x}, \mathbf{u}|x, u) = n(\mathbf{u}|u)Q_1(\mathbf{x}|u)$, for all $x \in \mathcal{X}$, $u \in \mathcal{U}$. The set of sequences $\mathbf{x} \in T_P(X)$, having conditional type Q_1 for given $\mathbf{u} \in \mathcal{U}^n$, is denoted by $T_{P,Q}(X|\mathbf{u})$. A sequence $\mathbf{u} \in \mathcal{U}^n$ has type $Q_2 \triangleq \{Q_2(u), u \in \mathcal{U}\}$ where $Q_2(u) = \sum_{x,v} Q(u, v|x)P(x)$, if $n(\mathbf{u}|u) = nQ_2(u)$, for all $u \in \mathcal{U}$. The set of sequences $\mathbf{u} \in \mathcal{U}^n$ having type Q_2 is denoted by $T_{P,Q}(U)$.

The family

$$\{T_{P,Q}(X|\mathbf{u}_j), \quad j = \overline{1, J}\}$$

is named a covering of $T_P(X)$, if

$$T_P(X) \subseteq \bigcup_{j=1}^J T_{P,Q}(X|\mathbf{u}_j), \text{ where } \mathbf{u}_j \in T_{P,Q}(U), \text{ for } j = \overline{1, J}.$$

It is clear that

$$T_{P,Q,W}(Y|\mathbf{u}) = \bigcup_{\mathbf{x} \in T_{P,Q}(X|\mathbf{u})} T_W(Y|\mathbf{x}).$$

A covering $\{T_{P,Q,W}(Y|\mathbf{u}_j), j = \overline{1, J}\}$ of $T_{P,Q}(Y)$ is called a -balanced [17] if for each $\mathbf{y} \in T_{P,W}(Y) \left| \{\mathbf{u}_j: \mathbf{y} \in T_{P,Q,W}(Y|\mathbf{u}_j)\} \right| \leq a$.

Now, we shall prove a modification of the Covering Lemma 3 from [17].

Lemma. For $\varepsilon > 0$ and any types P, Q , for large enough n there exists a covering $\{T_{P,Q}(X|\mathbf{u}_j), j = \overline{1, J(P, Q)}\}$ of $T_P(X)$ such that $\mathbf{u}_j \in T_{P,Q}(U)$

$$J(P, Q) = \exp\{nI_{P,Q}(X \wedge U) + 2\varepsilon n\}$$

and for all conditional types W there exist $a(P, Q, W)$ -balanced covering $\{T_{P,Q,W}(Y|\mathbf{u}_j), j = \overline{1, J(P, Q)}\}$ of $T_{P,W}(Y)$ with the same $\{\mathbf{u}_j, j = \overline{1, J(P, Q)}\}$ and $J(P, Q)$ such that

$$a(P, Q, W) = \exp\{n[I_{P,Q}(X \wedge U) - I_{P,Q,W}(Y \wedge U)] + 4\varepsilon n\}.$$

Proof. We prove the existence of coverings by the method of random selection. Let $\{\xi_j, j = \overline{1, J(P, Q)}\}$ be a sequence of random variables independent and uniformly distributed over $T_{P,Q}(U)$.

Let us denote by

$$\psi_1(x) \triangleq \begin{cases} 1, & \text{if } \mathbf{x} \notin \bigcup_{j=1}^{J(P,Q)} T_{P,Q}(X|\xi_j), \\ 0, & \text{otherwise;} \end{cases}$$

$$\psi_2(y, j) \triangleq \begin{cases} 1, & \text{if } \mathbf{y} \notin T_{P,Q,W}(Y|\xi_j), \\ 0, & \text{otherwise.} \end{cases}$$

Upperbound now the following expression

$$\Pr\left\{\sum_{\mathbf{x} \in T_P(X)} \psi_1(\mathbf{x}) \geq 1\right\} + \sum_W \Pr\left\{\sum_{\mathbf{y} \in T_{P,W}(Y)} \sum_{j=1}^{J(P,Q)} \psi_2(\mathbf{y}, j) \geq 1\right\} + \sum_W \sum_{\mathbf{y} \in T_{P,W}(Y)} \Pr\left\{\sum_{j=1}^{J(P,Q)} \psi_2(\mathbf{y}, j) \leq J(P, Q) - a(P, Q, W)\right\}.$$

As in the proof of Lemma 4.1 from Chapter 2 of [13] we can obtain

$$\Pr\left\{\sum_{\mathbf{x} \in T_P(X)} \psi_1(\mathbf{x}) \geq 1\right\} \leq |T_P(X)| \Pr\left\{\mathbf{x} \notin \bigcup_{j=1}^{J(P,Q)} T_{P,Q}(X|\xi_j)\right\} \leq |T_P(X)|(1 - |T_{P,Q}(X|\mathbf{u})||T_P(X)|^{-1})^{J(P,Q)} \leq |T_P(X)| \exp\{-J(P, Q) \exp\{-nI_{P,Q}(X \wedge U) + n\varepsilon\}\},$$

and

$$\Pr\left\{\sum_{j=1}^{J(P,Q)} \sum_{\mathbf{y} \in T_{P,W}(Y)} \psi_2(\mathbf{y}, j) \geq 1\right\} \leq |T_{P,W}(Y)| \exp\{-J(P, Q) \exp\{-n[I_{P,Q,W}(Y \wedge U) + \varepsilon]\}\}.$$

As it holds from the Covering Lemma 3 [17], we have for $\alpha > 0$

$$\Pr\left\{\sum_{j=1}^{J(P,Q)} \psi_2(\mathbf{y}, j) < J(P, Q) - a(P, W, Q)\right\} \leq \exp\{-\alpha(J(P, Q) - a(P, W, Q))\} \prod_{j=1}^{J(P,Q)} M \exp\{\alpha\psi_2(\mathbf{y}, j)\}.$$

Then we obtain

$$M \exp\{\alpha\psi_2(\mathbf{y}, j)\} = |T_{P,Q,W}(Y|\mathbf{u})||T_{P,W}(Y)|^{-1} + e^\alpha (|T_{P,W}(Y)| - |T_{P,Q,W}(Y|\mathbf{u})|)|T_{P,W}(Y)|^{-1}$$

and if we take

$$\alpha = \log[(J(P, Q) - a(P, Q, W))|T_{P,Q,W}(Y|\mathbf{u})|] - \log[a(P, Q, W)(|T_{P,W}(Y)| - |T_{P,Q,W}(Y|\mathbf{u})|)]$$

then

$$\Pr \left\{ \sum_{j=1}^{J(P,Q)} \psi_2(\mathbf{y}, j) < J(P, Q) - a(P, Q, W) \right\} \leq \leq \exp\{J(P, Q)[h(\lambda) - \lambda n I_{P,Q,W}(Y \wedge U)]\},$$

where $h(\lambda) = -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda)$, $\lambda = a(P, Q, W)/J(P, Q)$. Hence, if we choose

$$J(P, Q) = \exp\{n I_{P,Q}(X \wedge U) + 2\epsilon n\},$$

and

$$a(P, Q, W) = \exp\{n[I_{P,Q}(X \wedge U) - I_{P,Q,W}(Y \wedge U)] + 4\epsilon n\},$$

then the statement of the Lemma holds.

3. Proof of Theorem 1

We have $\mathcal{X}^n = \bigcup_P T_P(X)$, with P running over all possible types on \mathcal{X}^n . For each type P choose some conditional type Q such that $Q = \Phi(P)$ for some fixed $\Phi \in \mathfrak{M}(E, \Delta)$. Let, according to the Lemma,

$$\{\mathbf{u}_{j(P, \Phi(P))} \in T_{P, \Phi(P)}(U), \quad j = \overline{1, J(P, \Phi(P))}\}$$

be a covering for $T_P(X)$ and $a(P, Q, W)$ -balanced covering for $T_{P,W}(Y)$.

Consider the covering of $T_P(X)$, consisting of disjoint components

$$C_j(P, \Phi) = T_{P, \Phi(P)}(X | \mathbf{u}_{j(P, \Phi(P))}) \setminus \bigcup_{j'(P, \Phi(P)) < j(P, \Phi(P))} T_{P, \Phi(P)}(X | \mathbf{u}_{j'(P, \Phi(P))}), \quad j = \overline{1, J(P, \Phi(P))}.$$

For each type P and conditional type W define following sets

$$S_1(\mathbf{y}, P, \Phi, W) \triangleq \triangleq \{\mathbf{u}_{j(P, \Phi(P))} : \mathbf{y} \in T_{P, \Phi(P)}(Y | \mathbf{u}_{j(P, \Phi(P))}), \quad j = \overline{1, J(P, \Phi(P))}\}.$$

From the Lemma we have that

$$S_1(\mathbf{y}, P, \Phi, W) \leq \leq \exp\{n[I_{P, \Phi(P)}(X \wedge U) - I_{P, \Phi(P), W}(Y \wedge U)] + 4\epsilon n\}.$$

Further, by the Lemma, for every fixed $\mathbf{u}_{j(P, \Phi(P))} \in T_{P, \Phi(P)}(U)$ there exist coverings

$$\{T_{P, \Phi(P)}(X | \mathbf{u}_{j(P, \Phi(P))}, \mathbf{v}_{g(P, \Phi(P))}), g(P, \Phi(P)) = \overline{1, G(P, \Phi(P))}\}$$

for $T_{P, \Phi(P)}(X | \mathbf{u}_{j(P, \Phi(P))})$, and for each conditional type W there exist $\exp\{n[I_{P, \Phi(P)}(X \wedge V | U) - I_{P, \Phi(P), W}(Y \wedge V | U)] + 4\varepsilon\}$ -balanced coverings

$$\{T_{P, \Phi(P), W}(Y | \mathbf{u}_{j(P, \Phi(P))}, \mathbf{v}_{g(P, \Phi(P))}), g(P, \Phi(P)) = \overline{1, G(P, \Phi(P))}\}$$

of $T_{P, \Phi(P), W}(Y | \mathbf{u}_{j(P, \Phi(P))})$, where

$$T_{P, \Phi(P), W}(Y | \mathbf{u}, \mathbf{v}) = \bigcup_{\mathbf{x} \in T_{P, \Phi(P)}(X | \mathbf{u}, \mathbf{v})} T_W(Y | \mathbf{x}),$$

and

$$G(P, \Phi(P)) = \exp\{n[I_{P, \Phi(P)}(X \wedge V | U) + 2\varepsilon]\}.$$

Let

$$\begin{aligned} C_{j, g}(P, \Phi(P)) &\triangleq \\ &\triangleq C_j(P, \Phi(P)) \cap (T_{P, \Phi(P)}(X | \mathbf{u}_{j(P, \Phi(P))}, \mathbf{v}_{g(P, \Phi(P))}) \setminus \\ &\setminus \bigcup_{g'(P, \Phi(P)) < g(P, \Phi(P))} T_{P, \Phi(P)}(X | \mathbf{u}_{j(P, \Phi(P))}, \mathbf{v}_{g'(P, \Phi(P))})), \end{aligned}$$

and

$$\begin{aligned} S_2(\mathbf{y}, \mathbf{u}, P, \Phi, W) &\triangleq \\ &\triangleq \{\mathbf{v}_{g(P, \Phi(P))} : \mathbf{y} \in T_{P, \Phi(P), W}(Y | \mathbf{u}, \mathbf{v}_{g(P, \Phi(P))}), g(P, \Phi(P)) = \overline{1, G(P, \Phi(P))}\}. \end{aligned}$$

We have

$$|S_2(\mathbf{y}, \mathbf{u}, P, \Phi, W)| \leq \exp\{n[I_{P, \Phi(P)}(X \wedge V | U) - I_{P, \Phi(P), W}(Y \wedge V | U) + 4\varepsilon]\}.$$

Now, using the random coding method we shall prove the existence of code (f, F) satisfying the conditions of Theorem 1. Let $\{\zeta_1(\mathbf{u}), \mathbf{u} \in \mathcal{U}^n\}$ be the family of independent, identically distributed random variables with distribution $\Pr\{\zeta(\mathbf{u}) = l\} = L^{-1}(n)$, $l = \overline{1, L(n)}$ and $\{\eta(\mathbf{v}), \mathbf{v} \in \mathcal{V}^n\}$ be the family of independent, identically distributed random variables with distribution $\Pr\{\eta(\mathbf{v}) = k\} = K^{-1}(n)$, $k = \overline{1, K(n)}$.

Consider the following coding functions

$$\begin{aligned} f_1(\mathbf{x}) &= \zeta(\mathbf{u}_{j(P, \Phi(P))}), & \text{if } \mathbf{x} \in C_j(P, \Phi(P)), & \text{ and} \\ f_2(\mathbf{x}) &= \eta(\mathbf{v}_{g(P, \Phi(P))}), & \text{if } \mathbf{x} \in C_{j, g}(P, \Phi(P)), \end{aligned}$$

and decoding functions (with some fixed vectors \mathbf{u}_0 and \mathbf{v}_0)

$$F_1(l, \mathbf{y}) \triangleq \begin{cases} \mathbf{u}, & \text{if } \mathbf{u} \in S_1(\mathbf{y}, P, \Phi, W) \cap f_1^{-1}(l), \\ & \text{where } |S_1(\mathbf{y}, P, \Phi, W)| \geq |S_1(\mathbf{y}, P', \Phi, W')|; \\ \mathbf{u}_0, & \text{otherwise;} \end{cases}$$

and

$$F_2(l, k, \mathbf{y}) \triangleq \begin{cases} \mathbf{v}, & \text{if } \mathbf{v} \in S_2(\mathbf{y}, \mathbf{u}, P, \Phi, W) \cap f_2^{-1}(k), \\ & \text{where } \mathbf{u} = F_1(l, \mathbf{y}); \\ & |S_2(\mathbf{y}, \mathbf{u}, P, \Phi, W)| \geq |S_2(\mathbf{y}, \mathbf{u}, P', \Phi, W')|; \\ \mathbf{v}_0, & \text{otherwise.} \end{cases}$$

Define now

$$\varphi_1(\mathbf{x}, \mathbf{y}) \triangleq \begin{cases} 1, & \text{if } \zeta(\mathbf{u}_j) = \zeta(\mathbf{u}_{j'}), \mathbf{x} \in C_j(P, \Phi) \text{ and} \\ & \mathbf{u}_j, \mathbf{u}_{j'} \in S_1(\mathbf{y}, P, \Phi, W), \\ 0, & \text{otherwise;} \end{cases}$$

$$\varphi_2(\mathbf{x}, \mathbf{y}) \triangleq \begin{cases} 1, & \text{if } \eta(\mathbf{v}_g) = \eta(\mathbf{v}_{g'}), \mathbf{x} \in C_{j,g}(P, \Phi), \\ & \mathbf{v}_g, \mathbf{v}_{g'} \in S_2(\mathbf{y}, \mathbf{u}_j, P, \Phi, W), \mathbf{u}_j \in S_1(\mathbf{y}, P, \Phi, W), \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$e_1(\zeta, \eta) \triangleq \sum_{\mathbf{x}, \mathbf{y}} P^*(\mathbf{x}) W^*(\mathbf{y}|\mathbf{x}) \varphi_1(\mathbf{x}, \mathbf{y})$$

and

$$e_2(\zeta, \eta) \triangleq \sum_{\mathbf{x}, \mathbf{y}} P^*(\mathbf{x}) W^*(\mathbf{y}|\mathbf{x}) \varphi_1(\mathbf{x}, \mathbf{y}) \varphi_2(\mathbf{x}, \mathbf{y}).$$

Now, we upperbound

$$\Pr\{e_1(\zeta, \eta) > \exp(-nE_1)\} + \Pr\{e_2(\zeta, \eta) > \exp(-nE_2)\}.$$

As it is shown in [15] for large enough n

$$\Pr\{e_1(\zeta, \eta) > \exp(-nE_1)\} \leq \exp\left\{n \max_{P, W, B(E_1+\epsilon)} [I_{P, \Phi(P)}(X \wedge U) - I_{P, \Phi(P), W}(Y \wedge U) + E_1 - D(P \circ W \| P^* \circ W^*) + 2\epsilon] - \log L(n)\right\}.$$

Similarly, for large enough n

$$\Pr\{e_2(\zeta, \eta) > \exp(-nE_2)\} \leq \exp\{n(E_2 + \epsilon)\} \times \sum_{P, W \in B(E_2+\epsilon)} \sum_{\mathbf{x} \in T_P(X)} \sum_{\mathbf{y} \in T_W(Y|\mathbf{x})} P^*(\mathbf{x}) W^*(\mathbf{y}|\mathbf{x}) M \varphi_1(\mathbf{x}, \mathbf{y}) M \varphi_2(\mathbf{x}, \mathbf{y}) \leq \exp\left\{n \max_{P, W \in B(E_2+\epsilon)} [I_{P, \Phi(P)}(X \wedge UV) - I_{P, \Phi(P), W}(Y \wedge UV) + E_2 - D(P \circ W \| P^* \circ W^*) + 2\epsilon] - \log L(n) - \log K(n)\right\}.$$

Finally, we obtain that for

$$L(n) \geq \exp \left\{ n \max_{P, W \in B(E_1 + \epsilon)} [I_{P, \Phi(P)}(X \wedge U) - I_{P, \Phi(P), W}(Y \wedge U) + E_1 - D(P \circ W \| P^* \circ W^*) + 3\epsilon] \right\},$$

and

$$L(n)K(n) \geq \exp \left\{ n \max_{P, W \in B(E_2 + \epsilon)} [I_{P, \Phi(P)}(X \wedge UV) - I_{P, \Phi(P), W}(Y \wedge UV) + E_2 - D(P \circ W \| P^* \circ W^*) + 3\epsilon] \right\},$$

there exists a code (f, F) with rates satisfying the conditions of the Theorem.

4. Proof of Theorem 2

Let rates R_1, R_2 be (E, Δ) -achievable for some code (f, F) . Denote by

$$G_1 \triangleq \{(\mathbf{x}, \mathbf{y}) : d_1(\mathbf{x}; F_1(f_1(\mathbf{x}), \mathbf{y})) \leq \Delta_1\}.$$

Consider some type $P \circ W \in B(E_1 - \epsilon)$, for $\epsilon > 0$. Then

$$|G_1 \cap T_{P, W}(X, Y)| = P^* \circ W^*(G_1 \cap T_{P, W}(X, Y))(P^* \circ W^*(\mathbf{x}, \mathbf{y}))^{-1},$$

where $(\mathbf{x}, \mathbf{y}) \in T_{P, W}(X, Y)$. It follows from Section 2 and inequality $P(A \cap B) \geq P(A) + P(B) - 1$ that for types $P, W \in B(E_1 - \epsilon)$ and for sufficiently large n

$$|G_1 \cap T_{P, W}(X, Y)| \geq \exp\{n(H_{P, W}(X, Y) - \epsilon)\}. \tag{4.1}$$

Denote by $A(\mathbf{y})$ the set of those $\mathbf{u} \in \mathcal{U}^n$, which for given \mathbf{y} and some $\mathbf{x} \in \mathcal{X}^n$ satisfies the condition $\mathbf{u} = F_1(f_1(\mathbf{x}), \mathbf{y})$. Then

$$|G_1 \cap T_{P, W}(X, Y)| \leq \sum_{\mathbf{y}: \exists \mathbf{x}, (\mathbf{x}, \mathbf{y}) \in G_1} \sum_{\mathbf{u} \in A(\mathbf{y})} |\{\mathbf{x} : d_1(\mathbf{x}, \mathbf{u}) \leq \Delta_1\}|. \tag{4.2}$$

Let $Q' \triangleq \{Q'(u|x, y)\}$ be some conditional type in \mathcal{U}^n for given $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{y} \in \mathcal{Y}^n$.

If $(\mathbf{x}, \mathbf{y}, \mathbf{u}) \in T_{P, Q, W}(X, Y, U)$, then

$$\begin{aligned} d_1(\mathbf{x}, \mathbf{u}) &= \frac{1}{n} \sum_{i=1}^n d_1(x_i, u_i) = \\ &= \frac{1}{n} \sum_{\mathbf{x}, \mathbf{y}, \mathbf{u}} n(x, y, u|x, y, u) d_1(\mathbf{x}, \mathbf{y}) = M_{P, Q', W} d_1(X, U) \leq \Delta_1. \end{aligned}$$

For fixed \mathbf{y} and \mathbf{u} the set of those \mathbf{x} , for which $d_1(\mathbf{x}, \mathbf{u}) \leq \Delta_1$, can be covered by the collection of conditional types $T_{P, Q', W}(X|y, u)$. Hence, from (4.2) we obtain

$$\begin{aligned} |G_1 \cap T_{P, W}(X, Y)| &\leq \\ &\leq \sum_{\mathbf{y}: \exists \mathbf{x}, (\mathbf{x}, \mathbf{y}) \in G_1} \sum_{\mathbf{u} \in A(\mathbf{y})} (n+1)^{|\mathcal{X}||\mathcal{Y}|} \max_{Q': M_{P, Q', W} d_1(X, U) \leq \Delta_1} |T_{P, W, Q'}(X|y, u)| \leq \\ &\leq L_1 \exp \left\{ n \left[H_{P, W}(Y) + \max_{Q': M_{P, Q', W} d_1(X, U) \leq \Delta_1} H_{P, W, Q'}(X|Y, U) - \varepsilon/2 \right] \right\}. \end{aligned} \tag{4.3}$$

From (4.1) and (4.3) we have

$$\begin{aligned} \frac{1}{n} \log L_1 &\geq \\ &\geq H_{P, W}(X, Y) - H_{P, W}(Y) - \max_{Q': M_{P, Q', W} d_1(X, U) \leq \Delta_1} H_{P, Q', W}(X|Y, U) - \varepsilon = \\ &= \min_{Q': M_{P, Q', W} d_1(X, U) \leq \Delta_1} I_{P, Q', W}(X \wedge U|Y) - \varepsilon. \end{aligned} \tag{4.4}$$

Denote by

$$Q(\mathbf{u}|\mathbf{x}) = \sum_{\mathbf{y}} Q'(\mathbf{u}|\mathbf{x}, \mathbf{y}) W(\mathbf{y}|\mathbf{x}).$$

We have from (4.4) that

$$\begin{aligned} \frac{1}{n} \log L_1 &\geq \\ &\geq \max_{P, W \in B(E_1 - \varepsilon)} \min_{Q: M_{P, Q} d_1(X, U) \leq \Delta_1} I_{P, Q', W}(X \wedge U|Y) - \\ &- \max_{P, W \in B(E_1 - \varepsilon)} \min_{Q: M_{P, Q} d_1(X, U) \leq \Delta_1} [I_{P, Q}(X \wedge U) - I_{P, Q, W}(Y \wedge U) - \varepsilon]. \end{aligned} \tag{4.5}$$

Similarly, we obtain that

$$\begin{aligned} R_1 + R_2 &\geq \\ &\geq \max_{P, W \in B(E_2 - \varepsilon)} \min_{Q: M_{P, Q} d_2(X, V) \leq \Delta_2} I_{P, Q}(X \wedge UV) - I_{P, Q, W}(Y \wedge UV) - \varepsilon. \end{aligned} \tag{4.6}$$

Taking into account the continuity of the right-side of (4.5) and (4.6) by E_1 and E_2 , respectively, and the arbitrariness of ε , we obtain the statement of Theorem 2.

References

1. Gray, R. M., Wyner, A. D., Source Coding for a Simple Network. Bell System Technical Journal **58** (1974), 9, pp. 1681-1721.
2. El Gamal, A., Cover, T., Achievable rates for multiple descriptions, IEEE Trans. Inform. Theory **28** (1982), 6, pp. 851-857.
3. Ahlswede, R., The rate-distortion region for multiple descriptions without excess rate. IEEE Trans. Inform. Theory, **31** (1985), 6, pp. 721-726.
4. Heegard, C., Berger, T., Rate-distortion when side information may be absent. IEEE Trans. Inform. Theory, **31** (1985), 5, pp. 727-734.
5. Gelfand, S. I., Pinsker, M. S., Source coding without redundancy for network with two encoders and three receivers. Problems of Control and Information Theory, **14** (1985), 5, pp. 319-328.
6. Haroutunian, E. A., Mekaush, B., Estimates for optimal rates of codes with given error probability exponent for several sources. VIth Intern. symposium on information theory. Thesis of report. Moscow-Tashkent, (1984), Part I, pp. 22-23 (in Russian).
7. Haroutunian, E. A., Rate-reliability function. Journal of Information Processing and Cybernetics, (1989) (in Russian).
8. Marutian, R. Sh., Achievable rates for multiple descriptions with given exponent and distortion levels. "Problemy peredachi informatsii", (1989) (in Russian).
9. Slepian, D., Wolf, J. K., Noiseless coding of correlated information source. IEEE Trans. Inform. Theory, **25** (1973), 4, pp. 471-480.
10. Wyner, A. D., Ziv, J., A theorem on the entropy of certain binary sequences and application. Part 2. IEEE Trans. Inform. Theory, **19** (1973), 6, pp. 769-778.
11. Ahlswede, R., Körner, J., Source coding with side information at the decoder and a converse for degraded broadcast channels. IEEE Trans. Inform. Theory, **21** (1975), 6, pp. 629-657.
12. Gelfand, S. I., Pinsker, M. S., Source coding under observations with uncomplete information. Problemy peredachi informatsii, **15** (1979), 2, pp. 45-58 (in Russian).
13. Csiszár, I., Körner, J., Information theory. Coding theorems for discrete memoryless systems. Akadémiai Kiadó, Budapest, 1981.
14. Kolesnik, B. D., Poltirev, G. Sh., Handbook of Information Theory, Moscow, Nauka, 1982 (in Russian).
15. Haroutunian, E. A., Marutian, R. Sh., E -optimal rates of coding for randomly varying source, Problemy peredachi informatsii, (1989) (in Russian).
16. Haroutunian, E. A., Marutian, R. Sh., (E, Δ) -achievable rates multiple descriptions for randomly varying source, IXth All-Union conference on coding theory and information transmission, Odessa, 1988, pp. 6-9 (in Russian).
17. Ahlswede, R., Coloring hypergraphs: a new approach to multi-user source coding-II. Journal of Combin. Information and System Sciences, **5** (1980), 3, pp. 220-268.

(E, Δ) -достижимые скорости множественного описания случайно меняющегося источника

Е. А. АРУТЮНЯН, Р. Ш. МАРУТЯН

(Ереван)

Множественное описание источника — это его кодирование одновременно несколькими кодерами и декодирование соответственно несколькими декодерами, каждый из которых связан с частью из кодеров. Изучается задача множественного описания случайно меняющегося источника двумя кодерами и двумя декодерами, один из которых связан лишь с первым кодером, а второй — с обоими. Найдены внутренняя и внешняя границы (E, Δ) -достижимых скоростей, то есть скоростей, достижимых при заданной паре экспонент $E = (E_1, E_2)$ вероятностей превышения, соответственно, уровней искажения $\Delta = (\Delta_1, \Delta_2)$.

Е. А. Арутюнян,
Р. Ш. Марутян
СССР, 375044 Ереван-44,
ул. П. Севака, 1
ВЦ АН Арм. ССР и ЕГУ

Typesetting by TYPOT_EX Kft, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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 V. STREJC
UTIA ČSAV
18208 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

or to L. GYÖRFI
Technical University of Budapest
H-1111 Budapest, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

The papers, preferably in English or Russian, should be typed double spaced on one side of good-quality paper with wide margins (4–5 cm). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as $\text{diag}(d_1, d_2, \dots, d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском языке в трех экземплярах, на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и возвратить в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи принятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Kramosil I.</i> : Definition and recognition of classical sets by the rough ones (<i>Крамосил И.</i> Определение и распознавание классических множеств посредством грубых множеств)	77
<i>Korovin S. K., Nikitina M. G. and Nikitin S. V.</i> : Infinite-dimensional systems: Design of Sakawa controllers. Part II (<i>Коровин С. К., Никитина М. Г., Никитин С. В.</i> Бесконечномерные системы. Синтез регуляторов Сакавы. Часть II)	97
<i>Hulkó G.</i> : Lumped input and distributed output systems at the control of distributed parameter systems (<i>Хулко Г.</i> Системы со сосредоточенным входом и распределенным выходом при управлении системами с распределенными параметрами)	113
<i>Chentsov A. G.</i> : On the construction of solution to nonregular problems of optimal control (<i>Ченцов А. Г.</i> О конструкции решений нерегулярных задач оптимального управления)	129
<i>Lemos J. M.</i> : Long-range adaptive control of ARMAX plants with accessible disturbances (<i>Лемос Й. М.</i> Адаптивное управление широкого диапазона для систем ARMAX с доступными помехами)	145
<i>Naroutunian E. A., Maroutian R. Sh.</i> : (E, Δ) -achievable rates for multiple descriptions of random varying source (<i>Арутюнян Е. А., Марутян Р. Ш.</i> (E, Δ) -достижимые скорости множественного описания случайно меняющегося источника)	165

316920

9

VOL. 20 • NUMBER 3
TOM HOMEP

no

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



PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
ПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

no

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1991

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

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

- theory of control processes
- theory of adaptive systems
- theory of estimation and identification
- theory of controlling robot-technologic and flexible manufacturing systems
- information theory
- information-theoretic aspects of multiple access networks.

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

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

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

- теория процессов управления;
- теория адаптивных систем;
- теория оценивания и идентификации;
- теория управления робототехническими и гибкими производственными системами;
- теория информации;
- теория информации в области сетей с множественным доступом.

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

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czech and Slovak Federal Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia:

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

For all other countries:

PERGAMON PRESS PLC Headington Hill Hall, Oxford OX3 OBW, England

or

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

1991 Subscription Rate DM 627,— per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITOR

N. N. KRASOVSKII (USSR)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

K. V. FROLOV

A. B. KURZHANSKI

I. A. OVSEEVICH

E. D. TERYAEV

R. Z. KHASHMINSKII

HUNGARY

T. VÁMOS

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

L. GYÖRFI

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJČ

I. VAJDA

РЕДАКТОР ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

К. В. ФРОЛОВ

А. Б. КУРЖАНСКИЙ

И. А. ОВСЕЕВИЧ

Е. Д. ТЕРЯЕВ

Р. З. ХАСЬМИНСКИЙ

ВНР

Т. ВАМОШ

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

Л. КЕВИЦКИ

Л. ДЪЕРФИ

Я. КОЧИШ

ЧССР

Й. БЕНЕШ

В. СТРЕЙЦ

И. ВАЙДА



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

THE DISCRETE MAXIMUM PRINCIPLE AS A SUFFICIENT OPTIMALITY CONDITION

MARCIN STUDNIARSKI

(*Lódź*)

(Received October 22, 1991)

We present an example showing that the validity of the maximum principle in the subdifferential form for convex nondifferentiable discrete-time control problems is not sufficient for optimality. We also prove that the desired sufficiency property can be achieved under some additional assumption concerning the objective function.

1. Introduction

In some recent publications, e.g. [2-6, 9, 11], various kinds of generalized subdifferentials were used to formulate the maximum principle for certain classes of nonsmooth and nonconvex discrete-time optimal control problems. All those generalized subdifferentials, except for the ones considered in [4], reduce to the usual subdifferentials (cf. [7]) when the control problem is convex. In this case, the natural question arises whether the necessary optimality conditions formulated in the subdifferential form are also sufficient for optimality or not. Unfortunately, the answer is not as simple as in the case of convex programming problems.

It is known that if the functions appearing in the problem are both convex and differentiable, then the discrete maximum principle is actually a sufficient optimality condition (cf. [1, § 13]). For the nondifferentiable case, the authors of [5] claim (Theorem 5) that the same conclusion is true for the "weak" maximum principle. In this paper we present an example which shows that the validity of the "strong" maximum principle for convex problems is not sufficient for optimality. However, the same example can also be used to prove that the above-mentioned statement in [5] is false (see Remark 3.2 below). Further, we show that the desired sufficiency property can be achieved under some additional assumption concerning the objective function. This assumption is always satisfied when the function is differentiable.

In the paper we make use of some notions and theorems of convex analysis which can be found in Chapter V of [7]. In particular, we recall that the *subdifferential* of a convex function $f : \mathbf{R}^n \rightarrow \mathbf{R} \cup \{+\infty\}$ at x is defined by

$$\partial f(x) := \{z \in \mathbf{R}^n : f(x+h) - f(x) \geq \langle z, h \rangle, \forall h \in \mathbf{R}^n\} \quad (1.1)$$

(we assume that the value $f(x)$ is finite). Moreover, we have, by [7, Thm. 23.2],

$$\partial f(x) = \{z \in \mathbf{R}^n : f'(x; v) \geq \langle z, v \rangle, \forall v \in \mathbf{R}^n\} \quad (1.2)$$

where the *directional derivative* $f'(x; v)$ is defined by

$$f'(x; v) := \lim_{\lambda \rightarrow 0^+} (f(x + \lambda v) - f(x))/\lambda. \quad (1.3)$$

The above limit always exists and has the following property (cf. [7, Thm. 23.1]):

$$f'(x; v) = \inf_{\lambda > 0} (f(x + \lambda v) - f(x))/\lambda. \quad (1.4)$$

Let us now consider the Cartesian product $X = X_1 \times \dots \times X_k$ where $X_i = \mathbf{R}^{m_i}$, $i = 1, \dots, k$, and let $f : X \rightarrow \mathbf{R}$ be a convex function. The derivative of the function $f(\bar{x}_1, \dots, \bar{x}_{i-1}, \cdot, \bar{x}_{i+1}, \dots, \bar{x}_k)$ at \bar{x}_i in the direction v will be denoted by $f'_{x_i}(\bar{x}_1, \dots, \bar{x}_k; v)$, while its subdifferential at \bar{x}_i - by $\partial_{x_i} f(\bar{x}_1, \dots, \bar{x}_k)$ (this set will be called the *partial subdifferential* of f with respect to x_i). It is easy to prove (cf. [8, Lemma 2]) that

$$\text{pr}_i(\partial f(\bar{x}_1, \dots, \bar{x}_k)) \subset \partial_{x_i} f(\bar{x}_1, \dots, \bar{x}_k), \quad i = 1, \dots, k. \quad (1.5)$$

For a convex set $A \subset \mathbf{R}^n$, we shall denote by $N(\bar{x} | A)$ the *normal cone* to A at $\bar{x} \in A$, i.e.

$$N(\bar{x} | A) := \{z \in \mathbf{R}^n : \langle a - \bar{x}, z \rangle \leq 0, \forall a \in A\}.$$

It is easy to verify that if $A = A_1 \times \dots \times A_k \subset X$, then

$$\text{pr}_i N((\bar{x}_1, \dots, \bar{x}_k) | A) = N(x_i | A_i), \quad i = 1, \dots, k. \quad (1.6)$$

Throughout the paper, A^T will denote the transpose of a matrix A , and $\text{ri } U$ will denote the relative interior of a set U .

2. The discrete maximum principle for convex problems

In this section we consider a convex version of the discrete-time optimal control problem examined in [9]. Theorem 2.1 below contains a maximum principle

formulated under an additional regularity condition (a variant of the Slater condition) which ensures that the multiplier corresponding to the objective function can be chosen as 1. In the next section we shall show that even this strengthened maximum principle may not be sufficient for optimality.

Let us consider the following problem:

$$\text{minimize } J(\mathbf{x}, \mathbf{u}) := \sum_{i=0}^{N-1} f_i(x_i, u_i) \quad \text{subject to} \quad (2.1)$$

$$x_{i+1} = A_i x_i + B_i u_i + c_i, \quad i = 0, 1, \dots, N-1, \quad (2.2)$$

$$u_i \in U_i \subset \mathbb{R}^r, \quad i = 0, 1, \dots, N-1, \quad (2.3)$$

$$E_0 x_0 = d_0, \quad E_N x_N = d_N, \quad (2.4)$$

$$g_i(x_i) \leq 0, \quad i = 1, \dots, N-1, \quad (2.5)$$

where $\mathbf{x} = (x_0, x_1, \dots, x_N)$, $\mathbf{u} = (u_0, u_1, \dots, u_{N-1})$, $x_i \in \mathbb{R}^n$, $u_i \in \mathbb{R}^r$, A_i , B_i , E_0 and E_N are given matrices of dimensions $n \times n$, $n \times r$, $n \times m_0$ and $n \times m_N$, respectively, c_i are given vectors in \mathbb{R}^n , the sets U_i are convex and closed, while the functions $f_i : \mathbb{R}^n \times \mathbb{R}^r \rightarrow \mathbb{R}$ and $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex.

By the *optimal value* of problem (2.1)–(2.5) we shall mean the infimum of J over the set of all pairs (\mathbf{x}, \mathbf{u}) satisfying conditions (2.2)–(2.5) (such pairs will be called *admissible*).

THEOREM 2.1. Suppose that the optimal value of problem (2.1)–(2.5) is greater than $-\infty$. Let us define

$$I := \{i \in \{1, \dots, N-1\} : g_i \text{ is not affine}\}.$$

Suppose that there exist an admissible pair (\mathbf{x}, \mathbf{u}) such that

$$u_i \in \text{ri } U_i \quad \text{for } i = 0, 1, \dots, N-1, \quad (2.6)$$

$$g_i(x_i) < 0 \quad \text{for } i \in I. \quad (2.7)$$

If $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ is an optimal pair for problem (2.1)–(2.5), then there exist elements

$$p_i \in \mathbb{R}^n, \quad i = 1, \dots, N, \quad l_0 \in \mathbb{R}^{m_0}, \quad l_N \in \mathbb{R}^{m_N}, \quad \mu_i \geq 0, \quad i = 1, \dots, N-1, \quad (2.8)$$

such that

$$A_0^T p_1 - E_0^T l_0 \in \partial_{x_0} f_0(\bar{x}_0, \bar{u}_0), \quad (2.9)$$

$$A_i^T p_{i+1} - p_i \in \partial_{x_i} f_i(\bar{x}_i, \bar{u}_i) + \mu_i \partial g_i(\bar{x}_i), \quad i = 1, \dots, N-1, \quad (2.10)$$

$$p_N = -E_N^T l_N, \quad (2.11)$$

$$\begin{aligned} & (p_{i+1}, A_i \bar{x}_i + B_i \bar{u}_i + c_i) - f_i(\bar{x}_i, \bar{u}_i) = \\ & \max_{u_i \in U_i} \{ (p_{i+1}, A_i \bar{x}_i + B_i u_i + c_i) - f_i(\bar{x}_i, u_i) \}, \\ & i = 0, 1, \dots, N-1, \end{aligned} \quad (2.12)$$

$$\mu_i g_i(\bar{x}_i) = 0, \quad i = 1, \dots, N - 1. \quad (2.13)$$

Proof. We omit the proof since it is entirely analogous to that of [9, Theorem 4.5]. Let us only note two essential differences. Firstly, instead of applying the Fritz John optimality conditions for locally Lipschitzian programming problems, one should apply the Kuhn–Tucker optimality conditions for convex programming problems, described in [7, Theorems 28.1 and 28.2]. Secondly, one should use the partial subdifferentials instead of partial generalized gradients and, consequently, apply inclusions (1.5) instead of [9, Proposition 2.2]. ■

3. An example

We shall give here an example of a convex discrete-time control problem for which there exists an admissible but not optimal pair $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ satisfying the discrete maximum principle of Theorem 2.1.

Example 3.1. We consider the following particular case of problem (2.1)–(2.5) (with $N = 3$, $n = r = 1$):

$$\begin{aligned} \text{minimize } J(\mathbf{x}, \mathbf{u}) &= \max\{x_0, u_0\} + \max\{x_1, u_1\} \text{ subject to} \\ x_{i+1} &= x_i + u_i, \quad i = 0, 1, 2, \\ u_i &\in [-2, 2] \subset \mathbf{R}, \quad i = 0, 1, 2, \\ x_0 &= 0, \quad x_3 = 0. \end{aligned}$$

Since each admissible trajectory $\mathbf{x} = (x_0, \dots, x_3)$ is contained in the set $\{0\} \times [-2, 2] \times [-4, 4] \times \{0\}$, the set of all admissible pairs (\mathbf{x}, \mathbf{u}) is compact, and so, the optimal value of the problem is greater than $-\infty$. Furthermore, assumptions (2.6) and (2.7) are trivially satisfied (we may assume all g_i to be identically zero, hence affine).

Let $\bar{\mathbf{x}}$ and $\bar{\mathbf{u}}$ be zero vectors in \mathbf{R}^4 and \mathbf{R}^3 , respectively. Then the pair $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ is admissible. But it is not optimal since we can find another admissible pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{u}})$ with $\tilde{\mathbf{x}} = (0, -1, -2, 0)$, $\tilde{\mathbf{u}} = (-1, -1, 2)$, for which

$$J(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) = -1 < 0 = J(\bar{\mathbf{x}}, \bar{\mathbf{u}}).$$

Let us now verify that $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ satisfies conditions (2.9)–(2.12). We have

$$\partial_{x_i} f_i(\bar{x}_i, \bar{u}_i) = \partial(\max\{0, u_i\})(0) = [0, 1] \quad \text{for } i = 0, 1,$$

$$\partial_{x_2} f_2(\bar{x}_2, \bar{u}_2) = \{0\}.$$

Hence (2.9)–(2.12) reduce to

$$\begin{aligned} p_1 - l_0 &\in [0, 1], \\ p_2 - l_1 &\in [0, 1], \quad p_3 - p_2 = 0, \\ p_3 &= -l_3, \\ 0 &= \max_{u_i \in [-2, 2]} \{p_{i+1}u_i - \max\{0, u_i\}\}, \quad i = 0, 1, 2. \end{aligned}$$

To satisfy all these conditions, one can choose, for instance,

$$p_1 = p_2 = p_3 = l_0 = l_3 = 0$$

or

$$p_1 = p_2 = p_3 = 1, \quad l_0 = 0, \quad l_3 = -1.$$

Remark 3.2. It is not difficult to show that the pair $(\bar{\mathbf{x}}, \bar{\mathbf{u}}) = (0, 0)$ in Example 3.1 satisfies also the “weak” maximum principle of [5, Theorem 2] with $\beta^0 = -1$. Consequently, the sufficient optimality conditions stated in [5, Theorem 5] are false.

4. When is the maximum principle sufficient?

In this section we shall impose an additional assumption on the functions f_i occurring in (2.1). Under this assumption, the validity of the maximum principle will suffice for optimality in problem (2.1)–(2.5).

THEOREM 4.1. Let $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ be an admissible pair for problem (2.1)–(2.5). Suppose that there exist elements (2.8) such that conditions (2.9)–(2.13) are satisfied. Suppose also that, for each $i \in \{0, 1, \dots, N-1\}$ and for each $(\mathbf{x}_i, \mathbf{u}_i) \in \mathbf{R}^n \times \mathbf{R}^r$, we have

$$f'_i((\bar{\mathbf{x}}_i, \bar{\mathbf{u}}_i); (\mathbf{x}_i, \mathbf{u}_i)) \geq (f_i)'_{\mathbf{x}_i}(\bar{\mathbf{x}}_i, \bar{\mathbf{u}}_i; \mathbf{x}_i) + (f_i)'_{\mathbf{u}_i}(\bar{\mathbf{x}}_i, \bar{\mathbf{u}}_i; \mathbf{u}_i). \quad (4.1)$$

Then $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ is optimal for problem (2.1)–(2.5).

Proof. Problem (2.1)–(2.5) can be considered as a convex programming problem on the space $X := (\mathbf{R}^n)^{N+1} \times (\mathbf{R}^r)^N$. Let $L : X \rightarrow \mathbf{R} \cup \{+\infty\}$ be the “extended” Lagrange function for (2.1)–(2.5) with multipliers (2.8), i.e.

$$L = L_1 + L_2 + \delta(\cdot | A) \quad (4.2)$$

where

$$L_1(\mathbf{x}, \mathbf{u}) := \sum_{i=0}^{N-1} \langle p_{i+1}, x_{i+1} - A_i x_i - B_i u_i - c_i \rangle \quad (4.3)$$

$$+ \langle l_0, E_0 x_0 - d_0 \rangle + \langle l_N, E_N x_N - d_N \rangle,$$

$$L_2(\mathbf{x}, \mathbf{u}) := J(\mathbf{x}, \mathbf{u}) + \sum_{i=1}^{N-1} \mu_i g_i(x_i), \quad (4.4)$$

$$A := (\mathbf{R}^n)^{N+1} \times U_0 \times U_1 \dots \times U_{N-1}, \quad (4.5)$$

$$\delta((\mathbf{x}, \mathbf{u}) \mid A) := \begin{cases} 0 & \text{if } (\mathbf{x}, \mathbf{u}) \in A, \\ +\infty & \text{if } (\mathbf{x}, \mathbf{u}) \notin A. \end{cases} \quad (4.6)$$

In order to prove that $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ is an optimal pair, it suffices to verify that L attains its global minimum at $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$, which is equivalent, by (1.4), to the following condition:

$$L'((\bar{\mathbf{x}}, \bar{\mathbf{u}}); (\mathbf{x}, \mathbf{u})) \geq 0 \quad \text{for all } (\mathbf{x}, \mathbf{u}) \in X. \quad (4.7)$$

Since L_1 is affine, thus differentiable, we have, for all $(\mathbf{x}, \mathbf{u}) \in X$,

$$L'_1((\bar{\mathbf{x}}, \bar{\mathbf{u}}); (\mathbf{x}, \mathbf{u})) = \sum_{i=0}^N (L_1)'_{x_i}(\bar{\mathbf{x}}, \bar{\mathbf{u}}; x_i) + \sum_{i=0}^{N-1} (L_1)'_{u_i}(\bar{\mathbf{x}}, \bar{\mathbf{u}}; u_i). \quad (4.8)$$

Next, from (4.1) and (4.4) we obtain

$$\begin{aligned} L'_2((\bar{\mathbf{x}}, \bar{\mathbf{u}}); (\mathbf{x}, \mathbf{u})) &= \sum_{i=0}^{N-1} f'_i((\bar{x}_i, \bar{u}_i); (x_i, u_i)) + \sum_{i=0}^{N-1} \mu_i g'_i(\bar{x}_i; x_i) \\ &\geq \sum_{i=0}^{N-1} (f_i)'_{x_i}(\bar{x}_i, \bar{u}_i; x_i) + \sum_{i=0}^{N-1} (f_i)'_{u_i}(\bar{x}_i, \bar{u}_i; u_i) + \sum_{i=0}^{N-1} \mu_i g'_i(\bar{x}_i; x_i) \\ &= \sum_{i=0}^N (L_2)'_{x_i}(\bar{\mathbf{x}}, \bar{\mathbf{u}}; x_i) + \sum_{i=0}^{N-1} (L_2)'_{u_i}(\bar{\mathbf{x}}, \bar{\mathbf{u}}; u_i). \end{aligned} \quad (4.9)$$

Moreover, it is easy to verify that

$$(\delta(\cdot \mid A))'((\bar{\mathbf{x}}, \bar{\mathbf{u}}); (\mathbf{x}, \mathbf{u})) = \sum_{i=0}^{N-1} (\delta(\cdot \mid U_i))'(\bar{u}_i; u_i). \quad (4.10)$$

Conditions (4.8)–(4.10) imply

$$L'((\bar{\mathbf{x}}, \bar{\mathbf{u}}), (\mathbf{x}, \mathbf{u})) \geq \sum_{i=0}^N L'_{x_i}(\bar{\mathbf{x}}, \bar{\mathbf{u}}; x_i) + \sum_{i=0}^{N-1} L'_{u_i}(\bar{\mathbf{x}}, \bar{\mathbf{u}}; u_i). \quad (4.11)$$

By using the same method as in the proofs of [9, Theorems 3.1 and 4.5], it can be shown that conditions (2.9)–(2.12) are equivalent to

$$0 \in \partial_{x_i}(L_1 + L_2)(\bar{\mathbf{x}}, \bar{\mathbf{u}}), \quad i = 0, 1, \dots, N. \quad (4.12)$$

$$0 \in \partial_{u_i}(L_1 + L_2)(\bar{\mathbf{x}}, \bar{\mathbf{u}}) + N(\bar{u}_i \mid U_i), \quad i = 0, 1, \dots, N-1. \quad (4.13)$$

(The equivalence follows from the fact that, by the convexity of the problem, we

can use the Moreau–Rockafellar theorem [7, Theorem 23.8]. Consequently, the inclusions such as [9, (3.10)] can be replaced by equalities.) Further, observe that

$$\partial_{x_i}(\delta(\cdot | A))(\bar{x}, \bar{u}) = \{0\}, \quad i = 0, 1, \dots, N, \tag{4.14}$$

$$\partial_{u_i}(\delta(\cdot | A))(\bar{x}, \bar{u}) = \partial(\delta(\cdot | U_i))(\bar{u}_i) = N(\bar{u}_i | U_i), \quad i = 0, 1, \dots, N - 1. \tag{4.15}$$

It follows from (4.2), (4.12)–(4.15) and the Moreau–Rockafellar theorem that

$$\begin{aligned} 0 &\in \partial_{x_i} L(\bar{x}, \bar{u}), \quad i = 0, 1, \dots, N, \\ 0 &\in \partial_{u_i} L(\bar{x}, \bar{u}), \quad i = 0, 1, \dots, N - 1. \end{aligned}$$

This means, by (1.2), that the right-hand side of (4.11) is always nonnegative, and so, (4.7) holds. ■

Finally, we shall specify several simple conditions which ensure that assumption (4.1) is fulfilled.

PROPOSITION 4.2. Let $f : \mathbb{R}^n \times \mathbb{R}^r \rightarrow \mathbb{R}$ be a convex function, and let $(\bar{x}, \bar{u}) \in \mathbb{R}^n \times \mathbb{R}^r$. Suppose that one of the following conditions holds:

- (a) f is Gâteaux differentiable at (\bar{x}, \bar{u}) ;
- (b) for each $u \in \mathbb{R}^r$, the function $f'_u(\cdot, \bar{u}; u)$ is lower semicontinuous at \bar{x} ;
- (c) for each $x \in \mathbb{R}^n$, the function $f'_x(\bar{x}, \cdot; x)$ is lower semicontinuous at \bar{u} ;
- (d) $f(x, u) = g(x) + h(u)$ where g and h are convex functions on \mathbb{R}^n and \mathbb{R}^r , respectively.

Then

$$f'((\bar{x}, \bar{u}); (x, u)) \geq f'_x(\bar{x}, \bar{u}; x) + f'_u(\bar{x}, \bar{u}; u) \quad \text{for all } (x, u) \in \mathbb{R}^n \times \mathbb{R}^r.$$

Proof. Since cases (a) and (d) are obvious, and cases (c) is analogous to (b), we shall only prove case (b).

For each $(x, u) \in \mathbb{R}^n \times \mathbb{R}^r$ and each $\lambda > 0$ we have, by (1.4),

$$\begin{aligned} &(f(\bar{x} + \lambda x, \bar{u} + \lambda u) - f(\bar{x}, \bar{u}))/\lambda \\ &= (f(\bar{x} + \lambda x, \bar{u} + \lambda u) - f(\bar{x} + \lambda x, \bar{u}))/\lambda + (f(\bar{x} + \lambda x, \bar{u}) - f(\bar{x}, \bar{u}))/\lambda \tag{4.16} \\ &\geq f'_u(\bar{x} + \lambda x, \bar{u}; u) + f'_x(\bar{x}, \bar{u}; x). \end{aligned}$$

But the lower semicontinuity assumption implies

$$\liminf_{\lambda \rightarrow 0+} f'_u(\bar{x} + \lambda x, \bar{u}; u) \geq f'_u(\bar{x}, \bar{u}; u).$$

Hence, taking the upper limit of both sides of (4.16) as $\lambda \rightarrow 0+$, we get the desired inequality. ■

Remark 4.3. Sufficient optimality conditions for nonconvex discrete-time control systems under separability assumptions similar to that of Proposition 4.2(d) were considered in [10].

References

1. *Boltianskii, V. G.*, Optimal Control of Discrete Systems (in Russian). Nauka, Moskva, 1973 (English translation: Wiley, New York, 1978).
2. *Doležal, J.*, Non-smooth and non-convex problems in discrete optimal control. *Int. J. Systems Sci.* **13** (1982), pp. 969–978.
3. *Doležal, J.*, Necessary conditions for Pareto optimality in nondifferentiable discrete control problems. *Control Cybernet.* **17** (1988), pp. 213–223.
4. *Morduhovič, B. Š.*, Approximation Methods in Optimization and Control Problems (in Russian). Nauka, Moskva, 1988.
5. *Pytlak, R., Malinowski, K.*, Optimality conditions for nondifferentiable problems of discrete-time control (in Polish). *Arch. Automat. Telemekh.* **30** (1985), pp. 169–191.
6. *Pytlak, R.*, Discrete maximum principle for problems with lipschitzian functions. *Int. J. Control* **48** (1988), pp. 641–654.
7. *Rockafellar, R. T.*, Convex Analysis. Princeton University Press, 1970.
8. *Studniarski, M.*, Application of the Dubovitskii–Milyutin method to some locally convex extremal problems. *Bull. Soc. Sci. Lett. Łódź* **29** (1979), No. 6, pp. 1–8.
9. *Studniarski, M.*, Necessary optimality conditions for a nonsmooth discrete control problem. *Control Cybernet.* **11** (1982), pp. 109–119.
10. *Vidal, R. V. V.*, On the sufficiency of the linear maximum principle for discrete-time control problems. *J. Optimization Theory Appl.* **54** (1987), pp. 583–589.
11. *Vinter, R. B.*, Optimality and sensitivity of discrete time processes. *Control Cybernet.* **17** (1988), pp. 191–211.

Дискретный принцип максимума как необходимое условие оптимальности

М. СТУДНЯРСКИ

(Лодзь)

В работе представлен пример, который показывает, что принцип максимума в субдифференциальной форме для выпуклых недифференцируемых проблем управления с дискретным временем не всегда является достаточным условием оптимальности. Показано также, что это желаемое свойство достаточности может быть установлено при некотором дополнительном предположении о целевой функции.

Marcin Studniarski
Institute of Mathematics
University of Łódź
ul. S. Banacha 22
90–238 Łódź, Poland

STABILITY AND SENSITIVITY ANALYSIS OF DISCRETE OPTIMAL CONTROL PROBLEMS

K. MALANOWSKI

(*Warsaw*)

(Received November 1, 1990)

A family of discrete optimal control problem subject to state and control constraints is considered. All data of the problems depend on a parameter. Using the known sensitivity and stability results for mathematical programming problems, the conditions are formulated under which the solutions to optimal control problems are Lipschitz continuous and directionally differentiable functions of the parameter. The directional derivatives are characterized as the solutions to auxiliary quadratic optimal control problems.

1. Introduction

Mathematical models of numerous dynamic systems are built in discrete form, using difference equations. Many technological, economic, social or biological systems are discrete by their very nature. On the other hand, discrete models are often used for continuous control processes. It takes place, for example, if process measurements are performed or control action is executed at some sampled moments and we restrict ourselves to the analysis of system behaviour at these moments only. Such situations occur, almost as a rule, in computerized on-line control. Hence, control of discrete systems, including optimal control, has an important practical significance.

Usually we do not know the exact values of parameters of control systems, or these values are subject to perturbations. Therefore, it is important to know how the calculated control depend on the parameters of the model.

In optimal control, like in other optimization problems, we are interested in the stability and sensitivity of obtained solutions, i.e., in their continuity and differentiability with respect to parameters of the system. These properties can be investigated either for the optimal controls or for the so-called optimal value function, which to every value of the parameter assigns the optimal value of the cost functional.

This paper is devoted to the sensitivity and stability analysis of optimal control of discrete systems. This analysis is of importance not only for discrete but

also for continuous optimal control problems, since they can be approximated by discrete ones (see [4]).

It is well known (see [2]) that discrete optimal control problems can be reformulated as mathematical programs with a specific structure.

Accordingly, in our analysis, we shall use the known stability and sensitivity results for mathematical programming problems and specialize them for optimal control.

We are going to characterize a class of discrete optimal control problems, depending on a parameter, that have locally isolated local solutions, which are locally Lipschitz continuous and directionally differentiable functions of the parameter.

It is known (see [8, 11]) that mathematical programming problems possess the above properties if the linear independence and the strong second order sufficient conditions are satisfied, whereas the strict complementarity is not required. If additionally the strict complementarity holds, then the solution becomes a Fréchet differentiable function of the parameters [5].

Therefore, the main part of the paper (Sections 3 and 4) is devoted to the analysis of linear independence and strong second order sufficient conditions for discrete optimal control problems.

In Section 5 the principal results is formulated. In particular, the directional derivative of the optimal control is characterized as a solution to an auxiliary optimal control problem.

It seems that the above results are not only interesting from theoretical point of view, but, like in case of mathematical programming, they may find practical applications in the stability and sensitivity methodology approach to optimal control problems (see [5]). Some notation:

\mathbb{R}^n is an n -dimensional Euclidean space with the inner product denoted by $\langle \cdot, \cdot \rangle$ and the norm

$$|\mathbf{x}| = \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}}.$$

If $\mathbf{f} : H \rightarrow \mathbb{R}^k$, where H is a Banach space, is sufficiently regular, then

$$\partial_h \mathbf{f}(h, g) = \lim_{\alpha \downarrow 0} \frac{1}{\alpha} [\mathbf{f}(h + \alpha g) - \mathbf{f}(h)]$$

denotes the directional derivative of \mathbf{f} at h in the direction g .

If $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^k$ is sufficiently regular, then $D_{\mathbf{x}} \mathbf{f}(\mathbf{x}, \mathbf{u})$, $D_{\mathbf{x}\mathbf{u}}^2 \mathbf{f}(\mathbf{x}, \mathbf{u})$ denote the first and the second Fréchet derivatives, with respect to the appropriate variable.

Superscript T denotes transposition, c is a generic constant, not necessarily the same in two different places.

2. Problem statement and preliminary results

Let H denote an open set in a Banach space, which will be called the set of feasible parameters. Let $\bar{h} \in H$ be a fixed value of the parameter.

For each h belonging to a neighbourhood $G \subset H$ of \bar{h} we consider the following state and control constrained discrete optimal control problem:

(O_h) minimize

$$\Phi(\mathbf{x}, \mathbf{u}, h) := \sum_{i=0}^{N-1} \phi_i(\mathbf{x}_i, \mathbf{u}_i, h) + \psi(\mathbf{x}_N, h) \quad (2.1)$$

subject to

$$\mathbf{x}_{i+1} - \mathbf{x}_i = \mathbf{f}_i(\mathbf{x}_i, \mathbf{u}_i, h), \quad i = 0, \dots, N-1 \quad (2.2)$$

$$\mathbf{x}_0 = \mathbf{t}(h), \quad (2.2a)$$

$$\theta_i^j(\mathbf{u}_i, h) \leq 0, \quad i = 0, 1, \dots, N-1, \quad j = 1, 2, \dots, k, \quad (2.3)$$

$$\chi_i^j(\mathbf{x}_i, h) \leq 0, \quad i = 0, 1, \dots, N-1, N, \quad j = 1, 2, \dots, l. \quad (2.4)$$

where

$$\mathbf{x}^T = [\mathbf{x}_0^T, \mathbf{x}_1^T, \dots, \mathbf{x}_N^T] \in \mathbf{R}^{n(N+1)},$$

$$\mathbf{u}^T = [\mathbf{u}_0^T, \mathbf{u}_1^T, \dots, \mathbf{u}_{N-1}^T] \in \mathbf{R}^{mN},$$

$$\phi_i(\cdot, \cdot, \cdot) : \mathbf{R}^n \times \mathbf{R}^m \times G \rightarrow \mathbf{R}^1, \quad i = 0, 1, \dots, N-1,$$

$$\psi(\cdot, \cdot) : \mathbf{R}^n \times G \rightarrow \mathbf{R}^1,$$

$$\mathbf{f}_i(\cdot, \cdot, \cdot) : \mathbf{R}^n \times \mathbf{R}^m \times G \rightarrow \mathbf{R}^n, \quad i = 0, 1, \dots, N-1.$$

$$\mathbf{t}(\cdot) : G \rightarrow \mathbf{R}^n,$$

$$\theta_i^j(\cdot, \cdot) : \mathbf{R}^m \times G \rightarrow \mathbf{R}^1, \quad i = 0, 1, \dots, N-1, \quad j \in K := \{1, 2, \dots, k\},$$

$$\chi_i^j(\cdot, \cdot) : \mathbf{R}^n \times G \rightarrow \mathbf{R}^1, \quad i = 0, 1, \dots, N, \quad j \in L := \{1, 2, \dots, l\}.$$

Denote

$$\theta_i(\mathbf{u}_i, h) = [\theta_i^1(\mathbf{u}_i, h), \theta_i^2(\mathbf{u}_i, h), \dots, \theta_i^k(\mathbf{u}_i, h)]^T,$$

$$\chi_i(\mathbf{x}_i, h) = [\chi_i^1(\mathbf{x}_i, h), \chi_i^2(\mathbf{x}_i, h), \dots, \chi_i^l(\mathbf{x}_i, h)]^T,$$

$$\theta^T(\mathbf{u}, h) = [\theta_0^T(\mathbf{u}_0, h), \theta_1^T(\mathbf{u}_1, h), \dots, \theta_{N-1}^T(\mathbf{u}_{N-1}, h)],$$

$$\chi^T(\mathbf{x}, h) = [\chi_0^T(\mathbf{x}_0, h), \chi_1^T(\mathbf{x}_1, h), \dots, \chi_N^T(\mathbf{x}_N, h)].$$

It is well known (see e.g. [2]) that (O_h) is equivalent to a mathematical programming problem with a specific structure. Indeed, putting $\alpha^T = [\mathbf{x}_0^T, \mathbf{u}_0^T, \mathbf{x}_1^T, \mathbf{u}_1^T, \dots, \mathbf{x}_{N-1}^T, \mathbf{u}_{N-1}^T, \mathbf{x}_N^T] \in X := \mathbf{R}^{n(N+1)+mN}$ we can reformulate (O_h) as follows:

$$(\tilde{O}_h) \text{ minimize } \tilde{\Phi}(\alpha, h) \quad (2.5)$$

subject to

$$\mathbf{r}(\alpha, h) = \mathbf{0}, \quad (2.6)$$

$$\mathbf{s}(\alpha, h) \leq \mathbf{0}, \quad (2.7)$$

where

$$\begin{aligned} \tilde{\Phi}(\cdot, \cdot) &: X \times G \rightarrow \mathbf{R}^1, \\ \tilde{\Phi}(\alpha, h) &:= \Phi(\mathbf{x}, \mathbf{u}, h). \\ \mathbf{r}(\cdot, \cdot) &: X \times G \rightarrow \mathbf{R}^{n(N+1)} \quad \text{corresponds to state equation (2.2)} \\ &\quad \text{and initial condition (2.2a),} \\ \mathbf{s}(\cdot, \cdot) &: X \times G \rightarrow \mathbf{R}^{kN+l(N+1)} \quad \text{corresponds to control} \\ &\quad \text{and state constraints (2.3), (2.4).} \end{aligned}$$

Let us assume that for each $h \in G$ the feasible set of (O_h) is nonempty and that there exists a local minimizer $(\mathbf{x}(h), \mathbf{u}(h))$. We are interested in the stability and sensitivity analysis of $(\mathbf{x}(\cdot), \mathbf{u}(\cdot))$ in a neighbourhood of \bar{h} , i.e. in the continuity and differentiability properties of $(\mathbf{x}(\cdot), \mathbf{u}(\cdot))$ treated as a function of the parameter h .

In our analysis we shall use known sensitivity and stability results for mathematical programming problems obtained in [8, 11]. Using the specific structure of the functions $\tilde{\Phi}(\cdot, \cdot)$, $\mathbf{r}(\cdot, \cdot)$ and $\mathbf{s}(\cdot, \cdot)$ in (\bar{O}_h) we shall reformulate these results in terms of the original data of (O_h) .

We start with recalling some definitions. Let

$$I = \{1, \dots, n(N + 1)\},$$

and

$$J = \{1, \dots, kN + l(N + 1)\}$$

be the sets of indices of equality and inequality constraints, respectively.

Let $\alpha(h)$ be a local solution to (\bar{O}_h) . Denote by

$$J_h = \{j \in J \mid s^j(\alpha(h), h) = 0\}$$

the set of indices of all inequality type constraints active at h .

DEFINITION 2.1. Let assume that $\mathbf{r}(\cdot, h)$ and $\mathbf{s}(\cdot, h)$ are of class C^1 . We say that the linear independence condition (LI) is satisfied at $\alpha(h)$ if the gradients of all constraints active at $\alpha(h)$ are linearly independent.

It is well known (see [7] and also [1]) that (LI) is equivalent to the condition that the mapping

$$\begin{aligned} T(\alpha(h), h) &: \mathbf{R}^{n(N+1)+mN} \times \mathbf{R}^{kN+l(N+1)} \rightarrow \mathbf{R}^{kN+l(N+1)} \\ T(\alpha(h), h) &:= \begin{bmatrix} D_\alpha \mathbf{r}(\alpha(h), h), & 0 \\ D_\alpha \mathbf{s}(\alpha(h), h), & S(\alpha(h), h) \end{bmatrix} \end{aligned} \tag{2.8}$$

is surjective.

Here $S(\alpha(h), h)$ denotes a $[kN + l(N + 1)] \times [kN + l(N + 1)]$ dimensional diagonal matrix with $s^j(\alpha(h), h)$ as diagonal elements.

Let us introduce the Lagrangean associated with (\tilde{O}_h) :

$$\begin{aligned} \tilde{\mathcal{L}}(\cdot, \cdot, \cdot, \cdot) : X \times \mathbb{R}^{n(N+1)} \times \mathbb{R}^{kN+l(N+1)} \times G \rightarrow \mathbb{R}^1, \\ \tilde{\mathcal{L}}(\alpha, \xi, \zeta, h) := \tilde{\Phi}(\alpha, h) + \langle \xi, r(\alpha, h) \rangle + \langle \zeta, s(\alpha, h) \rangle, \end{aligned}$$

where $\xi \in \mathbb{R}^{n(N+1)}$ and $\zeta \in \mathbb{R}^{kN+l(N+1)}$ are Lagrange multipliers associated with (2.6) and (2.7), respectively.

It is well known (see e.g. [2]) that if the functions $\tilde{\Phi}(\cdot, h)$, $r(\cdot, h)$ and $s(\cdot, h)$ are C^1 in a neighbourhood of $\alpha(h)$ and linear independence condition hold at $\alpha(h)$, then there exist unique Lagrange multipliers $\xi(h)$ and $\zeta(h)$, such that the following Kuhn-Tucker conditions hold:

$$\begin{aligned} D_\alpha \tilde{\mathcal{L}}(\alpha(h), \xi(h), \zeta(h), h) = 0, \\ \langle \zeta(h), s(\alpha(h), h) \rangle = 0, \quad \zeta^j(h) \geq 0. \end{aligned} \quad (2.9)$$

DEFINITION 2.2. Let $\tilde{\Phi}(\cdot, h)$, $r(\cdot, h)$ and $s(\cdot, h)$ be C^2 in a neighbourhood of $\alpha(h)$. We say that the strong second order sufficient conditions (SC) is satisfied at $\alpha(h)$, if

$$\langle \beta, D_\alpha^2 \tilde{\mathcal{L}}(\alpha(h), \xi(h), \zeta(h), h) \beta \rangle > 0$$

for all non-zero

$$\begin{aligned} \beta \in \{ \beta \in X \mid \langle \beta, D_\alpha r^i(\alpha(h), h) \rangle = 0, \quad i \in I, \\ \langle \beta, D_\alpha s^j(\alpha(h), h) \rangle = 0, \quad j \in J_h^c \} \end{aligned}$$

where

$$J_h^c = \{ j \in J_h \mid \zeta^j(h) > 0 \}.$$

We can formulate now the stability and sensitivity results obtained in [8, 11]. For our purpose they will take on the form:

THEOREM 2.3. Assume that the functions $\tilde{\Phi}(\cdot, \cdot)$, $r(\cdot, \cdot)$ and $s(\cdot, \cdot)$ are C^2 in a neighbourhood of $\alpha(\bar{h}, \bar{h})$. Moreover, (LI) and (SC) hold at $\alpha(\bar{h})$. Then there exist neighbourhoods \bar{G} of \bar{h} in H and \bar{C} of $\alpha(\bar{h})$ in X , such that each $h \in \bar{G}$, $\alpha(h)$ is the unique minimizer of (\tilde{O}_h) in \bar{C} and $(\xi(h), \zeta(h))$ are the unique associated Lagrange multipliers. The functions $\alpha(\cdot)$, $\xi(\cdot)$, $\zeta(\cdot)$ are Lipschitz continuous on \bar{G} and directionally differentiable at \bar{h} . The directional derivatives $\partial_h \alpha(\bar{h}, g)$, $\partial_h \xi(\bar{h}, g)$, $\partial_h \zeta(\bar{h}, g)$ are given by the solution and the associated Lagrange multipliers of the following quadratic programming problem.

$(\bar{QO}_{\bar{h}, g})$ minimize

$$\frac{1}{2} \langle \beta, D_\alpha^2 \tilde{\mathcal{L}}[\bar{h}] \beta \rangle + \langle \beta, D_\alpha \tilde{\mathcal{L}}[\bar{h}] g \rangle \quad (2.10)$$

subject to

$$D_\alpha r^i(\alpha(\bar{h}), \bar{h}) \beta + D_h r^i(\alpha(\bar{h}), \bar{h}) g = 0 \quad i \in I, \quad (2.11)$$

$$D_{\alpha} s^j(\alpha(\bar{h}), \bar{h})\beta + D_h s^j(\alpha(\bar{h}), \bar{h})g \begin{cases} = 0, & \text{for } j \in J_h^c \\ \leq 0, & \text{for } j \in J_h \setminus J_h^c, \end{cases} \quad (2.12)$$

where

$$\tilde{\mathcal{L}}[\bar{h}] := \tilde{\mathcal{L}}(\alpha(\bar{h}), \xi(\bar{h}), \zeta(\bar{h}), h).$$

Remark 2.4. Note that by the strong second order sufficient condition, $(\widetilde{QO}_{\bar{h},g})$ has a unique solution, whereas by the linear independence condition the associated Lagrange multipliers are unique.

Remark 2.5. The result proved in [11] is stronger than in Theorem 2.3. Namely, it is shown there that $\mathbf{y}(\cdot)$ is Bouligand differentiable at \bar{h} , which is stronger than directional differentiability. However, to avoid technical definitions we restricted ourselves to directional derivatives.

We are going to apply Theorem 2.3 to discrete optimal control problem (O_h) . In order to do that we have to express all assumptions of Theorem 3.1 in terms of the original data of (O_h) .

It is obvious that the condition of regularity is satisfied if all involved functions are of class C^2 in a neighbourhood of $(\mathbf{x}(\bar{h}), \mathbf{u}(\bar{h}), \bar{h})$. Hence, it remains to analyse linear independence and strong second order sufficient conditions.

3. Linear independence

To simplify notation we put

$$A_i(h) := [D_{\mathbf{x}} \mathbf{f}_i(\mathbf{x}_i(h), \mathbf{u}_i(h), h)], \quad (3.1a)$$

$$B_i(h) := [D_{\mathbf{u}} \mathbf{f}_i(\mathbf{x}_i(h), \mathbf{u}_i(h), h)], \quad (3.1b)$$

$$C_i(h) := [D_h \mathbf{f}_i(\mathbf{x}_i(h), \mathbf{u}_i(h), h)], \quad (3.1c)$$

$$\Omega_i(h) := [D_{\mathbf{u}} \boldsymbol{\theta}_i(\mathbf{u}_i(h), h)], \quad (3.1d)$$

$$\Lambda_i(h) := [D_{\mathbf{x}} \boldsymbol{\chi}_i(\mathbf{x}_i(h), h)]. \quad (3.1e)$$

Let us introduce the following subsets of the active constraints indices:

$$K_i(h) := \{j \in K \mid \theta_i^j(\mathbf{u}_i(h), h) = 0\}, \quad i = 0, 1, \dots, N-1,$$

$$L_i(h) := \{j \in L \mid \chi_i^j(\mathbf{x}_i(h), h) = 0\}, \quad i = 0, 1, \dots, N.$$

It turns out that thanks to the structure of (O_h) the linear independence condition (LI) can be checked independently on each stage $i = 0, 1, \dots, N-1$, as it is formulated in the following

PROPOSITION 3.1. Assume that

$$\chi_0^j(t(h), h) < 0, \quad j \in L. \quad (3.2)$$

Then the linear independence condition (LI) for (O_h) is satisfied at $(\mathbf{x}(h), \mathbf{u}(h))$ if and only if the following mappings

$$\begin{aligned} \Gamma_i(h) : \mathbb{R}^{m+k+1} &\rightarrow \mathbb{R}^{k+1} & i = 0, 1, \dots, N-1 \\ \Gamma_i(h) &:= \begin{bmatrix} \Omega_i(h) & \theta_i(h) & 0 \\ \Lambda_{i+1}(h)B_i(h) & 0 & X_{i+1}(h) \end{bmatrix} \end{aligned} \quad (3.3)$$

are surjective.

Here $\theta_i(h)$ and $X_i(h)$ are $(k \times k)$ and $(l \times l)$ -diagonal matrices, whose diagonal elements are $\theta_i^j(u_i(h), h)$ and $\chi_i^j(u_i(h), h)$, respectively.

Proof. It follows from (2.8) that for discrete optimal control problem (O_h) the linear independence condition is satisfied if and only if the system of equations

$$\mathbf{y}_{i+1} - \mathbf{y}_i - A_i(h)\mathbf{y}_i - B_i(h)\mathbf{v}_i = \mathbf{a}_i \quad (3.4a)$$

$$\mathbf{y}_0 = \mathbf{b}_0 \quad (3.4b)$$

$$\Omega_i(h)\mathbf{v}_i + \theta_i(h)\boldsymbol{\mu}_i = \mathbf{c}_i \quad (3.4c)$$

$$\Lambda_i(h)\mathbf{y}_i + X_i(h)\boldsymbol{\nu}_i = \mathbf{d}_i \quad i = 0, 1, \dots, N-1 \quad (3.4d)$$

has a solution for arbitrary $\mathbf{a}_i \in \mathbb{R}^n$, $\mathbf{b}_0 \in \mathbb{R}^h$, $\mathbf{c}_i \in \mathbb{R}^k$, $\mathbf{d}_i \in \mathbb{R}^l$. From (3.4d) we have

$$\Lambda_0(h)\mathbf{y}_0 + X_0(h)\boldsymbol{\nu}_0 = \mathbf{d}_0 \quad (3.5)$$

and

$$\Lambda_{i+1}(h)\mathbf{y}_{i+1} - \Lambda_i(h)\mathbf{y}_i + X_{i+1}(h)\boldsymbol{\nu}_{i+1} - X_i(h)\boldsymbol{\nu}_i = \mathbf{d}_{i+1} - \mathbf{d}_i$$

or

$$\begin{aligned} &\Lambda_{i+1}(h)[\mathbf{y}_{i+1} - \mathbf{y}_i] + X_{i+1}(h)(\boldsymbol{\nu}_{i+1} - \boldsymbol{\nu}_i) = \\ &= (\mathbf{d}_{i+1} - \mathbf{d}_i) + [\Lambda_i(h) - \Lambda_{i+1}(h)]\mathbf{y}_i + [X_i(h) - X_{i+1}(h)]\boldsymbol{\nu}_i, \end{aligned} \quad (3.6)$$

$$i = 1, 2, \dots, N-1.$$

Multiplying (3.4a) by $\Lambda_{i+1}(h)$ and subtracting from (3.6) we obtain

$$\begin{aligned} &\Lambda_{i+1}(h)B_i(h)\mathbf{v}_i + X_{i+1}(h)(\boldsymbol{\nu}_{i+1} - \boldsymbol{\nu}_i) = \\ &= [\Lambda_i(h) - \Lambda_{i+1}(h) - \Lambda_{i+1}(h)A_i(h)]\mathbf{y}_i + [X_i(h) - X_{i+1}(h)]\boldsymbol{\nu}_i + \\ &\quad + [\mathbf{d}_{i+1} - \mathbf{d}_i - \Lambda_{i+1}(h)\mathbf{a}_i] = \\ &= \Delta_i^1(h)\mathbf{y}_i + \Delta_i^2(h)\boldsymbol{\nu}_i + \bar{\mathbf{d}}_i, \end{aligned} \quad (3.7)$$

where

$$\begin{aligned} \Delta_i^1(h) &:= \Lambda_{i+1}(h) - \Lambda_{i+1}(h) - \Lambda_{i+1}(h)A_i(h), \\ \Delta_i^2(h) &:= X_i(h) - X_{i+1}(h), \\ \bar{\mathbf{d}}_i &:= \mathbf{d}_{i+1} - \mathbf{d}_i - \Lambda_{i+1}(h)\mathbf{a}_i. \end{aligned}$$

Equations (3.4c) and (3.7) can be rewritten in the form

$$\begin{aligned} \Gamma_i(h) \begin{bmatrix} \mathbf{v}_i \\ \mathbf{y}_i \\ \boldsymbol{\nu}_{i+1} - \boldsymbol{\nu}_i \end{bmatrix} &:= \\ &:= \begin{bmatrix} \Omega_i(h) & \theta_i(h) & 0 \\ \Lambda_{i+1}(h)B_i(h) & 0 & X_{i+1}(h) \end{bmatrix} \begin{bmatrix} \mathbf{v}_i \\ \mathbf{y}_i \\ \boldsymbol{\nu}_{i+1} - \boldsymbol{\nu}_i \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{c}_i \\ \Delta_1^1(h)\mathbf{y}_i + \Delta_2^2\boldsymbol{\nu}_i + \bar{\mathbf{d}}_i \end{bmatrix}. \end{aligned} \tag{3.8}$$

It is easy to see that (3.8) has a solution for any right-hand side, i.e., for any $(\mathbf{c}_i, \mathbf{d}_i)$, if and only if $\Gamma_i(h)$ is surjective. In this case (3.8) is satisfied if we put

$$\begin{aligned} \begin{bmatrix} \mathbf{v}_i \\ \boldsymbol{\mu}_i \\ \boldsymbol{\nu}_{i+1} - \boldsymbol{\nu}_i \end{bmatrix} &= \Gamma_i^T(h)(\Gamma_i(h)\Gamma_i^T(h))^{-1} \begin{bmatrix} \mathbf{c}_i \\ \Delta_1^1(h)\mathbf{y}_i + \Delta_2^2(h)\boldsymbol{\nu}_i + \bar{\mathbf{d}}_i \end{bmatrix} := \\ &:= \begin{bmatrix} \mathbf{e}_i^1(h) + e_i^{11}(h)\mathbf{y}_i + e_i^{12}(h)\boldsymbol{\nu}_i \\ \mathbf{e}_i^2(h) + e_i^{21}(h)\mathbf{y}_i + e_i^{22}(h)\boldsymbol{\nu}_i \\ \mathbf{e}_i^3(h) + e_i^{31}(h)\mathbf{y}_i + e_i^{32}(h)\boldsymbol{\nu}_i \end{bmatrix}. \end{aligned} \tag{3.9}$$

From (3.4a) and (3.9) we obtain

$$\begin{aligned} \mathbf{y}_{i+1} - \mathbf{y}_i - [A_i(h) + B_i(h)e_i^{11}(h)]\mathbf{y}_i - B_i(h)e_i^{12}(h)\boldsymbol{\nu}_i &= \mathbf{a}_i + B_i(h)e_i^1(h) := \bar{\mathbf{a}}_i \\ \boldsymbol{\nu}_{i+1} - \boldsymbol{\nu}_i - e_i^{31}(h)\mathbf{y}_i - e_i^{32}(h)\boldsymbol{\nu}_i &= \mathbf{e}_i^3(h) := \bar{\mathbf{e}}_i. \end{aligned} \tag{3.10}$$

On the other hand, (3.4b), together with (3.5) and (3.2), yields

$$\begin{aligned} \mathbf{y}_0 &= \mathbf{b}_0, \\ \boldsymbol{\nu}_0 &= X_0^{-1}(h)[\mathbf{d}_0 - \Lambda_0(h)\mathbf{b}_0] := \bar{\mathbf{d}}_0. \end{aligned} \tag{3.10a}$$

It is obvious that (3.10) has a unique solution $(\mathbf{y}_i, \boldsymbol{\nu}_i)$ for any $\bar{\mathbf{a}}_i, \bar{\mathbf{e}}_i, \bar{\mathbf{d}}_0$, i.e. for any $\mathbf{a}_i, \mathbf{b}_0, \mathbf{c}_i, \mathbf{d}_i$. Having $\mathbf{y}_i, \boldsymbol{\nu}_i$ we find \mathbf{v}_i and $\boldsymbol{\mu}_i$ from (3.9). \square

Remark 3.2. Condition (3.3) is a discrete analogon of Hager's constraint qualifications for continuous optimal control problems [6]. It was first introduced in [10], however, the proof given there is completely different from the proof of Proposition 3.1.

4. Second order sufficient condition

In this section we are going to analyse the second order sufficient condition for (O_h) . Like in case of (\tilde{O}_h) let us introduce the Lagrangean

$$\mathcal{L}(\cdot, \cdot, \cdot, \cdot, \cdot, \cdot) : \mathbf{R}^{n(N+1)} \times \mathbf{R}^{mN} \times \mathbf{R}^n \times \mathbf{R}^{nN} \times \mathbf{R}^{kN} \times \mathbf{R}^{l(N+1)} \times G \rightarrow \mathbf{R}^1$$

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \mathbf{u}, \boldsymbol{\rho}, \boldsymbol{\lambda}, \boldsymbol{\mu}, h) := & \phi(\mathbf{x}, \mathbf{u}, h) + \langle \boldsymbol{\rho}, \mathbf{x}_0 - \mathbf{t}(h) \rangle + \\ & + \sum_{i=0}^{N-1} \langle \mathbf{p}_i, \mathbf{x}_{i+1} - \mathbf{x}_i - \mathbf{f}_i(\mathbf{x}_i, \mathbf{u}_i, h) \rangle + \\ & + \sum_{i=0}^{N-1} \langle \boldsymbol{\lambda}_i, \boldsymbol{\theta}_i(\mathbf{u}_i, h) \rangle + \sum_{i=0}^N \langle \boldsymbol{\mu}_i, \boldsymbol{\chi}_i(\mathbf{x}_i, h) \rangle. \end{aligned} \quad (4.1)$$

Let us assume that the linear independence condition (LI) holds for (O_h) i.e., (3.2) and (3.5) are satisfied. Then, in particular, the Lagrange multipliers $\boldsymbol{\rho}(h)$, $\mathbf{p}(h)$, $\boldsymbol{\lambda}(h)$ and $\boldsymbol{\mu}(h)$ are defined uniquely.

For the sake of simplicity we denote

$$P_i(h) = D_{\mathbf{x}, \mathbf{x}_i}^2 \mathcal{L}(\mathbf{x}(h), \mathbf{u}(h), \boldsymbol{\rho}(h), \mathbf{p}(h), \boldsymbol{\lambda}(h), \boldsymbol{\mu}(h), h), \quad (4.2a)$$

$$Q_i(h) = D_{\mathbf{x}, \mathbf{u}_i}^2 \mathcal{L}(\mathbf{x}(h), \mathbf{u}(h), \boldsymbol{\rho}(h), \mathbf{p}(h), \boldsymbol{\lambda}(h), \boldsymbol{\mu}(h), h), \quad (4.2b)$$

$$R_i(h) = D_{\mathbf{u}, \mathbf{u}_i}^2 \mathcal{L}(\mathbf{x}(h), \mathbf{u}(h), \boldsymbol{\rho}(h), \mathbf{p}(h), \boldsymbol{\lambda}(h), \boldsymbol{\mu}(h), h). \quad (4.2c)$$

It follows from Definition 2.2 that the strong second order sufficient condition (SC) for (O_h) amounts to

$$\sum_{i=0}^{N-1} [\langle \mathbf{z}_i, P_i(h)\mathbf{z}_i \rangle + 2\langle \mathbf{z}_i, Q_i(h)\mathbf{w}_i \rangle + \langle \mathbf{w}_i, R_i(h)\mathbf{w}_i \rangle] + \langle \mathbf{z}_N, P_N(h)\mathbf{z}_N \rangle > 0 \quad (4.3)$$

for all $(\mathbf{z}, \mathbf{w}) \neq 0$ such that

$$\mathbf{z}_{i+1} - \mathbf{z}_i = A_i(h)\mathbf{z}_i + B_i(h)\mathbf{w}_i, \quad i = 0, 1, \dots, N-1 \quad (4.4)$$

$$\mathbf{z}_0 = \mathbf{0}, \quad (4.4a)$$

$$\langle D_{\mathbf{u}} \theta_i^j(\mathbf{u}_i(h), h), \mathbf{w}_i \rangle = 0, \quad i = 0, 1, \dots, N-1, \quad j \in K_i^c(h) \quad (4.5a)$$

$$\langle D_{\mathbf{x}} \chi_i^j(\mathbf{x}_i(h), h), \mathbf{z}_i \rangle = 0, \quad i = 0, 1, \dots, N, \quad j \in L_i^c(h) \quad (4.5b)$$

where $A_i(h)$, $B_i(h)$ are given in (3.1) and

$$K_i^c(h) = \{j \in K_i \mid \lambda_i^j(h) > 0\}, \quad (4.6a)$$

$$L_i^c(h) = \{j \in L_i \mid \mu_i^j(h) > 0\}. \quad (4.6b)$$

LEMMA 4.1. Suppose that (3.2) and (3.5) hold. Then (SC) is satisfied if and only if the following quadratic optimal control problem:

$(QC_{h,a})$ minimize

$$\sum_{i=0}^{N-1} [\langle \mathbf{z}_i, P_i(h)\mathbf{z}_i \rangle + 2\langle \mathbf{z}_i, Q_i(h)\mathbf{w}_i \rangle + \langle \mathbf{w}_i, R_i(h)\mathbf{w}_i \rangle] + \langle \mathbf{z}_N, P_N(h)\mathbf{z}_N \rangle \quad (4.7)$$

subject to

$$\mathbf{z}_{i+1} - \mathbf{z}_i = A_i(h)\mathbf{z}_i + B_i(h)\mathbf{w}_i \tag{4.8}$$

$$\mathbf{z}_0 = \mathbf{a}, \tag{4.8a}$$

$$\langle D_{\mathbf{u}}\theta_i^j(\mathbf{u}_i(h), h), \mathbf{w}_i \rangle = 0, \quad i = 0, 1, \dots, N - 1, \quad j \in K_i^c(h), \tag{4.9a}$$

$$\langle D_{\mathbf{x}}\chi_i^j(\mathbf{x}_i(h), h), \mathbf{z}_i \rangle = 0, \quad i = 0, 1, \dots, N, \quad j \in L_i^c(h) \tag{4.9b}$$

has a unique solution for any $\mathbf{a} \in \mathbb{R}^n$.

Proof. First, let us consider the homogeneous case $\mathbf{a} = \mathbf{0}$. Problem $(QC_{h,\mathbf{0}})$ has a solution if and only if the cost functional is non-negative for all feasible controls. Indeed, suppose that there exists a feasible \mathbf{w} for which the cost functional is negative, then scaling \mathbf{w} we can make the cost functional arbitrary negative, i.e., $(QC_{h,\mathbf{0}})$ has no solution.

Hence, if $(QC_{h,\mathbf{0}})$ has a solution, then $\mathbf{w} \equiv \mathbf{0}$ is such a solution. If the solution is unique, then (SC) holds. Certainly, if (SC) holds then $(QC_{h,\mathbf{0}})$ has a unique solution $\mathbf{w} \equiv \mathbf{0}$.

Therefore, to complete the proof it is enough to show that (LI) and (SC) imply the existence and uniqueness of the solution of $(Q_{h,\mathbf{a}})$ for an arbitrary $\mathbf{a} \in \mathbb{R}^n$. To do that let us note that using the same argument as in the proof of Proposition 3.1, we find that for any $\mathbf{a} \in \mathbb{R}^n$ there exists a pair $(\bar{\mathbf{z}}(h), \bar{\mathbf{w}}(h))$ satisfying (4.8) and (4.9). Let us introduce new variables $\mathbf{y} = \mathbf{z} - \bar{\mathbf{z}}(h)$ and $\mathbf{v} = \mathbf{w} - \bar{\mathbf{w}}(h)$. Problem $(QC_{h,\mathbf{a}})$ formulated in terms of (\mathbf{y}, \mathbf{v}) has homogeneous constraints and the quadratic term in the cost functional is given by (4.7). Hence, by (SC) it has a unique solution.

□

By Lemma 4.1 the second order sufficient condition can be verified by studying the existence and uniqueness of the auxiliary quadratic optimal control problem $(QC_{h,\mathbf{a}})$. Unfortunately, this last problem is fairly complicated.

Below we are going to formulate an explicit sufficient (but not necessary) criterion of (SC).

Namely, (SC) is obviously satisfied if

$$(\overline{SC})$$

$$\sum_{i=0}^{N-1} [\langle \mathbf{z}_i, P_i(h)\mathbf{z}_i \rangle + 2\langle \mathbf{z}_i, Q_i(h)\mathbf{w}_i \rangle + \langle \mathbf{w}_i, R_i(h)\mathbf{w}_i \rangle] + \langle \mathbf{z}_N, P_N(h)\mathbf{z}_N \rangle > 0$$

for all $(\mathbf{z}, \mathbf{u}) \neq \mathbf{0}$ such that

$$\begin{aligned} \mathbf{z}_{i+1} - \mathbf{z}_i &= A_i(h)\mathbf{z}_i + B_i(h)\mathbf{w}_i, \\ \mathbf{z}_0 &= \mathbf{0}. \end{aligned}$$

The following proposition provides a criterion of (\overline{SC}) .

PROPOSITION 4.2. (\overline{SC}) is satisfied if the matrices E_i , $i = N, N-1, \dots, 2, 1, 0$ given by the following recursive formulas

$$\begin{aligned} E_N &= P_N, \\ E_{i-1} &= P_{i-1} + (I + A_{i-1})^T E_i (I + A_{i-1}) + [Q_{i-1} + (A_{i-1} + I)^T E_i B_{i-1}] \times \\ &\quad \times [R_{i-1} + B_{i-1}^T E_i B_{i-1}]^{-1} [Q_{i-1} + (I + A_{i-1})^T E_i B_{i-1}]^T \\ &\quad i = N-1, \dots, 0 \end{aligned} \quad (4.10)$$

are well defined. Here the arguments h are omitted for the sake of simplicity.

Proof. By Lemma 4.1 (\overline{SC}) is equivalent to the condition that the following quadratic optimal control problem

$(\overline{QC}_{h,\mathbf{a}})$ minimize

$$V_0(\mathbf{z}, \mathbf{w}) = \sum_{i=0}^{N-1} [\langle \mathbf{z}_i, P_i(h)\mathbf{z}_i \rangle + 2\langle \mathbf{z}_i, Q_i(h)\mathbf{w}_i \rangle + \langle \mathbf{w}_i, R_i(h)\mathbf{w}_i \rangle] + \langle \mathbf{z}_N, P_N(h)\mathbf{z}_N \rangle \quad (4.11)$$

subject to

$$\mathbf{z}_{i+1} - \mathbf{z}_i = A_i(h)\mathbf{z}_i + B_i(h)\mathbf{w}_i, \quad (4.12)$$

$$\mathbf{z}_0 = \mathbf{a}, \quad (4.12a)$$

has a unique solution for any $\mathbf{a} \in \mathbb{R}^n$.

It is well known in optimal control theory (see e.g. [3, 9]) that the above condition is satisfied if (4.10) holds.

The proof of this result is performed in a standard way using Bellman's principle of optimality and will not be repeated here. \square

5. Stability and sensitivity results

Now, we are in a position to specialize the stability and sensitivity results of mathematical programming problems, given in Theorem 2.1, for discrete optimal control problems.

Using Proposition 3.2, Lemma 4.1 and Proposition 4.2 we obtain:

THEOREM 5.1. Assume that

- (i) the functions $\phi_i(\cdot, \cdot, \cdot)$, $\psi_i(\cdot, \cdot, \cdot)$, $\mathbf{f}_i(\cdot, \cdot, \cdot)$, $\mathbf{t}(\cdot)$, $\boldsymbol{\theta}_I(\cdot, \cdot)$ and $\chi_i(\cdot, \cdot)$ are C^2 in a neighbourhood of $(\mathbf{x}(\bar{h}), \mathbf{u}(\bar{h}), \bar{h})$,
 - (ii) $\chi_0^j(\mathbf{t}(\bar{h}), \bar{h}) < 0$ for $j \in L$,
 - (iii) for $h = \bar{h}$ matrices (3.3) have full row rank,
 - (iv) quadratic optimal control problem $(\overline{QC}_{\bar{h},\mathbf{a}})$ has a unique solution for any $\mathbf{a} \in \mathbb{R}^n$,
- or in particular

(iv') for $h = \bar{h}$ matrices (4.10) are well defined.

Then there exist neighbourhoods \bar{G} of \bar{h} in H and U of $(\mathbf{x}(\bar{h}), \mathbf{u}(\bar{h}))$ in $\mathbb{R}^{n(N+1)+mN}$, such that for $h \in \bar{G}$, $(\mathbf{x}(\bar{h}), \mathbf{u}(\bar{h}))$ is the unique minimizer of (O_h) in U and $(\boldsymbol{\rho}(h), \mathbf{p}(h), \boldsymbol{\lambda}(h), \boldsymbol{\mu}(h))$ are the unique associated Lagrange multipliers.

The functions $\mathbf{x}(\cdot)$, $\mathbf{u}(\cdot)$, $\boldsymbol{\rho}(\cdot)$, $\mathbf{p}(\cdot)$, $\boldsymbol{\lambda}(\cdot)$, $\boldsymbol{\mu}(\cdot)$ are Lipschitz continuous on \bar{G} and directionally differentiable at \bar{h} .

The directional derivatives $\partial_h \mathbf{x}(\bar{h}, g)$, $\partial_h \mathbf{u}(\bar{h}, g)$, $\partial_h \boldsymbol{\rho}(\bar{h}, g)$, $\partial_h \mathbf{p}(\bar{h}, g)$, $\partial_h \boldsymbol{\lambda}(\bar{h}, g)$, $\partial_h \boldsymbol{\mu}(\bar{h}, g)$ are given by the solution and the associated Lagrange multipliers of the following quadratic optimal control problem:

(QO $_{\bar{h},g}$) minimize

$$\sum_{i=1}^{N-1} [\frac{1}{2} \langle \mathbf{y}_i, P_i(\bar{h}) \mathbf{y}_i \rangle + \langle \mathbf{y}_i, Q_i(\bar{h}) \mathbf{v}_i \rangle + \frac{1}{2} \langle \mathbf{v}_i, R_i(\bar{h}) \mathbf{v}_i \rangle + \langle \mathbf{y}_i, S_i(\bar{h}) g \rangle + \langle \mathbf{w}_i, T_i(\bar{h}) g \rangle] + \langle \mathbf{y}_N, P_N(\bar{h}) \mathbf{y}_N \rangle + \langle \mathbf{y}_N, S_N(\bar{h}) g \rangle$$

subject to

$$\mathbf{y}_{i+1} - \mathbf{y}_i = A_i(\bar{h}) \mathbf{y}_i + B_i(\bar{h}) \mathbf{v}_i + C_i(\bar{h}) g,$$

$$\mathbf{y}_0 = D_h \mathbf{t}(\bar{h}) g,$$

$$\langle D_{\mathbf{u}} \theta_i^j(\mathbf{u}_i(\bar{h}), \bar{h}), \mathbf{v}_i \rangle + \langle D_h \theta_i^j(\mathbf{u}_i(\bar{h}), \bar{h}), g \rangle \begin{cases} = 0, & \text{for } j \in K_i^c(\bar{h}) \\ \leq 0, & \text{for } j \in K_i(\bar{h}) \setminus K_i^c(\bar{h}), \end{cases}$$

$$i = 0, 1, \dots, N-1$$

$$\langle D_{\mathbf{x}} \chi_i^j(\mathbf{x}_i(\bar{h}), \bar{h}), \mathbf{y}_i \rangle + \langle D_h \chi_i^j(\mathbf{x}_i(\bar{h}), \bar{h}), g \rangle \begin{cases} = 0, & \text{for } j \in L_i^c(\bar{h}) \\ \leq 0, & \text{for } j \in L_i(\bar{h}) \setminus L_i^c(\bar{h}), \end{cases}$$

$$i = 0, 1, \dots, N-1.$$

where

$$S_i(h) = D_{\mathbf{x},h}^2 \mathcal{L}(\mathbf{x}(h), \mathbf{u}(h), \boldsymbol{\rho}(h), \mathbf{p}(h), \boldsymbol{\lambda}(h), \boldsymbol{\mu}(h), h),$$

$$T_i(h) = D_{\mathbf{u},h}^2 \mathcal{L}(\mathbf{x}(h), \mathbf{u}(h), \boldsymbol{\rho}(h), \mathbf{p}(h), \boldsymbol{\lambda}(h), \boldsymbol{\mu}(h), h),$$

and $P_i(h)$, $Q_i(h)$, $R_i(h)$ are given by (4.2).

Note that due to the equivalence of Problems (O_h) and (\tilde{O}_h) other sensitivity results known in mathematical programming can be also specialized for discrete optimal control problems. In particular, using the well-known result concerning continuous differentiability with respect to the parameter of the solutions to mathematical programming problems (see [5]), we obtain

COROLLARY 5.2. If in addition to assumptions (i)–(iv) of Theorem 5.1 the strict complementary holds at \bar{h} , i.e.,

$$(v) \quad K_i(\bar{h}) = K_i^c(\bar{h}), \quad i = 0, 1, \dots, N-1 \quad \text{and}$$

$$L_i(\bar{h}) = L_i^c(\bar{h}), \quad i = 1, 2, \dots, N$$

then $(\mathbf{x}(\cdot), \mathbf{u}(\cdot))$ is Fréchet differentiable on a neighbourhood of \bar{h} . The derivative is given by the solution of $(\text{QO}_{\bar{h},g})$ where inequality type constraints disappear.

References

1. Alt, W., Stability of Solutions to Control Constrained Nonlinear Optimal Control Problems. *Applied Mathematics and Optimization* **21** (1990), pp. 53–68.
2. Canon, M. D., Cullum, C. D. jr., Polak, E., *Theory of Optimal Control and Mathematical Programming*. McGraw-Hill, New York, 1970.
3. Clements, D. J., Anderson, B. D. O., *Singular Optimal Control: the Linear-Quadratic Problem*. Lecture Notes in Control and Information Sciences, vol. 5, Springer-Verlag, Berlin, 1978.
4. Dontchev, A. L., *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*. Lecture Notes in Control and Information Sciences, vol. 52, Springer-Verlag, Berlin, 1983.
5. Fiacco, A. V., *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. Academic Press, New York, 1983.
6. Hager, W. W., Lipschitz Continuity for Constrained Processes. *SIAM J. Control and Optimization* **17** (1979), pp. 321–333.
7. Hestenes, M. R., *Calculus of Variations and Optimal Control Theory*. J. Wiley and Sons, New York, 1969.
8. Jittomtrum, K., Solution Point Differentiability without Strict Complementary in Nonlinear Programming. *Math. Programming Study* **21** (1984), pp. 127–138.
9. Kwakernaak, H., Sivan, R., *Linear Optimal Control Systems*. Wiley, New York, 1972.
10. Malanowski, K., Differential Sensitivity of Solutions to Convex Constrained Optimal Control Problems for Discrete Systems. *JOTA* **53** (1987), pp. 429–449.
11. Robinson, S. M., Local Structure of Feasible Sets in Nonlinear Programming, Part III: Stability and Sensitivity. *Math. Programming Study* **30** (1987), pp. 45–66.

Анализ устойчивости и чувствительности дискретных задач оптимального управления

К. МАЛАНОВСКИ

(Варшава)

Рассматривается семейство дискретных задач оптимального управления при наличии ограничений управления и состояния. Показатель качества и ограничения зависят от параметра.

Используя известные результаты по устойчивости и чувствительности задач математического программирования находятся условия, при которых решения задач

оптимального управления являются непрерывными по Липшицу и дифференцируемыми по направлению функциями параметра. Производная по направлению характеризуется в виде решения вспомогательной линейно-квадратической задачи оптимального управления.

Kazimierz Malanowski
Systems Research Institute
Polish Academy of Sciences
Warsaw
Poland

RELAXABILITY AND WELL-POSEDNESS FOR INFINITE DIMENSIONAL OPTIMAL CONTROL PROBLEMS*

N. S. PAPAGEORGIOU**

(Athens)

(Received November 1, 1990)

In this paper we investigate the relation between the notions of "well-posedness in the sense of performance convergence" and "relaxability" for a large class of nonlinear infinite-dimensional optimal control problems. We show that relaxability implies well-posedness and the two are equivalent for semilinear systems. In doing this we also prove two new density results concerning the original and relaxed trajectories, which are of independent interest.

Keywords and -phrases: Relaxability, well-posedness, Arzelà-Ascoli theorem, evolution operator, compact embedding, monotone operator, measurable multifunction, weak norm, Aumann's selection theorem.

AMS Subject Classification (1980): 49 A 20, 49 C 15

1. Introduction

In a recent paper, Dontchev and Morduhovic [11] established the equivalence between "performance well-posedness" and "regularity in the sense of relaxation" (what we call in this paper "relaxability"), for a large class of finite-dimensional, non-linear, optimal control problems. The result of Dontchev and Morduhovic also can be viewed as a generalization of an earlier important work by Clarke [7].

The purpose of this paper is to extend the work of Dontchev and Morduhovic [11] to nonlinear, infinite-dimensional, control systems (distributed parameter systems). In this process, we obtain a density result concerning the original and relaxed trajectories, which corrects an earlier attempt for an analogous theorem by Ahmed [1], which appears to have a serious gap in the proof. Moreover, we extend the finite-dimensional work of Gamkrelidze [14].

*Revised version

**Research supported by N.S.F. Grant D.M.S.-8802688.

2. Preliminaries

Let (Ω, Σ, μ) be a σ -finite measure space and X a separable Banach space. Throughout this paper we will be using the following notations:

$$P_{f(c)}(X) = \{A \subseteq X : \text{nonempty, closed, (convex)}\}$$

$$\text{and } P_{\text{wk}(c)}(X) = \{A \subseteq X : \text{nonempty, weakly compact, (convex)}\}.$$

A multifunction $F : \Omega \rightarrow P_f(X)$ is said to be measurable if and only if for all $z \in X$, $\omega \rightarrow d(z, F(\omega)) = \inf\{\|z - x\| : x \in F(\omega)\}$ is measurable. Other equivalent definitions of measurability can be found in Wagner [21]. A multifunction $G : \Omega \rightarrow 2^X \setminus \{\emptyset\}$ is said to be graph measurable, if $\text{Gr } G = \{(\omega, x) \in \Omega \times X : x \in G(\omega)\} \in \Sigma \times B(X)$, with $B(X)$ being the Borel σ -field of X . For closed valued multifunctions, measurability implies graph measurability and the two are equivalent if Σ is μ -complete. By S_F^p , $1 \leq p \leq \infty$ we will denote the set of $L^p(X)$ -selectors of $F(\cdot)$, i.e. $S_F^p = \{f \in L^p(X) : f(\omega) \in F(\omega) \mu\text{-a.e.}\}$. It is easy to see that if $F(\cdot)$ is L^p -integrably bounded i.e. $\omega \rightarrow |F(\omega)| = \sup\{\|x\| : x \in F(\omega)\} \in L^p_+$, then $S_F^p \neq \emptyset$.

Next, let Y, Z be Hausdorff topological spaces. A multifunction $G : Y \rightarrow 2^Z \setminus \{\emptyset\}$ is said to be upper semicontinuous (u.s.c.) (resp. lower semicontinuous (l.s.c.)), if for all $U \subseteq Z$ open, $F^+(U) = \{y \in Y : F(y) \subseteq U\}$ (resp. $F^-(U) = \{y \in Y : F(y) \cap U \neq \emptyset\}$) is open. Other equivalent definitions of upper and lower semicontinuity can be found in Delahaye and Denel [8]. If Y is a metric space, then on $P_f(Y)$ we can define a metric $h(\cdot, \cdot)$, known as the Hausdorff metric, by setting $h(A, B) = \max\{\sup\{d(a, B) : a \in A\}, \sup\{d(b, A) : b \in B\}\}$. It is well known that if Y is complete, then so is the metric space $(P_f(Y), h)$.

Finally, recall that if V, W are Polish spaces (i.e. complete, separable, metrizable spaces), a function $f : \Omega \times V \rightarrow W$ is said to be "Carathéodory function" if and only if (i) $\omega \rightarrow f(\omega, v)$ is measurable and (ii) $v \rightarrow f(\omega, v)$ is continuous. It is a well-known fact from measure theory that such a function is jointly measurable.

Now, let us introduce the problem that we will be studying in this paper. Let $E \subseteq \mathbb{R}$ be a nonempty parameter set containing zero as a limit point. Let $T = [0, b]$ and let (X, H, X^*) be a Gelfand triplet of spaces, i.e. H is a separable Hilbert space and X is a dense subspace of H carrying the structure of a separable reflexive Banach space, which embeds continuously in H . Identifying H with its dual (pivot space), we have $X \hookrightarrow H \hookrightarrow X^*$, with all embeddings being continuous and dense and also assumed to be compact. By $\langle \cdot, \cdot \rangle$ we will denote the duality brackets for the pair (X, X^*) and by (\cdot, \cdot) the inner product in H . The two are compatible, in the sense that if $x \in X \subseteq H$ and $h \in H \subseteq X^*$, then $\langle x, h \rangle = (x, h)$. Also by $\|\cdot\|$ (resp. $|\cdot|, \|\cdot\|_*$), we will denote the norm in X (resp. in H, X^*). Also let Y be another separable Banach space, modelling the control space. Finally, by X_w we will denote X with the weak topology and by \xrightarrow{s} (resp. \xrightarrow{w}) we will

denote the strong (resp. weak) convergence. We will consider the following family of infinite-dimensional optimal control problems, with state and control constraints:

$$\left\{ \begin{array}{l} J(u, \varepsilon) = g(x(b)) \rightarrow \inf \\ \text{s.t. } \dot{x}(t) + A(t, x(t)) = f(t, x(t), u(t), \varepsilon) \text{ a.e.} \\ x(0) = x_0, \quad x(t) \in K(t, \varepsilon) \\ u(t) \in U(t) \text{ a.e., } u(\cdot) \text{ is measurable} \end{array} \right\} \quad (P(\varepsilon)).$$

These form a family of perturbed problems, the original one corresponding to $\varepsilon = 0$. Denote the value of $P(\varepsilon)$ by $m(\varepsilon)$ and the value of the original problem ($\varepsilon = 0$), by m . Following Dontchev and Morduhovic [11], we say that the family $\{P(\varepsilon) : \varepsilon \in E\}$ is “well-posed” in the sense of performance convergence if and only if $m(\varepsilon) \rightarrow m$ as $\varepsilon \rightarrow 0$.

To problem $P(0)$, we can associate a new, augmented system, with convexified dynamics, known as the “relaxed system” (see Clarke [6] and Warga [22]). This has the following form:

$$\left\{ \begin{array}{l} J_r(u) = g(x(b)) \rightarrow \inf \\ \text{s.t. } \dot{x}(t) + A(t, x(t)) \in \overline{\text{conv}} F(t, x(t)) \text{ a.e.} \\ x(0) = x_0, \quad x(t) \in K(t) \\ u(t) \in U(t) \text{ a.e., } u(\cdot) \text{ is measurable} \end{array} \right\} \quad (P_r).$$

where $F(t, x) = \bigcup\{f(t, x, u) = f(t, x, u, 0) : u \in U(t)\}$. We will denote the value of this problem by m_r . Both equations in $P(\varepsilon)$ and P_r should be interpreted in the distributional sense.

We say that problem $P(0)$ is “relaxable” (or in the terminology of Dontchev and Morduhovic [11], “regular in the sense of relaxation”), if $m = m_r$.

In this paper we investigate the relation between the notions of well-posedness and relaxability. To avoid trivialities, we will assume that all systems considered here have admissible “state-control” pairs, i.e. there exists a pair of functions $(x(\cdot), u(\cdot))$ satisfying the constraints in $P(\varepsilon)$ and P_r .

3. Main theorems

In the first theorem we prove that relaxability implies well-posedness. For this we will need the following set of hypotheses:

$H(A)_1$: $A : T \times X \rightarrow X^*$ is an operator s.t.

- (1) $t \rightarrow A(t, x)$ is measurable,
- (2) $x \rightarrow A(t, x)$ is monotone and sequentially weakly continuous,

(3) $\langle A(t, x), x \rangle \geq c\|x\|^p$ a.e. with $c > 0$ and $1 < p < \infty$,

(4) $\|A(t, x)\|_* \leq c_1(1 + \|x\|^{p-1})$ with $c_1 > 0$.

$H(f)_1$: $f : T \times H \times Y \times E \rightarrow X^*$ is a mapping s.t.

(1) $t \rightarrow f(t, x, u, \varepsilon)$ is measurable,

(2) $(x, u, \varepsilon) \rightarrow f(t, x, u, \varepsilon)$ is sequentially continuous from $H \times Y_w \times \{0\}$ into X_w^* ,

(3) $\langle f(t, x, u, \varepsilon), x \rangle \geq c'\|x\|^p$ a.e., $c' > 0$,

(4) $\|f(t, x, u, \varepsilon)\|_* \leq g_2(t) + c_2\|x\|^{p-1}$ a.e. with $g_2(\cdot) \in L^q_+$, $c_2 > 0$, $1/p + 1/q = 1$.

$H(U)$: $U : T \rightarrow P_{\text{wkc}}(Y)$ is integrably bounded, $U(t) \subseteq W$ a.e. with $W \in P_{\text{wkc}}(Y)$.

As it was illustrated with an example by Dontchev and Morduhovic [11] (section 3), since the notion of well-posedness is defined through performance convergence, we need the following additional hypothesis:

H_0 : There exists a minimizing sequence $\{u_n\}_{n \geq 1}$ for $P(0)$ s.t. if $x_n(\cdot, \varepsilon)$ and $x_n(\cdot)$ $n \geq 1$, solve the dynamics of $P(\varepsilon)$ and $P(0)$, respectively, with control $u_n(\cdot)$, then $x_n(t, \varepsilon) \in K(t, \varepsilon)$ for all $t \in T$ and $x_n(b, \varepsilon) \rightarrow x_n(b)$ as $\varepsilon \rightarrow 0$.

Also we need hypotheses for the viability domains $K(t, \varepsilon)$ and the cost functional $g(\cdot)$:

$H(K)_1$: $K : T \times E \rightarrow P_f(H)$ is u.s.c in the ε -variable.

$H(g)$: $g(\cdot)$ is continuous from H into \mathbf{R} .

Because of hypotheses $H(A)_1$ and $H(f)_1$ and using theorem 4.2, p. 167, of Barbu [4], we deduce that given an admissible control $u \in S_U^1$, both the perturbed and original problems have a unique trajectory $x(\cdot) \in W(T) = \{z \in L^p(X), \dot{z} \in L^q(X^*)\} \subseteq C(T, H)$. Also in [19] we proved that $x(\cdot) \in C(T, X_w)$ (X_w is the space with the weak topology).

Now, we are ready for our first theorem.

THEOREM 1. *If hypotheses $H(A)_1$, $H(f)_1$, $H(U)$, H_0 , $H(K)_1$ and $H(g)$ hold, then relaxability implies well-posedness.*

Proof. Let $\{x_n\}_{n \geq 1}$ be the minimizing sequence for problem $P(0)$ postulated by hypothesis H_0 . For any $n \geq 1$ we have $g(x_n(b, \varepsilon)) \rightarrow g(x_n(b))$ as $\varepsilon \rightarrow 0$. Also note that $m(\varepsilon) \leq g(x_n(b, \varepsilon))$. So we get

$$\overline{\lim}_{\varepsilon \rightarrow 0} m(\varepsilon) \leq m. \quad (1)$$

On the other hand, let $\varepsilon_n \in E$, $\varepsilon_n \rightarrow 0$. Choose admissible state-control pairs (x_n, u_n) s.t.

$$J(u_n, \varepsilon_n) \leq m(\varepsilon_n) + \frac{1}{n}. \quad (2)$$

From [19] (see also [3]), we know that $\{\dot{x}_n\}_{n \geq 1}$ is relatively sequentially weakly compact in $L^q(X^*)$ and $\{x_n\}_{n \geq 1}$ is relatively sequentially compact in $C(T, X_w)$. So,

by passing to a subsequence if necessary, we may assume that $x_n \rightarrow x$ in $C(T, X_w)$ and $\dot{x}_n \xrightarrow{w} z = \dot{x}$ in $L^q(X^*)$. Also because of hypotheses $H(A)$ (4) and $H(f)$ (4), we see that for every $t \in T$, $\{\dot{x}_n(t)\}_{n \geq 1}$ is bounded and because of the reflexivity of X^* , w -compact. Thus, we can apply theorem 3.1 of [16] and get that:

$$\dot{x}(t) \in \overline{\text{conv } w\text{-}\lim}(F(t, x_n(t)) - A(t, x_n(t))) \text{ a.e.}$$

where recall that $F(t, x) = f(t, x, U(t))$. Since $x_n \rightarrow x$ in $C(T, X_w)$ and by hypothesis $X \looparrowright H$ compactly, we have that for all $t \in T$ $x_n(t) \xrightarrow{s} x(t)$ in H . Also because of hypotheses $H(f)$ (2) and $H(U)$, through proposition 1, p. 47, of Aubin and Cellina [2], we get that $F(t, \cdot)$ is u.s.c. from H into X_w^* and so from Delahaye and Denel [8] we deduce that $w\text{-}\lim F(t, x_n(t)) \subseteq F(t, x(t))$. Furthermore, hypothesis $H(A)$ (2) tells us that $A(t, x_n(t)) \xrightarrow{w} A(t, x(t))$ in X^* . Finally, we have:

$$\begin{aligned} \dot{x}(t) &\in \overline{\text{conv}}(F(t, x(t)) - A(t, x(t))) \text{ a.e.} \\ \Rightarrow \dot{x}(t) + A(t, x(t)) &\in \overline{\text{conv}} F(t, x(t)) \text{ a.e.} \end{aligned}$$

Thus, $x(\cdot)$ satisfies the dynamics of the relaxed problem. Furthermore, using hypothesis $H(K)_1$ and the compact embedding of X in H , we have:

$$\begin{aligned} x(t) &\in \overline{\lim} K(t, \varepsilon_n) \subseteq K(t) \\ \Rightarrow x(\cdot) &\text{ is a viable, relaxed trajectory.} \end{aligned}$$

Now, note that by passing to the limit in (2), we get:

$$\underline{\lim} J(u_n, \varepsilon_n) = \lim g(x_n(b, \varepsilon_n)) = g(x(b)) \leq \underline{\lim} m(\varepsilon_n).$$

Since by hypothesis relaxability holds, we can write that:

$$\begin{aligned} m_r = m &\leq g(x(b)) \leq \underline{\lim} m(\varepsilon_n) \\ \Rightarrow m &\leq \underline{\lim} m(\varepsilon_n). \end{aligned} \tag{3}$$

From (1) and (3) above we conclude that $m(\varepsilon) \rightarrow m$ and so well-posedness holds.

Q.E.D.

Our goal now is to show that the converse of the above theorem also holds, namely that well-posedness implies relaxability.

To this end, we prove a density result of the trajectories of $P(0)$ in the trajectories of P_r . Another result of this kind for semilinear systems was obtained by Ahmed [1]. However, it appears that his proof has a serious gap, that makes his theorem incorrect. Namely, in the second half of page 292 of [1], the author claims that $y_n \xrightarrow{s} y$ in $L^p(E)$ (we use the notation of [1]). Unfortunately, such a conclusion is not justified by the hypotheses on the sequence $\{y_n\}_{n \geq 1}$. Additional

hypotheses of strong compactness and p -equi-integrability are needed in order to have the desired strong convergence of the y_n 's. Furthermore, it is well-known from the theory of differential inclusions, that in order to have a density result of the original trajectories into those of the convexified (relaxed) system, we need a Lipschitz hypothesis in the state variable on the orientor field (in the case of differential inclusions originating from control problems, a dissipativity hypothesis can also do the job).

Here we provide density results for both semilinear and strongly nonlinear distributed parameter control systems. Another such density theorem for infinite-dimensional differential inclusions was recently obtained by the author in [18]. However, the hypotheses there are such that limit its applicability to control systems satisfying stricter hypotheses.

Although the converse of theorem 1 will be stated only for semilinear systems, we prove density results for both the strongly nonlinear and semilinear cases.

We start with the nonlinear one.

$H(A)_2$: $A : T \times X \rightarrow X^*$ is an operator s.t.

- (1) $t \rightarrow A(t, x)$ is measurable,
- (2) $x \rightarrow A(t, x)$ is monotone and weakly sequentially continuous,
- (3) $\langle A(t, x), x \rangle \leq c_1 \|x\|^p$ a.e. with $c_1 > 0$ and $1 < p < \infty$,
- (4) $\|A(t, x)\|_* \leq c_2(1 + \|x\|^{p-1})$ with $c_2 > 0$.

$H(f)_2$: $f : T \times H \times Y \times E \rightarrow H$ is a mapping s.t.

- (1) $t \rightarrow f(t, x, u, \varepsilon)$ is measurable,
- (2) $(x, u, \varepsilon) \rightarrow f(t, x, u, \varepsilon)$ is sequentially continuous from $H \times Y_w \times \{0\}$ into H ,
- (3) $|f(t, x, u, \varepsilon) - f(t, x', u, \varepsilon)| \leq k(t)|x' - x|$ a.e.,
- (4) $\langle f(t, x, u, \varepsilon), x \rangle \geq c_3 \|x\|^p$ a.e. with $c_3 > 0$,
- (5) $|f(t, x, u, \varepsilon)| \leq r(\cdot) + c_4 \|x\|^{p-1}$ a.e., with $c_4 > 0$, $r(\cdot) \in L^2_+$.

Let S_0 be the set of trajectories of the original evolution and S_r the set of trajectories of the relaxed one. The next theorem relates those two sets.

THEOREM 2. *If hypotheses $H(A)_2$, $H(f)_2$ and $H(U)$ hold, then $\overline{S_0} = S_r$, the closure taken in $C(T, X_w)$.*

Proof. Set $f(t, x, u) = f(t, x, u, 0)$ and define:

$$F(t, x) = f(t, x, U(t)) \text{ and } F_c(t, x) = \overline{\text{conv}} F(t, x).$$

We have:

$$\begin{aligned} \text{Gr } F &= \{(t, x, z) \in T \times X \times H : z \in F(t, x)\} \\ &= \{(t, x, z) \in T \times X \times H : z = f(t, x, u), u \in U(t)\}. \end{aligned}$$

Set $k(t, x, z, u) = z - f(t, x, u)$ and $q(t, u) = d(u, U(t))$. Then we have:

$$\begin{aligned} \text{Gr } F &= \{(t, x, z) \in T \times X \times H : k(t, x, z, u) = 0, q(t, u) = 0\} \\ &= \text{proj}_{T \times X \times H} \{(t, x, z, u) \in T \times X \times H \times W : k(t, x, z, u) = 0, q(t, u) = 0\}. \end{aligned}$$

Note that both k and q are $B(T) \times B(X) \times B(H) \times B(W)$ -measurable. Also since by hypothesis Y is separable, is weakly compactly generated and so it admits a Kadec norm (see Diestel [10]). Thus, we can apply corollary 2.4 of Edgar [13] and get that $B(Y) = B(Y_w)$, where Y_w denotes the Banach space Y equipped with the weak topology. Hence, $B(Y) \cap W = B(Y_w) \cap W \Rightarrow B(W) = B(W_w)$. But recall that W_w is a compact, Polish space (see Dunford and Schwartz [12], theorem 3, p. 434). So, the Arsenin-Novikov theorem (see, for example, Dellacherie [9]), tells us that:

$$\begin{aligned} \text{proj}_{T \times X \times H} \{ & (t, x, z, u) \in T \times X \times H \times W : k(t, x, z, u) = 0, q(t, u) = 0 \} \\ & \in B(T) \times B(X) \times B(H) \\ \Rightarrow \text{Gr } F & \in B(T) \times B(X) \times B(H). \end{aligned}$$

So, $F(\cdot, \cdot)$ is graph measurable. Hence, we can apply theorem 2 of Chuong [5] and get that $S_{F(\cdot, x(\cdot))}^1$ is dense in $S_{F_c(\cdot, x(\cdot))}^1$ for the weak norm $\|\cdot\|_w$ defined by $\|h\|_w = \sup \left\{ \left| \int_t^{t'} h(s) ds \right| : t, t' \in T \right\}$. Next, let $x(\cdot) \in S_r$. Then $\dot{x}(t) + A(t, x(t)) = g(t)$ a.e., $x(0) = x_0$ with $g(\cdot) \in S_{F_c(\cdot, x(\cdot))}^1$. Let $g_n \in S_{F_c(\cdot, x(\cdot))}^1$ s.t. $\|g_n - g\|_w \rightarrow 0$. Because of $H(f)_2$ (5) we see that $\{g_n, g\}_{n \geq 1} \in L^2(H)$ and is bounded there. Let $v : T \rightarrow H$ be a step function i.e. $v(t) = \sum_{k=1}^m \chi_{[t_{k-1}, t_k]}(t) \cdot v_k$ and by $(\cdot, \cdot)_0$ denote the inner product in $L^2(H)$. We have:

$$(g_n - g, v)_0 \leq \sum_{k=1}^m \left| \int_{t_{k-1}}^{t_k} (g_n(s) - g(s)) ds \right| \cdot |v_k| \leq \|g_n - g\|_w \sum_{k=1}^m |v_k| \rightarrow 0.$$

Since step functions are dense in $L^2(H)$ and $\{g_n, g\}_{n \geq 1}$ is L^2 -bounded we get that $g_n \xrightarrow{w} g$ in L^2 (and, in particular $g_n \xrightarrow{w} g$ in $L^1(H)$).

Now, consider the multifunction $L_n : T \rightarrow 2^W \setminus \{\emptyset\}$ $n \geq 1$, defined by:

$$L_n(t) = \{u \in U(t) : g_n(t) = f(t, x(t), u)\}.$$

Clearly, $L_n \neq \emptyset$ for all $t \in T \setminus N_n$, where N_n is a Lebesgue null set. Set $L_n(t) = \{0\}$ for $t \in N_n$. Also let $\{x_m\}_{m \geq 1}$ be dense in H and consider the following functions:

$$h_m^n(t, u) = \begin{cases} (x_m, g_n(t) - f(t, x(t), u)) & \text{for } t \in T \setminus N_n \\ 0 & \text{for } t \in N_n. \end{cases}$$

For all $n, m \geq 1$, $h_m^n(\cdot, \cdot)$ is a Carathéodory function, hence $\hat{B}(T) \times B(W)$ -measurable, with $\hat{B}(T)$ being the Lebesgue completion of $B(T)$. The observe that:

$$\begin{aligned} \text{Gr } L_n &= \left[\bigcap_{m \geq 1} \{(t, u) \in T \times W : h_m^n(t, u) = 0\} \right] \cap \text{Gr } U \\ &\Rightarrow \text{Gr } L_n \in \hat{B}(T) \times B(W) \text{ for all } n \geq 1. \end{aligned}$$

Apply Aumann's selection theorem (see for example Wagner [21]), to get $u_n : T \rightarrow W$ $n \geq 1$ measurable s.t. $u_n(t) \in L_n(t)$ for all $t \in T$. So, we have:

$$g_n(t) = f(t, x(t), u_n(t)) \text{ a.e. and } u_n(t) \in U(t) \text{ a.e.}$$

Let $y_n(\cdot)$ be the unique trajectory of the original system corresponding to the control function $u_n(\cdot)$. Again, we may assume that $y_n \rightarrow y$ in $C(T, X_w)$. We have:

$$\begin{aligned} \frac{d}{dt}|x(t) - y_n(t)|^2 &= 2\langle \dot{x}(t) - \dot{y}_n(t), x(t) - y_n(t) \rangle \\ &= 2\langle -A(t, x(t)) + g(t) + A(t, y_n(t)) - f(t, y_n(t), u_n(t)), x(t) - y_n(t) \rangle \\ &\leq (g(t) - f(t, y_n(t), u_n(t)), x(t) - y_n(t)) \text{ (since } A(t, \cdot) \text{ is monotone)} \\ &\Rightarrow |x(t) - y_n(t)|^2 \leq \\ &\leq \int_0^t (g(s) - f(s, x(s), u_n(s)), x(s) - y_n(s)) ds + \\ &+ \int_0^t (f(s, x(s), u_n(s)) - f(s, y_n(s), u_n(s)), x(s) - y_n(s)) ds \\ &\leq \int_0^t (g(s) - g_n(s), x(s) - y_n(s)) ds + \int_0^t k(s)|x(s) - y_n(s)|^2 ds. \end{aligned}$$

Recall that $g_n \xrightarrow{w} g$ in $L^1(H)$. Also $y_n \rightarrow y$ in $C(T, X_w)$ and since X embeds compactly into H , we have $y_n(t) \xrightarrow{s} y(t)$ in H . Thus, by passing to the limit as $n \rightarrow \infty$ in the last inequality, we get

$$|x(t) - y(t)|^2 \leq \int_0^t k(s)|x(s) - y(s)|^2 ds.$$

Invoking Gronwall's inequality, we get that $x = y$. So, $x \in \bar{S}_0$ and since $S_0 \subseteq S_r$ and the latter is sequentially compact in $C(T, X_w)$ (see [19]), we conclude that indeed $\bar{S}_0 = S_c$, the closure taken in the $C(T, X_w)$ topology.

Q.E.D.

If the system is semilinear, we can improve our density result.

$H(A)_3$: $A : T \times X \times X^*$ is an operator s.t.

- (1) $t \rightarrow A(t)x$ is measurable,
- (2) $x \rightarrow A(t)x$ is linear, monotone,
- (3) $\|A(t')x - A(t)x\|_* \leq k|t' - t|^\alpha \cdot \|x\|$, $k \geq 0$, $\alpha \in (0, 1]$,
- (4) $\langle A(t)x, x \rangle \geq c_1\|x\|^2$ a.e., $c_1 > 0$ (i.e. $A(t)(\cdot)$ is strongly monotone),

(5) $\|A(t)x\|_* \leq c_2\|x\|$, $c_2 > 0$ (i.e. $A(t)(\cdot)$ is continuous).

Because of $H(A)_3$ (2) (3) the family of linear operators $\langle A(t)(\cdot) : t \in T \rangle$ generates a strongly continuous evolution operator $\Phi : \Delta = \{(t, s) : 0 \leq s \leq t \leq b\} \rightarrow \mathcal{L}(H)$, with respect to which a trajectory of $P(0)$ is a solution of $x(t) = \Phi(t, 0)x_0 + \int_0^t \Phi(t, s)f(s, x(s), u(s)) ds$, $t \in T$, $u \in S_U^1$ (see Tanabe [20]). We will need the following hypothesis on $\Phi(\cdot, \cdot)$:

H_c : $\Phi(t, s)$ is compact for $t - s > 0$.

THEOREM 3. *If hypotheses $H(A)_3$, $H(f)_2$ (with $p = 2$), $H(U)$ and H_c hold, then $S_r = \overline{S_0}$, the closure taken in $C(T, H)$.*

Proof. A straightforward application of Gronwall's inequality tells us that for every $x(\cdot) \in S_r \subseteq C(T, H)$, we have $\|x(t)\| \leq M$ for all $t \in T$.

Next, we will show that S_0 is relatively compact in $C(T, H)$. To this end, let $y(\cdot) \in S_0$ and let $t', t \in T$, $t < t'$. Setting as before $f(t, x, u) = f(t, x, u, 0)$, we have:

$$\begin{aligned} |y(t') - y(t)| &\leq |\Phi(t', 0)x_0 - \Phi(t, 0)x_0| + \int_t^{t'} \|\Phi(t', s)\| \cdot |f(s, y(s), u(s))| ds + \\ &\quad + \int_0^t \|\Phi(t', s) - \Phi(t, s)\| \cdot |f(s, y(s), u(s))| ds. \end{aligned}$$

Recalling that $t \rightarrow \Phi(t, 0)x_0$ is continuous, given $\varepsilon > 0$ we can find $\delta_1 > 0$ s.t. $\|\Phi(t', 0)x_0 - \Phi(t, 0)x_0\| < \varepsilon/3$ for $t' - t < \delta_1$. Also since $\|\Phi(t, s)\| \leq M_1$ for $(t, s) \in \Delta$ and $|f(s, y(s), u(s))| \leq g_2(s) + c_4\|y(s)\| \leq g_2(s) + c_4M = \psi(s)$ a.e., $\psi(\cdot) \in L^1_+$. So, we can find $\delta_2 > 0$ s.t.

$$\int_t^{t'} \|\Phi(t', s)\| \cdot |f(s, x(s), u(s))| ds \leq \int_t^{t'} M_1 \psi(s) ds < \varepsilon/3.$$

Finally, for $\varepsilon' > 0$, write

$$\begin{aligned} &\int_0^t \|\Phi(t', s) - \Phi(t, s)\| \cdot |f(s, y(s), u(s))| ds = \\ &= \int_0^{t-\varepsilon'} \|\Phi(t', s) - \Phi(t, s)\| \cdot |f(s, y(s), u(s))| ds + \\ &\quad + \int_{t-\varepsilon'}^t \|\Phi(t', s) - \Phi(t, s)\| \cdot |f(s, y(s), u(s))| ds. \end{aligned}$$

Note that $\int_{t-\varepsilon'}^t \|\Phi(t', s) - \Phi(t, s)\| \cdot |f(s, y(s), u(s))| ds \leq \int_{t-\varepsilon'}^t 2M_1\psi(s) ds$.

Pick $\varepsilon' > 0$ so that $\int_{t-\varepsilon'}^t 2M_1\psi(s) ds < \varepsilon/6$. Also because of hypothesis H_c , from proposition 2.1 of [17], we know that $t \rightarrow \Phi(t, s)$ is continuous in the operator norm topology, uniformly for all $s \in (0, t)$ s.t. $t - s$ is bounded away from zero. So we can find $\delta_3 > 0$ s.t. $\int_0^{t-\varepsilon'} \|\Phi(t', s) - \Phi(t, s)\| \cdot |f(s, y(s), u(s))| ds \leq \int_0^{t-\varepsilon'} \|\Phi(t', s) - \Phi(t, s)\| \psi(s) ds < \varepsilon/6$ for $t' - t < \delta_3$. Therefore, finally, for $\delta = \min(\delta_1, \delta_2, \delta_3)$ we have:

$$\begin{aligned} \|y(t') - y(t)\| &< \varepsilon \text{ for } t' - t < \delta \text{ and all } y(\cdot) \in S_0, \\ &\Rightarrow S_0 \text{ is equicontinuous in } C(T, H). \end{aligned}$$

Also note that if $B(\psi)(s) = \{v \in H : |v| \leq \psi(s)\}$, then $s \rightarrow S(t, s)B(\psi)(s)$ is measurable and by hypothesis H_c , $P_{kc}(H)$ -valued. Applying Radstrom's embedding theorem (see, for example, Klein and Thompson [15]), we have that $\Phi(t, 0)x_0 + \int_0^t \Phi(t, s)B(\psi)(s) ds \in P_{kc}(H)$. Thus, $\{y(t) : y(\cdot) \in S_0\} \in P_k(H)$ for all $t \in T$. Invoking the Arzelà-Ascoli theorem, we deduce that S_0 is relatively compact in $C(T, H)$ as claimed.

Next, let $x(\cdot) \in S_r$. Then by definition we have:

$$x(t) = \Phi(t, 0)x_0 + \int_0^t \Phi(t, s)g(s) ds, \quad t \in T, \quad g(\cdot) \in S_{F_c(\cdot, x(\cdot))}^2,$$

where $F_c(t, x) = \overline{\text{conv}} f(t, x, U(t))$. Since convergence of $L^2(H)$ -bounded sequences in the weak norm $\|\cdot\|_w$, implies weak convergence in $L^1(H)$, we can find $g_n \in S_{F_c(\cdot, x(\cdot))}^1$ s.t. $g_n \xrightarrow{w} g$ in $L^1(H)$. As in the proof of theorem 2, an application of Aumann's selection theorem, gives us $u_n \in S_U^1$ s.t. $g_n(s) = f(s, x(s), u_n(s))$. Let $y_n(\cdot)$ be the original trajectories corresponding to the control functions $u_n(\cdot)$. We have seen in the first part of the proof that $\{y_n\}_{n \geq 1}$ is relatively compact in $C(T, H)$. So, by passing to a subsequence if necessary we may assume that $y_n \rightarrow y$ in $C(T, H)$.

Now, note that:

$$\begin{aligned} \frac{d}{ds} |x(s) - y_n(s)|^2 &= \\ 2\langle \dot{x}(s) - \dot{y}_n(s), x(s) - y_n(s) \rangle &= \\ 2\langle -A(s)x(s) + g(s) + A(s)y_n(s) - f(s, y_n(s), u_n(s)), x(s) - y_n(s) \rangle. \end{aligned}$$

Exploiting the monotonicity of $A(s)(\cdot)$, we get that:

$$\frac{d}{ds} |x(s) - y_n(s)|^2 \leq 2 \langle g(s) - f(s, y_n(s), u_n(s)), x(s) - y_n(s) \rangle.$$

Integrating both sides we have:

$$\begin{aligned} & |x(t) - y_n(t)|^2 \leq \\ \leq & 2 \int_0^t \langle g(s) - g_n(s), x(s) - y_n(s) \rangle ds + \int_0^t \langle g_n(s) - f(s, y_n(s), u_n(s)), x(s) - y_n(s) \rangle ds. \end{aligned}$$

Since $g_n \xrightarrow{w} g$ in $L^1(H)$ and $y_n \rightarrow y$ in $C(T, H)$, we see that $2 \int_0^t \langle g(s) - g_n(s), x(s) - y_n(s) \rangle ds \rightarrow 0$. Also recalling that $g_n(s) = f(s, x(s), u_n(s))$ and using the Lipschitzness of the vector field $f(t, \cdot, u)$ we have

$$2 \int_0^t \langle g_n(s) - f(s, y_n(s), u_n(s)), x(s) - y_n(s) \rangle ds \leq 2 \int_0^t k(s) |x(s) - y_n(s)|^2 ds.$$

So, in the limit we get:

$$|x(t) - y(t)|^2 \leq 2 \int_0^t k(s) |x(s) - y(s)|^2 ds.$$

Apply Gronwall's inequality to get that $x = y \Rightarrow \bar{S}_0 = S_r$, the closure taken in $C(T, H)$.

Q.E.D.

Remark. Instead of the Lipschitzness of $f(t, \cdot, u)$ we could have assumed dissipativity for the vector field.

With these density results (which are actually interesting in their own), we can prove a converse of theorem 1. We were able to do this only for the semilinear case. It will be interesting to know whether our result can be extended to strongly nonlinear systems.

For the converse of theorem 1, we will need the following new hypotheses:

$H(K)_2$: $K : T \times E \rightarrow P_f(H)$ and $\text{int } K(t, \varepsilon) \neq \emptyset$.

H_1 : There exists a relaxed optimal trajectory $y(\cdot)$ s.t. $y(t) \in \text{int } K(t, \varepsilon)$ and $d_H(y(t), \text{bd } K(t, \varepsilon)) \geq a(\varepsilon) > 0$, for all $t \in T$ and all $\varepsilon \in E \setminus \{0\}$ is a neighbourhood of zero (here $d_H(\cdot, \cdot)$ stands for the distance function in the space H).

Now, we are ready for the converse of theorem 1 for semilinear systems.

THEOREM 4. *If hypotheses $H(A)_3$, $H(f)_2$, $H(U)$, H_c , $H(K)_2$, $H(g)$ and H_1 hold, then well-posedness implies relaxability.*

Proof. It is clear from the definitions that $m_r \leq m$. Suppose that strict inequality holds i.e. $0 < \delta = m - m_r$. Using hypothesis H_1 , we know that there exists relaxed viable trajectory $y(\cdot)$ s.t. $m_r = y(b)$ and $y(t) \in \text{int } K(t, \varepsilon)$, $d(y(t), \text{bd } K(t, \varepsilon)) \geq a(\varepsilon) > 0$ for all $t \in T$ and all $\varepsilon \in E \setminus \{0\}$ in a neighbourhood of zero.

Invoking theorem 3, we can find $x_n \in S_0$ s.t. $x_n \rightarrow y$ in $C(T, H)$. Then $x_n(t) \xrightarrow{s} y(t)$ in H , uniformly in t . Also $x_n(b) \xrightarrow{s} y(b)$ in H . Given the continuity of $g(\cdot)$, we deduce that there exists $x(\cdot) \in S$ s.t. $x(t) \in \text{int } K(t, \varepsilon)$, $d_H(x(t), \text{bd } K(t, \varepsilon)) \geq a(\varepsilon)/2$ for all $t \in T$ and all $\varepsilon \in E \setminus \{0\}$ near zero, while $|g(x(b)) - g(y(b))| < \delta/2$.

Let $u(\cdot) \in S_U^1$ be the control function generating $x(\cdot)$ and let $x(\cdot, \varepsilon)$ be the trajectory of the ε -perturbed evolution, also produced by control $u(\cdot)$. From proposition 5.5.1 of Tanabe [20], we know that:

$$x(t, \varepsilon) = \Phi(t, 0)x_0 + \int_0^t \Phi(t, s)f(s, x(s, \varepsilon), u(s), \varepsilon) ds$$

$$\text{and } x(t) = \Phi(t, 0)x_0 + \int_0^t \Phi(t, s)f(s, x(s), u(s)) ds.$$

Hence, we have:

$$\begin{aligned} |x(t, \varepsilon) - x(t)| &\leq \\ &M \int_0^t |f(s, x(s, \varepsilon), u(s), \varepsilon) - f(s, x(s), u(s))| ds \leq \\ &M \int_0^t |f(s, x(s, \varepsilon), u(s), \varepsilon) - f(s, x(s), u(s), \varepsilon)| ds + \\ &M \int_0^t |f(s, x(s), u(s), \varepsilon) - f(s, x(s), u(s))| ds \leq \\ &M \int_0^t k(s)|x(s, \varepsilon) - x(s)| ds + M \int_0^t |f(s, x(s), u(s), \varepsilon) - f(s, x(s), u(s))| ds. \end{aligned}$$

Note that $\int_0^t |f(s, x(s), u(s), \varepsilon) - f(s, x(s), u(s))| ds \rightarrow 0$ as $\varepsilon \rightarrow 0$. So, through Gronwall's inequality, we get that $x(\cdot, \varepsilon) \rightarrow x(\cdot)$ in $C(T, H)$ as $\varepsilon \rightarrow 0$. So, $x(t, \varepsilon) \in$

$K(t, \varepsilon)$ for all $t \in T$ and all $\varepsilon \in E$ in a neighbourhood of zero. Thus, for all $\varepsilon \in E$ sufficiently close to zero, we have:

$$\begin{aligned} m(\varepsilon) &\leq g(y(b)) + \delta/2 = m_r + \delta/2 = m - \delta/2 \\ &\Rightarrow \overline{\lim}_{\varepsilon \rightarrow 0} m(\varepsilon) \leq m - \delta/2. \end{aligned}$$

But because of the well-posedness hypothesis we have:

$$m = \lim_{\varepsilon \rightarrow 0} m(\varepsilon) \leq m - \delta/2$$

a contradiction. So, $m = m_r$ i.e. $P(0)$ is relaxable.

Q.E.D.

Acknowledgement

The author would like to thank the referee for his constructive remarks.

References

1. *Ahmed, N.*, Properties of the relaxed trajectories for a class of nonlinear evolution equations on a Banach space. *SIAM J. Control Optim.* **21** (1983), pp. 953–967.
2. *Aubin, J.-P., Cellina, A.*, *Differential Inclusions*. Springer, Berlin, 1984.
3. *Avgerinos, E., Papageorgiou, N. S.*, An existence theorem for an optimal control problem in Banach spaces. *Bull. Austr. Math. Soc.* **43** (1991), pp. 211–224.
4. *Barbu, V.*, *Nonlinear Semigroups and Differential Equations in Banach Spaces*. Noordhoff International Publishing, Leyden, The Netherlands, 1976.
5. *Chuong, P. V.*, Some results on density of extreme selections for measurable multifunctions. *Math. Nachr.* **126** (1986), pp. 312–326.
6. *Clarke, F. H.*, *Optimization and Nonsmooth Analysis*. Wiley, New York, 1983.
7. *Clarke, F. H.*, Admissible relaxation in variational and control problems. *J. Math. Anal. Appl.* **51** (1975), pp. 557–576.
8. *Delahaye, J.-P., Denel, J.*, The continuities of the point to set maps, definitions and equivalences. *Math. Progr. Study* **10** (1982), pp. 8–12.
9. *Dellacherie, C.*, Ensembles analytiques: Théorèmes de separation et applications, *Seminaire de Probabilités IX*, Univ. de Strasbourg. *Lecture Notes in Math.*, vol. 465, Springer, Berlin, 1975.
10. *Diestel, J.*, *Geometry of Banach Spaces – Selected Topics*. *Lecture Notes in Math.*, vol. 485, Springer, Berlin, 1975.
11. *Dontchev, A., Morduhovic, B.*, Relaxation and well-posedness of nonlinear optimal processes. *Systems and Control Letters* **3** (1983), pp. 177–179.
12. *Dunford, N., Schwartz, J.*, *Linear Operators I*. Wiley, New York, 1958.

13. *Edgar, G.*, Measurability in a Banach space II. *Indiana Univ. Math. Jour.* **28** (1979), pp. 559–579.
14. *Gamkrelidze, R.*, Principles of Optimal Control Theory. Plenum, New York, 1978.
15. *Klein, E., Thompson, A.*, Theory of Correspondences. Wiley, New York, 1984.
16. *Papageorgiou, N. S.*, Convergence theorems for Banach space valued integrable multifunctions. *Intern. Jour. Math. and Math. Sci.* **10** (1987), pp. 433–442.
17. *Papageorgiou, N. S.*, On multivalued evolution equations and differential inclusions in Banach spaces. *Comm. Math. Univ. Sancti Pauli* **36** (1987), pp. 21–39.
18. *Papageorgiou, N. S.*, A relaxation theorem for differential inclusions in Banach spaces. *Tohoku Math. Jour.* **39**, (1987), pp. 505–517.
19. *Papageorgiou, N. S.*, Properties of the relaxed trajectories of evolution equations and optimal control. *SIAM Jour. Control and Optim.* **27** (1989), pp. 267–288.
20. *Tanabe, H.*, Equations of Evolution. Pitman, London, 1979.
21. *Wagner, D.*, Survey of measurable selection theorems. *SIAM Jour. Control and Optim.* **15** (1977), pp. 857–903.
22. *Warga, J.*, Relaxed variational problems. *J. Math. Anal. Appl.* **4** (1962), pp. 111–128.

**О корректности и устойчивости оптимального значения
функционала качества
для систем с нелинейными распределенными параметрами**

Н. С. ПАПАГЕОРГИУ

(Афины)

В статье рассматривается взаимосвязь следующих двух понятий для нелинейных управляемых систем в гильбертовом пространстве: корректности (непрерывной зависимости оптимального значения функционала качества от параметра задачи) и устойчивости оптимального значения функционала качества по отношению к овыпуклению задачи.

В теоремах 1 и 3 указываются соответствующие условия, при которых из корректности следует устойчивость, и наоборот. В теоремах 2 и 2' указываются условия плотности множества траекторий исходной задачи во множестве траекторий овыпукленной задачи.

Nikolaos S. Papageorgiou
National Technical University
Department of Mathematics
Zografou Campus
Athens 157 73, Greece

NULL-CONTROLLABILITY OF INFINITE-DIMENSIONAL DISCRETE-TIME SYSTEM WITH RESTRAINED CONTROL

NGUYEN VAN SU

(*Budapest*)

(Received November 1, 1990)

The present paper is concerned with the null-controllability of the system

$$x_{k+1} = Ax_k + Bu_k,$$
$$x_k \in X, \quad u_k \in \Omega \subset U, \quad A \in L(X, X), \quad B \in L(U, X),$$

where X and U are Hilbert spaces, Ω is a convex set, $\text{int } \Omega \neq \emptyset$, $0 \in \Omega$.

Here we consider the case of semi-infinite operator A . For such systems, a necessary and sufficient condition is given. The obtained result can be applied to the investigation of null-controllability of delay system (or neutral systems) of the form

$$\dot{\mathbf{x}}(t) = L(\mathbf{x}_t) + B_0 \mathbf{u}(t),$$
$$\mathbf{x}(t) \in \mathbf{R}^n, \quad \mathbf{u}(t) \in \Omega \subset \mathbf{R}^m, \quad \mathbf{x}_t(\theta) = \mathbf{x}(t + \theta), \quad \theta \in [-h, 0].$$

It will be also shown that the exact and approximate null-controllability of delay systems with infinitely many commensurate delays are equivalent. This fact has been known only for systems with unconstrained control.

1. Introduction

This paper is concerned with the null-controllability of infinite-dimensional discrete-time linear system described by

$$(A, B, \Omega) \quad x_{n+1} = Ax_n + Bu_n,$$

where $x_n \in X$, $u_n \in \Omega \subset U$, X and U are Hilbert spaces, $A \in L(X, X)$, $B \in L(U, X)$, Ω is a convex set, $\text{int } \Omega \neq \emptyset$ and $0 \in \Omega$.

In recent years, the study of the linear discrete-time systems in infinite-dimensional spaces has attracted the attention of many authors (see [4-5], [9], [19-21], [23-28]).

The question of the null-controllability for systems of the form (A, B, Ω) without any further condition for A — even more for systems given in Banach spaces X ,

U — has not been solved yet. In [28], the authors have studied the local null-controllability of system (A, B, Ω) under the additional assumption that the operator A satisfies the so-called finite condition. In this paper, we consider the case when operator A satisfies the semi-finite condition. In the first part of this paper, a necessary and sufficient condition for local null-controllability of such systems is given. In the second part, the obtained result will be applied to the investigation of the null-controllability of delay system of the form

$$(L, B_0, \Omega) \quad \dot{\mathbf{x}}(t) = L(\mathbf{x}_t) + B_0 \mathbf{u}(t),$$

where $\mathbf{x}(t) \in \mathbb{R}^n$, $\mathbf{u}(\cdot)$ belongs to a given set of admissible controls, $\mathbf{x}_t(\theta) = \mathbf{x}(t + \theta)$, $\theta \in [-h, 0]$.

The controllability of delay systems was intensively investigated in the literature [1–2], [4–6], [12], [15–20], [22–24], [26], [29]. Most of them deals with delay system with unconstrained control; e.g. [1–2], [12], [15–19], [22–23]. The case of delay systems with constrained control has been studied in recent years by [4], [6], [24], [26]. In [4], the set of admissible control functions is the closed unit ball of the function space with zero in its interior, in [6] it is a closed convex cone also in the function space with vertex at zero. The papers [24], [26] consider admissible control functions, all components of which are positive.

In this paper, we shall give necessary and sufficient conditions for the local and global controllability of system (L, B_0, Ω) under much weaker restrictions for the set of admissible controls. Moreover, we shall show that the exact and approximate null-controllability of delay systems with finitely many commensurate delay are equivalent. This fact was known [5] for systems with unconstrained control.

2. Notations, definitions

Let X be a Hilbert space. The $\langle \cdot, \cdot \rangle$ denotes the inner product in X . If $H \subset X$, then \overline{H} is the closure of H , $\text{sp}\{K\}$ is the span of K , $\text{int } K$ is the interior of K and $\text{ri } K$ is the relative interior of K in X . If $A \in L(X, X)$, then A^* denotes the adjoint of A , $\text{Ker } A$ and $\text{Im } A$ are the kernel and the range of A , $\sigma(A)$ is the spectrum of A . Let \mathbb{C} be the set of complex numbers. We shall denote

$$D_r := \{\lambda \in \mathbb{C} : |\lambda| < r\},$$

$$\overline{D}_r := \{\lambda \in \mathbb{C} : |\lambda| \leq r\}.$$

B_ε denotes an open ball of radius ε centered at the origin.

Let $U^n = U \times U \times \dots \times U$, where the direct product is taken n -times, and let us consider the operator $F_n : U^n \rightarrow X$ defined by

$$F_n(u^{(n)}) = A^{n-1}Bu_0 + A^{n-2}Bu_1 + \dots + Bu_{n-1},$$

where $u^{(n)} = (u_0, u_1, \dots, u_{n-1}) \in U^n$. Clearly,

$$F_n(U^n) = A^{n-1}BU + A^{n-2}BU + \dots + BU,$$

and

$$F_n(\Omega^n) = A^{n-1}B\Omega + A^{n-2}B\Omega + \dots + B\Omega,$$

where $\Omega^n = \Omega \times \Omega \times \dots \times \Omega$. (Here the direct product is also taken n -times.)

We need some definitions in following.

DEFINITION 1. The set

$$S_n(\Omega^n) := \{x \in X : -A^n x \in F_n(\Omega^n)\}$$

is called the null-controllability set of (A, B, Ω) after n step. Furthermore, the set

$$S(\Omega) := \bigcup_{n=1}^{\infty} S_n(\Omega^n)$$

is called the null-controllability set after varying step.

DEFINITION 2. System (A, B, Ω) is said to be locally controllable (LC) if $0 \in \text{int } S(\Omega)$. System (A, B, Ω) is said to be globally controllable (GC) if $S(\Omega) = X$.

DEFINITION 3. We say that the operator A satisfies the spectrum decomposition condition if for some positive number $r < 1$ the set $\sigma_1 = \sigma(A) \setminus D_r$ consists of finite number of points and the corresponding eigenspace X_1 is finite-dimensional.

DEFINITION 4. We say that the operator A satisfies the semi-finite condition if there exists number m such that

$$\begin{aligned} \text{Ker } A^m &= \text{Ker } A^{m+1} = \text{Ker } A^{m+2} = \dots \\ \overline{\text{Im } A^m} &= \overline{\text{Im } A^{m+1}} = \overline{\text{Im } A^{m+2}} = \dots \end{aligned}$$

Now, let M be a closed A -variant subspace of X . We shall denote by \hat{X} the factor space X/M equipped with the usual factor norm (it is known that \hat{X} is a Hilbert space), and by P the canonical projection $P : X \rightarrow \hat{X}$. We shall define the factor system $(\hat{A}, \hat{B}, \Omega)$ of (A, B, Ω) with respect to M by

$$(\hat{A}, \hat{B}, \Omega) \quad \hat{x}_n = \hat{A}\hat{x}_n + \hat{B}u_n,$$

where $\hat{x}_n \in \hat{X}$, $u_n \in \Omega \subset U$, $\hat{x} := Px$, $\hat{A}\hat{x} := P(Ax)$ and $\hat{B}u := P(Bu)$.

3. Main results

To obtain the main result, we need the following lemma.

LEMMA 1. Consider the system (A, B, Ω) and suppose that $\text{Ker } A = \{0\}$, $\overline{\text{Im } A} = X$, $\text{int } \Omega \neq \emptyset$. If there exists number k such that $\text{Im } A^k \subset F_k(U^k)$, then $\text{int } S_k(\Omega^k) \neq \emptyset$.

Proof. Consider the factor space $U^k / \text{Ker } F_k$ equipped with the usual factor norm and let P the canonical projection $P : U^k \rightarrow U^k / \text{Ker } F_k$. Let us define the operator $\hat{F}_k : U^k / \text{Ker } F_k \rightarrow X$ as follows:

$$\text{for all } \hat{u} \in U^k / \text{Ker } F_k, \hat{u} = u + \text{Ker } F_k \text{ we take } \hat{F}_k \hat{u}; = F_k u.$$

Then F_k is well-defined, linear, bounded one-to-one operator and $F_k = \hat{F}_k P$. Moreover,

$$\text{Im } A^k \subset \hat{F}_k(U^k / \text{Ker } F_k). \tag{1}$$

In fact,

$$\text{Im } A^k \subset F_k(U^k) = \hat{F}_k(P(U^k)) = \hat{F}_k(U^k / \text{Ker } F_k).$$

By assumption $\overline{\text{Im } A} = X$, therefore

$$\overline{\hat{F}_k(U^k / \text{Ker } F_k)} = X.$$

Since \hat{F}_k is a linear, bounded, one-to-one operator and the range of \hat{F}_k is dense X , Halmos' problem 42 [11] and Theorem 16, pp. 254 [8] can be applied to verify the relation

$$U^k / \text{Ker } F_k \approx X, \tag{2}$$

where \approx is an isometric isomorphism between the two spaces.

On the other side, from (1), it follows by the factorization theorem of Douglas [7] that there exists a linear, bounded operator $C : X \rightarrow U^k / \text{Ker } F_k$ such that $A^k = \hat{F}_k C$.

Now, we consider the Hilbert space \overline{CX} and the restriction of operator \hat{F}_k to \overline{CX} ($\hat{F}_k : \overline{CX} \rightarrow X$).

Since

$$X = \overline{\text{Im } A^k} = \overline{\text{Im } \hat{F}_k C} \subset \overline{\hat{F}_k(\overline{CX})}$$

we have

$$\overline{\hat{F}_k(\overline{CX})} = X.$$

Applying again the same results as above, we obtain that \overline{CX} is isometrically isomorphic with X :

$$\overline{CX} \approx X. \tag{3}$$

From (2) and (3) it follows that

$$\overline{CX} \approx U^k / \text{Ker } F_k.$$

But $\overline{CX} \subset U^k / \text{Ker } F_k$, therefore, it can be easily seen that

$$\overline{CX} = U^k / \text{Ker } F_k.$$

Let $\hat{\Omega}^k := P(\Omega^k)$. Since $\text{int } \Omega \neq \emptyset$ and P is open, $\text{int } \hat{\Omega}^k$ is not empty in $U^k / \text{Ker } F_k$. Hence, $CX \cap \text{int } \hat{\Omega}^k \neq \emptyset$. Consequently, the inverse image $C^{-1}(\text{int } \hat{\Omega}^k)$ has a non-empty interior in X . It is easy to see that

$$-C^{-1}(\text{int } \hat{\Omega}^k) \subset S_k(\Omega^k). \tag{4}$$

In fact, if

$$x \in (-C^{-1}(\text{int } \hat{\Omega}^k)),$$

then

$$-Cx = \hat{u}^{(k)} \quad \text{for some } \hat{u}^{(k)} \in \text{int } \hat{\Omega}^k.$$

Hence,

$$-\hat{F}_k Cx = \hat{F}_k \hat{u}^{(k)} = F_k(u^{(k)})$$

for some $u^{(k)} \in \text{int } \Omega^k$. But $\hat{F}_k C = A^k$, therefore,

$$A^k x + F_k(u^{(k)}) = 0.$$

This means that $x \in S_k(\Omega^k)$ and (4) shows that

$$\text{int } S_k(\Omega^k) \neq \emptyset.$$

The proof of Lemma 1 is complete.

THEOREM 1. Consider the system (A, B, Ω) . Suppose that $\text{Ker } A \neq \{0\}$, $\overline{\text{Im } A} = X$, $0 \in \Omega$ and $\text{int } \Omega \neq \emptyset$. Then the system (A, B, Ω) is LC iff

- (a) There exists number k such that $\text{Im } A^k \subset F_k(U^k)$,
- (b) There is no eigenvector x^* of A^* , $A^*x^* = \lambda x^*$, $\lambda > 0$ such that $\langle x^*, B\Omega \rangle \geq 0$.

Proof. Necessity. If system (A, B, Ω) is LC, then the system (A, B, U) with unconstrained control is GC, i.e. $S(U) = \bigcup_{n=1}^{\infty} S_n(U^n) = X$. By the theorem of Fuhrman [9], we obtain that there exists number k such that

$$\text{Im } A^k \subset F_k(U^k).$$

In order to prove condition (b), we assume the contrary: let us suppose there exists $\lambda > 0$ such that $A^*x^* = \lambda x^*$ and

$$\langle x^*, B\Omega \rangle \geq 0. \tag{5}$$

Then, for all $x \in S(\Omega)$, there exists n and $u_0, u_1, \dots, u_{n-1} \in \Omega$ such that

$$\langle x^*, x \rangle = \frac{1}{\lambda^n} \langle \lambda^n x^*, x \rangle = \frac{1}{\lambda^n} \langle A^{*n} x^*, x \rangle = \frac{1}{\lambda^n} \langle x^*, A^n \rangle. \tag{6}$$

But

$$\begin{aligned} \langle x^*, A^n x \rangle &= \langle x^*, A^{n-1} B u_0 + \dots + B u_{n-1} \rangle \\ &= \langle A^{*n} x^*, B u_0 \rangle + \dots + \langle x^*, B u_{n-1} \rangle \\ &= \lambda^{n-1} \langle x^*, B u_0 \rangle + \dots + \langle x^*, B u_{n-1} \rangle. \end{aligned} \tag{7}$$

From (5), (6), (7) it follows that

$$\langle x^*, x \rangle \geq 0 \text{ for all } x \in S(\Omega).$$

This contradicts the condition $0 \in \text{int } S(\Omega)$.

Sufficiency. By Lemma 1 it has been proved that $\text{int } S_k(\Omega^k) \neq \emptyset$. For every $l \geq k$, let us define the set S'_l by

$$S'_l := \{x \in X : -A^k x \in F_l(\Omega^l)\}.$$

It is easy to see that $S'_l \subset S'_{l+1}$, S'_l is convex and $AS'_l \subset S'_{l+1}$. Since $S'_k = S_k(\Omega^k)$, it follows that $\text{int } S'_l \neq \emptyset$. Setting $S' = \bigcup_{l \geq k} S'_l$, we will show that $0 \in \text{int } S'$. Assuming

the contrary, we readily verify that the cone $C = \bigcup_{\lambda > 0} \lambda S'$ is convex, not dense in X

and A -invariant, i.e. $AC \subset C$. By the Krein–Rutman’s theorem [13], there exists $\lambda > 0$ and $x^* \in X^*$ such that $A^* x^* = \lambda x^*$ and $\langle x^*, c \rangle \leq 0$ for all $c \in C$. On the other hand, since $A^k B \Omega \subset F_{k+1}(\Omega^{k+1})$, it follows that $-B \Omega \subset S'_{k+1} \subset S' \subset C$. Hence, $\langle x^*, B u \rangle \geq 0$ for all $u \in \Omega$. This contradicts the assumption (b) of the theorem. Thus, $0 \in \text{int } S'$. In view of Lemma 1 of [25], there exists $m \geq k$ such that $0 \in \text{int } S'_m$.

If $m = k$ then the assertion of the theorem is immediate, since $S'_k = S_k(\Omega)$.

If $m > k$ we consider operator $A^{m-k} : X \rightarrow X$. Since $0 \in \text{int } S$, we have $0 \in \text{int}((A^{m-k})^{-1} S'_m)$. On the other hand, it is easy to see that $(A^{m-k})^{-1} S'_m \subset S_m(\Omega)$. Therefore, $0 \in \text{int } S_m(\Omega^m)$, and the sufficiency is proved. The proof of Theorem 1 is complete.

We remark that Theorem 1 can be strengthened by assuming only that Ω has non-empty relative interior. Now, we formulate the modified version of Theorem 1.

Corollary 1. Consider the system (A, B, Ω) . Assume that $\text{Ker } A = \{0\}$, $\overline{\text{Im } A} = X$, $0 \in \Omega$ and $\text{ri } \Omega \neq \emptyset$. Then the system (A, B, Ω) is LC iff $\overline{\text{Im } A^k} \subset F_k(V^k)$, where $V = \overline{\text{sp}\{\Omega\}}$.

- (a) There exists number k such that $\text{Im } A^k \subset F_k(V^k)$, where $V = \overline{\text{sp}\{\Omega\}}$.
- (b) There is no eigenvector x^* of A^* , $A^* x^* = \lambda x^*$, $\lambda > 0$ such that $\langle x^*, B \Omega \rangle \geq 0$.

LEMMA 2. Let M be a closed A -invariant subspace contained in the controllability set $S(\Omega)$ of system (A, B, Ω) . Then the system (A, B, Ω) is LC iff the factor system $(\hat{A}, \hat{B}, \Omega)$ with respect to M is LC.

Proof. Sufficiency is immediate from the fact that P is an open operator from X onto \hat{X} and the set $PS(\Omega)$ belongs to the controllability set $\hat{S}(\Omega)$ of the system $(\hat{A}, \hat{B}, \Omega)$.

In order to prove the necessity, we take $\varepsilon > 0$ such that $\hat{B}_\varepsilon \subset \hat{S}(\Omega)$ and set $\varepsilon_1 = \varepsilon/\|P\|$. We shall show that $B_{\varepsilon_1} \subset S(\Omega)$. In fact, for any $x \in X$ with $\|x\| < \varepsilon_1$, we have $Px \in \hat{S}(\Omega)$, therefore,

$$\hat{A}^k Px + \hat{A}^{k-1} \hat{B}u_0 + \dots + \hat{B}u_{k-1} = 0$$

for some k and some $u_i \in \Omega, i = 0, 1, \dots, k - 1$. From this it follows that

$$A^k x + A^{k-1} Bu_0 + \dots + Bu_{k-1} \in M \subset S(\Omega),$$

that is $x \in S(\Omega)$. The proof is complete.

THEOREM 2. Consider the system (A, B, Ω) . Assume that $0 \in \Omega, \text{int } \Omega \neq \emptyset, \Omega$ is convex and operator A satisfies the semi-finite condition. Then the system (A, B, Ω) is LC iff

- (a) There exists number k such that $\text{Im } A^k \subset F_k(U^k)$,
- (b) There is no eigenvector x^* of A^* , $A^*x^* = \lambda x^*, \lambda > 0$ such that $\langle x^*, B\Omega \rangle \geq 0$.

Proof. The proof of necessity is analogous to that of Theorem 1.

In order to prove the sufficiency we set $l = \max(m, k)$, where m is the number which is to be found in the definition of the semi-finite condition and let $M = \text{Ker } A^l$. It can be easily verified that M has the properties required in Lemma 2. Consider the factor system $(\hat{A}, \hat{B}, \Omega)$ with respect to M :

$$\hat{x}_{n+1} = \hat{A}\hat{x}_n + \hat{B}u_n, \quad \hat{x}_n \in \hat{X}, \quad u_n \in \Omega \subset U.$$

From the semi-finite condition concerning operator A , it follows that $\text{Ker } \hat{A} = \{0\}$ and $\text{Im } \hat{A} = \hat{X}$. In the same way as in the proof of the Theorem 2.3 [28], it can be shown that conditions (a) and (b) are fulfilled also for the factor system $(\hat{A}, \hat{B}, \Omega)$:

- (a) There exists number k such that $\text{Im } \hat{A}^k \subset \hat{F}_k(U^n)$, where

$$\hat{F}_k(U^k) = \hat{A}^{k-1} \hat{B}U + \dots + \hat{B}U,$$

- (b) There is no eigenvector \hat{x}^* of \hat{A}^* , $\hat{A}^*\hat{x}^* = \lambda \hat{x}^*, \lambda > 0$ such that $\langle \hat{x}^*, \hat{B}\Omega \rangle \geq 0$.

Thus, making use of Theorem 1, it follows that the system $(\hat{A}, \hat{B}, \Omega)$ is LC. Consequently, by Lemma 2, we conclude that the system (A, B, Ω) is LC. This completes the proof.

Remark. In the case, when $\text{ri } \Omega \neq \emptyset$, Theorem 2 remains true if we replace U^k in condition (a) by V^k , where $V = \text{sp}\{\Omega\}$.

In [27], it has been proved that, under certain condition detailed below, system (A, B, Ω) is GC iff it is LC and $\sigma(A) \subset \overline{D}_1$. Taking into consideration this result, we obtain the following corollary of Theorem 2.

Corollary 2. Assume that the conditions of Theorem 2 are fulfilled, Ω is bounded and the operator A satisfies the spectrum decomposition condition. Then the system (A, B, Ω) is GC iff

- (a) There exists number k such that $\text{Im } A^k \subset F_k(U^k)$,
- (b) There is no eigenvector x^* of A^* , $A^*x^* = \lambda x^*$, $\lambda > 0$ such that $\langle x^*, B\Omega \rangle \geq 0$,
- (c) $\sigma(A) \subset \overline{D}_1$.

4. Application

In this Section we shall use the result obtained in the first Section to examine the null-controllability of linear autonomous retarded functional differential equation of the form:

$$(L, B_0, \Omega) \quad \dot{\mathbf{Z}}(t) = L(\mathbf{Z}_t) + B_0 \mathbf{u}(t),$$

where $\mathbf{Z}(t) \in \mathbf{R}^n$, $\mathbf{Z}_t(\theta) = \mathbf{Z}(t + \theta)$, $\theta \in [-h, 0]$, L is a bounded linear operator from $C = C([-h, 0], \mathbf{R}^n)$ into \mathbf{R}^n given by

$$L(\phi) = \int_{-h}^0 d\eta(\theta)\phi(\theta), \quad \phi \in C,$$

where $\eta(\cdot)$ is an $n \times n$ matrix function of bounded variation such that $\eta(\theta) = 0$ for $\theta \geq 0$, $\eta(\theta) = \eta(-h)$ for $\theta \leq -h$ and η is left-sided continuous on $(-h, 0)$. B_0 is an $n \times m$ matrix.

Moreover, $\mathbf{u}(\cdot) \in \Omega$ is an admissible control if Ω has the following properties:

$$\Omega \subset \bigcup_{T>0} L_2([0, T], \mathbf{R}^m); \tag{8a}$$

$$\begin{aligned} &\Omega \cap L_2([0, h], \mathbf{R}^m) \text{ is convex} \\ &\text{and its interior relative to } L_2([0, h], \mathbf{R}^m) \text{ is non-empty;} \end{aligned} \tag{8b}$$

$$0 \in \Omega; \tag{8c}$$

$$\begin{aligned} &\text{If } \mathbf{u}(\cdot) \in \Omega \text{ then } \mathbf{u}_i(\cdot) \in L_2([0, h], \mathbf{R}^m) \\ &\text{defined by } \mathbf{u}_i(\theta) = \mathbf{u}(ih + \theta), \theta \in [0, h] \\ &\text{is such that } \mathbf{u}_i \in \Omega, \text{ for each } i \in \mathbf{N}. \end{aligned} \tag{8d}$$

Let $M_2 := \mathbf{R}^n \times L_2([-h, 0], \mathbf{R}^n)$. Clearly, M_2 is a Hilbert space.

If $M \subset M_2$ is a subset of M_2 , then the negative polar cone of M is defined by

$$M^0 = \{f \in M_2 : \langle f, x \rangle \leq 0 \quad \forall x \in M\}.$$

The inner product in \mathbb{R}^n is denoted by $(\cdot, \cdot)_{\mathbb{R}^n}$.

It is well known [3] that the homogeneous equation

$$\dot{\mathbf{Z}}(t) = L(\mathbf{Z}_t) \tag{9}$$

induces a strongly continuous semigroup $\{S(t), t \geq 0\}$ on M_2 , by means of which the solution of (9) with the initial condition $\mathbf{Z}(0) = \phi^0, \mathbf{Z}(\theta) = \phi^1(\theta), \theta \in [-h, 0]$, where $\phi = (\phi^0, \phi^1) \in M_2$ can be given as $S(t)\phi = (\mathbf{Z}(t), \mathbf{Z}_t) \in M_2$.

Let $\mathbf{Z}(t)$ be a solution of the equation

$$\dot{\mathbf{Z}}(t) = L(\mathbf{Z}_t) + B_0\mathbf{u}(t)$$

corresponding to the initial condition $\phi = (\phi^0, \phi^1) \in M_2$ and to some control $\mathbf{u}(\cdot) \in L_2([0, T], \mathbb{R}^m)$. Then $\mathbf{x}(t) = (\mathbf{Z}(t), \mathbf{Z}_t)$ will be the mild solution of the abstract differential equation

$$\begin{aligned} \dot{\mathbf{x}}(t) &= A\mathbf{x}(t) + B\mathbf{u}(t), & t \in [0, T] \\ \mathbf{x}(0) &= \phi, \end{aligned}$$

where A is the infinitesimal generator of $S(t)$, $B : \mathbb{R}^m \rightarrow M_2$ is a bounded, linear operator defined by

$$B\mathbf{u} = (B_0\mathbf{u}, 0).$$

This mild solution can be expressed by

$$\mathbf{x} = S(t)\phi + \int_0^t S(t-\theta)B\mathbf{u}(\theta) d\theta.$$

Let $\{S^+(t), t \geq 0\}$ denote the strongly continuous semigroup induced by the transposed equation:

$$\dot{\mathbf{Z}}(t) = L^+(\mathbf{Z}_t), \quad \mathbf{Z}(t) \in \mathbb{R}^n,$$

where

$$L^+\phi = \int_{-h}^0 d\eta^T(\theta)\phi(\theta), \quad \phi \in C.$$

The corresponding infinitesimal generator of $S^+(t)$ is denoted by A^+ .

For a given system (L, B_0, Ω) we define the following sets:

$$R_T := \left\{ \int_0^T S(T-\theta)B\mathbf{u}(\theta) d\theta : \mathbf{u}(\cdot) \in \Omega \cap L_2([0, T], \mathbb{R}^m) \right\},$$

$$N_T := \{ \phi \in M_2 : -S(t)\phi \in R_T \},$$

$$N := \bigcup_{T>0} N_T,$$

$$C_T := \{ \phi \in M_2 : -S(t)\phi \in \bar{R}_T \}.$$

DEFINITION 5. System (L, B_0, Ω) is said to be locally controllable (LC) if $0 \in \text{int } N$. System (L, B_0, Ω) is said to be globally controllable (GC) if $N = M_2$.

DEFINITION 6. System (L, B_0, Ω) is said to be approximately locally controllable (ALC) at time $T > h$ if $0 \in \text{int } C_T$.

Let $\Delta(\lambda)$ be the characteristic matrix of (9):

$$\Delta(\lambda) = \lambda I - \int_{-h}^0 d\eta(\theta) e^{\lambda\theta}.$$

Then $\sigma(A) = \sigma(A^+)$ is a point spectrum (see [10], [14], [23]) and it is given by

$$\sigma(A) = \{\lambda \in \mathbb{C} : \det \Delta(\lambda) = 0\}.$$

If $\lambda \in \sigma(A)$ then the corresponding eigenspace is given by

$$\text{Ker}(\lambda I - A^+) = \{(\phi^0, \phi^1) \in M_2 : \Delta^T(\lambda)\phi^0 = 0, \phi^1(\theta) = \phi^0 e^{\lambda\theta}, \theta \in [-h, 0]\}.$$

Now, we can give the discretization for the retarded system (L, B_0, Ω) , i.e. we are going to construct a linear discrete-time system which is equivalent to the retarded system (L, B_0, Ω) from the controllability point of view.

Let us denote

$$\begin{aligned} U &:= L_2([0, h], \mathbf{R}^m), \\ \tilde{\Omega} &:= \Omega \cap L_2([0, h], \mathbf{R}^m). \end{aligned}$$

We define the operator

$$\mathcal{A} : M_2 \rightarrow M_2 \text{ by } \mathcal{A} := S(h)$$

and

$$\mathcal{B} : U \rightarrow M_2 \text{ by } \mathcal{B}\mathbf{u}(\cdot) := \int_0^h S(h - \theta) B\mathbf{u}(\theta) d\theta.$$

It has been proved [10], [23] that the operator \mathcal{A} is compact, therefore, it satisfies the spectrum decomposition condition. Moreover, (see [23]) there exists $T_0 > 0$ such that for all $t \geq T_0$

$$\text{Ker } S(t) = \text{Ker } S(T_0)$$

and

$$\overline{\text{Im } S(t)} = \overline{\text{Im } S(T_0)}.$$

Hence, it follows that there exists number n_0 such that, for $n \geq n_0$, we have

$$\text{Ker } \mathcal{A}^n = \text{Ker } \mathcal{A}^{n+1} = \text{Ker } \mathcal{A}^{n+2} = \dots$$

and

$$\overline{\text{Im } \mathcal{A}^n} = \overline{\text{Im } \mathcal{A}^{n+1}} = \overline{\text{Im } \mathcal{A}^{n+2}} = \dots$$

It means that the operator \mathcal{A} satisfies the semi-finite condition.

Now, we consider the discrete-time system associated with the system (L, B_0, Ω)

$$(\mathcal{A}, \mathcal{B}, \tilde{\Omega}) \quad \mathbf{x}_{n+1} = \mathcal{A}\mathbf{x}_n + \mathcal{B}\mathbf{u}_n,$$

where $\mathbf{x}_n \in M_2, \mathbf{u}_n \in \tilde{\Omega} \subset U$.

LEMMA 3. The system (L, B_0, Ω) is LC or GC iff the system $(\mathcal{A}, \mathcal{B}, \tilde{\Omega})$ is LC or GC, respectively.

Proof. The reachable set of $(\mathcal{A}, \mathcal{B}, \tilde{\Omega})$ in time k is:

$$R_k^d := \{0\} \quad \text{for } k = 0$$

and

$$R_k^d := \mathcal{A}^{k-1}\mathcal{B}\tilde{\Omega} + \mathcal{A}^{k-2}\mathcal{B}\tilde{\Omega} + \dots + \mathcal{B}\tilde{\Omega} \quad \text{for } k \geq 1.$$

It suffices to show that

$$R_{kh} = R_k^d. \tag{10}$$

Let $\phi \in R_{kh}$, then there exists $\mathbf{u}(\cdot) \in \Omega \cap L_2([0, kh], \mathbb{R}^m)$ such that

$$\phi = \int_0^{kh} S(kh - \theta)B\mathbf{u}(\theta) d\theta.$$

Defining a control sequence $\mathbf{u}_i \in \tilde{\Omega}, i = 1, 2, \dots, k$ by $\mathbf{u}_i(\theta) = \mathbf{u}((i - 1)h + \theta), \theta \in [0, h]$, we can easily show that

$$\phi = \mathcal{A}^{k-1}\mathcal{B}\mathbf{u}_1 + \mathcal{A}^{k-2}\mathcal{B}\mathbf{u}_2 + \dots + \mathcal{B}\mathbf{u}_k,$$

thus $\phi \in R_k^d$. Conversely, let $\phi \in R_k^d$, then there exists a control sequence $\mathbf{u}_i \in \tilde{\Omega}, i = 1, 2, \dots, k$ such that

$$\phi = \mathcal{A}^{k-1}\mathcal{B}\mathbf{u}_1 + \mathcal{A}^{k-2}\mathcal{B}\mathbf{u}_2 + \dots + \mathcal{B}\mathbf{u}_k.$$

Taking $\mathbf{u}(t) = \mathbf{u}_i(t - (i - 1)h)$ for $(i - 1)h \leq t < ih, i = 1, 2, \dots, k$ we have $\mathbf{u}(\cdot) \in \Omega \cap L_2([0, kh], \mathbb{R}^m)$ and

$$\phi = \int_0^{kh} S(kh - \theta)B\mathbf{u}(\theta) d\theta.$$

Hence, $\phi \in R_{kh}$. The proof of Lemma 3 is complete.

By Theorem 2, Corollary 2 and Lemma 3, we obtain the following.

THEOREM 3. Assume that $\text{int } \tilde{\Omega} \neq \emptyset$. Then the system (L, B_0, Ω) is LC iff

- (a) The system $(L, B_0, \bigcup_{T>0} L_2([0, T], \mathbb{R}^m))$ is GC,
- (b) $\text{Ker}(\lambda I^* - S^*(h)) \cap (\mathcal{B}\tilde{\Omega})^0 = \{0\}, \forall \lambda > 0$.

THEOREM 4. If $\text{int } \tilde{\Omega} \neq \emptyset$ and $\tilde{\Omega}$ is bounded in $L_2([0, h], \mathbb{R}^m)$ then the system (L, B_0, Ω) is GC iff

- (a) The system $(L, B_0, \bigcup_{T>0} L_2([0, T], \mathbb{R}^m))$ is GC,
- (b) $\text{Ker}(\lambda I^* - S^*(h)) \cap (\mathcal{B}\tilde{\Omega})^0 = \{0\}, \forall \lambda > 0$,
- (c) $\{\lambda \in \mathbb{C} : \det \Delta(\lambda) = 0\} \subset \mathbb{C}^-$, where $\mathbb{C}^- := \{\lambda \in \mathbb{C} : \text{Re } \lambda \leq 0\}$.

Now, we introduce the operator $D : U \rightarrow M_2$ by

$$\begin{aligned} (D\mathbf{u})^0 &= 0, \\ (D\mathbf{u})^1(\theta) &= B_0\mathbf{u}(\theta), \quad -h \leq \theta \leq 0. \end{aligned}$$

In [26], pp. 16, it has been proved that the condition

$$\text{Ker}(\lambda I^* - S^*(h)) \cap (\mathcal{B}\tilde{\Omega})^0 = \{0\}, \quad \forall \lambda > 0$$

is equivalent to

$$\text{Ker}(\lambda I^* - A^+) \cap (D\tilde{\Omega})^0 = \{0\}, \quad \forall \lambda \in \mathbb{R},$$

where $(D\tilde{\Omega})^0$ is the negative polar cone of $D\tilde{\Omega}$. It follows from the definition of D that this condition can be expressed as follows:

- (b') If $\Delta(\lambda) = 0$ for some real λ , then there is no vector $\phi^0 \in \mathbb{R}^n$ such that

$$\Delta^T(\lambda)\phi^0 = 0$$

and

$$\int_{-h}^0 (\phi^0 e^{\lambda\theta}, B_0\mathbf{u}(-\theta))_{\mathbb{R}^n} d\theta \leq 0, \text{ for all } \mathbf{u}(\cdot) \in \tilde{\Omega}. \tag{11}$$

This version of condition (b) is more convenient for calculations.

Let us consider how to use this result for the verification of the controllability of concrete systems with retarded arguments.

Example 1. Let

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathbf{x}(t) + \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x}(t-1) + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \mathbf{x}(t-2) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mathbf{u}$$

[26], and let $\tilde{\Omega}$ be an arbitrary set which satisfies the requirements (8a)–(8d) with $m = 1$. An easy calculation shows that

$$\Delta(\lambda) = \begin{pmatrix} \lambda + e^{-\lambda} & 1 + e^{-2\lambda} \\ -1 & \lambda - e^{-\lambda} \end{pmatrix},$$

thus,

$$\det \Delta(\lambda) = \lambda^2 + 1.$$

Since there is no real λ satisfying equation $\det \Delta^T(\lambda) = 0$, condition (b') evidently holds.

It is known [5], [23], [30], that condition (a) holds iff

$$\text{rank}(\Delta(\lambda), B_0) = n, \quad \text{for all } \lambda \in \mathbb{C}.$$

In our case

$$(\Delta(\lambda), B_0) = \begin{pmatrix} \lambda + e^{-\lambda} & 1 + e^{-2\lambda} & 1 \\ -1 & \lambda - e^{-\lambda} & 0 \end{pmatrix},$$

which has full rank for all $\lambda \in \mathbb{C}$.

This example shows that the controllability property of some systems can be verified "almost" independently of the control restraint set $\tilde{\Omega}$ ($\tilde{\Omega}$ has to have only the properties (8a)–(8d)). Nevertheless, it should be noted that in [26], the approximate local controllability of this system was only obtained under more special control constraint set $\tilde{\Omega}$ (namely, $\tilde{\Omega}$ consists of functions, all component of which take values from the cone $K = \{\mathbf{a} \in \mathbb{R}^n : a_i \geq 0, i = 1, 2, \dots, n\}$).

Example 2. Let

$$\dot{\mathbf{x}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \mathbf{x}(t) + \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{x}(t-1) + \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} \mathbf{x}(t-2) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} \mathbf{u}(t).$$

Let $h = 2\pi$ and let us define the set Ω_T by

$$\Omega_T = \left\{ \mathbf{u}(\cdot) \in L_2([0, T], \mathbb{R}) : \int_{2(j-1)\pi}^{2j\pi} |u(\theta) - \sin \theta|^2 d\theta \leq \int_{2(j-1)\pi}^{2j\pi} \sin^2 \theta d\theta, \right. \\ \left. j = 1, 2, \dots, i; \quad 2(i-1)\pi \leq T \leq 2i\pi, \right. \\ \left. \int_{2(i-1)\pi}^T |u(t) - \sin \theta|^2 d\theta \leq \int_{2(i-1)\pi}^T \sin^2 \theta d\theta \right\}.$$

and let

$$\Omega = \bigcup_{T>0} \Omega_T.$$

Then

$$\tilde{\Omega} = \Omega_{2\pi}.$$

First of all we have to show that Ω has the required properties (8a)–(8d). Condition (8a) and (8c) is evident. We observe that $\Omega_{2\pi}$ is a closed ball in $L_2([0, 2\pi], \mathbb{R})$ around

the center u_0 , $u_0(t) = \sin t$, and with the radius $\sqrt{\pi}$. Thus, (8b) is also true. (It has to be emphasized that the origin does not belong to the interior of $\tilde{\Omega}$. Therefore, the controllability problem for this system could not be solved on the basis of the previous results.)

In consequence of the construction of Ω_T , the condition expressed by (8d) also holds true.

Condition (a) of Theorem 3 can be verified on the same way as in the previous Example.

Now, we shall examine the condition (b'). Since

$$\Delta(\lambda) = \begin{pmatrix} \lambda + e^{-\lambda} & -1 + e^{-2\lambda} \\ -1 & \lambda - e^{-\lambda} \end{pmatrix}$$

we have

$$\det \Delta(\lambda) = \lambda^2 - 1,$$

therefore,

$$\det \Delta(\lambda) = 0 \quad \text{iff} \quad \lambda_1 = 1, \quad \lambda_2 = -1.$$

An easy calculation shows that $\phi_1^0 = \begin{pmatrix} 1 \\ 1 + e^{-1} \end{pmatrix}$ satisfies equation

$$\Delta^T(\lambda_1)\phi_1^0 = 0$$

and $\phi_{-1}^0 = \begin{pmatrix} 1 \\ -1 + e \end{pmatrix}$ satisfies equation

$$\Delta^T(\lambda_2)\phi_{-1}^0 = 0.$$

Let us choose the control function $\bar{u} \in \tilde{\Omega}$

$$\bar{u}(\theta) = \begin{cases} \sin \theta; & \theta \in [0, 2\pi] \\ 0; & \theta \in [\pi, 2\pi]. \end{cases}$$

Then

$$\begin{aligned} \int_{-2\pi}^0 (\phi_1^0 e^{\lambda\theta}, B_0 \bar{u}(-\theta))_{\mathbb{R}^2} d\theta &= \int_{-2\pi}^0 e^{\lambda\theta} \bar{u}(-\theta) d\theta \\ &= - \int_0^{\pi} e^{-\lambda\theta} \sin \theta d\theta < 0. \end{aligned}$$

Let us now take the control function $\bar{\bar{u}} \in \tilde{\Omega}$

$$\bar{\bar{u}}(\theta) = \begin{cases} 0; & \theta \in [0, \pi] \\ \sin \theta; & \theta \in [\pi, 2\pi]. \end{cases}$$

In the same way as above we see that

$$\int_{-2\pi}^0 (\phi_{-1}^0 e^{\lambda\theta}, B_0 \bar{u}(-\theta))_{\mathbb{R}^2} d\theta > 0.$$

An analogous computation with the same choice for \bar{u} and \bar{u} shows that relation (11) does not hold for the pair (λ_2, ϕ_{-1}^0) i.e. condition (b') also holds. Therefore, the system of this example is locally null-controllable.

In the following, the case of delay systems is considered when the delays are commensurate, that is the system can be written as

$$(A(Z), B(Z), \Omega) \quad \dot{\mathbf{x}}(t) = A(Z)\mathbf{x}(t) + B(Z)\mathbf{u}(t)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$, the set of admissible controls is the same as in the general case, Z is the right-shift operator i.e. $Z\mathbf{x}(t) = \mathbf{x}(t - h)$ for delay duration $h > 0$, $A(Z) \in \mathbb{R}^{n \times n}[Z]$, $B(Z) \in \mathbb{R}^{n \times m}[Z]$ and $\mathbb{R}^{i \times j}[Z]$ denotes $i \times j$ matrices composed of real polynomials in Z .

In this case, the spectrum of system $(A(Z), B(Z), \Omega)$ is given by

$$\sigma(A) = \{\lambda \in \mathbb{C} : \det(\lambda I - A(e^{-\lambda h})) = 0\},$$

and

$$\Delta(\lambda) = \lambda I - A(e^{-\lambda h})$$

is the characteristic polynomial of the system.

From [5], [23], [30] we know that the system $(A(Z), B(Z), \bigcup_{T>0} L_2([0, T], \mathbb{R}^m))$ i.e. the system with unconstrained control is GC iff this is AGC. A necessary and sufficient condition for AGC is

$$\text{rank}[\lambda I - A(e^{-\lambda h}), B(e^{-\lambda h})] = n$$

for all $\lambda \in \mathbb{C}$. Thus, we have the following.

Corollary 3. Assume that $\text{int } \tilde{\Omega} \neq \emptyset$. Then the system $(A(Z), B(Z), \Omega)$ is ALC iff it is LC.

Proof. The sufficiency of the Corollary is immediate. Assume that the system $(A(Z), B(Z), \Omega)$ is ALC. Consider the discrete-time system $(\mathcal{A}, \mathcal{B}, \tilde{\Omega})$ associated with the system $(A(Z), B(Z), \Omega)$:

$$(\mathcal{A}, \mathcal{B}, \tilde{\Omega}) : \mathbf{x}_{n+1} = \mathcal{A}\mathbf{x}_n + \mathcal{B}\mathbf{u}_n$$

where $\mathbf{x}_n \in M_2$, $\mathbf{u}_n \in \tilde{\Omega} \subset U$.

Since the system $(A(Z), B(Z), \Omega)$ is ALC it follows that

- (a) The system $(\mathcal{A}, \mathcal{B}, U)$ is AGC,

(b) $\text{Ker}(\lambda I^* - S^*(h)) \cap (\mathcal{B}\tilde{\Omega})^0 = \{0\}$ for all $\lambda > 0$.

Condition (a) means that the system $(A(Z), B(Z), \bigcup_{T>0} L_2([0, T], \mathbb{R}^m))$ is AGC.

Hence, by [5] and [30] it is GC, too, thus the system $(\mathcal{A}, \mathcal{B}, U)$ is GC. By Theorem 3 we have that the system $(\mathcal{A}, \mathcal{B}, \tilde{\Omega})$ is LC, thus, the system $(A(Z), B(Z), \Omega)$ is LC. The proof is complete.

Acknowledgement

I would like to take this opportunity to thank my adviser, dr. Éva Gyurkóvics, for her excellent advice and support.

References

1. Banks, H. T., Jacobs, M. Q., Langenhop, C. E., Characterization of the controlled states in $w_2^{(1)}$ of linear hereditary systems. *SIAM J. Contr.* **13** (1975), pp. 611–649.
2. Bartosiewicz, Z., Approximate controllability of neutral system with delay in control. *J. Diff. Equat.* **51** (1984), pp. 295–325.
3. Bernier, C., Manitius, A., On the semigroup in $\mathbb{R}^n \times L^p$ corresponding to differential equation with delays. *Canad. J. Math.* **30** (1978), pp. 296–332.
4. Chuckwu, En., Function space null-controllability of linear delay systems with limited power. *J. Mat. Anal. Appl.* **124** (1987), pp. 293–304.
5. Colonius, F., On approximate and exact null-controllability of delay systems. *System & Control Letter* **5** (1984), pp. 209–211.
6. Colonius, F., Stable and regular reachability of relaxed hereditary differential systems. *SIAM J. Contr. Opt.* **23** (1985), pp. 803–807.
7. Douglas, R. G., On majorisation, factorization and range inclusion of operator in Hilbert space. *Proceed. Amer. Math. Soc.* **17** (1966), pp. 409–416.
8. Dunford, N., Schwartz, J. T., *Linear Operators, Part I.* Wiley, New York, 1958.
9. Fuhrman, P. A., On weak and strong reachability and controllability of infinite-dimensional linear systems. *J. Opt. Theory Appl.* **9** (1972), pp. 77–89.
10. Hale, J., *Theory of functional differential equation.* Springer-Verlag, New York, 1977.
11. Halmos, P. R., *A Hilbert Space Problem Book.* Van Nostrand, Toronto, London, 1967.
12. Jacobs, M. Q., Langenhop, C. E., Criteria for function space controllability of linear neutral systems. *SIAM J. Control. Opt.* **14** (1976), pp. 1009–1048.
13. Krein, M. G., Rutman, M. A., Linear operator leaving invariant a cone in a Banach space. *Uspekhi Mat. Nauk* **I, 23** (1948), pp. 3–9.
14. Manitius, A., Completeness and F-completeness of eigenfunctions associated with retarded functional differential equation. *J. Diff. Equat.* **35** (1980), pp. 1–29.
15. Manitius, A., Necessary and sufficient conditions of approximate controllability for general linear retarded systems. *SIAM J. Control. Opt.* **19** (1981), pp. 516–532.
16. Manitius, A., Triggiani, R., Function space controllability of linear retarded systems: a derivation from abstract operator conditions. *SIAM J. Control. Opt.* **16** (1978), pp. 599–646.

17. *Marchenko, V. M.*, On complete controllability of systems with delay. *Probl. Contr. Inf. Theory* **8** (1979), pp. 421–432.
18. *Olbro, A. W.*, Algebraic criteria of controllability to zero function for linear constant time lag systems. *Contr. & Cyber.* **2** (1973), pp. 59–77.
19. *Olbro, A. W.*, Controllability of retarded systems with function space constraint: 2. Approximate controllability. *Contr. & Cyber.* **6** (1977), pp. 5–31.
20. *Przyłuski, K. M.*, Infinite-dimensional discrete-time: Theory and application to linear hereditary control systems. IPE PW, Technical Report, Warsaw, 1979.
21. *Przyłuski, K. M.*, Arbitrary stabilizability of infinite-dimensional discrete-time systems with applications to linear hereditary systems. *Inst. of Math., PAN, Preprint 212* (1980).
22. *Salamon, D.*, On controllability and observability of time delay system. *IEEE Trans. Aut. Contr.* **AC-29** (1984), pp. 432–439.
23. *Salamon, D.*, On control and observation of neutral system. Pitman, Boston, 1984.
24. *Skłjar, B. S.*, Approximate controllability of retarded systems in a class of positive controls. *Diff. Uravn.* **21** (1985), pp. 2086–2096 (in Russian).
25. *Son, N. K.*, Controllability of linear discrete-time with restrained controls in Banach space. *Contr. & Cyber.* **10** (1981), pp. 5–16.
26. *Son, N. K.*, Approximate controllability with positive controls, Part I, II. *Univ. Bremen, Report 181* (1987).
27. *Son, N. K., Su, N. V., Chau, N. V.*, On the global null-controllability of linear discrete-time systems with restrained controls in Banach space. *Inst. Math. Hanoi, Preprint 20* (1984).
28. *Son, N. K., Thanh, L.*, On the null-controllability of infinite-dimensional discrete-time systems. *Acta Math. Vietnam* **10** (1985), pp. 3–14.
29. *Zmood, R. B.*, The Euclidean space controllability of control system with delay. *SIAM J. Contr.* **12** (1974), pp. 609–623.
30. *Watanabe, K.*, Finite spectrum assignment and observer for multivariable systems with commensurate delays. *IEEE Trans. Aut. Contr.* **6** (1986), pp. 543–550.

**Нуль-управляемость
линейных дискретных систем бесконечной размерности
с ограничениями на управление**

НГУЕН ВАН СУ

(Будапешт)

Настоящая работа посвящена вопросу нуль-управляемости системы

$$x_{k+1} = Ax_k + Bu_k, \quad x_k \in X, \quad u_k \in \Omega \subset U, \quad A \in L(X, X), \quad B \in L(U, X),$$

где X и U — гильбертовы пространства, Ω — выпуклое множество с непустой внутренней частью, содержащее 0 .

Здесь рассматривается случай полу-конечного оператора A . Для таких систем задаётся необходимое и достаточное условие нуль-управляемости. Полученный ре-

зультат используется для исследования нуль-управляемости систем с запаздыванием следующего вида:

$$\dot{x}(t) = L(x_t) + B_0 u(t), \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \Omega \subset \mathbb{R}^m, \quad x_t(\theta) = x(t + \theta), \quad \theta \in [-h, 0].$$

Доказывается, что нуль-управляемость в точном и приближенном смысле эквивалентны в случае конечного числа соизмеримых запаздываний. Этот факт был известен только для системы без ограничений относительно управлений.

Nguyen Van Su
Technical University of Budapest
Faculty of Mechanical Engineering
Department of Mathematics
Budapest, Stoczek u. 2.
H-1111, Hungary

OPTIMIZATION OF DYNAMICAL SYSTEMS WITH IDENTIFICATION OF INPUT PERTURBATIONS

R. GABASOV, F. M. KIRILLOVA

(Minsk)

(Received June 12, 1990)

A finite algorithm for the construction of optimal control of dynamic systems with perturbations that identifies acting perturbations with help of observations over resulting processes as it operates, is proposed.

1. Introduction

A real control system is functioning, as a rule, in the presence of noise. This leads to the necessity of introduction of uncertainties of corresponding optimization problems [1, 2] into the mathematical model. The models of stochastic optimal control [3-5] are the more developed ones for optimal control under conditions of uncertainty. Lately, in connection with the development of the modern theory of extremal problems it became possible to investigate models in which other representations on the nature of perturbations and on the principles of control are admitted, under conditions of uncertainty not emphasizing the relative frequency of arising of these or those values of perturbations. In the new approaches the structure of large numbers of the possible values of perturbations is taken into consideration in detail. The latter is practically ignored in probabilistic models. Appreciating the quality of control, it is necessary to perform the conditions not as a whole but in all possible situations [6, 7].

In this paper the authors, based on the results on the constructive theory of extremal problems [8, 9], substantiate a finite algorithm for optimal control construction by dynamic systems with perturbations.

2. Statement of the problem

Consider the family of q -vector function $\omega(t)$, $t \in T = [0, t^*]$:

$$\omega(t) = \omega_0(t) + \sum_{i=1}^q w_i \omega_i(t), \quad (1)$$

defined with the fixed piecewise continuous p -vector function $\omega_0(t), \omega_1(t), \dots, \omega_q(t)$, $t \in T$, and q -vector of parameters $w = (w_1, \dots, w_q)$, which may take any value from the set

$$\check{W} = \{w \in \mathbb{R}^q : Gw = f, d_* \leq w \leq d^*\} \quad (f \in \mathbb{R}^l). \quad (2)$$

We shall consider that the functions (1) describe perturbations acting to dynamic system of control

$$\dot{x}(t) = A(t)x + b(t)u + \mathcal{D}(t)\omega(t), \quad x(0) = x_0 \quad (x \in \mathbb{R}^n, u \in \mathbb{R}) \quad (3)$$

with piecewise continuous elements $A(t), b(t), \mathcal{D}(t), t \in T$.

To every piecewise control $u(\cdot) = (u(t), t \in T)$ it corresponds the only movement

$$\check{X}(t) = \check{X}(t | x_0, u(\cdot)) = \{x(t | x_0, u(\cdot), w), w \in \check{W}\}, \quad t \in T,$$

consisting of all the trajectories $x(t) = x(t | x_0, u(\cdot), w), t \in T$, of the system (3) generated by the fixed initial state x_0 , control $u(\cdot)$ and different parameter vectors $w \in \check{W}$.

In the following we shall call \check{W} a priori distribution of parameters, $\check{X}(t), t \in T$, the a priori movement of the system (3).

Let the terminal set in the space of states of the system (3) be given

$$X^* = \bigcap_{i=1}^m X_i^*, \quad X_i^* = \{x \in \mathbb{R}^n : h_i'x \geq g_i\}.$$

The control $\check{u}(t), t \in T$, is constrained by

$$|u(t)| \leq 1, \quad t \in T, \quad (4)$$

and motion $\check{X}(t | \check{u}(\cdot)), t \in T$, corresponding to it will be called a priori admissible if a terminal inclusion is fulfilled

$$\check{X}(t^* | \check{u}(\cdot)) \subset X^*. \quad (5)$$

The quality of the a priori admissible control $\check{u}^0(\cdot)$ will be estimated according to the functional value

$$J(\check{u}) = \min_{w \in \check{W}} h_0'x(t^* | x_0, \check{u}(\cdot), w). \quad (6)$$

$\check{u}^0(\cdot)$ will be called an a priori optimal control if

$$J(\check{u}^0) = \max_{\check{u}(\cdot)} J(\check{u}). \quad (7)$$

The efficiency (7) of control by the system under conditions of uncertainty (2) is less than the efficiency of control in the case when perturbation $\omega^*(t), t \in T$,

is known, i.e. if the uncertainty of a problem is absent. Therefore, to increase the control efficiency we shall introduce the procedure of observation over the control process.

Consider the following types of linear inertia-free measuring systems:

1) direct complete

$$y = Cw \quad (y \in \mathbf{R}^q, \det C \neq 0); \quad (8)$$

2) direct incomplete exact: (8) for $y \in \mathbf{R}^l$, $\text{rank } C = l < q$;

3) indirect incomplete exact

$$y = Kx \quad (y \in \mathbf{R}^l); \quad (9)$$

4) mixed inexact incomplete

$$y = Kx + Cw + \xi. \quad (10)$$

In case (10) we shall assume that in the process of dimensions any piecewise continuous function of errors of $\xi(t)$, $t \in T$, satisfying inequalities

$$\xi_* \leq \xi(t) \leq \xi^*, \quad t \in T, \quad (11)$$

may be implemented.

Let for the chosen control $u^*(\cdot)$ the measuring device have a registered signal $y^*(t)$, $t \in T$.

The set \hat{W} of vectors, $w \in \hat{W}$, that together with some possible error function of dimensions $\xi(t)$, $t \in T$, are able to generate signal $y^*(t)$, $t \in T$, will be called a posteriori distribution of parameters w . The a posteriori motion $\hat{X}(t|x_0, u^*(\cdot)) = \{x(t|x_0, u^*(\cdot), w), w \in \hat{W}\}$, $t \in T$, a posteriori admissible control $\hat{u}^*(t)$, $t \in T$, ($\hat{X}(t^*|x_0, \hat{u}^*(\cdot)) \subset X^*$) and a posteriori optimal control $\hat{u}^0(\cdot)(J(\hat{u}^0) = \max_{\hat{u}^*(\cdot)} J(\hat{u}^*))$

correspond to it. Since $\hat{W} \subset \check{W}$ then $J(\hat{u}^0) \geq J(\check{u}^0)$, i.e. the control efficiency while using observation is increasing. Below methods for constructing of optimal controls at different ways of data obtaining on control processes are presented.

3. Construction of a priori optimal controls

Investigate the control problem (1)–(7) without using observation over the control process.

In accordance with the control problem (1)–(7) calculate the following estimates of a priori distribution \check{W} :

$$\check{\alpha}_i = \min_{w \in \check{W}} h'_i x(t^*|x_0, u(\cdot), w), \quad i = \overline{0, m}. \quad (12)$$

In accord with Cauchy formula [10] we have

$$x(t^*|x_0, u(\cdot), w) = F(t^*)x_0 + \int_0^{t^*} F(t^*)F^{-1}(t)b(t)u(t) dt +$$

$$+ \int_0^{t^*} F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt + \sum_{j=1}^q w_j \int_0^{t^*} F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_j(t) dt.$$

Using (12), we obtain

$$\begin{aligned} \check{\alpha}_i &= \check{\gamma}_i + h'_i F(t^*)x_0 + \int_0^{t^*} h'_i F(t^*)F^{-1}(t)b(t)u(t) dt + \\ &+ \int_0^{t^*} h'_i F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt, \end{aligned} \quad (13)$$

$$\check{\gamma}_i = \min a'_i w, \quad Gw = f, \quad d_* \leq w \leq d^*, \quad i = \overline{0, m}$$

$$a_i = (a_{ij}, \quad j = \overline{1, q}), \quad a_{ij} = \int_0^{t^*} F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_j(t) dt.$$

Using estimates (12), the conditions of the a priori admissibility of control $\check{u}(\cdot)$ will be written in the form

$$\begin{aligned} &h'_i F(t^*)x_0 + h'_i \int_0^{t^*} F(t^*)F^{-1}(t)b(t)u(t) dt + \\ &+ h'_i \int_0^{t^*} F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt \geq \check{g}_i, \quad i = \overline{1, m}. \end{aligned}$$

Find the value of a quality criterion on the a priori admissible control $\check{u}(t)$, $t \in T$,

$$\begin{aligned} J(\check{u}) &= \check{\alpha}_0 = \check{\gamma}_0 + h'_0 F(t^*)x_0 + \int_0^{t^*} h'_0 F(t^*)F^{-1}(t)b(t)\check{u}(t) dt + \\ &+ \int_0^{t^*} h'_0 F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt. \end{aligned}$$

According to (7) the a priori optimal control $\tilde{u}_0(t)$, $t \in T$, is the solution of problem

$$\begin{aligned} & \int_0^{t^*} h'_0 F(t^*) F^{-1}(t) b(t) u(t) dt + h'_0 F(t^*) x_0 + \\ & + h'_0 \int_0^{t^*} F(t^*) F^{-1}(t) \mathcal{D}(t) \omega_0(t) dt - \max. \\ & h'_0 F(t^*) x_0 + h'_i \int_0^{t^*} F(t^*) F^{-1}(t) b(t) u(t) dt + \\ & + h'_i \int_0^{t^*} F(t^*) F^{-1}(t) \mathcal{D}(t) \omega_0(t) dt \geq \check{g}_i, \\ & i = \overline{1, m}; \quad |u(t)| \leq 1, \quad t \in T. \end{aligned}$$

In dynamical statement it has the form

$$\begin{aligned} J_0(u) &= h'_0 x(t^*) - \max \\ \dot{x} &= A(t)x + b(t)u(t) + \mathcal{D}(t)\omega_0(t); \quad x(0) = x_0. \\ h'_i x(t^*) &\geq \check{g}_i, \quad i = \overline{1, m}; \quad |u(t)| \leq 1, \quad t \in T. \end{aligned} \quad (14)$$

Problem (14) will be called a determined problem of optimal control accompanying problem (1)–(7) for construction of the a priori optimal control under uncertainty conditions.

So, to construct an a priori optimal control $u^0(t)$, $t \in T$, of problem (1)–(7) we need to solve:

- 1) $(m+1)$ -problems of linear programming (13).
- 2) one problem of optimal control (14). The value of quality criterion on $\tilde{u}^0(\cdot)$ equals to $J(\tilde{u}^0) = \check{\gamma}_0 + J_0(\tilde{u}^0)$.

4. Optimization of perturbed dynamical control systems by means of observation results

Let us add to the control procedure the operations on processing of output signals of measuring device (9) or (10). In Section 1 there was introduced the notion of the a posteriori distribution \tilde{W} of parameters, that in a general (i.e. non-constructive) form contains the complete results of the uncertain elements' removal from the set \tilde{W} by means of data in the observed signal $y(t)$, $t \in T$. It is sufficient to introduce only separate numerical characteristics (estimates) of the set \tilde{W} , both

for solving practical filtration problems, wherein the whole probability distribution function is seldom used, and they are restricted only by some simplest numerical characteristics (mathematical expectation, variance, etc.) and for problems (1)–(7), (9) or (10).

Count the following estimates

$$\hat{\alpha}_i = \min_{w \in \bar{W}} h'_i x(t^* | x_0, u^*(\cdot), w), \quad i = \overline{0, m}. \quad (15)$$

We call the calculation of estimates (15) identification problems accompanying the dynamic system optimization problem under uncertainty conditions.

Consider the case of indirect incomplete exact measuring (9). Let the control $u^*(t)$, $t \in T$, be given onto the input of system (3). It generates trajectory $x(t | x_0, u^*(\cdot), \omega(\cdot))$, $t \in T$ ($x_0, \omega_0(t)$, $t \in T$, are given) with some value of parameter $w \in \bar{W}$. The measuring device (9) yields the necessary signal $y^*(t)$, $t \in T$. All conditions being highly general [10], hence such moments $t_j \in T$, $j = \overline{1, p}$, and sets \mathcal{L}_j , $|\mathcal{L}_j| \leq l$, $j = \overline{1, p}$; $\sum_{j=1}^p |\mathcal{L}_j| = q$, will be found that the matrix is non-singular

$$P = \left(K(\mathcal{L}_j) F(t_j) \right)_{j=1, p}.$$

Here $K(\mathcal{L}_j)$ is a submatrix of matrix K containing rows with indices from \mathcal{L}_j . Form signal

$$v(t) = \int_0^t KF(t)F^{-1}(\tau)b(\tau)u(\tau) d\tau + KF(t)x_0 + \int_0^t KF(t)F^{-1}(\tau)\mathcal{D}(\tau)\omega_0(\tau) d\tau.$$

Assume

$$y^*(t) - v(t) = z(t).$$

Compose n -vector $z_{op} = (z_s(t_s), s \in \mathcal{L}_j, j = \overline{1, p})$ and find the unknown parameter $w^0 = P^{-1}z_{op}$.

Therefore, the a posteriori parameters' distribution degenerates into a fixed element w^0 (uncertainty of the problem (1)–(7) disappears) for exact dimensions (9) of support signals $y_s^*(t)$, $s \in \mathcal{L}_j$, $t \in T$, in support moments t_j , $j = \overline{1, p}$,

w^0 being constructed, the problem (1)–(7), (9) comes to the following one

$$\begin{aligned} h'_0 x(t^*) \rightarrow \max, \quad \dot{x} &= A(t)x + b(t)u(t) + \mathcal{D}(t)\omega^0(t), \\ \omega^0(t) &= \omega_0(t) + \sum_{j=1}^q w_j^0 \omega_j(t), \end{aligned} \quad (16)$$

$$x(0) = x_0, \quad h'_i x(t^*) \geq g_i, \quad i = \overline{1, m}; \quad |u(t)| \leq 1, \quad t \in T,$$

that will be called a determined problem of optimal control accompanying that of (1)–(7) with measuring device (9).

We shall call the solution of problem (16) an ideal optimal control and denote by $\hat{u}^0(t)$, $t \in T$.

The value $J(\hat{u}^0) - J(\tilde{u}^0)$ characterizes the increase of control efficiency while using measuring device (9).

Consider the most interesting case when problem (1)–(7) is connected with measuring device (10).

Let $u^*(t)$, $t \in T$, be some control restricted by (4), $x(t|x_0, u^*(\cdot), w)$, $t \in T$, some system trajectory generated by this control, given by initial state x_0 , its unknown value of parameter $w \in \tilde{W}$, $y^*(t)$, $t \in T$, be an observed signal of measuring device (10) caused by trajectory $x(t|x_0, u^*(\cdot), w)$, $t \in T$, and unknown error function of measuring $\xi(t)$, $t \in T$. Assume

$$z(t) = y^*(t) - \int_0^t KF(t)F^{-1}(\tau)b(\tau)u^*(\tau) d\tau - KF(t)x_0 - \\ - \int_0^t KF(t)F^{-1}(\tau)\mathcal{D}(\tau)\omega_0(\tau) d\tau.$$

For the a posteriori distribution of terminal states $\hat{X}(t^*)$ calculate the following estimates

$$\hat{\alpha}_i = \min_{w \in \tilde{W}} h'_i x(t^* | x_0, u^*(\cdot), w), \quad i = \overline{0, m}. \quad (17)$$

Using Cauchy formula [10] from (17) we obtain

$$\hat{\alpha}_i = \hat{\gamma}_i + h'_i F(t^*)x_0 + \int_0^{t^*} h'_i F(t^*)F^{-1}(t)b(t)u^*(t) dt + \\ + \int_0^{t^*} h'_i F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt, \quad i = \overline{0, m};$$

$$\hat{\gamma}_i = \min a'_i w, \quad w \in \tilde{W}.$$

Calculation of numbers $\hat{\gamma}_i$, $i = \overline{0, m}$, in a detailed notation comes to problems

$$\hat{\gamma}_i = \min a'_i w, \quad Gw = f, \quad d_* \leq w \leq d^*, \quad i = \overline{0, m}, \\ \xi_* \leq z(t) - Kw'd \leq \xi^*, \quad (18)$$

$$d = (d_j, j = \overline{1, q}), \quad d_j = \int_0^{t^*} F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_j(t) dt.$$

We call problem (18) the problem of identification, accompanying that of an optimal control (1)–(7) with measuring device (9).

The finite algorithm to solve linear problems (18) is described in [9].

Knowing estimates $\hat{\gamma}_i$, $i = \overline{1, m}$, write the conditions of control $u(t)$, $t \in T$, the a posteriori admissibility

$$h'_i F(t^*)x_0 + h'_i \int_0^{t^*} F(t^*)F^{-1}(t)b(t)u(t) dt + \\ + h'_i \int_0^{t^*} F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt \geq \hat{g}_i, \quad i = \overline{1, m},$$

where $\hat{g}_i = g_i - \hat{\gamma}_i$, $i = \overline{1, m}$.

The a posteriori optimal control $\hat{u}^0(t)$, $t \in T$, is the solution of the problem

$$J_0(u) = \int_0^{t^*} h'_0 F(t^*)F^{-1}(t)b(t)u(t) dt + h'_0 F(t^*)x_0 + \\ + \int_0^{t^*} h'_0 F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt \rightarrow \max, \\ h'_i F(t^*)x_0 + h'_i \int_0^{t^*} F(t^*)F^{-1}(t)b(t)u(t) dt + \\ + h'_i \int_0^{t^*} F(t^*)F^{-1}(t)\mathcal{D}(t)\omega_0(t) dt \geq \hat{g}_i, \quad i = \overline{1, m}, \quad |u(t)| \leq 1, \quad t \in T. \quad (19)$$

The dynamical form of problem (19) is as follows

$$J_0(u) = h'_0 x(t^*) \rightarrow \max, \quad \dot{x} = A(t)x + b(t)u + \mathcal{D}(t)\omega_0(t), \quad x(0) = x_0, \\ h'_i x(t^*) \geq \hat{g}_i, \quad i = \overline{1, m}; \quad |u(t)| \leq 1, \quad t \in T.$$

We call it a determined problem of optimal control, accompanying that of (1)–(7) with measuring device (10). This problem can also be solved by applying methods from [9].

The value of quality criterion on an a posteriori optimal control $\hat{u}^0(t)$, $t \in T$, is equal to $J(\hat{u}^0) = J_0(\hat{u}^0) + \hat{\gamma}_0$. Value $J(\hat{u}^0) - J(\check{u}^0)$ characterizes the increase of control efficiency at the expense of observation results by means of measuring device (10). Number $J(\hat{u}^0) - J(\check{u}^0)$ is equal to the loss of control efficiency because of errors in measuring device (10).

From (19) it can be seen that the problem of identification of parameters w does not depend on control, i.e. the problem of control and that of identification are separate ones.

5. Example*

Consider a problem on acceleration of a mass point on horizontal section of path when an unknown constant force affects it. The mathematical simulation of the problem is as follows:

$$x_2(1) \rightarrow \max, \dot{x}_1 = x_2, \dot{x}_2 = u + w, x_1(0) = x_2(0) = 0,$$

$$x_1(1) \leq 0.5; |u(t)| \leq 1, t \in T = [0, 1], 0 \leq w \leq 1.$$

It presents a special case of problem (3):

$$n = 2, h_0 = (0, 1), A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, b = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, w = \check{W} = \{w \in \mathbb{R}^1 : 0 \leq w \leq 1\}$$

$$\check{\Omega}(\cdot) = \{\omega(t), t \in T, \omega(t) \equiv w\}, g = 1, \omega_0(t) \equiv 0, \omega_1(t) \equiv 1, t \in T;$$

$$p = 1, m = 1, h_1 = (-1, 0), g_1 = -0.5.$$

First, construct the a priori optimal control. Since a fundamental matrix of solutions $F(t)$, $t \in T$, has the form $F(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$ then

$$x_1(t) = \int_0^t (t - \tau)u(\tau) d\tau + w \frac{t^2}{2}, \quad x_2(t) = \int_0^t u(\tau) d\tau + wt.$$

According to (12) the a priori estimates $\check{\alpha}_1, \check{\alpha}_0$ are equal to

$$\check{\alpha}_1 = \int_0^1 (1 - t)u(t) dt + \check{\gamma}_1, \quad \check{\alpha}_0 = \int_0^1 u(t) dt + \check{\gamma}_0,$$

$$\check{\gamma}_1 = \max_{0 \leq w \leq 1} \frac{w}{2} = 0.5, \quad \check{\gamma}_0 = \min_{0 \leq w \leq 1} w = 0.$$

Hence, $\check{g}_1 = -0.5 + 0.5 = 0$ and an a priori optimal control $\check{u}^0(t)$, $t \in T$, is the solution of the accompanying (14) determined problem

$$x_2(1) \rightarrow \max, \dot{x}_1 = x_2, \dot{x}_2 = u, x_1(0) = x_2(0) = 0, \\ x_1(1) \leq 0, |u(t)| \leq 1, t \in T.$$

*Example is calculated by S. V. Prishchepova

Solution of the problem has the form

$$\ddot{u}^0(t) = -1, \quad t \in [0, 1 - \sqrt{2}/2], \quad \dot{u}^0(t) = 1, \quad t \in [1 - \sqrt{2}/2, 1].$$

The value of quality criterion is equal to $J(\ddot{u}^0) = \sqrt{2} - 1 \approx 0.414$. Assume a signal $y^*(t) = \int_0^t (t - \tau + 1)u(\tau) d\tau$, $t \in [0, 1]$, is written on a measuring device $y = x_1 + x_2$.

Then $\dot{w} = 0$ and control $\dot{u}^0(t)$, $t \in T$, is the solution of the determined problem

$$x_2(1) \rightarrow \max, \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad x_1(0) = x_2(0) = 0,$$

$$x_1(1) \leq 0.5, \quad |u(t)| \leq 1, \quad t \in T,$$

and has the form $\dot{u}^0(t) = 1$, $t \in T$. The value of quality criterion on it is equal to $J(\dot{u}^0) = 1$.

Now, consider measuring device

$$y = x_1 + x_2 + \xi, \tag{20}$$

operating with errors $\xi(t)$, $t \in T$, satisfying restrictions $-0.2 \leq \xi(t) \leq 0.1$, $t \in [0, 1]$.

Assume that a signal $y^*(t) = \int_0^t (t - \tau + 1)u(\tau) d\tau$, $t \in T$, is written again. In this case the accompanying identification problems have the form

$$\hat{\gamma}_1 = \max_{0 \leq w \leq 1} \frac{w}{2}, \quad -0.2 \leq \left(\frac{t^2}{2} + t\right)w \leq 0.1, \quad t \in [0, 1];$$

$$\hat{\gamma}_0 = \min w, \quad -0.2 \leq \left(\frac{t^2}{2} + t\right)w \leq 0.1, \quad t \in [0, 1]; \quad 0 \leq w \leq 1.$$

Hence, $\hat{\gamma}_1 = 0.0333$, $\hat{\gamma}_0 = 0$. Then $\hat{g}_1 = -0.5 + 0.0333 \approx 0.47$. The a posteriori optimal control $\dot{u}^0(t)$, $t \in T$, is the solution of the accompanying determined problem

$$x_2(1) \rightarrow \max, \quad \dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad x_1(0) = x_2(0) = 0,$$

$$x_1(1) \leq 0.47, \quad |u(t)| \leq 1, \quad t \in [0, 1],$$

and has the form

$$\dot{u}^0(t) = -1, \quad t \in [0, \tau_1]; \quad \dot{u}^0(t) = 1, \quad t \in [\tau_1, 1],$$

$$\tau_1 = 1 - \sqrt{0.97} \approx 1 - 0.9849 = 0.0151.$$

The quality criterion on it takes the value $J(\hat{u}^0) = 2\sqrt{0.97} - 1 = 0.9698$. The increase of control efficiency at the expense of observation by means of the measuring device (20) is as follows:

$$J(\hat{u}^0) - J(\check{u}^0) = 0.9698 - 0.414 = 0.5558.$$

The control efficiency loss because of dimension error is equal to

$$J(\check{u}^0) - J(\hat{u}^0) = 1 - 0.9698 = 0.0302.$$

References

1. *Lanig, J. H. Battin, R. H.*, Random Processes in Automatic Control. McGraw-Hill, New York, 1956.
2. *Pugachev, V. C.*, Theory of Random Functions and its Applications to Automatic Control Problems. Moscow, Gostekhizdat, 1957.
3. *Feldbaum, A. A.*, Foundation of Theory of Optimal Automatic Systems. Moscow, Gosizdat FML, 1963.
4. *Astrom, K. J.*, Introduction to Stochastic Control Theory. Academic Press. New York, 1969.
5. *Kalman, R., et al.*, On General Theory of Control Systems, Proceedings of the 1st IFAC Congress, Vol. 2, Moscow, Academy of Sc., 1961, pp. 521-547 (in Russian).
6. *Krasovskii, N. N.*, Control by Dynamic System. Moscow, Nauka, 1977.
7. *Kurzhanskii, A. B.*, Control and Observation under Conditions of Uncertainty. Moscow, Nauka, 1977.
8. *Gabasov, R., Kirillova, F. M.*, Linear Programming Methods. P. I-III. Minsk, BGU Publishing House, 1977, 1978, 1980.
9. *Gabasov, R., et al.*, Constructive Methods of Optimization. P. I-V. Minsk, BGU Publishing House, 1984, 1986, 1988.
10. *Gabasov, R., Kirillova, F. M.*, The Qualitative Theory of Optimal Processes. Marcel Dekker Inc., New York and Basel, USA, 1976.

Оптимизация динамических систем с идентификацией входных воздействий

Р. ГАБАСОВ, * М. КИРИЛЛОВА

(Минск)

В работе исследуется задача оптимального управления линейной системой при условии действия на нее неизвестных возмущений. Для эффективного выбора управления системой проводится наблюдение за динамикой системы; при этом рассматриваются четыре возможных типа линейных «безинерционных измерительных устройств». Для рассматриваемых случаев, с учетом линейности задачи, строятся

конструктивные алгоритмы управления. Работоспособность предлагаемых алгоритмов проиллюстрирована на содержательном примере.

Р. Габасов
Белорусский госуниверситет
СССР, Минск, Университетский городок

Ф. М. Кириллова
Институт математики АН БССР
СССР, Минск, Сурганова, 11

Typesetting by TYPOT_EX Kft, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

or to L. GYÖRFI
Technical University of Budapest
H-1111 Budapest, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

The papers, preferably in English or Russian, should be typed double spaced on one side of good-quality paper with wide margins (4–5 cm). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as $\text{diag} (d_1, d_2, \dots, d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском языке в трех экземплярах, на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Studniarski, M.</i> : The discrete maximum principle as a sufficient optimality condition (<i>Студнярски М.</i> Дискретный принцип максимума как необходимые условия оптимальности)	179
<i>Malanowski, K.</i> : Stability and sensitivity analysis of discrete optimal control problems (<i>Малановски К.</i> Анализ устойчивости и чувствительности дискретных задач оптимального управления)	187
<i>Papageorgiou, S.</i> : Relaxability and well-posedness for infinite dimensional optimal control problems (<i>Папагеоргиу Н. С.</i> О корректности и устойчивости оптимального значения функционала качества для систем с нелинейными распределенными параметрами)	201
<i>Nguyen Van Su</i> : Null-controllability of infinite-dimensional discrete-time system with restrained control (<i>Нгуйен Ван Су</i> Нуль-управляемость линейных дискретных систем бесконечной размерности с ограничениями на управление)	215
<i>Gabasov, R., Kirillova, F. M.</i> : Optimization of dynamical systems with identification of input perturbations (<i>Габасов, Р., Кириллова, Ф. М.</i> Оптимизация динамических систем с идентификацией входных воздействий)	233

316920

VOL. 20 • NUMBER 4
TOM HOMEP

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

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
ПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1991

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

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

- theory of control processes
- theory of adaptive systems
- theory of estimation and identification
- theory of controlling robot-technologic and flexible manufacturing systems
- information theory
- information-theoretic aspects of multiple access networks.

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

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

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

- теория процессов управления;
- теория адаптивных систем;
- теория оценивания и идентификации;
- теория управления робототехническими и гибкими производственными системами;
- теория информации;
- теория информации в области сетей с множественным доступом.

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

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czech and Slovak Federal Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia:

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

For all other countries:

PERGAMON PRESS PLC Headington Hill Hall, Oxford OX3 0BW, England
or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1991 Subscription Rate DM 627,— per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITOR

N. N. KRASOVSKII (USSR)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

K. V. FROLOV

A. B. KURZHANSKI

I. A. OVSEEVICH

E. D. TERYAEV

R. Z. KHASMINSKI

HUNGARY

T. VÁMOS

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

L. GYÖRFI

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJC

I. VAJDA

РЕДАКТОР ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

К. В. ФРОЛОВ

А. Б. КУРЖАНСКИЙ

И. А. ОВСЕЕВИЧ

Е. Д. ТЕРЯЕВ

Р. З. ХАСЬМИНСКИЙ

ВНР

Т. ВАМОШ

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

Л. КЕВИЦКИ

Л. ДЬЕРФИ

Я. КОЧИШ

ЧССР

Й. БЕНЕШ

В. СТРЕЙЦ

И. ВАЙДА



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Typesetting by TYPOT_EX Kft, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

ON THE OPTIMAL CONTROL AND RELAXATION OF FINITE DIMENSIONAL SYSTEMS DRIVEN BY MAXIMAL MONOTONE DIFFERENTIAL INCLUSIONS*

N. S. PAPAGEORGIOU**

(Athens)

(Received November 1, 1990)

In this paper we examine finite dimensional optimal control problems driven by maximal monotone differential inclusions and having state dependent control constraints. First, with the help of a convexity hypothesis, we prove the existence of optimal admissible pairs. Then we drop convexity hypothesis and we look at the relaxed system. For that system we establish the existence of optimal solutions under minimal hypotheses. Finally, by strengthening our hypotheses we show that the original trajectories are dense in the relaxed ones for the topology of uniform convergence and that the two problems relaxed and original have the same value.

Keywords and -phrases: Monotone operator, orientor field, optimal pair, minimizing sequence, transition probabilities, relaxed system, selection theorem, Hausdorff metric, density result, relaxability.

AMS Subject Classification (1980): 49 A 20

1. Introduction

In this paper we examine the following finite dimensional optimal control problem:

$$\left\{ \begin{array}{l} J(x, u) = \int_0^b L(t, x(t), u(t)) dt \rightarrow \inf = m \\ \text{s. t. } \dot{x}(t) \in Ax(t) + f(t, x(t), u(t)) \text{ a. e.} \\ x(0) = x_0, \quad u(t) \in U(t, x(t)) \text{ a. e.} \end{array} \right. \quad (*)$$

where $A : \mathbf{R}^n \rightarrow 2^{\mathbf{R}^n}$ is a maximal monotone operator. First, with the aid of a convexity hypothesis on an appropriate orientor field, we establish the existence

*Revised version

**Research supported by N.S.F. Grant D.M.S.-8802688.

of optimal solutions for (*). Then we remove this convexity hypothesis. Now, the system may fail to have an optimal solution. Nevertheless, it is important to study the asymptotic behaviour of the minimizing sequences. This leads us to the introduction of a larger system, which is known in the literature as "relaxed system" and which captures the asymptotic behaviour of the minimizing sequences. For this augmented system, we prove the existence of optimal solutions, we show that its set of trajectories is the closure in the topology of uniform convergence of the set of trajectories of the original system and, finally, we prove that under mild hypotheses original and relaxed problems have equal values.

The problem studied here can be viewed as an extension of the works of Berkovitz [5], Cesari [6], Gamkrelidze [9], Pappas [15] and Warga [19], where $A = 0$, in their study of the relaxed problem the control constraint set was state independent (open-loop) and the hypotheses on the data were stronger.

An important special case of the problem studied here is when $A = \partial\delta_K$, where $\delta_K(\cdot)$ is the indicator function of a nonempty, closed convex set K (i.e. $\delta_K(x) = 0$ if $x \in K$ and $+\infty$ if $x \notin K$) and $\partial\delta_K(\cdot)$ denotes the subdifferential of $\delta_K(\cdot)$ in the sense of convex analysis. Recall the $\partial\delta_K(x) = N_K(x)$, $x \in K$, the normal cone to K at x . Such differential inclusions are called "differential variational inequalities" (see Aubin-Cellina [1]) and play an important role in mathematical economics, in the study of planning procedures (see Aubin-Cellina [1], chapter 5). Another important class of systems covered by our work are the gradient systems.

Throughout this work by $P_{f(c)}(\mathbb{R}^n)$ we will denote the family of nonempty, closed (convex) subsets of \mathbb{R}^n . Let (Ω, Σ) be a measurable space. A multifunction $F : \Omega \rightarrow P_f(\mathbb{R}^n)$ is said to be measurable, if for all $x \in \mathbb{R}^n$, $\omega \rightarrow d(x, F(\omega)) = \inf\{\|x - z\| : z \in F(\omega)\}$ is measurable. A multifunction $F : \Omega \rightarrow 2^{\mathbb{R}^n} \setminus \{\emptyset\}$ is said to be graph measurable, if $\text{Gr } F = \{(\omega, x) \in \Omega \times \mathbb{R}^n : x \in F(\omega)\} \in \Sigma \times B(\mathbb{R}^n)$, with $B(\mathbb{R}^n)$ being the Borel σ -field of \mathbb{R}^n . For closed-valued multifunctions, measurability implies graph measurability. The converse is true if Σ is complete with respect to a given measure $\mu(\cdot)$. For more details we refer to Wagner [18]. By S_F^1 we will denote the set of integrable selectors of $F(\cdot)$ i.e. $S_F^1 = \{g \in L^1(\mathbb{R}^n) : g(\omega) \in F(\omega) \mu\text{-a. e.}\}$. This set may be empty. It is nonempty if and only if $F(\cdot)$ is measurable and $\omega \rightarrow \inf\{|z| : z \in F(\omega)\} \in L_+^1$.

2. Existence theorem

For (*) let $T = [0, b]$ be the time horizon, \mathbb{R}^n the state space and \mathbb{R}^k the control space. We will need the following hypotheses on the data of (*).

H(A): $A : \mathbb{R}^n \rightarrow 2^{\mathbb{R}^n}$ is a maximal monotone operator.

H(f): $f : T \times \mathbb{R}^k \times \mathbb{R}^n$ is a map s. t.

(1) $t \mapsto f(t, x, u)$ is measurable,

- (2) $(x, u) \mapsto f(t, x, u)$ is continuous,
- (3) $|f(t, x, u)| \leq a(t) + b(t)(|x| + |u|)$ a. e. with $a(\cdot), b(\cdot) \in L^1_+$.

$H(U): U : T \times \mathbb{R}^n \times \mathbb{R}^k \rightarrow P_{fc}(\mathbb{R}^k)$ is a multifunction s. t.

- (1) $U(\cdot, \cdot)$ is graph measurable,
- (2) for every $t \in T$, $\text{Gr}U(t, \cdot) = \{(x, v) \in \mathbb{R}^n \times \mathbb{R}^k : v \in U(t, x)\}$ is closed,
- (3) $|U(t, x)| = \sup\{|v| : v \in U(t, x)\} \leq M$ for all $(t, x) \in T \times \mathbb{R}^n$.

$H(L): L : T \times \mathbb{R}^n \times \mathbb{R}^k \rightarrow \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ is an integrand s. t.

- (1) $(t, x, u) \mapsto L(t, x, u)$ is measurable,
- (2) $(x, u) \mapsto L(t, x, u)$ is l. s. c.,
- (3) $\phi(t) - M(|x| + |u|) \leq L(t, x, u)$ a. e. with $\phi(\cdot) \in L^1, M > 0$.

As we already mentioned in the introduction, in order to get an existence result we need a convexity hypothesis on an appropriate orientor field (recall “property Q ” of Cesari [6]). So, we introduce the following hypothesis:

$H_c: Q(t, x) = \{(v, \eta) \in \mathbb{R}^n \times \mathbb{R} : v \in Ax + f(t, x, u), u \in U(t, x), L(t, x, u) \leq \eta\}$ is convex for all $(t, x) \in T \times \mathbb{R}^n$.

Also in order to avoid trivial situations, we will need the following admissibility hypothesis:

H_a : There exists an admissible “state-control” pair (x, u) s. t. $J(x, u) < +\infty$.

THEOREM 2.1. *If hypotheses $H(A), H(f), H(U), H(L), H_c$ and H_a hold, then (*) admits an optimal admissible pair.*

Proof. Let $M(t, x, v) = \{u \in U(t, x) : v \in Ax + f(t, x, u)\}$. It is easy to see that this set is for almost all $t \in [0, b]$ compact, may be empty in \mathbb{R}^k . Set $p(t, x, v) = \inf_{u \in \mathbb{R}^k} [L(t, x, u) + \delta_{M(t, x, v)}(u)]$ (recall that by convention $\delta_\emptyset(\cdot) = +\infty$).

Hence, $p(t, x, v)$ represents the minimum cost needed to produce velocity v at time $t \in T$, state $x \in \mathbb{R}^n$ and using all admissible controls $U(t, x)$.

Claim #1: $(t, x, v) \rightarrow p(t, x, v)$ is measurable.

Given $\lambda \in \mathbb{R}$ we need to show that $\Lambda_\lambda = \{(t, x, v) \in T \times \mathbb{R}^n \times \mathbb{R}^n : p(t, x, v) \leq \lambda\} \in B(T) \times B(\mathbb{R}^n) \times B(\mathbb{R}^n)$. To this end note that $\Lambda_\lambda = \text{proj}_{T \times \mathbb{R}^n \times \mathbb{R}^n} \{(t, x, v, u) \in T \times \mathbb{R}^n \times \mathbb{R}^n \times B_M : L(t, x, u) \leq \lambda, u \in M(t, x, v)\}$, where $B_M = \{u \in \mathbb{R}^k : |u| \leq M\}$. Also $\text{Gr} M = \{(t, x, u, v) \in T \times \mathbb{R}^n \times \mathbb{R}^n \times B_M : v \in Ax + f(t, x, u), u \in U(t, x)\} = \{(t, x, u, v) \in T \times \mathbb{R}^n \times \mathbb{R}^n \times B_M : (x, v - f(t, x, u)) \in \text{Gr} A, (t, x, u, v) \in \text{Gr} \hat{U}\}$, where $\hat{U}(t, x) = U(t, x) \times \mathbb{R}^n$. Observe that $(t, x, v, u) \rightarrow (x, v - f(t, x, u))$ is measurable (hypothesis $H(f)$ (1)) and $\text{Gr} A$ is closed (since A is maximal monotone, see Barbu [4]). So $\{(t, x, v, u) \in T \times \mathbb{R}^n \times \mathbb{R}^n \times B_M : (x, v - f(t, x, u)) \in \text{Gr} A\} \in B(T) \times B(\mathbb{R}^n) \times B(\mathbb{R}^n) \times B(B_M)$. Also since $U(\cdot, \cdot)$ is graph measurable (hypothesis $H(U)$ (1)), $\hat{U}(\cdot, \cdot)$ is, too, and of course $(t, x, v, u) \rightarrow (t, x, u, v)$ is measurable. Thus, $\{(t, x, v, u) \in T \times \mathbb{R}^n \times \mathbb{R}^n \times B_M : (u, v) \in \hat{U}(t, x)\} \in B(T) \times B(\mathbb{R}^n) \times B(\mathbb{R}^n) \times B(B_M)$. Therefore, we deduce that $\text{Gr} M \in B(T) \times B(\mathbb{R}^n) \times B(\mathbb{R}^n) \times B(B_M)$. Using this

fact, hypothesis $H(L)$ and Novikov's theorem (see Levin [10], lemma 2.2), we get that $\Lambda_\lambda \in B(T) \times B(\mathbf{R}^n) \times B(\mathbf{R}^n)$, establishing claim #1.

Claim #2: $(x, v) \rightarrow p(t, x, v)$ is l. s. c..

To prove this claim, we need to show that for every $\lambda \in \mathbf{R}$, the set $K_\lambda = \{(x, v) \in \mathbf{R}^n \times \mathbf{R}^n : p(t, x, v) \leq \lambda\}$ is closed. So, let $\{(x_n, v_n)\}_{n \geq 1} \subseteq K_\lambda$ and assume that $(x_n, v_n) \rightarrow (x, v)$ as $n \rightarrow \infty$. Since $\lambda < \infty$, $M(t, x_n, v_n) \neq \emptyset$ and so by Weierstrass theorem we can find $u_n \in M(t, x_n, v_n)$ $n \geq 1$ s. t. $p(t, x_n, v_n) = L(t, x, u_n)$. By passing to a subsequence if necessary, we may assume that $u_n \rightarrow u$ as $n \rightarrow \infty$. Then using hypothesis $H(L)$ (2), we have $L(t, x, u) \leq \liminf L(t, x_n, u_n) = \liminf p(t, x_n, v_n) \leq \lambda$. So, to prove our claim, it suffices to show that $u \in M(t, x, v)$. Note that for every $n \geq 1$ $v_n \in Ax_n + f(t, x_n, u_n) \Rightarrow (x_n, v_n - f(t, x_n, u_n)) \in \text{Gr} A$ and $(x_n, v_n - f(t, x_n, u_n)) \rightarrow (x, v - f(t, x, u))$ as $n \rightarrow \infty$ (hypothesis $H(f)$ (2)). Since $\text{Gr} A$ is closed $(x, v - f(t, x, u)) \in \text{Gr} A \Rightarrow v \in Ax + f(t, x, u)$. Also because of hypothesis $H(U)$ (2), we have $u \in U(t, x)$. Hence, $u \in M(t, x, v)$ and so $p(t, x, v) \leq L(t, x, u) \leq \lambda \Rightarrow K_\lambda$ is closed and the claim is proved.

Claim #3: $p(t, x, \cdot)$ is convex.

Note that $\text{epi} p(t, x, \cdot) = Q(t, x)$. Then the claim follows from hypothesis H_c .

Let $\{S(t)\}_{t \in T}$ be the semigroup of nonlinear contractions generated by A . Then from theorem 2.1, p. 124 of Barbu [4], for any trajectory $x(\cdot) \in C(T, \mathbf{R}^n)$, we have

$$|x(t) - S(t)x_0| \leq \int_0^t |f(s, x(s), u(s))| ds \Rightarrow |x(t)| \leq |S(t)x_0| + \int_0^t |f(s, x(s), u(s))| ds \leq$$

$$N + \int_0^t (a(s) + b(s))(|x(s)| + |u(s)|) ds \leq N + \int_0^t (a(s) + b(s)|x(s)|) ds, \text{ where } \hat{a}(s) =$$

$a(s) + b(s)M$. Then invoking Gronwall's inequality, we get $|x(t)| \leq N_1$, $N_1 > 0$.

Thus, for every admissible pair (x, u) we have $|f(t, x, u)| \leq a(t) + b(t)(N_1 + M)$ (hypotheses $H(f)$ (3) and $H(U)$ (3)). So, corollary 2.3.1, p. 67 of Vrabie [17], tells us that the set of admissible trajectories of (*) is relatively compact in $C(T, \mathbf{R}^n)$.

So, if $\{(x_n, u_n)\}_{n \geq 1}$ is a minimizing sequence for (*), by passing to a subsequence if necessary, we may assume that $x_n \rightarrow x$ in $C(T, \mathbf{R}^n)$ as $n \rightarrow \infty$. Also from lemma 3.1 of Colombo-Fonda-Ornelas [7] and the Dunford-Pettis compactness criterion, we have that $\dot{x}_n \xrightarrow{w} \dot{x}$ in $L^1(\mathbf{R}^n)$ as $n \rightarrow \infty$. Then because of claims #1, #2, #3 and because of hypothesis $H(L)$ (3), we can apply theorem 2.1 of Balder [3] and get that

$$\begin{aligned} \int_0^t p(t, x(t), -\dot{x}(t)) dt &\leq \liminf \int_0^b p(t, x_n(t), -\dot{x}_n(t)) dt \\ &\leq \liminf \int_0^b L(t, x_n(t), u_n(t)) dt = m > \infty \text{ (hypotesis } H_a), \\ &\Rightarrow p(t, x(t), -\dot{x}(t)) < \infty \text{ a. e..} \end{aligned}$$

By modifying the function on a set of measure zero, we can say that $p(t, x(t), -\dot{x}(t))$ is finite for every $t \in T$, hence $M(t, x(t), -\dot{x}(t)) \neq \emptyset$ for every $t \in T$. Let $R(t) = \{u \in M(t, x(t), -\dot{x}(t)) : p(t, x(t), -\dot{x}(t)) = L(t, x(t), u)\}$. From Weierstrass theorem $R(t) \neq \emptyset$ for all $t \in T$. Also because of hypothesis $H(L)$ (1), claim #1 and the graph measurability of $M(\cdot, \cdot, \cdot)$ we have $\text{Gr } R \in B(T) \times B(\mathbf{R}^k)$. Applying the Lusin–Yankov–Aumann selection theorem (see Levin [10] theorem 1 and Wagner [18] theorem 5.8), we get $u : T \rightarrow \mathbf{R}^k$ measurable set $u(t) \in R(t)$ for all $t \in T$. Then $p(t, x(t), -\dot{x}(t)) = L(t, x(t), u(t))$ a. e. $\Rightarrow \int_0^b L(t, x(t), u(t)) dt \leq m$. But $u(t) \in$

$M(t, x(t), -\dot{x}(t))$ a. e.. So, (x, u) is admissible, thus $J(x, u) = \int_0^b L(t, x(t), u(t)) dt = m$. Hence, (x, u) is optimal.

Q.E.D.

3. Relaxed problem

Now, we remove hypothesis H_c . Then in order to be able to guarantee optimal solution, we need to pass to a larger system, known as the “relaxed system”. This is the following:

$$\left\{ \begin{array}{l} J_r(x, \lambda) = \int_0^b \int_{B_M} L(t, x(t), u) \lambda(t) (du) dt = m_r \\ \text{s. t. } -\dot{x}(t) \in Ax(t) + \int_{B_M} f(t, x(t), u) \lambda(t) (du) \text{ a. e.} \\ x(0) = x_0, \quad \lambda(\cdot) \in S_{\Sigma(\cdot, x(\cdot))} \end{array} \right. \quad (*)_r$$

Here $\Sigma(t, x) = \{\mu \in M_+^1(B_M) : \mu(U(t, x)) = 1\}$, with $M_+^1(B_M)$ being the space of probability measures on the compact metric space $B_M = \{u \in \mathbf{R}^k : |u| \leq M\}$. Also $S_{\Sigma(\cdot, x(\cdot))}$ is the set of measurable selectors of $\Sigma(\cdot, x(\cdot))$. Thus, the elements of $S_{\Sigma(\cdot, x(\cdot))}$ are transition probabilities. Note that $(*)$ embeds into $(*)_r$ by sending the original control $u(\cdot)$ in $\delta_{u(\cdot)}(\cdot)$ the Dirac transition probability at $u(\cdot)$. Finally, by $M(B_M)$ we will denote the space of bounded regular Borel measures, endowed with the weak (narrow) topology. From the Dinculeanu–Foias theorem we get $L^1(T, C(B_M))^* = L^\infty(T, M(B_M))$ (see theorem 18, p. 268 of Warga [19]).

THEOREM 3.1. *If hypotheses $H(A)$, $H(f)$, $H(U)$, $H(L)$ and H_a hold, then $(*)_r$ admits an optimal admissible pair.*

Proof. Let $\{(x_n, \lambda_n)\}_{n \geq 1}$ be a minimizing sequence for problem $(*)_r$. As before, by passing to a subsequence if necessary, we may assume that $x_n \rightarrow x$ in $C(T, \mathbf{R}^n)$ and $\dot{x}_n \xrightarrow{w} \dot{x}$ in $L^1(T, \mathbf{R}^n)$. Also from Alaoglu’s theorem, we may

assume that $\lambda_n \xrightarrow{w^*} \lambda$ in $L^\infty(T, M(B_M))$. Then since $L(\cdot, \cdot, \cdot)$ is a normal integrand (hypothesis H(L)), as in the proof of theorem 3.2 of [8], we have

$$\int_0^b \int_{B_M} L(t, x(t)u)\lambda(t) (du) dt \leq \varliminf_{n \rightarrow \infty} \int_0^b \int_{B_M} L(t, x_n(t), u)\lambda_n(t) (du) dt = m_r.$$

Also for every $h \in L^1(T, C(B_M))$ we have

$$\langle \lambda, h \rangle \leq \overline{\lim} \int_0^b \sigma_{\Sigma(t, x_n(t))}(h(t)) dt \leq \int_0^b \overline{\lim} \sigma_{\Sigma(t, x_n(t))}(h(t)) dt$$

where $\langle \cdot, \cdot \rangle$ denotes the duality brackets for the pair $(L^1(T, C(B_M)), L^\infty(T, M(B_M)))$ and $\sigma_{\Sigma(t, x_n(t))}(\cdot)$ the support functions of the set $\Sigma(t, x_n(t))$.

We claim that $\overline{\lim} \Sigma(t, x_n(t)) \subseteq \Sigma(t, x(t))$ for all $t \in T$. To this end let $\mu \in \overline{\lim} \Sigma(t, x_n(t))$. By definition we can find $\mu_{n_k} \in \Sigma(t, x_{n_k}(t))$ $k \geq 1$ s. t. $\mu_{n_k} \xrightarrow{w} \mu$ in $M_+^1(B_M)$. Then from theorem 2 of Lucchetti-Salinetti-Wets [11], we have $\overline{\lim} \mu_{n_k}(U(t, x_n(t))) \leq \mu(U(t, x)) \Rightarrow \mu(U(t, x)) = 1 \Rightarrow \mu \in \Sigma(t, x(t))$. Thus, the claim follows and from it we get (see proposition 3.1 in [12])

$$\overline{\lim} \sigma_{\Sigma(t, x_n(t))}(h(t)) \leq \sigma_{\Sigma(t, x(t))}(h(t)) \Rightarrow \langle \lambda, h \rangle \leq \int_0^b \sigma_{\Sigma(t, x(t))}(h(t)) dt.$$

Since $h \in L^1(T, C(B_M))$ was arbitrary, from the last inequality we deduce that $\lambda(\cdot) \in S_{\Sigma(\cdot, x(\cdot))}$.

Note that because of hypothesis H(f) (2) and since $\lambda_n \xrightarrow{w^*} \lambda$ in $L^\infty(T, M(B_M))$, $\eta_n(t) = \int_{B_M} f(t, x_n(t), u)\lambda_n(t) (du) \rightarrow \eta(t) = \int_{B_M} f(t, x(t)u)\lambda(t) (du)$. Let \hat{A} the realization of A on $L^1(T, \mathbf{R}^n)$. We know that \hat{A} is maximal monotone (see Barbu [4]). So, $\text{Gr } \hat{A}$ is semiclosed. Also note that $(x_n, -\dot{x}_n - \eta_n) \xrightarrow{s \times w} (x, -\dot{x} - \eta)$ in $L^1(T, \mathbf{R}^n) \times L^1(T, \mathbf{R}^n)$. Hence, $(x_n, -\dot{x}_n - \eta_n) \in \text{Gr } \hat{A}$. Therefore, $-\dot{x}(t) \in Ax(t) + \int_{B_M} f(t, x(t), u)\lambda(t) (du)$, $x(0) = x_0 \Rightarrow (x, \lambda)$ is an admissible relaxed pair. So, $J_r(x, \lambda) = m_r$ i.e. (x, λ) is optimal for $(*)_r$.

Q.E.D.

4. A density result

In this section we compare the sets of trajectories of $(*)$ and $(*)_r$. Denote the first by $P(x_0)$ and the latter by $P_r(x_0)$. Our goal is to show that $P(x_0)$ is dense

in $P_r(x_0)$ for the $C(T, \mathbf{R}^n)$ -topology. For this we will need the following stronger hypotheses.

$H(f)_1$: $f : T \times \mathbf{R}^n \times \mathbf{R}^k \rightarrow \mathbf{R}^k$ is a map s. t.

- (1) $t \rightarrow f(t, x, u)$ is measurable,
- (2) $|f(t, x, u) - f(t, y, v)| \leq k(t)(|x - y| + |u - v|)$ a. e. with $k(\cdot) \in L^1_+$,
- (3) $|f(t, x, u)| \leq a(t) + b(t)(|x| + |u|)$ a. e. with $a(\cdot), b(\cdot) \in L^1_+$.

$H(U)_1$: $U : T \times \mathbf{R}^n \rightarrow P_{fc}(\mathbf{R}^k)$ is a multifunction s. t.

- (1) $t \rightarrow U(t, x)$ is measurable,
- (2) $h(U(t, x), U(t, y)) \leq r(t)|x - y|$ a. e. with $r(\cdot) \in L^\infty_+$ and $h(\cdot, \cdot)$ being the Hausdorff metric on $P_{fc}(\mathbf{R}^k)$,
- (3) $|U(t, x)| \leq M$.

THEOREM 4.1. *If hypotheses $H(A)$, $H(f)_1$ and $H(U)_1$ hold, then $P_r(x_0) = \overline{P(x_0)}$, the closure taken in $C(T, \mathbf{R}^n)$.*

Proof. Let $\eta : L^1(T, \mathbf{R}^n) \rightarrow C(T, \mathbf{R}^n)$ be the map that each $h(\cdot) \in L^1(T, \mathbf{R}^n)$ assigns the unique solution of the evolution $-\dot{x}(t) \in Ax(t) + h(t)$ a. e., $x(0) = x_0$. We know (see Vrabie [17], corollary 2.3.1, p. 67), that $\eta(\cdot)$ is weakly-strongly continuous. Let $F(t, x) = f(t, x, U(t, x))$. Clearly, $F(\cdot, \cdot)$ is compact-valued in \mathbf{R}^n . Fix $x \in \mathbf{R}^n$ and let $u_n : T \rightarrow \mathbf{R}^k$ s. t. $U(t, x) = \overline{\{u_n(t)\}_{n \geq 1}}$. Such a sequence exists because of $H(U)_1$ (1) (see Wagner [18]). So, $F(t, x) = \overline{\{f(t, x, u_n(t))\}_{n \geq 1}}$ and for every $n \geq 1$, $t \rightarrow f(t, x, u_n(t))$ is measurable (hypotheses $H(f)$ (1) and (2)). Thus, again by Wagner [18] (theorem 4.2), we deduce that $t \rightarrow F(t, x)$ is measurable. Now, fix $t \in T$. Let $x, y \in \mathbf{R}^n$ and $z \in F(t, x)$. We have $z = f(t, x, v)$, $v \in U(t, x)$. Let $w \in U(t, y)$ s. t. $d(v, U(t, y)) = |v - w|$. Then we can write

$$\begin{aligned} d(z, F(t, y)) &\leq |f(t, x, v) - f(t, y, w)| \\ &\leq k(t)(|x - y| + |v - w|) \text{ a. e. (hypothesis } H(f)_1 \text{ (2))} \\ &\leq k(t)(|x - y| + h(U(t, x), U(t, y))) \text{ a. e.} \\ &\leq k(t)(|x - y| + r(t)|x - y|) \text{ a. e. (hypothesis } H(U)_1 \text{ (2))} \\ &\leq l(t)|x - y| \text{ a. e., where } l(t) = k(t)(1 + \|r\|_\infty). \end{aligned}$$

Using the Lusin-Yankov-Aumann selection theorem we can easily check that every trajectory of the differential inclusion $-\dot{x}(t) \in Ax(t) + F(t, x(t))$ a. e. $x(0) = x_0$ is an admissible trajectory of (*) and, of course, vice versa.

Finally, from [14] we know that $\overline{\text{conv}} F(t, x) = \left\{ \int_{B_M} f(t, x, u) \lambda(du) : \lambda \in \Sigma(t, x) \right\}$ and $S^1_{\overline{\text{conv}} F(\cdot, x(\cdot))} = \left\{ \int_{B_M} f(t, x(t), u) \lambda(t)(du) : \lambda(\cdot) \in S_{\Sigma(\cdot, x(\cdot))} \right\}$.

Next, let $x(\cdot) \in P_r(x_0)$. Then $x = \eta(g)$ with $g \in S^1_{\overline{\text{conv}} F(\cdot, x(\cdot))}$. Let $\epsilon > 0$. We can find \mathcal{U} , a balanced convex weak neighbourhood of the origin in $L^1(T, \mathbf{R}^n)$ s. t. if $g_1 \in L^1(T, \mathbf{R}^n)$, $g - g_1 \in \mathcal{U}$, then $\|x - z_1\|_\infty < \epsilon$, where $z_1 = \eta(g_1)$. This

is possible since $\eta(\cdot)$ is weakly-strongly continuous. From proposition 4.1 of [13], we know that we can choose $g_1 \in S_{F(\cdot, x(\cdot))}^1$. Through the Lusin-Yankov-Aumann selection theorem, we can find $g_2 : T \rightarrow \mathbf{R}^n$ measurable s. t. $d(g_1(t), F(t, z_1(t))) = |g_1(t) - g_2(t)|$, $g_2(\cdot) \in S_{F(\cdot, z_1(\cdot))}^1$. We have $|g_1(t) - g_2(t)| \leq h(F(t, x(t)), F(t, z_1(t))) \leq l(t)|x(t) - z_1(t)| \leq l(t)\epsilon$ a. e.. So, if $z_2(t) = \eta(g_2)$, we have $|z_2(t) - x(t)| \leq |z_2(t) - x_1(t)| + |z_1(t) - x(t)| \leq \int_0^t |g_1(s) - g_2(s)| ds + \epsilon \leq \epsilon \left(\int_0^t l(s) ds + 1 \right)$.

Suppose that we obtained $g_1, \dots, g_n \in L^1(T, \mathbf{R}^n)$ s. t. $|g_{k+1}(t) - g_k(t)| \leq \epsilon l(t) \frac{1}{(k-1)!} \left(\int_0^t l(s) ds \right)^{k-1}$, $g_{k+1}(t) \in F(t, z_k(t))$ a. e. $z_k = \eta(g_k)$ $k = 1, 2, \dots, n-1$.

Then we can write

$$\begin{aligned} |z_{k+1}(t) - z_k(t)| &\leq \int_0^t |g_{k+1}(s) - g_k(s)| ds \leq \epsilon \int_0^t \frac{l(s)}{(k-1)!} \left(\int_0^s l(r) dr \right)^{k-1} ds \\ &= \frac{\epsilon}{k!} \int_0^t d \left(\int_0^s l(r) dr \right)^k = \frac{\epsilon}{k!} \int_0^t d \left(\int_0^s l(s) ds \right)^k. \end{aligned}$$

Hence, we have

$$|z_{k+1} - x(t)| \leq \epsilon \sum_{q=1}^{k+1} \frac{1}{q!} \left(\int_0^t l(s) ds \right)^q \leq \epsilon \exp \|l\|_1.$$

Once again the Lusin-Yankov-Aumann selection theorem gives us $g_{n+1} \in S_{F(\cdot, z_n(\cdot))}^1$ s. t.

$$\begin{aligned} |g_{n+1}(t) - g_n(t)| &\leq h(F(t, z_n(t)), F(t, z_{n-1}(t))) \leq l(t)|z_n(t) - z_{n-1}(t)| \\ &\leq \frac{\epsilon}{(n-1)!} l(t) \left(\int_0^t l(s) ds \right)^{n-1} \end{aligned}$$

and so the induction is completed.

It is clear from this construction that $g_n \xrightarrow{s} \hat{g}$ in $L^1(T, \mathbf{R}^n)$, $\hat{g} \in L^1(T, \mathbf{R}^n)$. Then $z_n = \eta(g_n) \xrightarrow{s} z = \eta(\hat{g})$ and $\hat{g}(t) \in \lim F(t, z_n(t)) = F(t, z(t)) \Rightarrow z(\cdot) \in \overline{P(x_0)}$. In the limit as $n \rightarrow \infty$ we have $\|x - z\|_\infty \leq \epsilon \exp \|l\|_1$. Since $\epsilon > 0$ was arbitrary and from the observation at the beginning of the proof, we conclude that $P_r(x_0) \subseteq \overline{P(x_0)}$, the closure in $C(T, \mathbf{R}^n)$. But it is easy to check that $P_r(x_0)$ is closed. So, $P_r(x_0) = \overline{P(x_0)}$.

Q.E.D.

Remark. There is a counter-example due to Pliss (see Aubin-Cellina [4]), which illustrates that simple continuity of the orientor field $x \rightarrow F(t, x) = f(t, x, U(t, x))$ is not enough to give us the above density result.

5. Relation between original and relaxed problems

The aim in this Section is to prove that $m = m_r$. We were able to prove this only for open-loop systems (i.e. $U(t, x) = U(t)$, independent of x). It will be very interesting to know whether it can be also proved by closed-loop (feedback) systems.

We will need the following stronger hypothesis on the cost integrand $L(t, x, u)$.

$H(L)_1$: $L : T \times \mathbf{R}^n \times \mathbf{R}^k \rightarrow \mathbf{R}$ is an integrand s. t.

- (1) $t \rightarrow L(t, x, u)$ is measurable,
- (2) $(x, u) \rightarrow L(t, x, u)$ is continuous,
- (3) $|L(t, x, u)| \leq a_1(t) + b_1(t)(|x| + |u|)$ a. e. with $a_1(\cdot), b_1(\cdot) \in L^1_+$.

Also the hypothesis on the control constraints set has now the following simpler form

$H(U)_2$: $U : T \rightarrow P_{fc}(\mathbf{R}^k)$ is a measurable multifunction s. t. $|U(t)| \leq M$ for all $t \in T$.

THEOREM 5.1. *If hypotheses $H(A), H(f)_1, H(U)_1, H(L)_1$ and H_a hold, then $m = m_r$.*

Proof. Let (x, λ) be an optimal admissible pair for $(*)_r$. From theorem 3.1 we know that such a pair exists. Invoking corollary 3 of Balder [2], we can find $u_n \in S^1_U$ s. t. $\delta_{u_n} \xrightarrow{w^*} \lambda$ in $l^\infty(T, M(B_M))$, where δ_{u_n} denotes the Dirac transition probability concentrated at $u_n(\cdot)$. Let $x_n(\cdot) \in C(T, \mathbf{R}^n)$ be the unique original trajectory generated by control $u_n(\cdot)$. Let $\hat{L}_n : T \rightarrow L^1(T, C(B_M))$ be defined by $\hat{L}_n(t)(\cdot) = L(t, x_n(t), \cdot)$ and $\hat{L} : T \rightarrow L^1(T, C(B_M))$ by $\hat{L}(t)(\cdot) = L(t, x(t), \cdot)$. Using hypothesis $H(L)$, it is easy to check that $\hat{L}_n \xrightarrow{s} L$ in $L^1(T, C(B_M))$. Then, if as before by $\langle \cdot, \cdot \rangle$ we denote the duality brackets for the pair $(L^1(T, C(B_M)), L^\infty(T, C(B_M)))$, we have $\langle \hat{L}_n, \delta_{u_n} \rangle = \int_0^b L(t, x_n(t), u_n(t)) dt \rightarrow \langle \hat{L}, \lambda \rangle = \int_0^b \int_{B_M} L(t, x(t), u) \lambda(t) (du) dt = m_r \Rightarrow m \leq m_r$. Since we always have $m_r \leq m$, we conclude $m = m_r$.

Q.E.D.

Acknowledgement

The author would like to thank the referee for his constructive criticism.

References

1. Aubin, J.-P., Cellina, A., *Differential Inclusions*. Springer, Berlin, 1984.
2. Balder, E., A general denseness result for relaxed control theory. *Bull. Austr. Math. Soc.* **30** (1984), pp. 463–475.
3. Balder, E., Necessary and sufficient conditions for L_1 -strong-weak lower semicontinuity of integral functionals. *Nonl. Anal. TMA* **11** (1987), pp. 1399–1404.
4. Barbu, V., *Nonlinear Semigroups and Differential Equations in Banach Spaces*. Noordhoff International Publishing, Leyden, the Netherlands, 1976.
5. Berkovitz, L., *Optimal Control Theory*. Springer-Verlag, New York, 1974.
6. Cesari, L., *Optimization — Theory and Applications*. Springer-Verlag, New York, 1983.
7. Colombo, G., Fonda, A., Ornelas, A., Lower semicontinuous perturbations of maximal monotone differential inclusions. *Israel J. Math.* **61** (1988), pp. 211–218.
8. Flytzanis, E., Papageorgiou, N. S., On the existence of optimal controls for a class of nonlinear infinite dimensional systems. *Math. Nachrichten* **150** (1991), pp. 203–217.
9. Gamkrelidze, R., *Principles of Optimal Control Theory*. Plenum Press, New York, 1978.
10. Levin, V., Borel sections of many-valued maps. *Siberian Math. Jour.* **19** (1979), pp. 434–438.
11. Lucchetti, R., Salinetti, G., Wets, R., Uniform convergence of probability measures; topological criteria. *Annals of Statistics* (to appear)
12. Papageorgiou, N. S., Convergence theorems for Banach space valued integrable multifunctions. *Intern. J. Math. and Math. Sci.* **10** (1987), pp. 433–442.
13. Papageorgiou, N. S., Measurable multifunctionals and their applications to convex integral functionals. *Intern. J. Math. and Math. Sci.* **12** (1989), pp. 175–192.
14. Papageorgiou, N. S., Optimal control of nonlinear evolution inclusions. *J. Optim. Theory Appl.* **67** (1990), pp. 321–354.
15. Pappas, G., An approximation result for normal integrands and applications to relaxed control theory. *J. Math. Anal. Appl.* **93** (1983), pp. 132–141.
16. Parthasarathy, K., *Probability Measures on Metric Spaces*. Academic Press, New York, 1967.
17. Vrabie, I., *Compactness Methods for Nonlinear Evolutions*. Longman Scientific and Technical, London, 1987.
18. Wagner, D., Survey of measurable selection theorems. *SIAM J. Control and Optim.* **15** (1977), pp. 859–903.
19. Warga, J., *Optimal Control of Differential and Functional Equations*. Academic Press, New York, 1970.

**Об оптимальном управлении и релаксации конечномерных систем
с максимально монотонным оператором,
воздействующем на фазовые скорости систем**

Н. С. ПАПАГЕОРГИУ

(Афины)

В статье рассматривается круг вопросов, связанных с решением задач оптимального управления дифференциальными включениями в пространство конечной

размерности. Многозначность в уравнении динамики управляемой системы вызвана многозначным максимально монотонным (вообще говоря, неограниченным) оператором, воздействующим на фазовые скорости системы.

Nikolaos S. Papageorgiou
National Technical University
Department of Mathematics
Zografou Campus
Athens 157 73
Greece

ON FINITE DIMENSIONAL FILTERING IN DISCRETE TIME

M. FERRANTE

(Trieste)

(Received June 12, 1990)

In this paper we prove that, under suitable regularity assumptions, a necessary condition for the existence of a finite-dimensional filter in discrete time is the possibility of finding a convenient decomposition of the filter system $s(z, y)$ into a sum of a term depending only on the first variable plus one depending only on the second one. We also give additional results concerning the observation, prediction and filtering distributions.

Keywords: Nonlinear filtering, finite dimensional filters, exponential class of distributions.

1. Introduction

Let $(X_n, Y_n)_{n \in \mathbb{N}}$ be a discrete time stochastic process, where $X_n \in X \subset \mathbb{R}^p$, $p \in \mathbb{N}$, and $Y_n \in Y \subset \mathbb{R}^m$, $m \in \mathbb{N}$. The component Y_n can only be observed, while the component X_n (signal or state process) is unobservable. We assume

A.1: (X_n) is a Markov process with transition kernel $P_{X_n}(\cdot | X_{n-1} = x_{n-1})$ and initial distribution $P_{X_0}(\cdot)$;

A.2: (Y_n) satisfies the following "conditional independence property":

$$P_{Y_n}(\cdot | X^{n-1} = x^{n-1}, Y^{n-1} = y^{n-1}, X_n = x_n) = P_{Y_n}(\cdot | X_n = x_n)$$

where $X^{n-1} := (X_1, \dots, X_{n-1})$ and, analogously, for $x^{n-1}, Y^{n-1}, y^{n-1}$.

We shall assume that all distributions have strictly positive densities with respect to the same dominating measure which, to fix ideas, we shall take as Lebesgue measure. We shall denote by $p(x_n | x_{n-1}), p(x_0), p(y_n | x_n)$ the densities corresponding to $P_{X_n}(\cdot | X_{n-1} = x_{n-1}), P_{X_0}(\cdot), P_{Y_n}(\cdot | X_n = x_n)$, respectively.

We shall consider the filtering problem for $(X_n, Y_n)_{n \in \mathbb{N}}$, which in its most complete form, consists in computing, for each step n , the conditional density $p(x_n | y^n)$ of x_n given y^n . By Bayes rule, we have the following relation

$$p(x_n|y^n) = \frac{p(y_n|x_n)p(x_n|y^{n-1})}{\int_X p(y_n|x_n)p(x_n|y^{n-1}) dx_n}. \quad (1)$$

We shall call $p(x_n|y^n)$ the *filtering*, $p(x_n|y_n)$ the *observation* and $p(x_n|y^{n-1})$ the *prediction density*. Moreover,

$$P(x_n|y^{n-1}) = \int_X p(y_n|x_{n-1})p(x_{n-1}|y^{n-1}) dx_{n-1}. \quad (2)$$

DEFINITION 1.1 (e.g. Sawitzki (1981), Van Schuppen (1979)). Let $\{p(x; z), z \in Z \subset \mathbf{R}^k\}$ $k > 0$, be a family of densities on X parametrized by $z \in Z$. We say that a sequence of measurable functions

$$\varphi_n : Z \times Y \rightarrow Z$$

is a *finite dimensional* (k -dimensional) *filter system* for (X_n, Y_n) if $\forall n \geq 1$

$$\{p(x_{n-1}|y^{n-1}) = p(x_{n-1}; z) \text{ for some } z \in Z\} \Rightarrow \{p(x_n|y^n) = p(x_n|\varphi_n(z, y_n))\}. \quad (3)$$

We say that $\{\varphi_n\}_{n \in \mathbf{N}}$ is *minimal* if k is the minimum positive integer for which the previous condition holds.

Remark 1.1. Notice that, if for all n the filtering densities belong to the same family on X parametrized by $z \in Z$, then by (2) also the prediction densities belong, for all n , to the same family of densities on X parametrized by $z \in Z$. We denote this latter family by

$$\{p'(x; z), z \in Z \subset \mathbf{R}^k\}.$$

We give also the following

DEFINITION 1.2 (e.g. Barndorff-Nielsen (1978)). We shall say that a family $\{p(x; z), z \in Z \subset \mathbf{R}^k\}$ of densities on X is of *exponential class of order q* if q is the smallest positive integer such that there exist $q + 1$ pairs of functions $(a_i(x), b_i(z))$, $i = 0, \dots, q$, such that for all $z \in Z$ we have the representation

$$p(x; z) = a_0(x)b_0(z) \exp\left\{\sum_{i=1}^q a_i(x)b_i(z)\right\}. \quad (4)$$

Remark 1.2. Notice that if $\{p(x; z), z \in Z \subset \mathbf{R}^k\}$ is of exponential class of order q , then the functions $a_i(x)$, $b_i(z)$ of the previous representation are not constant for every $i \in (1, \dots, q)$.

It is possible to prove (e.g. Ferrante (1989) and Ferrante, Runggaldier (1990), which extend to the multidimensional case results contained in Sawitzki (1981))

that a necessary condition for the existence of a finite-dimensional filter system in discrete time, is that, under suitable regularity assumptions, the filtering, observation and prediction densities are all of exponential class. The regularity assumptions mentioned above can be summarized in the following two:

B.1 Z and Y are connected sets with nonempty interior;

B.2 $\varphi_n(\cdot, \cdot)$ is differentiable and, at every step n , there exist $z_0 \in \text{int}(Z)$ and $y_0 \in \text{int}(Y)$ such that

$$\begin{aligned} \det\left(\frac{\partial}{\partial z}\varphi_n(z, y)\right)\Bigg|_{z=z_0} &\neq 0 & \forall y \in Y' \text{ open and dense in } Y \\ \text{rank}\left(\frac{\partial}{\partial z}\varphi_n(z, y)\right)\Bigg|_{y=y_0} &= k & \forall z \in Z' \text{ open and dense in } Z \end{aligned}$$

where $k = \dim(Z)$. (For a generic subset X of a linear space we denote by $\dim(X)$ the dimension of its affine hull.)

Remark 1.3. If $\dim(Z) = k > m = \dim(Y)$, the assumption B.2 is replaced by a slightly stronger assumption (see Ferrante (1989)), since it does not make sense in this case to ask that $\text{rank}\left(\frac{\partial}{\partial z}\varphi_n(z, y)\right)\Big|_{y=y_0} = k$, being the matrix of rank at most $m < k$.

More precisely, the above mentioned results lead to the following representation for the prediction, observation and filtering densities, respectively:

$$\begin{aligned} p'(x; z) &= a(x)b(z) \exp\left\{\sum_{i=1}^{q_1} \alpha_i(x)\beta_i(z)\right\} \\ p(y; x) &= c(x)d(y) \exp\left\{\sum_{i=1}^{q_2} \gamma_i(x)\delta_i(y)\right\} \\ p(x; \varphi(z, y)) &= t(x)s(\varphi(z, y)) \exp\left\{\sum_{i=1}^{q_3} \tau_i(x)\sigma_i(\varphi(z, y))\right\} \end{aligned} \tag{5}$$

where, letting $k = \dim Z$, $m = \dim Y$,

$$p := \max\{k, \nu m\} \quad \text{and} \quad \nu := \begin{cases} 0, & \text{if } m \geq k \\ \min\{n \in \mathbf{N} : nm > k\}, & \text{if } m < k \end{cases}$$

we have $q_1 \leq p$, $q_2 \leq k$, $q_3 \leq p$.

Remark 1.4. Notice that if we assume that the filter system is minimal, then we have $q_3 \geq k$ (otherwise we could choose $(\sigma_1(\varphi(z, y)), \dots, \sigma_{q_3}(\varphi(z, y)))$ as filter system and k would not be minimal). Moreover, if $\nu m = k$ then we have that $q_3 = k$.

In the present paper we are interested in giving further characterizations of the relations (5) based always only on the assumption that there exists a finite-dimensional filter as well as on the assumptions B.1 and B.2. In particular, we shall prove that every $\sigma_i(\varphi(z, y))$ can be written as the sum of a function of y plus a function of z (except for a constant) and that $\alpha_i(x)$, $\gamma_i(x)$ are functions of $\tau(x) = (\tau_1(x), \dots, \tau_{q_3}(x))$ only. In this way we prove that every partially observable stochastic process admitting a finite-dimensional filter, has a filter system consisting of functions that can be decomposed in the way described above.

The interest of our results consists in having found a necessary condition on the filter system itself, assuming just that there exists a finite-dimensional filter in addition to some regularity assumption. Moreover, it is important to note that the technical results obtained here are exactly what one can expect from this sort of problem (similar results have been obtained for particular models in Sawitzki (1979)).

It is interesting to note that the natural assumption, made in Bather (1965) and Spizzichino (1988), namely that the filter system φ_n can be decomposed into a sum of two terms, both depending on a single variable, is now, under suitable regularity assumptions, a necessary condition for the existence of a finite-dimensional filter and thus the only one possible if one tries to find finite-dimensional filters in discrete time.

To conclude notice that the most important example of a finite-dimensional filter in discrete time, namely the Kalman-Bucy filter, valid for a linear Gaussian model, satisfies all our assumptions.

2. Main results

We prove the following

THEOREM 2.1. Let f be a function from $X \times Z$ into \mathbb{R} and suppose that there exist $a(x)$, $\alpha_1(x), \dots, \alpha_{q_1}(x) : X \rightarrow \mathbb{R}$ and $\beta_1(z), \dots, \beta_{q_1}(z) : Z \rightarrow \mathbb{R}$, $q_1 \geq 1$, with $a(x) > 0 \forall x$ such that

$$f(x, z) = a(x) \exp \left\{ \sum_{i=1}^{q_1} \alpha_i(x) \beta_i(z) \right\} \quad \forall x \in X \text{ and } \forall z \in Z \quad (6a)$$

with q_1 the minimum positive integer for which the previous property holds.

Let g be a function from $X \times Y$ into \mathbb{R} and suppose that there exist $c(x)$, $\gamma_1(x), \dots, \gamma_{q_2}(x) : X \rightarrow \mathbb{R}$ and $\delta_1(y), \dots, \delta_{q_2}(y) : Y \rightarrow \mathbb{R}$, $q_2 \geq 1$, with $c(x) > 0 \forall x$ such that

$$g(x, y) = c(x) \exp \left\{ \sum_{i=1}^{q_2} \gamma_i(x) \delta_i(y) \right\} \quad \forall x \in X \text{ and } \forall y \in Y \quad (6b)$$

with q_2 the minimum positive integer for which the previous property holds.

If there exist functions $t(x)$ and $\tau_i(x)$ from X into \mathbf{R} , $i \in (1, \dots, q_3)$ with $q_3 \geq 1$, and functions $r(z, y)$ and $\sigma_i(z, y)$ from $Z \times Y$ into \mathbf{R} , $i \in (1, \dots, q_3)$, for which

$$f(x, z)g(x, y) = t(x)r(z, y) \exp\left\{\sum_{i=1}^{q_3} \tau_i(x)\sigma_i(z, y)\right\} \quad \forall(x, y, z) \in X \times Y \times Z \quad (7)$$

with q_3 the minimum positive integer for which that holds, then there exist an invertible $q_3 \times q_3$ matrix A , a $q_3 \times q_1$ matrix B , a $q_3 \times q_2$ matrix C and a constant vector $P \in \mathbf{R}^{q_3}$ such that

$$\begin{aligned} \text{i)} \quad {}^t(\sigma_1(z, y), \dots, \sigma_{q_3}(z, y)) &= A^{-1}B^t(\beta_1(z), \dots, \beta_{q_1}(z)) + \\ &+ A^{-1}C^t(\delta_1(y), \dots, \delta_{q_2}(y)) + A^{-1}P \end{aligned}$$

Moreover, for all $x \in X$

$$\begin{aligned} \text{ii)} \quad {}^t(\alpha_1(x), \dots, \alpha_{q_1}(x)) &= {}^t(A^{-1}B)^t(\tau_1(x), \dots, \tau_{q_3}(x)) + \Delta \\ {}^t(\gamma_1(x), \dots, \gamma_{q_2}(x)) &= {}^t(A^{-1}C)^t(\tau_1(x), \dots, \tau_{q_3}(x)) + \Omega \end{aligned}$$

with Δ and Ω appropriate constant vectors.

Remark 2.1. Notice that the opposite implication of Theorem 2.1 is always true, namely, if there exist two functions f and g satisfying (6a) and (6b) respectively, and $\sigma_i(z, y)$, $\tau_i(x)$ for which i) and ii) hold, then by a convenient choice of $t(x)$ and $r(z, y)$ equation (7) holds.

Remark 2.2. The requirement that q_3 is the minimum positive integer for which (7) holds implies that none of the functions $\tau_1(x), \dots, \tau_{q_3}(x)$ is constant (and so for q_1 and q_2).

Proof. Evaluating (7) for a fixed value $\bar{x} \in X$, we obtain

$$\begin{aligned} a(\bar{x}) \exp\left\{\sum_{i=1}^{q_1} \alpha(\bar{x})\beta_i(z)\right\} c(\bar{x}) \exp\left\{\sum_{i=1}^{q_2} \gamma_i(\bar{x})\delta_i(y)\right\} &= \\ &= t(\bar{x})r(z, y) \exp\left\{\sum_{i=1}^{q_3} \tau_i(\bar{x})\sigma_i(z, y)\right\}. \end{aligned}$$

Let us now divide (7) by the previous relation and pass to the logarithm obtaining

$$\begin{aligned} \log \frac{a(x)c(x)t(\bar{x})}{a(\bar{x})c(\bar{x})t(x)} + \sum_{i=1}^{q_1} (\alpha_i(x) - \alpha_i(\bar{x}))\beta_i(z) + \sum_{i=1}^{q_2} (\gamma_i(x) - \gamma_i(\bar{x}))\delta_i(y) &= \\ &= \sum_{i=1}^{q_3} (\tau_i(x) - \tau_i(\bar{x}))\sigma_i(z, y). \end{aligned} \quad (8)$$

Putting

$$P(x) = \log \frac{\alpha(x)c(x)t(\bar{x})}{\alpha(\bar{x})\gamma(\bar{x})t(x)}$$

$\alpha(x) = (\alpha_1(x), \dots, \alpha_{q_1}(x))$ and, analogously for $\beta(z)$, $\gamma(x)$, $\delta(y)$, $\tau(x)$, $\sigma(z, y)$, relation (8) becomes

$$P(x) + \langle \alpha(x) - \alpha(\bar{x}), \beta(z) \rangle + \langle \gamma(x) - \gamma(\bar{x}), \delta(y) \rangle = \langle \tau(x) - \tau(\bar{x}), \sigma(z, y) \rangle. \quad (9)$$

Let us now prove that there exist $x_1, \dots, x_{q_3} \in X \setminus \{\bar{x}\}$ such that $(\tau(x_1) - \tau(\bar{x}), \dots, \tau(x_{q_3}) - \tau(\bar{x}))$ are linearly dependent for every possible choice of x_1, \dots, x_{q_3} , there exist an orthogonal matrix Q and a vector $b \in \mathbb{R}^{q_3}$ such that

$$Q(\tau(x) - \tau(\bar{x})) + b = \tilde{\tau}(x) \quad \text{with} \quad \tilde{\tau}_{q_3}(x) = 0 \quad \forall x \in X.$$

But this leads to a contradiction: in fact we then have (recall that $\bar{x} \in X$ is fixed)

$$\begin{aligned} & t(x)r(z, y) \exp \{ \langle \tau(x), \sigma(z, y) \rangle \} = \\ & = t(x)r(z, y) \exp \{ \langle \tau(x) - \tau(\bar{x}), \sigma(z, y) \rangle \} \exp \{ \langle \tau(\bar{x}), \sigma(z, y) \rangle \} = \\ & = t(x)r_1(z, y) \exp \{ \langle \tau(x) - \tau(\bar{x}), {}^t Q Q \sigma(z, y) \rangle \} \times \\ & \quad \exp \{ \langle b, Q \sigma(z, y) \rangle \} \exp \{ \langle -b, Q \sigma(z, y) \rangle \} = \\ & = t(x)r_2(z, y) \exp \{ \langle Q(\tau(x) - \tau(\bar{x})) + b, Q \sigma(z, y) \rangle \} = \\ & = t(x)r_2(z, y) \exp \{ \langle \tilde{\tau}(x), Q \sigma(z, y) \rangle \} = \\ & = t(x)r_2(z, y) \exp \left\{ \sum_{i=1}^{q_3-1} \langle \tilde{\tau}_i(x), Q \sigma(z, y) \rangle \right\} \end{aligned}$$

where

$$\begin{aligned} r_1(z, y) &= r(z, y) \exp \{ \langle \tau(\bar{x}), \sigma(z, y) \rangle \} \\ r_2(z, y) &= r_1(z, y) \exp \{ \langle -b, Q \sigma(z, y) \rangle \}. \end{aligned}$$

But we required q_3 to be the minimum positive integer for which two functions $t(x)$ and $r(y, z)$ satisfying (7) exist.

Let, therefore, $A = {}^t(\tau(x_1) - \tau(\bar{x}), \dots, \tau(x_{q_3}) - \tau(\bar{x}))$ be a $q_3 \times q_3$ matrix with linearly independent columns. Then $\det(A) \neq 0$ and so A is invertible.

Defining now $B = {}^t(\alpha(x_1) - \alpha(\bar{x}), \dots, \alpha(x_{q_3}) - \alpha(\bar{x}))$, $C = {}^t(\gamma(x_1) - \gamma(\bar{x}), \dots, \gamma(x_{q_3}) - \gamma(\bar{x}))$ and $P = {}^t(P(x_1), \dots, P(x_{q_3}))$, from (9) we have

$$P + B\beta(z) + C\delta(y) = A\sigma(z, y) \quad (10)$$

and so

$$\sigma(z, y) = A^{-1}B\beta(z) + A^{-1}C\delta(y) + A^{-1}P \quad (11)$$

which gives i).

To obtain ii), replacing $\sigma(z, y)$ in (9) by its expression (11), we have

$$\begin{aligned}
 P(x) + \langle \alpha(x) - \alpha(\bar{x}), \beta(z) \rangle + \langle \gamma(x) - \gamma(\bar{x}), \delta(y) \rangle = \\
 = \langle \tau(x) - \tau(\bar{x}), A^{-1}B\beta(z) + A^{-1}C\delta(y) + A^{-1}P \rangle.
 \end{aligned}
 \tag{12}$$

From (12) we immediately obtain that

$$\begin{aligned}
 \langle \alpha(x) - \alpha(\bar{x}) - {}^t(A^{-1}B)(\tau(x) - \tau(\bar{x})), \beta(z) \rangle + \\
 + \langle \gamma(x) - \gamma(\bar{x}) - {}^t(A^{-1}C)(\tau(x) - \tau(\bar{x})), \delta(y) \rangle + P_1(x) = 0 \quad \forall x, y, z
 \end{aligned}
 \tag{13}$$

where $P_1(x) = P(x) - \langle \tau(x) - \tau(\bar{x}), A^{-1}P \rangle$.

Recall that, by Remark 2.2, $\beta(z)$ and $\delta(y) \neq$ constant in every component (that means that $\beta_i(z) \neq$ constant $\forall i$ and $\delta_i \neq$ constant $\forall i$). We now show ii) again by contradiction: let us suppose that there exists $x_1 \in X$ such that

$$\alpha(x_1) - \alpha(\bar{x}) - {}^t(A^{-1}B)(\tau(x_1) - \tau(\bar{x})) \neq (0, \dots, 0).$$

Calling (R_1, \dots, R_{q_1}) the previous vector and fixing in (13) and $y \in Y$, we have that

$$R_1\beta_1(z) + \dots + R_{q_1}\beta_{q_1}(z) + \Xi = 0 \text{ with } R_i \neq 0 \text{ for some } i \in (1, \dots, q_1).$$

Then, for that i

$$\begin{aligned}
 \beta_i(z) = \\
 = -\frac{1}{R_i}R_1\beta_1(z) - \dots - \frac{1}{R_i}R_{i-1}\beta_{i-1}(z) - \frac{1}{R_i}R_{i+1}\beta_{i+1}(z) - \dots - \frac{1}{R_i}R_{q_1}\beta_{q_1}(z) - \frac{1}{R_i}\Xi
 \end{aligned}
 \tag{14}$$

and we obtain that

$$\begin{aligned}
 f(x, z) = a(x) \exp \left\{ \sum_{i=1}^{q_1} \alpha_i(x) \beta_i(z) \right\} = \\
 = a(x) \exp \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{q_1} \left(\alpha_j(x) - \frac{1}{R_i}R_j\alpha_i(x) \right) \beta_j(z) \right\} \exp \left\{ -\alpha_i(x) \frac{1}{R_i}\Xi \right\}.
 \end{aligned}$$

But this is a contradiction; in fact q_1 was supposed to be the minimum positive integer for which the previous representation would be possible.

So, we obtain that

$$\alpha(x) = {}^t(A^{-1}B)\tau(x) + \Delta \quad \text{with} \quad \Delta = \alpha(\bar{x}) - {}^t(A^{-1}B)\tau(\bar{x}).$$

Analogously,

$$\gamma(x) = {}^t(A^{-1}C)\tau(x) + \Omega \quad \text{with} \quad \Omega = \gamma(\bar{x}) - {}^t(A^{-1}C)\tau(\bar{x})$$

and this concludes the proof.

COROLLARY 2.1. Let (X_n, Y_n) be a discrete time, partially observable stochastic process for which A.1 and A.2 hold. If there exists a finite-dimensional filter system for which B.1 and B.2 are satisfied at every step n , then we have that, calling $\varphi_n : Z \times Y \rightarrow Z$ the filter system, there exists $\sigma_n : Z \rightarrow \mathbf{R}^{q_3}$ of class C^1 such that

$$\sigma_n(\varphi_n(z, y)) = M_n \beta_n(z) + N_n \delta_n(y) + C_n, \quad \forall (z, y) \in Z \times Y$$

with M_n a $q_3 \times q_1$ matrix, N_n a $q_3 \times q_2$ matrix and C_n a vector in \mathbf{R}^{q_3} .

Moreover, the function $\alpha(x)$ and $\gamma(x)$ in (5) depend both linearly on $\tau(x)$.

Proof. Under our assumptions we have from Ferrante (1989) and Ferrante Runggaldier (1989) that, at every step n , relations (5) hold, from which, by the Bayes' rule we obtain

$$\begin{aligned} t(x)r(\varphi(z, y)) \exp\{\langle \tau(x), \sigma(\varphi(z, y)) \rangle\} = \\ \alpha(x) \exp\{\langle \alpha(x), \beta(z) \rangle\} c(x) \exp\{\langle \gamma(x), \delta(y) \rangle\} \end{aligned}$$

$$\text{with } r(\varphi(z, y)) = \frac{s(\varphi(z, y)) \cdot \int p(y; x)p(x; z) dx}{\int b(z)d(y)}.$$

Letting q_1, q_2, q_3 in (5) represent their minimal possible values, Theorem 2.1 then implies

$$\sigma(\varphi(z, y)) = M\beta(z) + N\delta(y) + C, \quad \forall (z, y) \in Z \times Y$$

with M a $q_3 \times q_1$ matrix, N a $q_3 \times q_2$ matrix and C a vector in \mathbf{R}^{q_3} . Moreover, we can prove, in force of the last part of Theorem 2.1, that

$$\alpha(x) = M\tau(x) + \Delta; \quad \gamma(x) = N\tau(x) + \Omega$$

and this concludes the proof.

Remark 2.3. Notice that Corollary 2.1 states that every discrete time, partially observable stochastic process that admits a finite-dimensional filter, also admits a filter system consisting of functions that can be decomposed into the sum of two functions, each depending only on a single variable.

Moreover, if the minimal filter system has dimension k and $k \bmod m = 0$, then the filter system, decomposed as above, is still minimal.

References

1. Bather, J. A., (1965), Invariant Conditional Distributions. Ann. Math. Stat. 36, pp. 829-846.

2. *Barndoff-Nielsen, O.*, (1978), *Information and Exponential Families*. John Wiley, New York.
3. *Ferrante, M.*, (1989), *Condizioni necessarie all'esistenza di filtri a dimensione finita (Necessary conditions for the existence of finite dimensional filters)*. Thesis, University of Padova.
4. *Ferrante, M., Runggaldier, W. J.*, (1990), *On Necessary Conditions for the Existence of Finite-Dimensional Filters in Discrete Time*. *Syst. & Contr. Letters* **14**, pp. 63-69.
5. *Sawitzki, G.*, (1981), *Finite Dimensional Filter System in Discrete Time*. *Stochastics* **5**, pp. 107-114.
6. *Sawitzki, G.*, (1979), *Exact Filtering in Exponential Families: Discrete Time*. In: *M. Kohlmann and W. Vogel, Eds., Stochastic Control Theory and Stochastic Differential Systems (Lect. Notes in Control and Info. Sci. 16)* pp. 554-558.
7. *Spizzichino, F.*, (1988), *Nonlinear dynamic models in discrete time admitting finite dimensional conjugate families (preprint)*.
8. *Van Schuppen, J. H.*, (1979), *Stochastic Filtering Theory: A Discussion on Concepts, Methods and Results*. In: *M. Kohlmann and W. Vogel, Eds., Stochastic Control Theory and Stochastic Differential Systems (Lect. Notes in Control and Info. Sci. 16)* pp. 209-226.

О конечномерной фильтрации в дискретном времени

М. ФЕРРАНТЕ

(Триест)

В статье доказывается, что при подходящих предположениях о регулярности необходимым условием существования конечномерного фильтра в дискретном времени является возможность нахождения удобного разложения системы фильтра $s(z, y)$ в сумму из члена, зависящего только от первой переменной плюс члена, зависящего только от второй. Даются также дополнительные результаты для распределений наблюдения, предсказания и фильтрации.

Marco Ferrante
International School for Advanced Studies (ASAS)
Strada Costiera 11
34014 Trieste, Italy

VIABLE CONTROL TRAJECTORIES IN LINEAR SYSTEMS¹

B. MARTOS

(*Budapest*)

(Received 11 December, 1990)

We consider a linear differential system with decoupled control, the system being asymptotically stable. The state is constrained to a proper subset of the state space, the viability set, with non-empty interior. We determine a subset of the control space, the set of viable controls, such that any control trajectory taking its course within this subset generates viable state trajectory (sufficient viability conditions). Finally, we apply these results to interval shaped viability sets and determine viable control sets of maximum radius.

Keywords. Multivariable control system; viability theory; linear differential equations; matrix algebra; state-space methods.

1. Introduction

Motivation

In controlled systems it is a frequent occurrence, that the state trajectory is restricted to a subset of the state space, the viability set [6], [1]. The known results of viability theory refer to conditions under which the existence of viable trajectories can be decided. In the analysis of linear controlled economic systems [3], [5] the necessity emerged to go beyond existence and determine a subset of control trajectories that generate viable state trajectories. The state constraints defining the viability set were given usually in the form of inequalities (e.g. non-negativity constraints, budget constraints) which implied nonlinearity of the system

¹Abridged version of this paper was originally published in the Proceedings of the IFAC Symposium on Dynamic Modelling and Control of National Economies, 27-29 June 1989, Edinburgh, UK, copyright IFAC 1990, published by Pergamon Press, Oxford. Valuable comments on an earlier draft by V. Kertész, A. Lee and A. Simonovits are gratefully acknowledged.

and prevented the use of standard (e.g. Laplace-transform) techniques. Although the motivation came from economics, it is expected that the present viability results prove to be applicable to physical or biological systems, too.

The problem setting

Consider the time-invariant linear inhomogeneous differential equation system:

$$\dot{x}(t) = Ax(t) + u(t), \quad x(0+) = x_0, \quad (1)$$

where

$$\begin{aligned} t \in \mathbf{R}_+ &:= (0, \infty), && \text{(positive) time} \\ x : \mathbf{R}_+ &\rightarrow \mathbf{R}^n, && \text{state vector} \\ u : \mathbf{R}_+ &\rightarrow \mathbf{R}^n && \text{control vector} \\ A \in \mathbf{R}^{n \times n}, &&& \text{system matrix.} \end{aligned}$$

The time functions $t \rightarrow x(t)$, $t \rightarrow u(t)$ will be called *state trajectory* and *control trajectory* and denoted as $x(\cdot)$ and $u(\cdot)$, respectively. A trajectory pair $u(\cdot)$, $x(\cdot)$ is called a *solution* if they jointly satisfy (1) for all $t \in \mathbf{R}_+$, and then we also say that the control trajectory $u(\cdot)$ *generates* the state trajectory $x(\cdot)$.

Let us denote by

$$\begin{aligned} \tilde{x} &:= \{x(t) : t \in \mathbf{R}_+\} \\ \tilde{u} &:= \{u(t) : t \in \mathbf{R}_+\} \end{aligned} \quad (2)$$

the range of a particular state (control) trajectory, these are subsets of the state (control) space, respectively.

Viability concepts

Let $\mathcal{X} \subset \mathbf{R}^n$ denote a given proper subset of the state space, the *viability set*. A state trajectory is said to be *viable* if $x(t) \in \mathcal{X}$, $\forall t$, or more concisely: $\tilde{x} \subset \mathcal{X}$. A control trajectory $u(\cdot)$ is said to be *viable* if it generates a viable state trajectory. Then we also say that the solution is *viable*.

We give sufficient but not necessary conditions for a solution to be viable. More precisely, we determine $\mathcal{U} \subset \mathbf{R}^n$, a subset of the control space, called *viable control set*, such that any control trajectory whose range is contained in it, is a *viable control trajectory*, i.e. it generates viable state trajectory. We are not able to determine all viable solutions, but we try to define a possibly large viable control set.

2. Notations and assumptions

Vector and matrix norms ([4] Chaps 6 and 7)

For any (real or complex) n -vector $g \in \mathbb{C}^n$ and square matrix $G \in \mathbb{C}^{n \times n}$ let us denote by g^+ and G^+ , respectively, the non-negative vector and matrix consisting of the absolute values (moduli) of the components or entries of g and G

$$\begin{aligned} g^+ &:= [|g_j|], & \forall g = [g_j] \in \mathbb{C}^n \\ G^+ &:= [|G_{i,j}|], & \forall G = [G_{i,j}] \in \mathbb{C}^{n \times n}. \end{aligned} \quad (3)$$

Let us denote by $\|g\|$ and $\|G\|$ a vector norm and a matrix norm, respectively. Throughout the paper the following properties will be assumed of these norms:

a) The vector norm is an *absolute* norm, i.e.:

$$\|g\| = \|g^+\|, \quad \forall g \in \mathbb{C}^n. \quad (4)$$

b) The matrix norm is *induced* by the vector norm:

$$\|G\| := \max \{ \|Gg\| : \|g\| = 1 \}. \quad (5)$$

(5) implies ([4], Th. 6.3.1.)

$$\|Gg\| \leq \|G\| \cdot \|g\|, \quad \forall g \in \mathbb{C}^n, \quad \forall G \in \mathbb{C}^{n \times n}. \quad (6)$$

For any *diagonal matrix* $H := \text{diag}\{\chi_1, \chi_2, \dots, \chi_n\}$ (4) and (5) imply ([4], Th. 6.4.1.):

$$\|H\| = \max_i |\chi_i|. \quad (7)$$

The interval norm

A special norm which will be applied in Chap. 4 of the paper is the *interval norm* or q -norm. Let $q \in \mathbb{R}_+^n$ be a positive vector and define for any $g \in \mathbb{C}^n$, $G \in \mathbb{C}^{n \times n}$:

$$\|g\|_q := \max_i \frac{|g_i|}{q_i}, \quad (8)$$

$$\|G\|_q := \max_i \sum_j \frac{q_j}{q_i} |G_{ij}|. \quad (9)$$

It is easy to see that $\|g\|_q$ satisfies (4) and $\|G\|_q$ satisfies (5). (Moreover, $\|G\|_q$ is an absolute matrix norm since $\|G\| = \|G^+\|$.) The interval norm is nothing else than a scale-transformed l_∞ norm.

The spectrum of the system matrix and the stability of the system

Let us denote by $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ the *eigenvalues* of A , the set of the eigenvalues is called the *spectrum* of A and any diagonal matrix $\Lambda \in \mathbb{C}^{n \times n}$ formed from the eigenvalues (in different successions) is called a *spectral matrix* of A : $\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_n\}$. In connection with the spectrum of the system matrix the following two parameters shall be needed in the sequel:

$$\rho := \max_j |\lambda_j| \quad (10)$$

the *spectral radius* of A , and

$$\mu := \min_j (-\text{Re } \lambda_j) \quad (11)$$

the *degree of stability* of A ($\text{Re } x = \text{real part of } x$).

ASSUMPTION "A": The system (1) is *asymptotically stable* i.e.

$$\mu > 0. \quad (12)$$

This stability assumption seems to be essential for the subsequent analysis, since it guarantees that the state range is bounded whenever the control range is such. ([7], Chap. 7.)

The set of modal matrices and the simplicity of A

Let us denote the set of diagonal matrices by \mathcal{H} . Let us consider now the (possibly empty) set of such non-singular matrices (the *modal matrices* of A) that applied as similarity transformation on A result in a diagonal matrix

$$\mathcal{L} := \{L \in \mathbb{C}^{n \times n} : L^{-1}AL \in \mathcal{H}\}. \quad (13)$$

ASSUMPTION "B": The system matrix A is *simple* (i.e. diagonalizable by a similarity transformation)

$$\mathcal{L} \neq \emptyset. \quad (14)$$

This simplicity assumption is rather technical. I guess that it can be dispensed with on the account of more complicated formulae.

Under Assumption "B" any modal matrix $L \in \mathcal{L}$ produces a *spectral decomposition* of A :

$$A = L\Lambda L^{-1}, \quad (15)$$

where Λ is a spectral matrix of A .

The modal condition number

Besides the two parameters ρ and μ characterizing A a third, perhaps less known parameter, the *modal condition number* κ , plays important role in the sequel ([4], pp. 222, 232).

$$\kappa := \inf\{\|L\| \cdot \|L^{-1}\| : L \in \mathcal{L}\} \quad (16)$$

Since $\|L\| \cdot \|L^{-1}\| \geq \|LL^{-1}\| = \|E\| = 1$, we have

$$\kappa \geq 1 \quad (17)$$

under Assumption "B".

The numerical value of κ depends also on the choice of the matrix norm. While the parameters ρ and μ are easy to calculate, it is hard to solve the minimum problem (16) in this generality. But for the interval norm (9) and for system matrices of distinct eigenvalues we have the following simple formula:

$$\kappa = \|L^+ K^+\|_q, \quad (18)$$

for any $L \in \mathcal{L}$, and $K = L^{-1}$. (This formula is a direct consequence of [2, p. 80, Theorem II. a.])

Assumptions on the viability set

ASSUMPTION "C": The interior \mathcal{X}^0 of the viability set is non-empty: $\mathcal{X}^0 \neq \emptyset$.

ASSUMPTION "D": The initial point x_0 belongs to the interior of \mathcal{X} : $x_0 \in \mathcal{X}^0$.

It is to be noted that no closedness assumption is made on \mathcal{X} at this point, in contrast to the common practice in viability theory.

Clearance, nest, nestpoints

For any $d \in \mathbf{R}^n$, $\delta > 0$ let us denote by $\mathcal{B}(d, \delta)$ the *open ball* and by $\overline{\mathcal{B}}(d, \delta)$ the *closed ball* with *center* d and *radius* δ . (The shape of the ball depends on the chosen vector norm.)

$$\mathcal{B}(d, \delta) := \{y \in \mathbf{R}^n : \|y - d\| < \delta\} \quad (19)$$

$$\bar{\mathcal{B}}(d, \delta) := \{y \in \mathbf{R}^n : \|y - d\| \leq \delta\}. \quad (20)$$

If $d \in \mathcal{X}^0$ we can take the supremum of such radii δ that $\mathcal{B}(d, \delta)$ is contained in \mathcal{X} , denote it by $\theta(d)$ and call it the *clearance* of d :

$$\theta(d) := \sup \{\delta : \mathcal{B}(d, \delta) \subset \mathcal{X}\}. \quad (21)$$

If we additionally assume that the viability set is *closed*, then we also use the following equivalent definition:

$$\theta(d) = \max \{\delta : \bar{\mathcal{B}}(d, \delta) \subset \mathcal{X}\}. \quad (22)$$

The clearance is thus the distance of d from the closest boundary point of \mathcal{X} . The larger is $\theta(d)$ the “more interior” is d to \mathcal{X} .

The following subset \mathcal{D} of \mathcal{X}^0 will be called the *nest*

$$\mathcal{D} := \{d \in \mathcal{X}^0 : \kappa \|x_0 - d\| \leq \theta(d)\}, \quad (23)$$

i.e. d is a *nestpoint* ($d \in \mathcal{D}$) if its κ -fold distance from the initial point x_0 does not exceed its clearance. The nest is non-empty since $x_0 \in \mathcal{D}$.

3. Sufficient viability conditions

Two lemmas

Before wording the basic results of the present paper we establish two lemmas.

LEMMA 1. For any $d \in \mathbf{R}^n$

$$\|x(t) - d\| < \theta(d), \forall t \implies \tilde{x} \subset \mathcal{B}(d, \theta(d)) \implies \tilde{x} \subset \mathcal{X}, \quad (24)$$

and if \mathcal{X} is closed, then

$$\|x(t) - d\| \leq \theta(d), \forall t \implies \tilde{x} \subset \tilde{\mathcal{B}}(d, \theta(d)) \implies \tilde{x} \subset \mathcal{X}. \quad (25)$$

Proof. The lemma follows immediately from definitions (2), (19), (20), (21) and (22). ■

This lemma bears the promise, that if we can construct control trajectories such that the generated state trajectory satisfies the premissa for some d , then a sufficient viability condition is established. As we will see soon the nest is just the set of such points d .

LEMMA 2. Under Assumptions “A” and “B” the following estimations hold true for all $t \in \mathbf{R}_+$:

$$\|\exp(At)\| \leq \kappa \exp(-\mu t) \tag{26}$$

$$\int_0^t \|\exp [A(t - \tau)]\| d\tau \leq \frac{\kappa}{\mu} [1 - \exp(-\mu t)] \tag{27}$$

$$\int_0^t \|A \exp [A(t - \tau)]\| d\tau \leq \frac{\kappa \rho}{\mu} [1 - \exp(-\mu t)]. \tag{28}$$

Proof. It is well known that if $A = L\Lambda L^{-1}$ (15) then $\exp(At) = L \exp(\Lambda t)L^{-1}$, hence

$$\|\exp(At)\| = \|L \exp(\Lambda t)L^{-1}\| \leq \|L\| \cdot \|\exp(\Lambda t)\| \cdot \|L^{-1}\|. \tag{29}$$

Since this inequality holds for all $L \in \mathcal{L}$, we get from definition (16) that

$$\|\exp(At)\| \leq \kappa \|\exp(\Lambda t)\|. \tag{30}$$

Applying (7) and (11)

$$\|\exp(\Lambda t)\| = \max_j |\exp(\lambda_j t)| = \max_j \exp(\operatorname{Re} \lambda_j t) = \exp(-\mu t) \tag{31}$$

results. Eq. (26) is the consequence of (30) and (31).

By a similar reasoning we establish that

$$\|A \exp(At)\| \leq \kappa \rho \exp(-\mu t). \tag{32}$$

Namely,

$$\begin{aligned} \|A \exp(At)\| &= \|(L\Lambda L^{-1})[L \exp(At)L^{-1}]\| = \|L\Lambda \exp(\Lambda t)L^{-1}\| \leq \\ &\leq \kappa \|\Lambda\| \cdot \|\exp \Lambda t\| = \kappa \rho \exp(-\mu t), \end{aligned} \tag{33}$$

where $\|\Lambda\| = \rho$ is the consequence of (7) and (10).

Now, we simply substitute $(t - \tau)$ for t into (26) and (32) and integrate on both sides of both inequalities from $\tau = 0$ to $\tau = t$ whereupon (27) and (28) result, respectively. ■

THEOREM 1: VIABLE CONTROL TRAJECTORIES. Let assumptions "A" to "D" hold and let $d \in \mathcal{D}$ be any nestpoint (23). Then $u(\cdot)$ is a viable control trajectory (i.e. it generates a viable state trajectory) if its range \tilde{u} satisfies $\tilde{u} \subset \mathcal{U}_1(d)$ or $\tilde{u} \subset \mathcal{U}_2(d)$, where

$$\mathcal{U}_1(d) := \mathcal{B}\left(-Ad, \frac{\mu}{\kappa}\theta(d)\right) \quad (34)$$

$$\mathcal{U}_2(d) := \left\{ u : -A^{-1}u \in \mathcal{B}\left(d, \frac{\mu}{\kappa\rho}\theta(d)\right) \right\}. \quad (35)$$

If the viability set \mathcal{X} is closed, then $\mathcal{B}(\cdot, \cdot)$ can be replaced by $\overline{\mathcal{B}}(\cdot, \cdot)$ in (34) and (35).

Interpretation of Theorem 1. The set $\mathcal{U}_1(d)$ is constructed in the following way. We select a nestpoint d in the state space and via premultiplication by $(-A)$ transfer it to the control space. Furthermore, we determine the clearance $\theta(d)$ of d and multiply it by μ/κ . The former will be the center, the latter the radius of a ball in the control space, this ball is $\mathcal{U}_1(d)$. For forming the set $\mathcal{U}_2(d)$ we also take a nestpoint, and form a ball around it in the state space, whose radius is its clearance reduced by the factor $\mu/\kappa\rho$. This ball is then mapped into the control space, the matrix of the mapping is $(-A)$. In this way we obtain $\mathcal{U}_2(d)$ which need not be a ball.

Proof of Theorem 1. Let us write system (1) in the following equivalent form:

$$\frac{d}{dt}[x(t) - d] = A[x(t) - d] + [u(t) + Ad], \quad x(0+) = x_0, \quad (36)$$

then the explicit solution of this differential equation can be written as

$$x(t) - d = \exp(At)(x_0 - d) + \int_0^t \exp[A(t - \tau)][u(\tau) + Ad] d\tau. \quad (37)$$

With the help of (6), the triangle inequality, and the integral inequality

$$\left\| \int_0^t f(\tau) d\tau \right\| \leq \int_0^t \|f(\tau)\| d\tau, \quad (38)$$

(where $t \in \mathbf{R}_+$ and $f : \mathbf{R}_+ \rightarrow \mathbf{R}^n$) from (37) we get:

$$\|x(t) - d\| \leq \|\exp(At)\| \cdot \|x_0 - d\| + \int_0^t \|\exp[A(t - \tau)]\| \cdot \|u(\tau) + Ad\| d\tau. \quad (39)$$

From assumption $d \in \mathcal{D}$ (23) yields:

$$\|x_0 - d\| \leq \frac{1}{\kappa} \theta(d), \tag{40}$$

while conditions $\tilde{u} \subset \mathcal{U}_1(d)$, $\tilde{u} \subset \mathcal{U}_2(d)$ can be expressed as

$$\|u(\tau) + Ad\| < \frac{\mu}{\kappa} \theta(d), \quad \forall \tau \tag{41}$$

and

$$\|A^{-1}u(\tau) + d\| < \frac{\mu}{\kappa\rho} \theta(d), \quad \forall \tau, \tag{42}$$

respectively.

Considering first the case $\tilde{u} \subset \mathcal{U}_1(d)$, we apply (40), (41), (26) and (27) to inequality (39):

$$\|x(t) - d\| < \frac{1}{\kappa} \theta(d) \cdot \kappa \exp(-\mu t) + \frac{\mu}{\kappa} \theta(d) \cdot \frac{\kappa}{\mu} [1 - \exp(-\mu t)] = \theta(d). \tag{43}$$

(43) and (24) gives $\tilde{x} \subset \mathcal{X}$, i.e. viability. If \mathcal{X} is closed and we relax \mathcal{B} to $\tilde{\mathcal{B}}$ in (34) then inequalities (41) and (43) become non-strict and the application of (25) instead of (24) leads to the same result.

The other case: $\tilde{u} \subset \mathcal{U}_2(d)$ proves the same way except that the last term of (37) has to be written in the form

$$\int_0^t A \exp [A(t - \tau)] \cdot [A^{-1}u(\tau) + d] d\tau$$

and the reference to (27) and (41) must be changed to (28) and (42), respectively, in the proof. ■

The problem of the best nestpoint

As seen from Theorem 1, the larger is $\theta(d)$ the larger will be both $\mathcal{U}_1(d)$ and $\mathcal{U}_2(d)$. Thus, in order to enlarge the available set of viable controls we may want to find the best nestpoint, i.e. the one with the largest clearance. Hence, the following optimization problem is to be solved:

$$\begin{aligned} &\text{“Find } d^* \in \mathcal{D} \text{ and } \theta^* := \theta(d^*) \text{ such that} \\ &\theta^* = \max\{\theta(d) : d \in \mathcal{D}\} \text{”}. \end{aligned} \tag{44}$$

This problem, of course, can not be solved without the specification of the viability set.

Having finished the discussion of Theorem 1 we have to emphasize that all the parameters which occur in it are calculable relatively simply. One exception is the clearance $\theta(d)$ whose value depends on the unspecified shape of the viability set. Furthermore, we still have a freedom in the choice of the norm. By an appropriate choice which takes the shape of \mathcal{X} into account we can highly improve upon the "roughness" of our estimations, which is caused just by the estimation in norm. In the sequel we deal with a specification of the viability set, where not only $\theta(d)$ will be easily calculable, but a best nestpoint, an explicit solution of problem (44) will also be found.

4. Interval shaped viability sets

Interval

We discuss the case when the viability set is a closed interval (parallelepiped).

Let b be any point in $\mathbf{R}^n : b \in \mathbf{R}^n$, and p a non-negative vector in $\mathbf{R}^n : p \geq 0$. We define the closed *interval* (with center b and half-diagonal p) as

$$\mathcal{I}(b, p) := \{y \in \mathbf{R}^n : b - p \leq y \leq b + p\}. \quad (45)$$

The viability set

ASSUMPTION "E": The viability set is an interval:

$$\mathcal{X} = \mathcal{I}(c, q), \quad (46)$$

where $q > 0$.

$q > 0$ implies that $\mathcal{I}(c, q)$ has a non-empty interior, hence Assumption "E" implies Assumption "C".

In the sequel we will use the interval vector-norm $\|\cdot\|_q$ as defined in (8) where q will be the same positive n -vector that appears in the definition (46) of \mathcal{X} . The following lemma clarifies the connection between intervals and the closed balls in interval norm.

LEMMA 3. In the interval norm $\|\cdot\|_q$ the closed balls of center y and radius δ are intervals with center y and half-diagonal δq :

$$\overline{B}(y, \delta) = \mathcal{I}(y, \delta q). \quad (47)$$

Proof. By the virtue of definitions (20), (8) and (45) we have:

$$\begin{aligned} z \in \bar{B}(y, \delta) &\iff \|z - y\|_q \leq \delta \iff \\ \max_i \frac{|z_i - y_i|}{q_i} \leq \delta &\iff |z_i - y_i| \leq \delta q_i, \forall i \iff \\ &\iff y - \delta q \leq z \leq y + \delta q \iff z \in \mathcal{I}(y, \delta q). \end{aligned} \tag{48}$$

■

Substituting $y = c$, $\delta = 1$ into (47) from (46) we get

$$\mathcal{X} = \bar{B}(c, 1) \tag{49}$$

i.e. the viability set is represented as a *unit ball* with center c in the q -norm.

The calculation of the clearance

LEMMA 4. The clearance of a point $y \in \mathcal{I}(c, q)$ is

$$\theta(y) = 1 - \|c - y\|_q. \tag{50}$$

Proof. By the virtue of (22), (47), (45) and (8) we have

$$\begin{aligned} \bar{B}(y, \delta) \subset \mathcal{X} &\iff \mathcal{I}(y, \delta q) \subset \mathcal{I}(c, q) \iff \\ &\iff \left\{ \begin{array}{l} y - \delta q \geq c - q \\ y + \delta q \leq c + q \end{array} \right\} \iff (1 - \delta)q \geq \left\{ \begin{array}{l} c - y \\ y - c \end{array} \right\} \iff \\ &\iff (1 - \delta)q_i \geq |c_i - y_i|, \forall i \iff 1 - \delta \geq \|c - y\|_q \iff \delta \leq 1 - \|c - y\|_q. \end{aligned}$$

Hence, we get (50) for $\theta(y) = \max \delta$.

■

The calculation of the maximum clearance of nestpoints

For sake of brevity let us introduce the following notation for the initial clearance

$$\theta_0 := \theta(x_0) = 1 - \|c - x_0\|_q, \tag{51}$$

which can be calculated from the given data of \mathcal{X} and x_0 .

We now turn to the solution of the problem of the best nestpoint (44) for the case of interval shaped viability set.

LEMMA 5. For the viability set $\mathcal{X} = \mathcal{I}(c, q)$ the maximum clearance of a nestpoint is

$$\theta^* = \min \left\{ 1, \frac{\kappa \theta_0}{\kappa - 1} \right\} \quad (52)$$

which is taken at a best nestpoint

$$d^* = x_0 + \frac{\theta^*}{\kappa - (\kappa - 1)\theta^*} (c - x_0). \quad (53)$$

Interpretation of Lemma 5. Two cases are to be distinguished.

a) If

$$\theta_0 \geq 1 - \frac{1}{\kappa}, \quad (54)$$

then $\theta^* = 1$ from (52) and $d^* = c$ from (53). Thus, in this case the center of the viability set is the (unique) best nestpoint. This case occurs if $\|c - x_0\| \leq 1/\kappa$, i.e. if the initial point is close enough to the center of \mathcal{X} .

b) If, on the contrary, $\theta_0 < 1 - 1/\kappa$, then

$$\theta^* = \frac{\kappa \theta_0}{\kappa - 1} \quad \text{and} \quad d^* = x_0 + \frac{\theta_0}{(\kappa - 1)(1 - \theta_0)} (c - x_0). \quad (55)$$

In this case an optimal nestpoint occurs in the interior of the line segment connecting the initial point x_0 with the center c . d^* of (55) will usually not be an unique best nestpoint since the interval norm is not strictly monotonic in the modulus of the components.

Proof of Lemma 5. Let us first assume $c \in \mathcal{D}$, i.e. (23)

$$\kappa \|c - x_0\|_q \leq \theta(c) = 1. \quad (56)$$

In this case we have for all $d \in \mathcal{D}$ from (50):

$$\theta(d) = 1 - \|c - d\|_q \leq 1 = \theta(c). \quad (57)$$

Hence, $d^* = c$ is an optimal nestpoint and $\theta^* = 1$ its clearance. Uniqueness follows from the fact that equality in (57) occurs only at $d = c$. On the other hand, (56) can also be written in the alternative forms $\|c - x_0\|_q \leq 1/\kappa$ or $1 \leq \kappa \theta_0 / (\kappa - 1)$, and hence, (52)–(53) yield the same result.

Consider now the other case $c \notin \mathcal{D}$, which implies that $1 > \kappa \theta_0 / (\kappa - 1)$ holds. The following triangle inequality

$$\|c - d\|_q + \|d - x_0\|_q \geq \|c - x_0\|_q \quad (58)$$

can be written in the following form in view of notations (50) and (51)

$$\|d - x_0\|_q \geq \theta(d) - \theta_0. \tag{59}$$

On the other hand, $d \in \mathcal{D}$ implies (23)

$$\|d - x_0\|_q \leq \frac{1}{\kappa} \theta(d) \tag{60}$$

which with (59) gives $\theta(d) - \theta_0 \leq \theta(d)/\kappa$ or rearranged:

$$\theta(d) \leq \frac{\kappa}{\kappa - 1} \theta_0 \tag{61}$$

for all $d \in \mathcal{D}$. (N.B. $\kappa > 1$, since for $\kappa = 1 : c \in \mathcal{D}$ holds.) Substituting from (55) $y = d^*$ into (50) we see that the r.h.s. upper bound of $\theta(d)$ is assumed at this point d^* . Furthermore, $d = d^*$ satisfies (60), hence d^* of (55) is an optimal nestpoint with clearance $\theta^* = \kappa\theta_0/(\kappa - 1)$ as the lemma tells us. ■

Combining Lemma 3 and Lemma 5 with Theorem 1 we get the following viability conditions for $\mathcal{X} = \mathcal{I}(c, q)$.

THEOREM 2: VIABLE CONTROL TRAJECTORIES FOR INTERVAL SHAPED VIABILITY SETS. Under assumptions "A", "B", "D" and "E" $u(\cdot)$ is a viable control trajectory if its range satisfies $\tilde{u} \subset \mathcal{U}_1^*$ or $\tilde{u} \subset \mathcal{U}_2^*$, where

$$\mathcal{U}_1^* = \begin{cases} \mathcal{I}\left(-Ac, \frac{\mu}{\kappa}q\right), & \text{if } \theta_0 \geq 1 - 1/\kappa \\ \mathcal{I}\left(-Ad^*, \frac{\mu\theta_0}{\kappa - 1}q\right), & \text{otherwise,} \end{cases} \tag{62}$$

$$\mathcal{U}_2^* = \begin{cases} \left\{ u : -A^{-1}u \in \mathcal{I}\left(c, \frac{\mu}{\kappa\rho}q\right) \right\}, & \text{if } \theta_0 \geq 1 - 1/\kappa \\ \left\{ u : -A^{-1}u \in \mathcal{I}\left(d^*, \frac{\mu\theta_0}{\rho(\kappa - 1)}q\right) \right\}, & \text{otherwise,} \end{cases} \tag{63}$$

and

$$\theta_0 = 1 - \|c - x_0\|_q, \tag{64}$$

$$d^* = x_0 + \frac{\theta_0}{(\kappa - 1)(1 - \theta_0)}(c - x_0). \tag{65}$$

Concluding remarks

Extensions to coupled controls, control space constraints, polyhedral or compact-convex viability sets, and, finally, to viable feedback rules will be dealt with in subsequent papers.

References

1. Aubin, J. P. and Cellina, A., *Differential Inclusions*. Berlin, Springer, 1983.
2. Bauer, F. L., *Optimally scaled matrices*, *Numerische Mathematik*. vol. 5 (1963), pp. 73–87.
3. Kornai, J. and Martos, B. (Eds.), *Non-Price Control*. Amsterdam, North-Holland, 1981.
4. Lancaster, P., *Theory of Matrices*. New York, Academic Press, 1969.
5. Martos, B., *Economic Control Structures*. Amsterdam, North-Holland, 1990.
6. Nagumo, M., *Über die Lage der Integralkurven gewöhnlicher Differentialgleichungen*. *Proc. Phys. Math. Soc. Japan*, vol. 24 (1942), pp. 551–559.
7. Zadeh, L. A. and Desoer, C. A., *Linear System Theory*. New York, McGraw-Hill, 1963.

Траектории управления выживаемости в линейных системах

В. МАРТОШ

(Будапешт)

Рассматривается система линейных дифференциальных уравнений с несвязанным управлением, являющаяся асимптотически устойчивой. Состояния ограничиваются собственным подмножеством пространства состояний (множеством выживаемости) с непустой внутренностью. Подмножество пространства управлений (множество управлений выживаемости) определяется таким образом, что любая траектория управления, берущая свое начало из этого множества, генерирует фазовую траекторию выживаемости (достаточные условия выживаемости). В заключении эти результаты были применены для множеств выживаемости, образованных интервалами, и были определены множества управления максимального радиуса.

Béla Martos
Institute of Economics
Hungarian Academy of Sciences
P.O.B. 262. H-1502, Budapest, Hungary

SUBOPTIMAL CONTROL ALGORITHM FOR DISCRETE SYSTEMS

J. KORBICZ, V. PODLADCHIKOV AND P. BIDYUK

(Zielona Góra, Kiev)

(Received 11 December, 1990)

An approach for the realization of a suboptimal control for linear discrete-time systems is proposed, when a control interval is divided into two sub-intervals. For the first sub-interval the vector of feedback coefficients with regard to state is assumed as constant, and then for the other one the exact optimal control law is used. The moment of switching the control modes is determined on the basis of acceptable deviations of the control system quality index from the optimal control. Finally, a simple example is used to demonstrate some features of the present approach for controller design.

1. Introduction

The problem of controller design for linear systems has been considered by a large numbers of researchers in the past thirty years. The analytical controller design problem was considered primarily by Letov (1960) for continuous time-invariant systems and solved simultaneously for nonstationary systems by Kalman (1960).

Today there exist a large number of analytical controller design procedures. For example, Repin-Tretyakov's method (Repin and Tretyakov, 1963), Newton-Raphson's method (Wonham, 1979), the method of diagonalization (Kwakernaak and Sivan, 1972) and some other authors (Yosida and Lopara, 1989) found a very wide range of applications in different fields.

The Repin-Tretyakov's method supposes the matrix Riccati equation solution, then required matrix $P > 0$ can be found as $P_\infty = \lim_{t \rightarrow \infty} P(t)$. The Newton-Raphson's method or quasi-linearization method represents an iterational procedure for matrix $P > 0$ computation. At each step of this procedure the Lyapunov's equation is solved and then P can be found as $P_\infty = \lim_{\lambda \rightarrow \infty} P^\lambda$, $\lambda = 1, 2, \dots$. Krasovsky (1967) introduced the following generalized work functional

$$J = \int_0^{\infty} \left[\sum_{i,j=1}^n q_{ij} x_i x_j + \sum_{k=1}^m u_k^2 + \frac{1}{4} \sum_{k=1}^m \left(\sum_{i=1}^n \frac{\partial V}{\partial x_i} b_{ki} \right)^2 \right] dt,$$

where $V = \sum_{i,j=1}^n p_{ij} x_i x_j$ is the chosen Lyapunov function, x is the state vector, b_{ki} are the elements of control matrix B of the state space model, u_k denote controls, q_{ij} are elements of non-negative definite matrix Q , which satisfies the observability condition of the system, and P_{ij} are elements of P . The controls are calculated from expression

$$u_k = - \sum_{i=1}^n \frac{\partial V}{\partial x_i} b_{ki}, \quad k = 1, \dots, m.$$

This approach is known as Lyapunov's functions based controller synthesis or analytical controller design using generalized work functional.

Several iterational controller design procedures were proposed by Aleksandrov (1986). They are shown to be applicable to continuous and discrete-time SISO and MIMO systems. A number of computationally efficient controller design procedures for discrete-time systems are presented by Iserman (1981). Most of them are ready for real-time applications using microprocessor hardware.

Our goal is to develop a computationally efficient suboptimal control algorithm, preserving the required control quality. When developing controllers for constant parameter systems, in many cases a variation of the feedback coefficients only over some insignificant part of the control interval in vicinity of the final moment of time is taken into consideration. Such a choice of the coefficients may be regarded as expedient from the point of view of practical applications, because for many control systems the feedback coefficient can be considered practically constant over the entire interval and the realization of the variable coefficient is rather complicated.

It is known that the optimal feedback coefficient is constant over the entire interval if the weighing matrix of final state coincides with a steady-state solution of Riccati equation (Andreyev, 1976). However, for large deviations of the above values, the realization of the constant coefficient may lead to a considerable change of the system state vector in comparison with the optimal law. Therefore, it is necessary to study the degree of these variations, when changing the optimal control law into a stationary one, since allowances for the control system quality characteristics determining the possibility of achieving the control objective can be found in advance.

We present an analytical solution of Riccati equations as an explicit function of system parameters and find the controller transition matrix. Then these solutions are used to obtain an expression for the relative variation of the quality index due to suboptimal control over the part of the whole control interval. The expression for

the relative variation of the quality index can be used for determining the minimum length of the controls sequence based on varying optimal feedback gains.

2. Problem formulation

Consider a linear discrete-time system with constant parameters

$$X_{k+1} = \phi X_k + BU_k, \quad k = 1, 2, \dots, K_1 - 1, \quad (1)$$

where X_k is the state vector, U_k is the controls vector, ϕ and B are the transient matrices of discrete states and control, respectively.

Problem solution for constructing of an optimal linear controller for the quality index of the following form

$$J = X_{K_1}^T S X_{K_1} + \sum_{i=0}^{K_1-1} \{X_{i+1}^T Q X_{i+1} + U_i^T R U_i\}, \quad (2)$$

is determined by the relationship (Kwakernaak and Sivan, 1972)

$$U_k = -L_k X_k, \quad (3)$$

where $L_k = [R + B^T(Q + P_{k+1})B]^{-1} B^T(Q + P_{k+1})\phi$, and P_k is a solution of the discrete Riccati equation

$$P_k = \phi^T(Q + P_{k+1})(\phi - BL_k)$$

with the final condition $P_{K_1} = S$.

We consider here the stationary controller problem solution for system (1), defined over a semi-infinite interval. In this case the control that minimizes the criterion

$$J_1 = \sum_{i=0}^{\infty} \{X_{i+1} Q X_{i+1} + \bar{U}_i^T R \bar{U}_i\}, \quad (4)$$

has the form

$$\bar{U}_k = -L_{\infty} X_k, \quad (5)$$

where $L_{\infty} = [R + B^T(Q + P_{\infty})B]^{-1} B^T(Q + P_{\infty})\phi$ and P_{∞} is the solution of the Riccati algebraic equation $P_{\infty} = \phi^T(Q + P_{\infty})(\phi - PL_{\infty})$.

The problem has been set here to study an effect of the substitution of the optimal control law (3) for system (1) by a suboptimal one on the control system properties. Thus, we can formulate our problem as follows:

- a) to study analytically the variations of values of the quality index (2) and the state vector X at the final moment of the control interval K_1 due to the

- non-optimality of control signals over the entire control interval, i.e. when the control is performed with constant gain L_∞ ;
- b) to accomplish a suboptimal controller synthesis accounting for an application of the stationary control law (5) with subsequent switching to the optimal law (3) in the vicinity of the final moment of the control interval so that to minimize the state vector deviation from the optimal value;
- c) as a result of the suboptimal controller application to reduce the computer time necessary for the control actions computation in comparison to the optimal controller.

3. Characteristics of the suboptimal controller

Let us define state variation X_{K_1} , if the control \bar{U}_k is determined according to the suboptimal law (5) over the entire interval. As shown in the Appendix the transient matrices of the optimal closed control system $\psi_{K_1,0}$ and the stationary one Ω^{K_1} are connected by the relationship

$$\psi_{K_1,0} = (S - P_\infty)^{-1} \Omega^{T(-K_1)} (P_0 - P_\infty), \quad (6)$$

where $\Omega = \phi - BR^{-1}B^T\phi^{-1}(P_\infty - Q)$.

Multiplying both parts of formula (6) on the left-hand side by the matrix

$$\Omega^{K_1} (P_0 - P_\infty)^{-1} \Omega^{T(K_1)} (S - P_\infty)$$

we obtain

$$\Omega^{K_1} (P_0 - P_\infty)^{-1} \Omega^{T(K_1)} (S - P_\infty) \psi_{K_1,0} = \Omega^{K_1}. \quad (7)$$

It is shown in the Appendix that

$$(P_k - P_\infty)^{-1} = \Omega^{k-K_1} (S - P_\infty) \Omega^{T(k-K_1)} + \Omega^k \Xi(k, K_1) \Omega^{T(k)} \quad (8)$$

where $\Xi(k, K_1) = \sum_{i=k+1}^{K_1} \Omega^{-i} B (R + B^T P_\infty B)^{-1} B^T \Omega^{T(-i)}$, $k = 0, 1, \dots, K_1$.

Substituting expression (8), when $k = 0$, into equation (7) we obtain the following relationship connecting the final state vectors X_{K_1} for the optimal control law and \bar{X}_{K_1} for the stationary one

$$[I + \Omega^{K_1} \Xi(0, K_1) \Omega^{T(K_1)} (S - P_\infty)] X_{K_1} = \bar{X}_{K_1}.$$

Using formulas (6) and (8) after some transformations we obtain the expression defining the final state deviation due to the unoptimality of the controls \bar{U}_i ($i = 0, 1, \dots, K_1 - 1$)

$$\bar{X}_{K_1} - X_k = \Omega^{K_1} [\Omega^{-K_1} (S - P_\infty)^{-1} \Omega^{T(-K_1)} \Xi^{-1}(0, K_1) + I]^{-1} X_0. \quad (9)$$

Let us determine the value of the quality index under unoptimal control

$$\bar{J} = \bar{X}_{K_1}^T S \bar{X}_{K_1} + \sum_{i=0}^{K_1-1} \{ \bar{X}_{i+1}^T Q \bar{X}_{i+1} + \bar{U}_i^T R \bar{U}_i \}. \quad (10)$$

Since $\sum_{i=0}^{K_1-1} \{ \cdot \} = \sum_{i=0}^{\infty} \{ \cdot \} - \sum_{i=K_1}^{\infty} \{ \cdot \}$, then

$$\bar{J} = \bar{X}_0^T S \bar{X}_{K_1} + X_0^T P_{\infty} X_0 + \bar{X}_{K_1}^T P_{\infty} \bar{X}_{K_1}.$$

Taking into consideration that $\bar{X}_{K_1} = \Omega^{K_1} X_0$, the quality index can be obtained as follows

$$\bar{J} = \bar{X}_{K_1}^T [\Omega^{T(K_1)} (S - P_{\infty}) \Omega^{K_1} + P_{\infty}] X_0.$$

The deviation of the actual value of the quality index from its minimal value can be found in the following way

$$\bar{J} - J = X_0^T \bar{M} X_0,$$

where $\bar{M} = \Omega^{T(K_1)} (S - P_{\infty}) \Omega^{K_1} - (P_0 - P_{\infty})$.

Changing in the latter expression $(P_0 - P_{\infty})$ in accordance with formula (8) and applying the lemma on the matrix inversion, we obtain

$$\bar{M} = \Omega^{T(K_1)} (S - P_{\infty}) \Omega^{K_1} [\Omega^{-K_1} (S - P_{\infty})^{-1} \Omega^{T(-K_1)} \Xi(0, K_1) + I]^{-1}. \quad (11)$$

The relative variation of the quality index will be $\delta J = (\bar{J} - J)/J$.

4. Suboptimal controller

If the feedback coefficient is chosen to be constant over the whole interval, then as it follows from the relationships (9)–(11), the actual quality characteristics may essentially differ from the optimal ones. Let us analyse characteristics of the controller, the feedback coefficient of which is chosen to be constant up to some moment of time k , when the quality index is less sensitive to the non-optimality of the control law and then switching to the optimal control is performed. Let the stationary control law be realized until the moment $k - 1$ ($0 < k < K_1$) is reached. Starting with the moment k the sequence of controls U_i ($i = k, k + 1, \dots, K_1 - 1$) is optimal. Determine the final state deviation $\bar{X}_{K_1} - X_{K_1}$ due to the non-optimality of the controls \bar{U}_i ($i = 0, 1, \dots, k - 1$).

Using formula (9), we obtain

$$\bar{X}_{K_1} - X_{K_1} = \psi_{k, K_1} \Omega^k [\Omega^{-k} (S - P_\infty)^{-1} \Omega^T \Xi^{-1}(0, k) + I]^{-1} X_0,$$

or taking into account expressions (6) and (8) we have

$$\begin{aligned} \bar{X}_{K_1} - X_{K_1} = & \{ [\Omega^{-k} (S - P_\infty) \Omega^{T(-k)} \Xi(0, k) + I] \times \\ & \times [\Omega^{-k} + \Xi(k, K_1) \Omega^{T(k)} (S - P_\infty)] \}^{-1} X_0. \end{aligned} \quad (12)$$

The quality index for the controller under consideration has the form

$$\bar{J} = \bar{X}_{K_1}^T S \bar{X}_{K_1} + \sum_{i=0}^{K_1-1} \{ \bar{U}_i^T R \bar{U}_i + \bar{X}_i^T Q \bar{X}_i \} + \sum_{i=k}^{K_1-1} \{ U_i^T R U_i + X_i^T Q X_i \},$$

where \bar{U}_i is the sequence of suboptimal controls defined by expression (5) and U_i is the sequence of optimal controls defined by formula (3). The final expression can be written as

$$\bar{J} = \bar{X}_k^T (P_k - P_\infty) \bar{X}_k + X_0^T P_\infty X_0.$$

The quality index deviation from its minimum value is equal to

$$\bar{J} - J = \bar{X}_k^T (P_k - P_\infty) \bar{X}_k + X_0^T (P_0 - P_\infty) X_0.$$

Taking into account that $\bar{X}_k = \Omega^k X_0$, we obtain

$$\bar{J} - J = X_0^T \bar{M} X_0, \quad (13)$$

where

$$\bar{M} = \Omega^{T(k)} (P_k - P_\infty) \Omega^k - (P_0 - P_\infty).$$

Using formula (8) we get

$$\begin{aligned} \bar{M} = & [\Omega^{-K_1} (S - P_\infty)^{-1} \Omega^{T(-K_1)} + \Xi(k, K_1)]^{-1} - \\ & - [\Omega^{K_1} (S - P_\infty) \Omega^{T(-K_1)} + \Xi(0, k)]^{-1}. \end{aligned}$$

Denote $Z = \Omega^{-K_1} (S - P_\infty)^{-1} \Omega^{T(-K_1)} + \Xi(k, K_1)$. Taking into consideration that $\Xi(0, K_1) = \Xi(0, k) + \Xi(k, K_1)$ and applying the lemma on matrix inversion to expression (14), it can be written

$$\bar{M} = [Z \Xi^{-1}(0, k) Z + Z]^{-1}. \quad (15)$$

The relative variation of the quality index can be found as

$$\delta J = \frac{X_0^T \bar{M} X_0}{X_0^T [\Omega^{T(K_1)} (S - P_\infty) \Omega^{K_1} + P_\infty - \bar{M}] X_0}.$$

The obtained formulas (12), (13) and (15) define the final state and quality index variation of the controller suboptimal over an interval due to the non-optimality of a sequence of controls \bar{U}_i ($i = 0, 1, \dots, k-1$) in the form of explicit functions of the parameters of system (1), weighing matrices S , Q and R in index (2) and steady-state solution of the Riccati equation P_∞ . The given expressions can be used for a non-recursive algebraic solution of the problem of determining the minimum length $K_1 - k$ of the sequence of optimal controls U_i ($i = k, k+1, \dots, K_1-1$) with varying coefficient L_k , which is to be realized for ensuring acceptable quality of control based on the assigned in advance allowances for deviations of quality characteristics from optimal.

5. Example

Let us consider a discrete form of the angular velocity stabilization problem (Kwakernaak and Sivan, 1972). The system is described by the difference equation

$$\xi_{k+1} = e^{-\alpha\Delta t}\xi_k + \frac{\kappa}{\alpha}(1 - e^{-\alpha\Delta t})\mu_k,$$

where Δt is the sampling period. The optimal control μ_k minimizes the criterion

$$J = \pi_1\xi_{K_1}^2 + \sum_{i=0}^{K_1-1} \{\xi_i^2 + \rho\mu_i^2\}.$$

If values of parameters are equal to $\alpha = 0.5 \text{ s}^{-1}$, $\kappa = 150 \text{ rad/s}^2$, $\rho = 1000$, $K_1 = 20$, $\Delta t = 0.1 \text{ s}$, $\pi_1 = 8.65$, then the steady solution of Riccati equation is the following: $P_\infty = 3.257758$. The transient matrix scalar in this case of the closed-loop control system

$$\Omega = e^{-\alpha\Delta t} - \frac{\kappa^2}{\alpha^2}(1 - e^{-\alpha\Delta t})^2 \frac{(1 - P_\infty)e^{-\alpha\Delta t}}{\rho + \kappa^2/\alpha^2(1 - e^{-\alpha\Delta t})^2(1 + P_\infty)}$$

with the above parameter values equals $\Omega = 0.8044$.

In accordance with formula (9) the final state deviation $\bar{\xi}(20)$ from the optimal value $\xi(20)$ is given by

$$\bar{\xi}(20) - \xi(20) = 0.004693 \xi_0.$$

Let us assume that the control quality is acceptable only in the case when the final state deviation does not exceed $0.003 \xi_0$. In this case the stationary controller can not be realized over the entire control interval. It is necessary to use the controller suboptimal over an interval considered in Section 3. Now, determine the

minimum number of the variable coefficient realizations. According to expression (12) we obtain the deviation $\bar{\xi}(20) - \xi(20) = 0.002987 \xi(0)$ with $k = 11$, and with $k = 12$ the deviation is equal to $\bar{\xi}(20) - \xi(20) = 0.0030047 \xi(0)$.

In such a way, the suboptimal controller allows for realization of the stationary law of control \bar{U}_i ($i = 0, 1, 2, \dots, 10$) and optimal one U_i ($i = 11, 12, \dots, 19$).

Similarly, a relative variation of the quality index due to the unoptimality of control can be found. Using expressions (10) and (11) we obtain $\delta J = 0.506$. Suppose that the system quality is acceptable if the quality index deviation does not exceed 10 % of its optimal value, i.e. it is necessary to satisfy the condition $\delta J < 0.1$. From formulas (13) and (15) it follows that the required condition is satisfied when $k = 19$, i.e. switching to optimal control is to be done only when $k = 19$.

6. Conclusion

The analytical relationships characterizing a decrease of quality indices of a control system were obtained when the constant feedback coefficient over the whole control interval is used. These relationships make it possible to select the maximum interval, based on assigned allowances of quality characteristics, over which it is acceptable to use the stationary law of control with subsequent switching to the optimal one, when approaching the final moment of time. Such a suboptimal controller allows to decrease substantially the required computing time, preserving optimality of control at the end of the control interval.

Appendix

An explicit form of the Riccati equation

A solution of the discrete Riccati equation in the problems of optimal filtering and the transient matrix of the optimal filter is obtained by Ortanidis (1982) in the form of explicit functions of steady-state solution of Riccati equation. The approach suggested here is based on constructing and solving an equation for the difference of Riccati equation solutions P_k and its steady-state solution value P_∞ .

To obtain the Riccati equation in control problems let us consider the difference of the feedback coefficient L_k and its steady-state value L_∞

$$\begin{aligned} \bar{L}_k = L_k - L_\infty = & [R + B^T(Q + P_{k+1})B]^{-1} B^T(Q + P_{k+1})\phi - \\ & - [R + B^T(Q + P_\infty)B]^{-1} B^T(Q + P_\infty)\phi. \end{aligned}$$

After simple transformations it is possible to get an expression for \bar{L}_k in the form of the feedback coefficient of the system allowing an explicit solution of the Riccati equation of transient matrix

$$\bar{L}_k = (\bar{R} + B^T \bar{P}_{k+1} B)^{-1} B^T \bar{P}_{k+1} \Omega, \quad (\text{A.1})$$

where $\Omega = \phi - BL_\infty$ is the transient matrix of a closed control system in steady-state mode, $\bar{P}_k = P_k - P_\infty$; $\bar{R} = R + B^T \bar{P}_{k-1} B$.

The difference of the Riccati equation solution and its steady state value will be written as follows

$$\begin{aligned} \bar{P}_k &= \phi^T (Q + \bar{P}_{k+1} + P_\infty) (\phi - B\bar{L}_k - BL_\infty) - \phi^T (Q + P_\infty) (\phi - BL_\infty) = \\ &= \phi^T \bar{P}_{k+1} (\phi - BL_k) - \phi^T (Q + P_\infty) B \bar{L}_k. \end{aligned}$$

From relationship (A.1) and taking into consideration that $\phi^T (Q + P_\infty) B = L_\infty^T \bar{R}$, after transformations we obtain

$$\bar{P}_k = \Omega^T \bar{P}_{k+1} (\Omega - B\bar{L}_k).$$

Applying the lemma on matrix inversion, put down for \bar{P}_k^{-1}

$$\bar{P}_k^{-1} = \Omega^{-1} (\bar{P}_{k+1}^{-1} + B\bar{R}^{-1} B^T)^{-1}, \quad (\text{A.2})$$

on solving of which we get

$$(P_k - P_\infty)^{-1} = \Omega (S - P_\infty)^{-1} \Omega^{T(k-K_1)} - \Omega^k \Xi(k, K_1) \Omega^{T(k)}, \quad (\text{A.3})$$

where $\Xi(k, K_1) = \sum_{i=k+1}^{K_1} \Omega^{-i} B [R + B^T (Q + P_\infty) B]^{-1} B^T \Omega^{T(-i)}$.

Using equation (A.2) the transient matrix of a closed-loop can be written as

$$\psi_{K_1, k} = (S - P_\infty)^{-1} \Omega^{T(k-K_1)} (P - P_\infty). \quad (\text{A.4})$$

Relationships (A.3) and (A.4) define the Riccati equation solution and transient matrix in the form of explicit functions of the steady-state solution of the Riccati equation.

References

1. Alexandrov, A. G., Controller Synthesis for MIMO Systems. Mashinostroyeniye, Moscow, 1986 (in Russian).
2. Anderson, B. D. O. and Liu, Y., Controller reduction: concepts and approaches. IEEE Trans. Automat. Control, 1989, vol. AC-34, No. 8, pp. 802-812.
3. Andreyev, Y. N., Control of Finite-Dimensional Dynamic Objects. Nauka, Moscow, 1976 (in Russian).
4. Bolnokin, V. E. and Chinayev, P. I., Computer Analysis and Synthesis of Automatic Control System. Radio and Svyaz, Moscow, 1986 (in Russian).
5. Iserman, R., Digital Control Systems, Springer-Verlag, New York, 1981.
6. Kalman, R. E., Contributions to the theory of optimal control. Bull. Soc. Math. Mec., 1960, vol. 5.
7. Krasovskiy, N. N., The integral estimates of moments and linear systems synthesis. Automat. and Remote Control, 1967, No. 10, pp. 53-71 (in Russian).
8. Krutko, P. D., Synthesis of Digital Control System Using Variational Techniques. Sov. Radio, Moscow, 1967 (in Russian).

9. Kwakernaak, H. and Sivan, R., *Linear Optimal Control Systems*, Wiley, New York, 1972.
10. Letov, A. M., Analytical controller design. *Automat. and Remote Control*, 1960, vol. 21, pp. 303-306 (in Russian).
11. Ortanidis, S., An exact solution of the time-invariant discrete Kalman filter, *IEEE Trans. Autom. Control*, 1982, vol. AC-27, No. 1, pp. 240-242.
12. Repin, Y. M. and Tretyakov, V. E., Analytical design of controllers using electronic models. *Automat. and Remote Control*, 1963, No. 6, pp. 738-743 (in Russian).
13. Wonham, M. M., *Linear Multivariable Control: A Geometric Approach*. Springer-Verlag, New York, 1979.
14. Yosida, T. and Loparo, K. A., Quadratic regulatory theory for analytic nonlinear systems with additive controls. *Automatica*, 1989, vol. 25, No. 4, pp. 531-544.

Алгоритм субоптимального управления дискретных систем

Ю. КОРБИЧ, В. Н. ПОДЛАДЧИКОВ, П. И. БИДЮК

(Зелёна-Гура, Киев)

Рассматривается проблема аналитического конструирования субоптимального линейного регулятора. Предложен подход к аналитическому исследованию изменения функционала качества и вектора состояния линейной дискретной системы при использовании неоптимального управления с постоянным коэффициентом обратной связи. Конструируется субоптимальный регулятор, в котором на первом подинтервале используется постоянный коэффициент обратной связи, а на втором — оптимальный, вычисляемый с помощью решения уравнения Риккати. Предлагается методика вывода момента переключения системы управления от субоптимального на оптимальный регулятор.

Józef Korbicz
Higher College of Engineering in Zielona Góra
Department of Applied Mathematics and Computer Sciences
ul. Podgórna 50,
65-246 Zielona Góra, Poland

Vladimir Podladchikov and Peter Bidyuk
Kiev Polytechnical Institute
Department of Computation Technique and Computer Sciences
252056 Kiev, prosp. Pobedy 37, USSR

A DECOUPLING POLE-PLACEMENT CONTROLLER FOR A CLASS OF MULTIVARIABLE SYSTEMS

J. MIKLEŠ, A. MÉSZÁROS

(*Bratislava*)

(Received December 4, 1990)

The design problem of controller for discrete-time linear MIMO systems is discussed. A pole-placement problem formulation and its solution is given. The resulting closed-loop system will be stable and decoupled. The controller design requires no *a priori* information either about stability or minimum phase of controlled plants. The designed controller consists of a feedback part, a feedforward part and a new precompensator and is suitable for control of systems with differing dead-times in each path.

1. Introduction

The deterministic tracking problem is one of the most significant ones in optimal control. The presented work is devoted to control algorithm design for multivariable discrete-time systems. The algorithm is based on explicit pole-placement design which, in addition to its other advantages, results in a decoupled control system. The autonomous state is achieved by using a suitably designed feedforward compensator part.

Recently, several works have appeared being dedicated to multi-input, multi-output (MIMO) control systems and taking into account also the decoupling effect. Deterministic decoupling design problems are discussed e.g. in [1], [5], [6], [8], [10], [12], [13] and [17] while the authors of [4], [7], [9] and [11] deal with decoupling aspects for stochastic systems. In [1], a deterministic discrete-time linear minimum phase system is considered and a regulator with decoupling influence to the overall system is proposed. A decoupled control system can be designed as well on the basis of state space approach as it has been shown in [5]. The decoupling effect can be also guaranteed by a suitable design of the system precompensator as a result of the left matrix denominator factorization of the controlled system [6]. The regulator proposed in [8] results from the assumption that the left matrix denominator of the controlled system transfer function is a diagonal polynomial matrix. The authors of [10] suggest a feedback controller which yields a decoupled closed-loop

system with higher order tracking. In [12] the decoupling is ensured by means of the feedforward part of the controller which has a transfer matrix, such as it is in [3], in the form of a simple polynomial matrix. To achieve a decoupled control system [13], like [8], came out from the assumption that the left matrix denominator of the controlled system description is a diagonal polynomial matrix. In [14] the problem of decoupling a linear continuous system by dynamic compensation into multi-input, multi-output subsystem is solved by the transfer matrix method. The decoupling is ensured, similarly as in [12], by means of the feedforward part of the controller. When comparing this with our results it is obvious that the regulator proposed here covers both the decoupling effect and the tracking problem for discrete systems (see Note 1). In [17] the closed-loop system is decoupled via internal loop while tracking is ensured using external decoupling loops. In comparison with [1], [7], [12], [13] and [14] the regulator designed herein has a precompensator which changes basically the solvability conditions of the decoupling tracking problem (see Note 2).

The controller introduced in this paper consists of a feedback part, a feedforward part and a precompensator. The feedback part is carried out on the basis of explicit pole-placement design (of course, implicit approach defined by a criterion function can be used as well) as it has been shown in [2] and [18]. The precompensator ensures usually integral action. The decoupling behaviour is created by a suitable feedforward part. The introduced precompensator has significantly simplified the solvability conditions of the problem [18]. Finally, a simple example is given.

2. Formulation

The configuration illustrated in Fig. 1 is considered; S is the plant to be controlled, C_B is the feedback part of the controller, C_F is the feedforward part of it, and P_C is the precompensator.

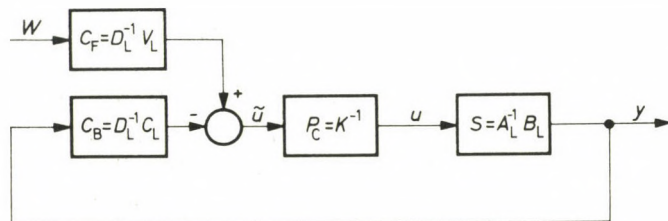


Fig. 1. System configuration

Consider a controllable and observable linear discrete-time invariant system so that it is minimal, modelled by equation

$$A_L(z^{-1})\mathbf{y} = B_L(z^{-1})\mathbf{u} \tag{1}$$

where \mathbf{y} is the r -vector output sequence and \mathbf{u} is the r -vector input sequence. A_L and B_L are polynomial matrices in z^{-1} , which is to be interpreted as the delay operator. The A_L is an $r \times r$ matrix with $A_L(0)$ invertible and the B_L is an $r \times r$ matrix with $B_L(0) = 0$. The A_L and B_L are of arbitrary relative degrees. $A_L^{-1}B_L$ is a left matrix fraction representation of the transfer matrix and reflects the input-output properties of the system. A_L and B_L are relatively left prime. The assumption $B_L(0) = 0$ means that the present value of \mathbf{u} can not affect the present value of \mathbf{y} . All time-delays of the controlled system are absorbed into the $B_L(z^{-1})$.

Further, consider a reference r -vector sequence \mathbf{w} modelled by the equation

$$K(z^{-1})\mathbf{w} = L_L(z^{-1}) \tag{2}$$

where K and L_L are $r \times r$ and $r \times 1$ polynomial matrices in the delay operator z^{-1} , respectively. It is assumed that $K(0)$ is invertible, $K(z^{-1}) = \text{diag}\{k\}$. K and L_L are relatively left prime, $A_L K$ and B_L are relatively left prime.

For future purposes, let us define relatively right prime polynomial matrices A_R and B_R of dimensions $r \times r$ such that

$$A_L^{-1}B_L = B_R A_R^{-1}. \tag{3}$$

The controller is another dynamical system, so that it is minimal, of the form

$$D_L(z^{-1})\tilde{\mathbf{u}} = -C_L(z^{-1})\mathbf{y} + V_L(z^{-1})\mathbf{w} \tag{4}$$

$$K(z^{-1})\mathbf{u} = \tilde{\mathbf{u}} \tag{5}$$

where the pairs of matrices D_L, C_L and D_L, V_L are relatively left prime polynomial matrices in the delay operator z^{-1} . D_L, C_L and V_L are $r \times r$ polynomial matrices, with $D_L(0)$ invertible.

Further, let us define relatively right prime polynomial matrices D_R and C_R being, both, of dimension $r \times r$, by

$$D_L^{-1}C_L = C_R D_R^{-1}. \tag{6}$$

Finally, let us introduce a stable polynomial matrix $M_r = \text{diag}\{m\}$ of dimension $(r \times r)$.

Then, the decoupling tracking problem using pole-placement design can be formulated as follows.

Given a system (1) and a class of references (2), it is desired to find a linear control law (4), (5) so as to make the sequence $\tilde{\mathbf{u}}$ and the tracking error $\mathbf{e} = \mathbf{w} - \mathbf{y}$, both being vectors of stable sequences, independently of L_L . Simultaneously, the decoupling of the overall system must be ensured in accordance with the equation

$$\mathbf{y} = G(z^{-1})\mathbf{w} \quad (7)$$

where $G(z^{-1})$ is a diagonal matrix.

3. Solution

THEOREM. The decoupling pole-placement tracking problem has a solution if and only if $A_L K$ and B_L are relatively left prime. Then D_L and C_L (or equivalently D_R and C_R) are as a solution of the equation

$$D_L K A_R + C_L B_R = M_r \quad (8)$$

(or equivalently

$$A_L K D_R + B_L C_R = M_r). \quad (9)$$

V_L is given in the form

$$V_L = \text{adj } B_R \text{diag} \left\{ \frac{1}{b_{rj}} \right\} \text{diag} \{r_j\} \quad (10)$$

where b_{rj} is the greatest common divisor of j -th column elements of the matrix $\text{adj } B_R$, $j = 1, 2, \dots, r$, and r_j is given by the equation

$$\text{diag}\{k\} \text{diag}\{s_j\} + B_R V_L = \text{diag}\{m\}. \quad (11)$$

Proof. The proof will be divided into two parts. First, we shall construct the controller, provided it exists, and then we shall establish the solvability.

Concerning the first part, we start by deriving $\tilde{\mathbf{u}}$ and \mathbf{e}

In accordance with equations (1), (2), (4), (5) the $\tilde{\mathbf{u}}$ sequence is given by

$$\tilde{\mathbf{u}} = (I_r + D_L^{-1} C_L K^{-1} A_L^{-1} B_L)^{-1} D_L^{-1} V_L K^{-1} L_L. \quad (12)$$

Using (3) and (6) and considering the assumption $K = \text{diag}\{k\}$ we get

$$\tilde{\mathbf{u}} = A_R (D_L K A_R + C_L B_R)^{-1} V_L L_L. \quad (13)$$

Similarly, through elementary algebraic operations, it can be derived

$$\mathbf{e} = (I_r - B_R(D_L K A_R + C_L B_R)^{-1} V_L) K^{-1} L_L. \quad (14)$$

If equation (8) holds true, it yields

$$\tilde{\mathbf{u}} = A_R M_r^{-1} V_L L_L \quad (15)$$

$$\mathbf{e} = (I_r - M_r^{-1} B_R V_L) K^{-1} L_L \quad (16)$$

$$\mathbf{y} = M_r^{-1} B_R V_L \mathbf{w}. \quad (17)$$

The simultaneous satisfaction of equations (8) and (9) is a result of the general Bezout identity. In equation (15), $\tilde{\mathbf{u}}$ is a vector of stable sequences, independently of L_L . Since (11) holds true, \mathbf{e} is a vector of stable sequences, independently of L_L .

The introduction of $\det B_R/b_{rj} = b_{r1j}$ yields

$$M_r \mathbf{y} = B_R \operatorname{adj} B_R \operatorname{diag} \left\{ \frac{1}{b_{rj}} \right\} \operatorname{diag}\{r_j\} \mathbf{w} \quad (18)$$

$$M_r \mathbf{y} = \det B_R \operatorname{diag} \left\{ \frac{1}{b_{rj}} \right\} \operatorname{diag}\{r_j\} \mathbf{w} \quad (19)$$

$$M_r \mathbf{y} = \operatorname{diag}\{b_{r1j} r_j\} \mathbf{w}. \quad (20)$$

From equation (20) it follows that the overall system is decoupled.

The second part of the proof is connected with the solvability of the problem.

The decoupling pole-placement tracking problem has a solution if the general pole-placement tracking problem given by equations (8) and (9) and extended by the equation

$$KS + B_R V_L = M_r \quad (21)$$

is solvable. The solutions of (21), V_L and S have no constraints. The only condition having to be fulfilled is that M_r must be a stable matrix. The general pole-placement tracking problem has a solution if and only if $A_L K$ and B_L are relatively left prime [18]. This completes the proof of the theorem.

Note 1. In [14] a regulator is proposed the feedforward part of which is given by the right inverse of the left matrix numerator of the continuous controlled system transfer matrix. The regulator given by eqs. (8), (9), (10) solves besides the decoupling problem also the discrete tracking problem for the class of reference sequences given by (2). All the polynomial matrices of the regulator are solutions of polynomial matrix Diophantine equations.

Note 2. In the case of a controller without a precompensator, i.e. when $\mathbf{u} = \tilde{\mathbf{u}}$, there exists an additional solvability condition in the tracking problem: that K must be a right divisor of A_L [15], [16]. Using a precompensator of the form [5] this additional condition disappears.

4. Example

Let us consider a controlled system described by matrices

$$A_L = \begin{pmatrix} 1 + 2z^{-1} & 0.1z^{-1} \\ 0.3z^{-1} & 1 + 0.1z^{-1} \end{pmatrix}, \quad B_L = \begin{pmatrix} 0.2z^{-1} & 0.1z^{-1} \\ 0.1z^{-1} & 0.3z^{-1} \end{pmatrix}.$$

Further, it is given

$$k = 1 - z^{-1}, \quad M_r = \begin{pmatrix} 1 + 0.25z^{-1} & 0 \\ 0 & 1 + 0.25z^{-1} \end{pmatrix}.$$

After a short calculation procedure the following matrices can be determined

$$A_R = \begin{pmatrix} -1 + 0.2z^{-1} & -0.6 - 1.14z^{-1} \\ 2 + 0.1z^{-1} & 0.2 + 0.28z^{-1} \end{pmatrix}, \quad B_R = \begin{pmatrix} 0 & -0.1z^{-1} \\ 0.5z^{-1} & 0 \end{pmatrix}$$

The feedback part of the controller can be calculated from (8) which in this case takes the form

$$\begin{aligned} & \begin{pmatrix} d_1 & d_2 \\ d_3 & d_4 \end{pmatrix} \begin{pmatrix} 1 - z^{-1} & 0 \\ 0 & 1 - z^{-1} \end{pmatrix} \begin{pmatrix} -1 + 0.2z^{-1} & -0.6 - 1.14z^{-1} \\ 2 + 0.1z^{-1} & 0.2 + 0.28z^{-1} \end{pmatrix} + \\ & + \begin{pmatrix} c_1 + c_2z^{-1} & c_3 + c_4z^{-1} \\ c_5 + c_6z^{-1} & c_7 + c_8z^{-1} \end{pmatrix} \begin{pmatrix} 0 & -0.1z^{-1} \\ 0.5z^{-1} & 0 \end{pmatrix} = \\ & = \begin{pmatrix} 1 + 0.25z^{-1} & 0 \\ 0 & 1 + 0.25z^{-1} \end{pmatrix}. \end{aligned}$$

r_1 and r_2 result from (10) as follows

$$(1 - z^{-1})s_1 - 0.1z^{-1}r_1 = 1 + 0.25z^{-1}$$

$$(1 - z^{-1})s_2 + 0.5z^{-1}r_2 = 1 + 0.25z^{-1}$$

$$r_1 = -12.5, \quad r_2 = 2.5$$

$$V_L = \begin{pmatrix} 0 & 2.5 \\ -12.5 & 0 \end{pmatrix}.$$

5. Conclusion

From the presented theory as well as the illustrated example, it follows that a matrix Diophantine equation and r scalar Diophantine equations have to be solved in order to obtain a regulator which insures a stable and decoupled overall

system. By extending the right side of equation (1) by vector $\epsilon(t)$, a nonzero initial condition influence on the control system as well as both, deterministic or stochastic disturbances can be treated. However, each of these cases needs a special analysis. Furthermore, the designed algorithm can serve as a good basis for multivariable self-tuning control.

References

1. Tade, M. O., Bayoumi, M. M. and Bacon, D. W., Adaptive decoupling of a class of multivariable dynamic systems using output feedback. IEE Proceedings, vol. 133 (1986), Pt. D., pp. 265–275.
2. Mikleš, J. Hutla, V., Control Theory (in Czech). ALFA-SNTL, Bratislava, 1986.
3. Mikleš, J., Simple multivariable self-tuning controllers (in Czech). Automatizace, vol. 29 (1986), pp. 149–154.
4. Lang, S. J., Gu, X. Y. and Chai, T. Y., A Multivariable Generalized Self-tuning Controller with Decoupling Design. IEEE Trans. Automatic Control, vol. AC-31 (1986), pp. 474–477.
5. Schumann, R., Entkoppelte Abtastregler für Mehrgrößenprozesse mit Totzeiten. Automatisierungstechnik, vol. 35 (1987), pp. 202–208.
6. Kinnaert, M., Hanus, R. and Henrotte, J. L., A New Decoupling Precompensator for Indirect Adaptive Control of Multivariable Linear Systems. IEEE Trans. Automatic Control, vol. AC-32 (1987), pp. 455–459.
7. Wang, W. J., Lee, T. T., Poles-zeros placement and decoupling in discrete LQG systems. IEE Proceedings, vol. 134 (1987), Pt. D., pp. 388–394.
8. Janiszovski, K., Unbehauen, H., Entwurf eines neuartigen diskreten Entkopplungsreglers für Mehrgrößenregelstrecken. Automatisierungstechnik, vol. 35 (1987), pp. 56–65.
9. Heroh, M. A., Multiloop self-tuning control: decoupling and pole assignment regulators for n-input n-output stochastic systems. Int. J. Control, vol. 45 (1987), pp. 33–45.
10. Chen, B. S. Lin, Ch. M., Multipurpose adaptive control in deterministic multivariable system. IEE Proceedings, vol. 135 (1988), Pt. D., pp. 282–288.
11. Chai, T. Y., A Self-Tuning Decoupling Controller for a Class of Multivariable Systems and Global Convergence Analysis. IEEE Trans. Automatic Control, vol. AC-33 (1988), pp. 767–771.
12. Bayoumi, M. M. Mo, L., Adaptive decoupling control of MIMO system. In: Preprints 8th IFAC/IFORS Symposium on Identification and system parameter estimation. Beijing, Pergamon Press, vol. 1, 1988, pp. 110–114.
13. Zhang, W. Y., Wang, Z. J. and Zhang, X. H., Decoupling-adaptive control of a class of multivariable systems. In: Preprints IFAC Symposium on Adaptive systems in control and signal processing, Glasgow, 1989, pp. 171–174.
14. Kučera, V., Block decoupling by dynamic compensation with internal properness and stability. Problems of Control and Information Theory, vol. 12 (1983), pp. 379–389.
15. Šebek, M., Multivariable deadbeat servo problem. Kybernetika, vol. 16 (1980), pp. 443–453.
16. Šebek, M. Kučera, V., Polynomial approach to quadratic tracking in discrete linear systems. IEEE Trans. Aut. Control, vol. AC-27 (1982), pp. 1248–1250.
17. Mikleš, J. Šandor, J., A decoupling pole-placement self-tuning controller for a class of multivariable processes (in Czech). Automatizace, vol. 32 (1989), pp. 29–32.

18. Mikleš, J., A Multivariable Self-tuning Controller Based on Pole-placement Design. *Automatica*, vol. 26 (1990), pp. 293–302.

**О разнесении полюсов
одного класса многомерных систем управления**

Й. МИКЛЕШ, А. МЕШАРОШ

(Братислава)

Рассматривается проблема конструкции контроллера для дискретной линейной ММО системы. Дается формулировка проблемы распределения полюсов и ее решений. Результирующая система замкнутой цепи будет стабильна и разорвана. Для конструкции контроллера не нужна априорная информация ни об устойчивости, ни о минимальной фазе управляемой системы. Конструированный контроллер состоит из части обратной связи, из части прямой связи и из нового прекомпенсатора, и пригоден для управления системы с разыми постоянными времени в контурах.

J. Mikleš and A. Mészáros
Department of Process Control
STU-CHTF, Radlinského 9
812 37 Bratislava
Czechoslovakia

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent to

E.D. TERYAEV coordinating editor	or to	V. STREJC
Department of Mechanics and Control Processes		UTIA ČSAV
Academy of Sciences of the USSR		182 08 Prague 8
Leninsky Prospect 14, Moscow V-71, USSR		Pod vodárenskou věží 4, Czechoslovakia

or to L. GYÖRFI
Technical University of Budapest
H-1111 Budapest, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

The papers, preferably in English or Russian, should be typed double spaced on one side of good-quality paper with wide margins (4–5 cm). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as $\text{diag} (d_1, d_2, \dots, d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статье должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском языке в трех экземплярах, на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Papageorgiou, N. S.:</i> On the optimal control and relaxation of finite dimensional systems driven by maximal monotone differential inclusions (<i>Папагеоргиу Н. С. Об оптимальном управлении и релаксации конечномерных систем с максимально монотонным оператором, воздействующем на фазовые скорости систем</i>)	245
<i>Ferrante, M.:</i> On finite dimensional filtering in discrete time (<i>Ферранте М. О конечномерной фильтрации в дискретном времени</i>)	257
<i>Martos, B.:</i> Viable control trajectories in linear systems (<i>Мартош Б. Траектории управления выживаемости в линейных системах</i>)	267
<i>Korbicz, J., Podladchikov, V., Bidiuk, P.:</i> Suboptimal control algorithm for discrete systems (<i>Корбич Ю., Подладчиков В. Н., Бидюк П. И. Алгоритм субоптимального управления дискретных систем</i>)	281
<i>Mikleš, J., Mészáros, A.:</i> A decoupling pole-placement controller for a class of multivariable systems (<i>Миклеши Й., Месарош А. О разнесении полюсов одного класса многомерных систем управления</i>)	291

316920

VOL. 20 • NUMBER 5
TOM HOMEP

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

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

ПРОБЛЕМЫ
ПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

АКАДЕМИЯ НАУК С С С Р
ВЕНГЕРСКАЯ АКАДЕМИЯ НАУК
ЧЕХОСЛОВАЦКАЯ АКАДЕМИЯ НАУК

1991

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

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

- theory of control processes
- theory of adaptive systems
- theory of estimation and identification
- theory of controlling robot-technologic and flexible manufacturing systems
- information theory
- information-theoretic aspects of multiple access networks.

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

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

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

- теория процессов управления;
- теория адаптивных систем;
- теория оценивания и идентификации;
- теория управления робототехническими и гибкими производственными системами;
- теория информации;
- теория информации в области сетей с множественным доступом.

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

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czech and Slovak Federal Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia:

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

For all other countries:

PERGAMON PRESS PLC Headington Hill Hall, Oxford OX3 OBW, England
or

PERGAMON PRESS INC, Maxwell House, Fairview Park, Elmsford, NY 10523, USA
1991 Subscription Rate DM 627,— per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITOR

N. N. KRASOVSKII (USSR)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

K. V. FROLOV

A. B. KURZHANSKI

I. A. OVSEEVICH

E. D. TERYAEV

R. Z. KHASHMINSKII

HUNGARY

T. VÁMOS

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

L. GYÖRFI

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJC

I. VAJDA

РЕДАКТОР ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

К. В. ФРОЛОВ

А. Б. КУРЖАНСКИЙ

И. А. ОВСЕВИЧ

Е. Д. ТЕРЯЕВ

Р. З. ХАСЬМИНСКИЙ

ВНР

Т. ВАМОШ

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

Л. КЕВИЦКИ

Л. ДЪЕРФИ

Я. КОЧИШ

ЧССР

Й. БЕНЕШ

В. СТРЕЙЦ

И. ВАЙДА



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

A METHOD OF DESIGNING OF OBSERVABLE OUTPUT ENSURING GIVEN ZEROS LOCATION

YE. M. SMAGINA

(Tomsk)

(Received June 12, 1990)

This paper deals with the problem of choosing an observable (output) vector in linear time-invariant system ensuring arbitrary zeros location. The conditions required for the solution of this problem under natural conditions on the matrix of the output: rank-fullness and system observability are given. The simple analytic method of the solution which can easily be programmed for computer is presented.

1. Introduction

In some problems of estimation and filtering it is assumed that it is possible to choose sensors in the systems forming the desirable vector of observations (output) [1]. It is known that system zeros greatly influence the dynamic behaviour of any control or estimation (filtering) system [2]. Zeros are invariant relative to both the state and output feedback and they can be shifted only by proper choice of the input or output system-matrix. Therefore, in filtering (estimate, control) systems the problem of choosing the vector of an observation (output) ensuring arbitrary zero location arises. For the first time conditions of the solution of this problem were proposed by H. H. Rosenbrock [3]. But there the restrictions on the choice of the output (observable) matrix were not taken into account, i.e. fullness-rank and the observability of the system. This problem with such restrictions was considered lately [4], where an iterative method of the solution was proposed. This paper presents further developments of the results in [5] which enable us to formulate and prove the sufficient conditions of the solution and propose a simple computation method.

2. Statement of the problem

Consider a linear time-invariant system described by the differential equations

$$\dot{x} = Ax + Eu, \quad (1)$$

$$y = Hx, \quad (2)$$

where x is an n -vector of system state, u is an r -vector of arbitrary input, y is an r -vector of observable output, A , E , H are real constant matrices of appropriate sizes.

The behaviour of any dynamic system depends both on its poles and zeros. The zeros of system (1), (2) are defined as complex numbers $s = s_i$ satisfying the following rank-inequality

$$\text{rank } P(s)|_{s=s_i} = \text{rank} \begin{bmatrix} s_i I - A & -E \\ H & 0 \end{bmatrix} < n + r. \quad (3)$$

It is assumed that $\text{rank } E = r$ and the pair (A, E) is completely controllable. Let $\varphi(s)$ be a zero polynomial of the system (1), (2)*. It is obvious that $\varphi(s) = \det P(s)$. We denote the desirable zero polynomial as

$$\psi^*(s) = \prod_{i=1}^{\mu} (s - s_i), \quad (4)$$

where s_i ($i = \overline{1, \mu}$) are given distinct real or complex-conjugate numbers, μ is the number of zeros. As the maximum number of zeros in an n -order system with r inputs and outputs is $n - r$, we put $\mu = n - r$. We consider the problem of defining of such output matrix H ensuring the coincidence of the zero polynomial system (1), (2) with polynomial (4) and at same the time satisfying the following conditions:

$$\text{pair } (A, H) \text{ is observable}; \quad (5a)$$

$$\text{rank } H = r; \quad (5b)$$

$$\text{rank}(HE) = r. \quad (5c)$$

The necessity of the conditions a) and b) is clear. Condition c) follows from the proposed method of the zero assignment. This condition ensures that n -order system with r inputs and r outputs have exactly $n - r$ zeros [2].

Usually, this problem is considered for systems with the same number of inputs and outputs since otherwise the system "almost always" has no zeros and the problem of zero assignment does not arise.

3. The main result

First, we note that securing the condition (5a) depends on the location of the desirable zeros on the complex plane and the controllability of the pair (A, E) ,

* Matrix transfer function of the system (1), (2) from $y(s)$ to $u(s)$ is given by $H(sI - A)^{-1}E$.

THEOREM. The problem of defining matrix H ensuring both setting of zero polynomial for system (1), (2) and the simultaneous fulfilment of condition (5a), has a solution if the pair of matrices (A, E) is controllable and any given set of distinct zeros s_i does not coincide with every set of eigenvalues of the matrix A .

This result is proved in [4]. It implies that a matrix H ensuring condition (5a), always exists if the conditions of the Theorem are true.

Now, it is necessary to define the form of the matrix H , satisfying conditions (5b), (5c). For this purpose we transform the system (1), (2) into canonical Yokoyama, Kinnen [5] form. This transformation always exists if the pair (A, E) is completely controllable. The nonsingular transformation matrix is denoted by N . The canonical system output matrix C will be

$$C = HN^{-1}. \tag{6}$$

Let us carry out the partition of the matrix N^{-1} into ν blocks P_i of dimension $n \times l_i$

$$N^{-1} = [P_1, P_2, \dots, P_\nu], \tag{7}$$

where ν is the smallest integer ($\nu \leq n$) such that $\text{rank}[E, AE, \dots, A^{\nu-1}E] = n$ and integer numbers l_i ($l_1 \leq l_2 \leq \dots \leq l_\nu = r, l_1 + l_2 + \dots + l_\nu = n$) are characteristic for the controllability of the pair (A, E) :

$$l_i = \text{rank}[E, AE, \dots, A^{\nu-i} - E] - \text{rank}[E, AE, \dots, A^{\nu-i-1}E], \quad l_\nu = r, \quad i = \overline{1, \nu - 1}.$$

Let us carry out the partition of matrix C into ν blocks

$$C = [C_1, C_2, \dots, C_\nu], \tag{8}$$

where

$$C_i = HP_i, \quad i = \overline{1, \nu}. \tag{9}$$

If $\det C_\nu \neq 0$ then system zeros of (1), (2) are eigenvalues of the $((n - r) \times (n - r))$ matrix [6]:

$$Z = \begin{bmatrix} 0 & [0, I_{l_1}] & 0 & \dots & 0 \\ 0 & 0 & [0, I_{l_2}] & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & [0, I_{l_{\nu-2}}] \\ -Q_1 & -Q_2 & -Q_3 & \dots & -Q_{\nu-1} \end{bmatrix} \begin{matrix} \} l_1 \\ \} l_2 \\ \vdots \\ \} l_{\nu-2} \\ \} l_{\nu-1} \end{matrix} \tag{10}$$

$\underbrace{\hspace{1.5cm}}_{l_1} \quad \underbrace{\hspace{1.5cm}}_{l_2} \quad \underbrace{\hspace{1.5cm}}_{l_3} \quad \dots \quad \underbrace{\hspace{1.5cm}}_{l_{\nu-1}}$

where

$$Q_i = C_\nu^{-1} C_i, \quad l_{\nu-1} = r, \quad (11a)$$

$$Q_i = [0, I_{l_{\nu-1}}] C_\nu^{-1} C_i, \quad l_{\nu-1} < r, \quad (11b)$$

I_{l_i} is the unity ($l_i \times l_i$) matrix.

ASSERTION 1. Matrix C_ν of dimension $r \times r$ is nonsingular if and only if $\det(HE) \neq 0$.

Proof. Using (7), (9) we define the matrix C_ν

$$C_\nu = HP_\nu \quad (12)$$

From canonical Yokoyama, Kinnen form [5] we can write

$$NE = \begin{bmatrix} 0 \\ \tilde{G}_\nu \end{bmatrix},$$

where \tilde{G}_ν is an $(r \times r)$ nonsingular matrix. Defining the matrix E and using the block partition N^{-1} (7) we get

$$E = P_\nu \tilde{G}_\nu.$$

Multiplication of both sides of the last expression by H from the left, together with (12) give the following rank equalities

$$\text{rank}(HE) = \text{rank}(HP_\nu \tilde{G}_\nu) = \text{rank}(C_\nu \tilde{G}_\nu) = \text{rank } C_\nu,$$

which prove the assertion.

ASSERTION 2. For any given polynomial $\varphi^*(s)$ of order $n - r$ one can always find an $(l_{\nu-1} \times (n - r))$ submatrix $Q = [-Q_1, -Q_2, \dots, -Q_{\nu-1}]$ such that roots of polynomials $\det(sI_{n-r} - Z)$ and $\varphi^*(s)$ coincide.

Proof. If $l_{\nu-1} = 1$ and $Q = q = [-q_1, -q_2, \dots, -q_{n-r}]$ is a vector row then the following equality holds

$$\begin{aligned} \det(sI_{n-r} - Z) &= \det \begin{bmatrix} s & -1 & 0 & \dots & 0 \\ 0 & s & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & -1 \\ q_1 & q_2 & q_3 & \dots & s + q_{n-r} \end{bmatrix} = \\ &= s^{n-r} + q_{n-r} s^{n-r-1} + \dots + q_1. \end{aligned}$$

That is, the vector row $[-q_1, -q_2, \dots, -q_{n-r}]$ always exists for which the right-side of the last expression is a desirable polynomial.

Consider the case $l_{\nu-1} > 1$. We set the first $l_{\nu-1} - 1$ rows of the submatrix Q in such a way that every row has only one unit element and the rest are zeros. Unit elements are situated in such a way that the first $n - r - 1$ rows of Z form a submatrix with one unit element in the every but the first column. The other elements of this submatrix are zeros. The last row elements of the submatrix Z are uncertain. Denote these elements as $-q_1, -q_2, \dots, -q_{n-r}$. One can easily verify that q_i can be defined so that the Assertion 2 is fulfilled. Indeed, this matrix Z can be transformed by permutation of the first row to the companion form Z^* . Therefore, the following equalities are true

$$\begin{aligned} \det(sI_{n-r} - Z) &= (-1)^\alpha \det(sI_{n-r} - Z^*) = \\ &= (-1)^\alpha \begin{vmatrix} s & -1 & 0 & \dots & 0 \\ 0 & s & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ q_1 & q_2 & q_3 & \dots & s + q_{n-r} \end{vmatrix} = \\ &= (-1)^\alpha (s^{n-r} + q_{n-r}s^{n-r-1} + \dots + q_1), \end{aligned}$$

where α is the number of row permutations. It is obvious that elements q_i ($i = \overline{1, n-r}$) can be fixed so that the roots of the polynomial $\det(sI_{n-r} - Z)$ coincide with the roots of $\varphi^*(s)$, i.e.

$$\det(sI_{n-r} - Z) = \varphi^*(s). \tag{13}$$

The proof is complete.

So, by making use of Assertion 2 we can always find a submatrix Q ensuring the validity of (13). With the help of Q we can define matrix C .

Case 1. $l_{\nu-1} = r$. From (11a) we get $C_i = C_\nu Q_i$ ($i = \overline{1, \nu-1}$) and it implies the following structure of the matrix

$$C = C_\nu [Q_1, Q_2, \dots, Q_{\nu-1}, I_r] = C_\nu [-Q, I_r]. \tag{14}$$

In (14) the $(r \times r)$ submatrix C_ν is chosen according to the condition that $\text{rank } C_\nu = r$. By Assertion 1 it implies condition (5c). Moreover, it is obvious that this matrix C (14) has the full rank.

The matrix output H of the system (1), (2) will satisfy conditions (5a)–(5c) and it is defined by (6) as

$$H = CN \tag{15}$$

Case 2. $l_{\nu-1} < r$. From expression (11b) it follows that the upper blocks $\overline{Q}_i = [I_{r-l_{\nu-1}}, 0]C_\nu^{-1}C_i$ of the submatrix $C_\nu^{-1}C_i$ are arbitrary. Combining (11b) and the last equality

$$Q_i^* = \begin{bmatrix} \overline{Q}_i \\ Q_i \end{bmatrix} = C_\nu^{-1}C_i$$

we consider the matrix

$$Q^* = \begin{bmatrix} \overline{Q} \\ Q \end{bmatrix} = [-Q_1^*, -Q_2^*, \dots, -Q_{\nu-1}^*].$$

Since $Q^* = C_{\nu}^{-1}[C_1, \dots, C_{\nu-1}]$, then the matrix C has the form

$$C = C_{\nu}[Q_1^*, Q_2^*, \dots, Q_{\nu-1}^*, I_r] = C_{\nu}[-Q^*, I_r]. \quad (16)$$

Matrix H is defined by (15).

So, we come to the following summarizing algorithm for zero placement:

1. Calculation of the eigenvalues of matrix A ,
2. Setting of the desirable zeros $\overline{s}_1, \overline{s}_2, \dots, \overline{s}_{n-r}$ ($\overline{s}_i \neq \overline{s}_j$), which do not coincide with the eigenvalues of matrix A ,
3. Verification of the controllability of the pair (A, E) . If pair (A, E) is not controllable then the problem has no solution,
4. Definition of integers $\nu, l_1, l_2, \dots, l_{\nu}$ and matrices N, N^{-1} ,
5. From condition (13) finding the $(l_{\nu-1} \times (n-r))$ submatrix Q , if $l_{\nu-1} < r$ that forming matrix Q^* ,
6. From the condition $\text{rank } C_{\nu} = r$ designing submatrix C_{ν} ,
7. Calculation of matrix C from (14) or (16) and matrix H from (15).

Remark. The problem of finding submatrix Q ensuring the fulfilment of condition (13) coincides with that of a state feedback eigenvalue assignment (modal control by state feedback) [6]. To satisfy condition (13) only one row of the submatrix is needed. Therefore, if $r > 1$ then Q has $((r-1) \times (n-r))$ free elements. These elements can be used to satisfy supplementary requirements to the system (as in [7]). For example, the problems of minimization of a performance $I = \text{tr } HH^T$ or ensuring the structure restriction on the matrix H can be considered in a similar way.

4. Numerical examples

Example 1. Consider system (1), (2) with $n = 4, r = 2$ and

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \quad E = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}. \quad (17)$$

Let the desired zeros be $\overline{s}_1 = -1, \overline{s}_2 = -2, (\varphi^*(s) = s^2 + 3s + 2)$.

Check the conditions of the theorem. One can verify that the pair (17) is completely controllable and eigenvalues of matrix A do not coincide with $\overline{s_1}, \overline{s_2}$. Since the rank $[E, AE] = 4$ then for this system we get $\nu = 2, l_1 = l_2 = 2$.

We find the matrix of transformation N and N^{-1} as

$$N = \begin{bmatrix} 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad N^{-1} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 1 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

Since in this case $n - r = 2, l_{\nu-1} = l_1 = r = 2$, then the upper block in matrix Z is absent. This matrix has following form

$$Z = Q = -Q_1 = \begin{bmatrix} -q_1 & -q_2 \\ -q_3 & -q_4 \end{bmatrix}.$$

We set $q_1 = 1, q_2 = 1$ and define a polynomial as

$$\det(sI - Q) = \det \begin{bmatrix} s+1 & 1 \\ q_3 & s+q_4 \end{bmatrix} = s^2 + s(1+q_4) - q_3 + q_4.$$

By making comparison of the right-hand part of the last expression with the polynomial $\varphi^*(s)$ we obtain the following equations: $1 + q_4 = 3, -q_3 + q_4 = 2$. Hence $q_3 = 0, q_4 = 2$. Thus, we get

$$Q = \begin{bmatrix} -1 & -1 \\ 0 & -2 \end{bmatrix}.$$

Putting $C_\nu = I_2$ and substituting these C_ν and Q into (14) yields

$$C = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 2 & 0 & 1 \end{bmatrix}.$$

Matrix H can be found by (15)

$$H = \begin{bmatrix} 1 & 0 & -1 & 1 \\ 0 & 3 & 0 & 1 \end{bmatrix}. \quad (18)$$

Substituting (18) into correlation (3) shows that zero polynomial coincides with the desirable one.

Example 2. Consider the system in Example 1 with the matrix

$$E^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Let the desired zero polynomial be as follows $\varphi^*(s) = s^2 + 3s + 2$. One can easily verify that for this system the conditions of the Theorem holds. But since the rank $[E, AE, A^2E] = 4$, then $\nu = 3$, $l_1 = l_2 = 1$, $l_3 = 2$. The transformation matrix N and N^{-1} are as follows [6]

$$N = \begin{bmatrix} 0 & 0 & 0.5 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \quad N^{-1} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix}.$$

As in this case $n - r = 2$, $l_{\nu-1} = l_2 = 1 < r = 2$ and, therefore, the (2×2) matrix Z has the form

$$Z = \begin{bmatrix} 0 & 1 \\ -q_1 & -q_2 \end{bmatrix}, \quad \text{where } [-q_1, -q_2] = Q$$

Substituting q_1 and q_2 in (13) yields the equation

$$\det(sI_2 - Z) = \det \begin{bmatrix} s & -1 \\ q_2 & s + q_2 \end{bmatrix} = s^2 + sq_2 + q_1 = \varphi^*(s).$$

From here it follows that $q_1 = 2$, $q_2 = 3$. Thus, we obtain $Q = [-2, -3]$. Put $\bar{Q} = [1 \ 1]$ and form the matrix

$$Q^* = \begin{bmatrix} 1 & 1 \\ -2 & -3 \end{bmatrix}.$$

Putting $C_\nu = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and substituting C_ν and Q^* in (16) yields

$$C_\nu = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 2 & 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 \\ 2 & 3 & 0 & 1 \end{bmatrix}.$$

Using (15) we find the matrix

$$H = CN = \begin{bmatrix} 1 & 1 & 0.5 & 0 \\ 0 & 4 & 1 & 1 \end{bmatrix}.$$

5. Conclusion

The problem of system zero placement in linear system using observable (output) vectors is considered. Conditions for complete zero placement are given. A computational algorithm for zero placement has been developed.

References

1. Arbel, A., Sensor placement in optimal filtering and smoothing problems, *IEEE Trans. Autom. Control*, 1982, **AC-27**, I, pp. 94-98
2. Smagina, Ye. M., Zeros of multi-dimensional systems: Definitions, classification, applications. *Avtomatika i Telemekhanika*, 1985, **12**, pp. 5-33.
3. Rosenbrock, H. H., *State-space and multivariable theory*. Wiley, 1970.
4. Смагина Е. М., Обеспечение заданных нулей линейной многомерной системы. В сб. Автоматическое управление объектами с переменными характеристиками. Новосибирск, 1986, с. 145-151.
5. Yokoyama, R., Kinnen, E., Phase-variable canonical form for linear multi-input, multi-output systems, *Intern. J. Control*, 1973, **17**, 6, pp. 1297-1312.
6. Smagina, Ye. M., Computing and specification of zeros in a linear multidimensional systems. *Avtomatika i Telemekhanika*, 1987, **12**, pp. 165-173.
7. Смагина Е. М., Синтез систем оптимального модального управления. Известия ВУЗов. Приборостроение, 1981, **7**, с. 32-36.

Метод проектирования наблюдаемого выхода, обеспечивающего заданные нули

Е. М. СМАГИНА

(Томск)

Для системы оценивания и/или фильтрации рассматривается проблема выбора вектора наблюдения (выхода), обеспечивающего заданные нули. Предлагается аналитический метод решения данной задачи, основанный на определении нулей в терминах собственных чисел специальной матрицы, сформированной на основе матрицы выхода системы. Получены простые условия разрешимости проблемы задания нулей при следующих естественных ограничениях на матрицы выхода: полнота ранга, наблюдаемость системы оценивания или фильтрации. Предложен алгоритм решения выбора матрицы наблюдения (выхода), легко реализуемый на ЭЦВМ. Метод иллюстрируется двумя числовыми примерами.

Е. М. Смагина
Сибирский физико-технический институт
им. В. Д. Кузнецова
СССР, 634050, г. Томск, пл. Революции, 1.

MINIMAX FILTERING IN REAL TIME OF MULTISTAGE SYSTEMS

V. I. SHIRYAEV

(Tchelyabinsk)

(Received September 9, 1990)

For nonlinear multistage systems with geometric restrictions in statistically uncertain situations there has been developed an estimation algorithm realized in real time. For this purpose the basic calculations are being made on the basis of a priori and a posteriori data separation before the results of the next measurement are obtained. The volume of calculations performed in real time does not much exceed the Kalman filter.

1. Introduction

This paper is devoted to the problems of dynamics estimation of phase vector of linear control multistage systems functioning in jerks, by the results of process parameter measurement under additive disturbances influence [1, 2]. Actual for the supplements [3, 4] are the problems when jerks and disturbance information are almost unavailable and reduced to either determining their coordinate measurement areas or the whole class of feasible distribution functions determining the deviation realization if the latter are of statistical origin.

The above mentioned situations resulted in the filtration theory development in a game-like organization [5-11]. In [6-8] minimax filtration determinative correlations are given for systems containing deviations and disturbances of both feasible and uncertain character at the same time. In this case, filtration algorithm is reduced to making information sets and their Tchebyshev centers, taken for phase vector estimation. However, constructive procedures of making both information sets and their Tchebyshev centers are not indicated for real time. Therefore, realization of the algorithm in real time is possible only for symmetrical areas of jerks and disturbance alterations of uncertain character.

In this paper a minimax filtration algorithm is given, when deviations and disturbance alteration areas are either convex polygons or can be approximated by them. Our paper develops the approach [6-9], [13-15] and adjoins [10] on convex polygon use in the optimal systems theory.

2. Preliminary problem formulation

Let the process dynamics be described by a linear n -vector system

$$\mathbf{x}_{k+1} = A_k \mathbf{x}_k + B_k \mathbf{u}_k + C_k \xi_k, \quad k = 0, 1, \dots \quad (2.1)$$

After each stage in the system the r -vector parameter y_k measurement is performed, where y_k -parameter is connected with x_k -vector by a linear correlation

$$\mathbf{y}_k = G_k \mathbf{x}_k + H_k V_k + \eta_k, \quad k = 1, 2, \dots \quad (2.2)$$

Here ξ_k, η_k are independent Gaussian sequences, at that

$$M \xi_k = M \eta_k = 0, \quad M \xi_k \xi_i' = Q_k \delta_{ki}, \quad M \eta_k \eta_i' = R_k \delta_{ki},$$

where M is the expectation, the prime ($'$) denotes the transpose, δ_{ki} is the Kronecker symbol, Q_k and R_k are the assigned $n \times n$ and $r \times r$ positively defined covariations matrices. The $u_k \in U_k, v_k \in V_k$ determinative influences are not known a priori.

Let us assume that the (2.1) system starting point x_0 is a Gaussian vector independent of ξ_k, η_k with the known positively defined covariation matrix

$$M(x_0 - Mx_0)(x_0 - Mx_0)' = P_0,$$

but with the a priori unknown average value $Mx_0 \in X_0$, the known convex compact from \mathbb{R}^n . The matrices A_k, B_k, C_k, G_k, H_k of corresponding measure and the convex compacts U_k, V_k are assumed to be known.

On the known realizations of observations $y_N(\cdot) = \{y_1, \dots, y_N\}$ it is necessary for each $N \geq 1$ to find the estimate $x^*(\cdot) = x_N^*(y_N(\cdot))$, being the Tchebyshev center of the set [6]

$$\bar{X}_N = \{\bar{x}_N = M[x_N | y_N(\cdot), \xi_N(\cdot)] : \xi_N \in \Xi_n\},$$

where $M[\cdot | y_N(\cdot), \xi_N(\cdot)]$ is the operator of the conditional expectation taken under fixed $\zeta_n(\cdot)$, $\bar{x}_N = M[x_N | y_N(\cdot), \zeta_N(\cdot)]$ is a solution of the known problem of the optimal average quadratic linear filtration, $\Xi_N = \{\zeta_N(\cdot) : u_j \in U_j, j = \overline{0, N-1}; v_j \in V_j, j = \overline{1, N}; Mx_0 \in \bar{X}_0\}$.

3. Minimax filter determinative equation

The solution of the problem formulation includes the description of \bar{X}_N sets change dynamics, determining the estimate x_N^* of the vector x_N . For the sets \bar{X}_k , starting from the known \bar{X}_0 there are [6, 8, 15] equations

$$\left. \begin{aligned} \bar{X}_{k+1} &= \hat{X}_{k+1} + \Lambda_{k+1} y_{k+1}; \\ \hat{X}_{k+1} &= \bar{A}_k \bar{X}_k + W_{k+1}; \\ W_{k+1} &= \bar{B}_k U_k + \bar{H}_{k+1} (-V_{k+1}). \end{aligned} \right\} \quad (3.1)$$

Here matrices $\bar{A}_k, \bar{B}_k, \bar{H}_{k+1}, \Lambda_{k+1}$ are of the form $\bar{A}_k = F_k A_k, \bar{B}_k = F_k B_k, \bar{H}_k = \Lambda_{k+1} H_k, F_k = I - \Lambda_{k+1} G_{k+1}, \Lambda_{k+1} = P_{k+1} G_{k+1} R_{k+1}^{-1}, I$ is a single $n \times n$ matrix, the matrix of covariations P_k satisfies the recurrent correlations of Riccati type [6].

In (3.1) the set sum is understood in the sense of Minkovsky [10] and finding it in real time (at the rate of obtaining the measurement results y_{k+1}) is difficult.

Note that the set W_k is completely identified by the a priori data. Operations over the sets [6-9] are performed with the help of the basic functions. This poses very strict requirements for the use of high speed computers realizing the algorithm of minimax estimation in real-time systems and, therefore, restricts their utilization except for some cases. Thus, for symmetrical sets \bar{X}_0, U_k, V_k equation for [6, 7] estimation has the form

$$x_{k+1}^* = \bar{A}_k x_k^* + \bar{B}_k u_k^* - \bar{H}_{k+1} v_{k+1}^* + \Lambda_{k+1} y_{k+1}, \quad k = 0, 1, \dots, \quad (3.2)$$

where u_k^*, v_{k+1}^* are the Tchebyshev centers of the sets U_k, V_{k+1} , respectively.

Let us consider the minimax algorithm development under the non-symmetrical sets \bar{X}_0, U_k, V_k . Let \hat{x}_{k+1}^* be the center of the set \hat{X}_{k+1} , then, for the estimation of x_{k+1}^* from (3.1), we directly get

$$x_{k+1}^* = \hat{x}_{k+1}^* + \Lambda_{k+1} y_{k+1}, \quad (3.3)$$

where \hat{x}_{k+1}^* can be calculated before the measurement results are obtained. In this case the estimation problem is reduced to developing the set \hat{X}_{k+1} and finding its center \hat{x}_{k+1}^* . If, for the period of time $[k, k+1]$, these problems can not be solved let us present the system (3.1) in the form of [6, 15]

$$\left. \begin{aligned} \bar{X}_{k+1} &= \check{X}_{k+1} + L_{k+1}, & k = 0, 1, \dots; \\ L_{k+1} &= \bar{A}_k L_k + \Lambda_{k+1} y_{k+1}, & L_0 = 0; \\ \check{X}_{k+1} &= \bar{A}_k \check{X}_{k+1} + W_{k+1}, & \check{X}_0 = \bar{X}_0. \end{aligned} \right\} \quad (3.4)$$

From (3.4) it follows that the set \check{X}_{k+1} is completely specified by the a priori data. And we get the informational set \bar{X}_{k+1} by displacing the set \check{X}_{k+1} on the vector L_{k+1} , which is specified above the measurement results. It is obvious that this statement is also true for the Tchebyshev centers $x_{k+1}^*, \check{x}_{k+1}^*$ of the sets $\bar{X}_{k+1}, \check{X}_{k+1}$, respectively

$$x_{k+1}^* = \check{x}_{k+1}^* + L_{k+1}. \quad (3.5)$$

Formulation of the results:

THEOREM 3.1. If system (2.1) is given, calculating equalities (2.2) and execute supposition p. 310. Then follows the optimal estimation of the expression correction (3.5), and for informational set, the true recurrent correlation (3.4).

The advantages of the presentation of the filter in the form of (3.4), (3.5) lie in the fact that the sequences of the sets \tilde{X}_i ($i = 0, 1 \dots$) and their centers \tilde{x}_i^* ($i = 0, 1 \dots$) do not depend on the signal realized and are completely identified by the a priori data. Consequently, from (3.4), the sequences \tilde{X}_i , \tilde{x}_i^* ($i = 0, 1 \dots$) can be determined before designing the filter, and thus, the volume of calculations being performed in real time can be significantly reduced. The sequence of centers is to be entered into the computer, by which the filter will be realized. Then the realization of the filter functioning in real-time is reduced to the calculation of $A_k L_k$, and before obtaining the results of the next calculation y_{k+1} we shall get the estimate x_{k+1}^* directly from (3.5).

Thus, to realize algorithms (3.1)–(3.5), the function development of the sets \hat{X}_k or \tilde{X}_k and their centers must be found. The solution of these problems we shall consider later.

4. Design of the sets \tilde{X}_k and finding their Tchebyshev centers

To realize algorithms (3.1), (3.3), and functions (3.4), (3.6), the sets \hat{X}_{k+1} and \tilde{X}_{k+1} are to be developed as well as their Tchebyshev centers are to be found. Consider only the development of the set \tilde{X}_{k+1} , since the development of \hat{X}_{k+1} is similar. Let us present the sets \bar{X}_0, U_k, V_k ($k = 0, 1 \dots$) in practical calculations in the form of polyhedra or ellipsoids.

The set \tilde{X}_k is a Minkovsky-type sum of the sets $\bar{A}_k \tilde{X}_k, \bar{B}_k U_k, \bar{H}_{k+1} V_{k+1}$. The ways of development of the set sum for convex polyhedra are given in [10, 15], for the ellipsoid in [12]. If the polyhedra are defined by their apexes the problem is reduced to designing a convex cover [10, 15, 16]. The number of the vertex points of the polyhedron \tilde{X}_{k+1} grows enormously with the increase of k and the initial polyhedron is to be approximated by the one with less apexes or by an ellipsoid.

The Tchebyshev center of the set X_k under the known vertex points x_l ($l = \overline{1, L}$), is the point x^* for which it is true that

$$\min_{z \in \tilde{X}_k} \max_{l = \overline{1, L}} \|x_l - z\|^2 = \|x_l - x^*\|^2, \quad (4.1)$$

where x_l ($l = \overline{1, m}$) are the extreme vertex points. From (4.1) it follows that the point x^* is the center of the surrounded hypersphere of the minimal radius $\delta_k = \|x_l - x^*\|$ going through the extreme vertex points. This property allows us to develop a recurrent algorithm for finding x^* and δ_k .

5. Example

Let us consider the model (2.1), (2.2) employed in the navigational and radio-locational information processing systems, where the matrices and the vectors are of the form

$$A_k = \begin{bmatrix} 1 & 10 \\ 0 & 1 \end{bmatrix}, B_k = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, G'_k = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, P_0 = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 0.11 \end{bmatrix},$$

$C_k = B_k, H_k = 1, Q_k = 9 \cdot 10^{-2}, R_k = 1, x_k \in \mathbb{R}^2; u_k, y_k, v_k, \xi_k, \eta_k \in \mathbb{R}^1.$

The set \bar{X}_0 (Fig. 1) is defined by its vertex points $x^{(1)} = [1, 2]'$, $x^{(2)} = [1, -1]'$, $x^{(3)} = [-1, -1]'$. The sets U_k, V_k are the segments of the form $U_k \in [-0.913; 0.913]$, $V_k \in [-1.075; 1.075]$. To simplify the calculations of the matrices Λ_k, \bar{A}_k it was assumed that all k are equal to

$$\Lambda_4 = \begin{bmatrix} 0.928 \\ 0.0804 \end{bmatrix}, \text{ and } \bar{A}_4 = \begin{bmatrix} 0.007 & 0.7 \\ -0.08 & 0.2 \end{bmatrix},$$

and $x_k \equiv 0, V_k \equiv 1.$

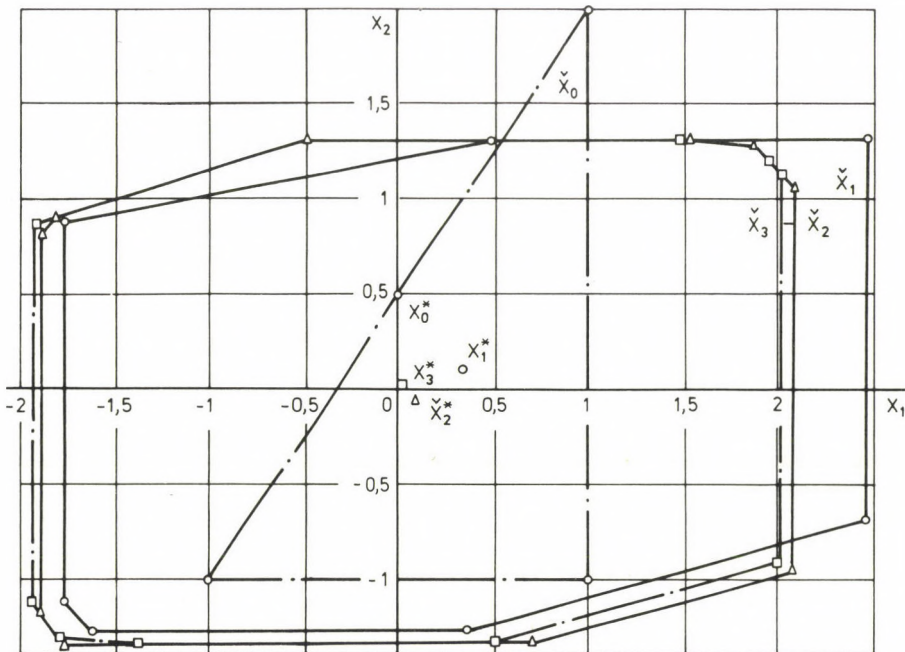


Fig. 1.

The sets \check{X}_k ($k = \overline{0,3}$) and their centers \check{x}_k^* are given on Fig. 1. In this example both the influence of the set \check{X}_0 on \check{X}_3 and the difference between \check{X}_2 and \check{X}_3 are not relatively large, although the number of the vertex points in the set \check{X}_k increases with k . The set \check{X}_k , according to (3.4), is obtained by shifting the sets \check{X}_k on the vector L_k . In Table 1 the centers \check{x}_k^* of the sets \check{X}_k , one of the realizations of measurement y_k , vectors L_k and their coordinate estimates $x_k^* = \check{x}_k^* + L_k$ ($k = \overline{0,3}$) are shown.

Table 1

k	0	1	2	3
\check{x}_k^*	0	0,35	0,1	0,01
	0,5	0,11	-0,05	0,015
y_k	-	2,38	0,23	0,52
L_k	0	2,209	0,501	0,433
	0	0,191	-0,12	-0,022
x_k^*	0	2,559	0,601	0,443
	0,5	0,301	-0,17	-0,07

6. Conclusions

An algorithm of minimax estimation has been developed. Its peculiar feature is that, on the basis of splitting the a priori and a posteriori data, the most labor consuming, as far as calculations are concerned, operations of designing the sets \check{X}_k , finding their Tchebyshev centers are performed by means of the a priori data beforehand (in the process of designing the filter, before the results of the next measurement are obtained). This enables us to realize the algorithm in real-time.

References

1. Krasovsky, N. N., Theory of motion control. Moscow, Nauka, 1968 (in Russian).
2. Krasovsky, N. N., Game-problems on motions encounter. Moscow, Nauka, 1970 (in Russian).
3. Kirichenko, N. F., Slabospitsky, A. S., Minimax filter in problems of state estimation, parameters identification and images recognition. Kibernetika i vychislitel'naja tekhnika. Kiev, 1985, No. 65 (in Russian).
4. Karlov, V. I., Krasilshchikov, M. N., Malyshev, V. V., Observation process control under statistical uncertainty conditions. (Review). Izvestiya Akademii Nauk SSSR. Tekhn. kibernetika. 1989, No. 2 (in Russian).
5. Kurzhan'sky, A. B., Control and observation under uncertainty. Moscow, Nauka, 1977 (in Russian).

6. Kats, I. Ya., Kurzhansky, A. B., Minimax multistage filtering in statistically uncertain situations. *Avtomatika i telemekhanika*, 1978, No. 11 (in Russian).
7. Kats, I. Ya., Minimax stochastic estimation problems in multistage systems. *Otsenivaniye v usloviyah neopredelyonnosty*. Sverdlovsk. The Ural Scientific Center of the USSR Academy of Sciences, 1982 (in Russian).
8. Kats, I. Ya., Asymptotic properties of information sets in problem of minimax stochastic filtering. Evolution systems in estimation problems. Sverdlovsk. The Ural Scientific Center of the USSR Academy of Sciences, 1985 (in Russian).
9. Koshcheyev, A. S., Kurzhansky, A. B., Adaptive estimation of multistage systems evolution under uncertainty. *Izvestiya Akademii Nauk SSSR. Tekhn. kibernetika*, 1983, No. 3 (in Russian).
10. Kuntsevitch, V. M., Determination of state vectors guaranteed estimates and linear dynamic systems parameters under restricted jerks. *Dokl. Akad. Nauk SSSR*, 1986, v. 288, No. 3 (in Russian).
11. Boguslavsky, I. A., Filtering and control applied problems. Moscow, Nauka, 1983 (in Russian).
12. Tchernousko, F. L., On ellipsoid equations which approximate attainability areas. *Problems of Control and Information Theory*, 1983, v. 12, No. 3
13. Anan'ev, B. I., Shiryayev, V. I., Determination of the worst signals in guaranteed estimation problems. *Avtomatika i telemekhanika*, 1987, No. 3, (in Russian).
14. Shiryayev, V. I., Minimax methods of observation systems synthesis in statistically uncertain situations. Optimization methods and their application. Proceedings of International school-seminar. Irkutsk, SEI, 1989 (in Russian).
15. Shiryayev, V. I., Minimax estimation of linear dynamic systems state with discrete time. *Izvestiya vuzov SSSR. Priborostroyeniye*, 1990, v. 33, No. 7 (in Russian).
16. Tchernykh, O. L., Convex spreadsheet design of ultimate set under approximated calculations. *Zhurnal vychnislitelnoy matematiki & matematicheskoy tekhniki* (The computational mathematics & mathematical physics journal). 1988, v. 28, No. 9.

Минимаксная фильтрация в реальном времени многшаговых систем

В. И. ШИРЯЕВ

(Челябинск)

Рассматривается минимаксная задача оценивания фазового состояния для линейных многшаговых систем вида (2.1), подверженных воздействию как случайных, так и неопределенных возмущений. Процесс наблюдения сопровождается помехами как случайного, так и неопределенного характера. Случайные возмущения и помехи являются независимыми гауссовскими последовательностями с нулевым средним и известными матрицами ковариаций. Относительно возмущений неопределенного характера известны только выпуклые компакты, которым они принадлежат.

За минимаксную оценку принимается чебышевский центр информационного множества, удовлетворяющий соотношению (3.3). На основе разделения априорных и апостериорных данных для информационных множеств получены соотношения (3.1),

(3.4). Это позволяет произвести основной объем вычислений, связанный с построением информационных множеств заранее, до поступления результатов очередного измерения как в задаче фильтрации, так и в задаче сглаживания. Получены оценки для чебышевских радиусов множеств, что позволяет оценить и ошибки оценивания по априорным данным.

Рассмотрены способы построения информационного множества, которое является суммой по Минковскому составляющих множеств в задаче фильтрации и геометрической разностью множеств в задаче сглаживания. Выпуклые компакты, определяющие геометрические ограничения, предлагается аппроксимировать выпуклыми многогранниками при задании многогранников угловыми точками. Чебышевский центр многогранника является центром описывающей его гиперсферы минимаксного радиуса. Это свойство позволяет построить алгоритм нахождения чебышевских центра и радиуса многогранника.

Предлагаемый подход позволяет построить минимаксные алгоритмы оценивания для систем с геометрическими ограничениями, реализуемые в реальном времени.

В. И. Ширяев

Челябинский политехнический институт
СССР, 454 080 Челябинск, пр. Ленина, 76

VISCOSITY SOLUTIONS OF THE BELLMAN EQUATION ON AN ATTAINABLE SET

HITOSHI ISHII, JOSE-LUIS MENALDI, LESZEK ZAREMBA

(Tokio, Detroit, St. John's)

(Received February 8, 1991)

By an appropriate modification of the viscosity solution concept, we introduce a notion solution of a PDE that is applicable, among others, to the Bellman equation and first general classes of optimal control problems with the only restriction on a payoff functional that the stopping time is bounded by a fixed number T . We consider this PDE on the attainable set from Ω_0 , a set of given initial conditions. We prove both existence and uniqueness results for optimal control problems. The approach is illustrated with several examples and comments.

1. Introduction

Consider an object whose dynamics are described by a system of differential equations

$$\dot{x}(t) = f(t, x(t)), \lambda(t), \quad (t, x) \in [0, T] \times \mathbf{R}^n, \quad \lambda(t) \in U(t), \quad x(t_0) = x_0. \quad (1)$$

The values of each piecewise continuous control $\lambda(\cdot)$ are selected by an agent who tries to make the cost

$$C = C[x(\cdot, t_0, x_0, \lambda(\cdot))] = g(\tau, x(\tau)) + \int_{t_0}^{\tau} h(t, x(t), \lambda(t)) dt \quad (2)$$

of transferring the object from a given initial state $(t_0, x_0) \in \Omega_0$ to a given terminal set $\Gamma \subset \mathbf{R}^{n+1}$ as small as possible; here $x(t, t_0, x_0, \lambda(\cdot))$ stands for the value of the unique trajectory $x(\cdot, t_0, x_0, \lambda(\cdot))$ at time $t \geq t_0$ that results from the control $\lambda(\cdot)$ and the initial point x_0 at t_0 . The stopping time is the first moment $t \geq t_0$ for which $x(t, t_0, x_0, \lambda(\cdot)) \in \Gamma$. We consider the optimal cost function

$$u(t_0, x_0) = \inf\{C[x(\cdot, t_0, x_0, \lambda(\cdot))] : \lambda(\cdot) \in \Lambda\} \quad (3)$$

on Ω , the attainable set form Ω_0 , i.e., on the set

$$\Omega = \{(\bar{t}, \bar{x}) : \bar{x} = x(\bar{t}, t_0, x_0, \lambda(\cdot)), (t_0, x_0) \in \Omega_0, t_0 \leq \bar{t} \leq \tau, \lambda(\cdot) \in \Lambda\}, \quad (4)$$

with Λ standing for the space of all piecewise continuous controls $\lambda(\cdot)$. It follows from this definition that Ω is invariant under the flow generated by equation (1), i.e., $\Omega = \{(\bar{t}, \bar{x}) : \bar{x} = x(\bar{t}, t, x, \lambda(\cdot)), (t, x) \in \Omega, \lambda(\cdot) \in \Lambda\}$. It has been known [6] that $u(t, x) : \Omega \rightarrow \mathbb{R}$ satisfies the Bellman equation provided it is differentiable. Since usually this is not the case, the natural question arises how to overtake this difficulty. Several authors worked out around this problem in the 60's and the 70's [13, 16, 17, 18, 24, 29].

From the general theory of PDE point of view, this problem was practically solved by M. G. Crandall and P. L. Lions who did a pioneering work (see, for example, [8–11, 25–28]) by developing a new approach based on their viscosity solution concept, and by their continuators; see, e.g. Ishii [19–23] and Sauganidis [30–32]. Independently of Crandall and Lions, a different approach to the problem of solving the Bellman equation has been found by Subbotin [34–38] who proved in [37] that his solution concept is equivalent to that of Crandall and Lions, when applied to fixed time duration problems; an interesting feature of viscosity solutions was discovered in [38] where Subbotin pointed out a close relationship between differential games and viscosity solutions.

The viscosity solution theory was applied in [5] to prove that the value of a differential game with fixed time duration is the only viscosity solution to the Isaacs equation. The natural question arises how to extend this result to control problems and differential games with a variable time of duration. A few authors have already started to work around this problem. It has been proved, for example, that given a partial differential equation $w_t + H(t, x, Dw) = 0$ with $w(0, x) = g(x)$, one can find a differential game with the property that its upper value is the viscosity solution to this PDE [15]; see also [14], where a similar result was obtained. In the latter paper the authors were able to apply the viscosity solution concept to differential games with variable times of duration. However, the terminal set was assumed to coincide with the boundary of Ω , what considerably restricts the applicability of these results to differential games.

Some other results in this direction have been obtained quite recently. For example, time-optimal control problems have been studied in [33] (a linear case) and [2] (a nonlinear case). So-called generalized time-optimal control problems, as well as generalized pursuit-evasion games ($h(t, x, \lambda) \geq h_0 > 0$) have been studied in the viscosity solution framework by Bardi and Soravia [3]; see also another paper by the same authors [4], and a paper by Berkovitz [7], both dealing with differential games.

In this paper we introduce a modification of the viscosity solution concept (Definitions 2.1, 2.2) that is suitable for a broad class of nonlinear problems occurring in optimal control theory with the only restriction on a payoff functional of the form (2) that the stopping time τ is bounded from above by a given number T .

The main results of Section 3 are Theorem 3.1 (an existence result) and Theorem 3.2 (a uniqueness result) referring to the partial differential equation (12) that encompasses the Bellman equation, defined on the attainable set Ω . We do not assume Ω is open or closed (Comment 3.1); also, the continuity of a viscosity solution is not required (Example 3.1) and the terminal set Γ need not be the boundary of Ω (Example 2.1) or even its part (Example 2.2). A preliminary version of these results was announced in [39].

2. Assumptions and the viscosity solution concept

In Sections 2 and 3 we need the following assumptions; below $Q = [0, T] \times \mathbb{R}^n$.

The multifunction $U : [0, T] \rightarrow \mathbb{R}^k$ is continuous (in the Hausdorff metrics sense), all closed-valued sets $U(t)$ lie in a fixed ball $U \subset \mathbb{R}^k$ and, for each $\bar{\lambda} \in U(\bar{t})$, there is a selection $\lambda(t)$ from $U(t)$ that is continuous at \bar{t} and satisfies $\lambda(\bar{t}) = \bar{\lambda}$. (5)

The function $f : Q \times U \rightarrow \mathbb{R}^n$, $g : Q \rightarrow \mathbb{R}$ and $h : Q \times U \rightarrow \mathbb{R}$ are continuous. (6)

The function $f(t, x, \lambda)$ is Lipschitz in x , i.e. for all $x \in \mathbb{R}^n$, $\bar{x} \in \mathbb{R}^n$, $\lambda \in U(t)$ one has

$$\|f(t, x, \lambda) - f(t, \bar{x}, \lambda)\| \leq k(t)\|x - \bar{x}\|, \quad \int_0^T k(t) dt < \infty. \tag{7}$$

Γ is a closed subset of \mathbb{R}^{n+1} . (8)

There exists a $T > 0$ such that, for each $x(\cdot, t_0, x_0, \lambda(\cdot))$ with $(t_0, x_0) \in \Omega_0$, one has $\tau((\cdot, t_0, x_0, \lambda(\cdot))) \leq T$ and $\inf\{t : (t, x) \in \Omega_0\} = 0$. (9)

It is well known that under assumptions (5)–(7) equation (1) has a unique solution on $[t_0, T]$ for each $(t_0, x_0) \in Q$ and any control function $\lambda(\cdot) \in \Lambda$. Besides [12, pp. 14–16], the solutions of equation (1) starting from any bounded domain remain uniformly bounded and equicontinuous, which implies the optimal cost function is locally bounded. In particular, it yields the following two properties:

- (*) For each $(\bar{t}, \bar{x}) \in \Omega$ there is a constants $K > 0$ such that $|u(t, x)| \leq K$ on the set $\{(t, x) \in \Omega : \text{dist}[(t, x), (\bar{t}, \bar{x})] \leq 1\}$; and

(**) For each $\delta > 0$ there is an $\varepsilon > 0$ such that if $(t, x) \in \Omega$, $\text{dist}[(t, x), \Gamma] \geq \delta$ and $\|(t, x)\| \leq 1/\delta$ then $\text{dist}[(t', x(t', t, x, \lambda(\cdot))), \Gamma] > \delta/2$ for each control function $\lambda(\cdot) \in \Lambda$ and $t \leq t' \leq t + \varepsilon$.

Property (**) will be used in the proofs of Lemma 3.1 and Theorem 3.1. Let us note that condition (5) implies [1, p. 53] the continuity of the Bellman function

$$H(t, x, p) = \inf_{\lambda \in U(t)} \{f(t, x, \lambda)p + h(t, x, \lambda)\}. \tag{10}$$

Observe also that because of the continuous dependence of solutions $x(\cdot, t_0, x_0, \lambda(\cdot))$ on (t_0, x_0) , the set

$$\Omega_T = \{(t, x) \in Q : \tau(x(\cdot, t, x, \lambda(\cdot))) \leq T, \lambda(\cdot) \in \Lambda, 0 \leq t \leq T\}$$

is closed; actually, it is the largest set of initial conditions satisfying (9).

The function $w^* : \bar{\Omega} \rightarrow \mathbf{R}$ satisfying

$$w^*(t, x) = \limsup_{\varepsilon \rightarrow 0} \{w(t', x') : (t', x') \in \Omega, |t' - t| < \varepsilon, \|x' - x\| < \varepsilon\} < \infty \tag{11}$$

is said to be the upper semi-continuous (usc) envelope of a function $w : \Omega \rightarrow \mathbf{R}$. In a similar fashion we define the lsc envelope $w_*(\cdot)$ of a function $w(\cdot)$. Let us note that property (*) implies that the optimal cost function has both the usc and lsc envelope.

DEFINITION 2.1a. Let $H : Q \times \mathbf{R}^n \rightarrow \mathbf{R}$ be a locally bounded function. A function $w : \Omega \rightarrow \mathbf{R}$ is said to be a *viscosity subsolution* of the equation

$$\begin{aligned} w_t(t, x) + H(t, x, w_x(t, x)) &= 0, & (t, x) \in \bar{\Omega} \setminus \Gamma, \\ w(t, x) &= g(t, x), & (t, x) \in \Gamma \subset \bar{\Omega} \end{aligned} \tag{12}$$

if $w(t, x) \leq g(t, x)$ on Γ and, for each $C^1(\bar{\Omega} \setminus \Gamma)$ function φ , one has $\varphi_t(\bar{t}, \bar{x}) + H^*(\bar{t}, \bar{x}, \varphi_x(\bar{t}, \bar{x})) \geq 0$ at each point $(\bar{t}, \bar{x}) \in \bar{\Omega} \setminus \Gamma$ which is a local maximum of the function $w^*(t, x) - \varphi(t, x) : \bar{\Omega} \setminus \Gamma \rightarrow \mathbf{R}$, where $H^*(t, x, p) = \limsup_{\varepsilon \rightarrow 0} \{H(s, y, q) : (s, y) \in \Omega, |s - t| + \|y - x\| + \|q - p\| \leq \varepsilon\}$.

DEFINITION 2.1b. With $H(t, x, p)$ being as previously, a function $w(\cdot)$ is said to be a *viscosity supersolution* of equation (12) if $w(t, x) \geq g(t, x)$ on Γ and, for each $C^1(\bar{\Omega} \setminus \Gamma)$ function φ , one has $\varphi_t(\bar{t}, \bar{x}) + H_*(\bar{t}, \bar{x}, \varphi_x(\bar{t}, \bar{x})) \leq 0$ at each point $(\bar{t}, \bar{x}) \in \bar{\Omega} \setminus \Gamma$ which is a local minimum of $w_*(t, x) - \varphi(t, x) : \bar{\Omega} \setminus \Gamma \rightarrow \mathbf{R}$ with $H_*(t, x, p) = \liminf_{\varepsilon \rightarrow 0} \{H(s, y, q) : (s, y) \in \Omega, |s - t| + \|y - x\| + \|q - p\| \leq \varepsilon\}$.

DEFINITION 2.2. A function $w : \bar{\Omega} \rightarrow \mathbf{R}$ is said to be a *viscosity solution* of (12) if $w(\cdot)$ is both a viscosity subsolution and a viscosity supersolution of (12).

The assumption below will be used in our uniqueness theorem only.

- (i) $H \in C(\bar{\Omega} \times \mathbb{R}^n)$;
- (ii) for each real $R > 0$ there is a function $w_R(\cdot) \in C[0, \infty)$ such that $w_R(0) = 0$ and for each p, q with $\|p\| \leq R, \|q\| \leq R$ the “Hamiltonian” H appearing in (12) satisfies the inequality $|H(t, x, p + q) - H(t, x, p)| \leq w_R(\|q\|)$;
- (iii) there is a function $m(\cdot) \in C[0, \infty), m(0) = 0$, for which $|H(t, x, p) - H(t, y, p)| \leq m(\|x - y\|(1 + \|q\|))$.

Observe that each local extremum of $w^* - \varphi$ ($w_* - \varphi$) is always a global extremum of $w^* - \varphi'$ ($w_* - \varphi'$ with, possible, another $C^1(\bar{\Omega} \setminus \Gamma)$ function φ' (and, of course, vice versa).

Remark 2.1 Although our setting has one basic restriction (the Hamiltonian $H(t, x, w, w_x)$ does not depend in our paper on w , as in some other publications, it does cover the basic equations in control theory and differential games, such as the Bellman and Isaacs equations. In our approach the terminal set Γ may be either a subset of $\partial\Omega$, the boundary of Ω (Example 2.1) or not (Example 2.2).

Example 2.1 Consider an optimal control problem given by $\dot{x} = u, x(0) = 0 \in \Omega_0 = \{0\}, u \in U = \{(u_1, u_2) : u_1^2 + u_2^2 \leq 1\}$ and $\Gamma = \{(t, x_1, x_2) : t = 1, x_1^2 + x_2^2 \leq 1\}$ with any payoff functional. It is obvious the attainable set Ω is the truncated cone with vertex $0 \in \mathbb{R}^3$ and the base Γ , therefore, $\Gamma \subset \partial\Omega$, although $\Gamma \neq \partial\Omega$.

Example 2.2 (a cone with an interior “stick”). Let everything will be the same as in Example 2.1 except for the terminal set Γ^1 that is now defined as follows; $\Gamma^1 = \Gamma \cup [\frac{1}{2}d, d]$ with $d = (0, 0, 1)$. It is obvious that each interior point of the segment $[\frac{1}{2}d, d]$ does not belong to $\partial\Omega$ (it does belong to the interior of Ω), so Γ^1 is not contained in $\partial\Omega$. If we set $g = 0, h = 1$ then the value function $u(t, x) = \text{dist}[(t, x), \Gamma^1], (t, x) \in \Omega$, and the resulting Hamiltonian (cf. (10)) is given by the formula $H(t, x, q) = 1 - \|q\|$, which implies condition (13) is satisfied. As we shall see later (Theorem 3.3), this value function is the unique bounded, continuous solution of the Bellman equation $w_t(t, x) + 1 - \|(\partial w / \partial x)(t, x)\| = 0$ with the boundary condition $w(t, x) = 0$ on Γ^1 .

Remark 2.2. We intentionally consider extrema on $\bar{\Omega} \setminus \Gamma$ rather than on $\bar{\Omega}$ or Ω because, otherwise, we would have no continuous viscosity solution if Ω were a compact set and $H \in C(\bar{\Omega} \times \mathbb{R}^n)$.

Proof. Assume on the contrary that a continuous solution $w : \bar{\Omega} \rightarrow \mathbb{R}$ exists. Set $A = \max\{|H(t, x, 0)| : (t, x) \in \bar{\Omega}\}$ and consider the “test” function $\Phi : \bar{\Omega} \times \bar{\Omega} \rightarrow \mathbb{R}$ given by

$$\Phi(t, x, s, z) = w(t, x) - w(s, z) + \left(A + \frac{1}{2}\right)(t + s) \tag{14}$$

which must attain the global maximum at some point (t_0, x_0, s_0, z_0) . Therefore, the function $w - \varphi = w^* - \varphi : \bar{\Omega} \rightarrow \mathbb{R}$ with the $C^\infty(\bar{\Omega})$ function $\varphi(t, x) = w(s_0, z_0) - (A + 1/2)(t + s_0)$ attains the maximum in $\bar{\Omega}$ at (t_0, x_0) . According to Definition 2.1, we derive the inequality: $-(A + 1/2) + H(t_0, x_0, 0) \geq 0$. On the other hand, the function $w - \psi = w_* - \psi : \bar{\Omega} \rightarrow \mathbb{R}$ with $\psi(s, z) = w(t_0, x_0) + (A + 1/2)(t_0 + s)$ attains the minimum in $\bar{\Omega}$ at (s_0, z_0) , which implies $(A + \frac{1}{2}) + H(s_0, z_0, 0) \leq 0$. Combining the last two inequalities, and taking into account that $|H(t, x, 0) - H(s, z, 0)| \leq 2A$, we easily obtain a contradiction $(2A + 1 \leq 2A)$.

3. Basic results

We start with a lemma, known as the optimality principle of dynamic programming, that was originally formulated and proved by R. Bellman [6] under strong regularity assumptions. Given $(t, x) \in \Omega \setminus \Gamma$ one can always find a δ such that $\|(t, x)\| \leq 1/\delta$ and $\text{dist}[(t, x), \Gamma] \geq \delta$; denote by $\delta_{(t,x)}$ the largest δ with the properties above.

Lemma 3.1. If conditions (5)–(8) hold then, for each $(t, x) \in \Omega \setminus \Gamma$, the equality

$$\inf_{\lambda(\cdot) \in \Lambda} \{u(t + \varepsilon, x(t + \varepsilon, t, x, \lambda(\cdot))) + \int_t^{t+\varepsilon} h(s, x(s), \lambda(s)) ds\} = u(t, x) \tag{15}$$

is satisfied for all positive $\varepsilon \leq \bar{\varepsilon}$, where $\bar{\varepsilon}$ depends on $\delta_{(t,x)} = \delta$, as specified in property (**).

Proof. For $0 \leq \varepsilon \leq \bar{\varepsilon}$, let $\Lambda_\varepsilon(\Lambda^\varepsilon)$ be the space of all portions of control functions $\lambda(\cdot) \in \Lambda$ on the segment $[t, t + \varepsilon]$ (resp. $[t + \varepsilon, T]$). By $\lambda_\varepsilon(\cdot), \lambda^\varepsilon(\cdot)$ we denote generic elements of Λ_ε (resp. Λ^ε) so that one can write down any control $\lambda(\cdot)$ in the form $\lambda(\cdot) = (\lambda_\varepsilon(\cdot), \lambda^\varepsilon(\cdot))$. We thus have $u(t, x) = \inf\{C(x(\cdot, t, x, \lambda_\varepsilon(\cdot), \lambda^\varepsilon(\cdot))) : \lambda_\varepsilon(\cdot) \in \Lambda_\varepsilon, \lambda^\varepsilon(\cdot) \in \Lambda^\varepsilon\}$ and, consequently, (15) because the right hand side of the last equality equals

$$\inf_{\lambda_\varepsilon(\cdot)} \{u(t + \varepsilon, x[t + \varepsilon, t, x, \lambda_\varepsilon(\cdot)]) + \int_t^{t+\varepsilon} h(s, x(s), \lambda_\varepsilon(s)) ds,$$

where, clearly, $\lambda_\varepsilon(\cdot)$ may be replaced by $\lambda(\cdot) \in \Lambda$.

THEOREM 3.1. If conditions (5)–(8) hold then $u(t, x)$ is a solution of equation (12) with $H(t, x, p)$ given by (10).

Proof. We shall show that $u(\cdot)$ is a viscosity subsolution of (12) with $H(t, x, p)$ given by (10). The remaining part of the proof may be carried out analogously. First of all, observe that $u(t, x) = g(t, x)$ for $(t, x) \in \Gamma$. Assume now that $u^* - \varphi : \bar{\Omega} \setminus \Gamma \rightarrow \mathbb{R}$, $\varphi \in C^1(\bar{\Omega} \setminus \Gamma)$, attains its maximum at $(\bar{t}, \bar{x}) \in \bar{\Omega} \setminus \Gamma$.

Consider first the more difficult case when $(\bar{t}, \bar{x}) \notin \Omega$. Since $(u^* - \varphi)(\bar{t}, \bar{x})$ is finite, we may assume $(u^* - \varphi)(\bar{t}, \bar{x}) = 0$ so that, locally in $\Omega \setminus \Gamma$, one $u^* = u \leq \varphi$. By the definition of $u^*(t, x)$, we can choose a sequence $(t_k, x_k) \in \Omega \setminus \Gamma$ such that $(t_k, x_k) \rightarrow (\bar{t}, \bar{x})$ as $k \rightarrow \infty$ and $(u^* - \varphi)(\bar{t}, \bar{x}) < (u - \varphi)(t_k, x_k) + 1/k$ for each natural number k ($k \in \mathbb{N}$ for short). Since, for some $\delta \geq 0$, $\|(t_k, x_k)\| \leq 1/\delta$, $k = 1, 2, \dots$, it follows from property (***) that there is an $\bar{\varepsilon} > 0$ for which $x(t, t_k, x_k, \lambda(\cdot)) \in \Omega \setminus \Gamma$ for each $\lambda(\cdot) \in \Lambda$, each $k \in \mathbb{N}$ and $t_k \leq t \leq t_k + \varepsilon$, $0 \leq \varepsilon \leq \bar{\varepsilon}$. Fix $\bar{\lambda} \in U(\bar{t})$ and, by virtue of (5), choose a control $\lambda(\cdot)$ that is continuous at \bar{t} with $\lambda(\bar{t}) = \bar{\lambda}$. By Lemma 3.1, we have

$$u(t_k, x_k) \leq u(t_k + \varepsilon, x(t_k + \varepsilon)) + \int_{t_k}^{t_k + \varepsilon} h(t, x(t), \lambda(t)) dt,$$

where $x(t) = x(t, t_k, x_k, \lambda(\cdot))$. Using the fact $(u^* - \varphi)$ attains its maximum in $\bar{\Omega} \setminus \Gamma$ at (\bar{t}, \bar{x}) , we easily derive

$$\varphi(t_k, x_k) - \frac{1}{k} < \varphi(t_k + \varepsilon, x(t_k + \varepsilon)) + \int_{t_k}^{t_k + \varepsilon} h(t, x(t), \lambda(t)) dt$$

and consequently

$$-\frac{1}{k} < \int_{t_k}^{t_k + \varepsilon} \varphi(t, x(t)) + \langle \varphi_x(t, x(t)), f(t, x(t), \lambda(t)) \rangle + h(t, x(t), \lambda(t)) dt$$

for $k \in \mathbb{N}$. Letting k go to ∞ and ε go to zero we obtain $\varphi_t(\bar{t}, \bar{x}) + \langle \varphi_x(\bar{t}, \bar{x}), f(\bar{t}, \bar{x}, \bar{\lambda}) \rangle + h(\bar{t}, \bar{x}, \bar{\lambda}) \geq 0$. Since $\bar{\lambda} \in U(\bar{t})$ was arbitrary, we conclude $\varphi_t(\bar{t}, \bar{x}) + H^*(\bar{t}, \bar{x}, \varphi_x(\bar{t}, \bar{x})) \geq 0$ (here $H^*(\cdot) = H(\cdot)$), as required.

In the easy case $(\bar{t}, \bar{x}) \in \Omega \setminus \Gamma$ the inequality $(u - \varphi)(\bar{t}, \bar{x}) \geq (u - \varphi)(\bar{t} + \varepsilon, x(\bar{t} + \varepsilon))$ holds locally for any trajectory $x(\cdot, \bar{t}, \bar{x}, \lambda(\cdot))$ of equation (1) with a control $\lambda(\cdot)$ satisfying $\lambda(\bar{t}) = \bar{\lambda}$. We use Lemma 3.1 again and proceed similarly as in the first part of the proof.

As an illustration of this theorem we present the following example.

Example 3.1. Let $f(t, x, \lambda) = \lambda$, $\lambda \in \{1\}$, $(t_0, x_0) \in \Omega_0 =$ parallelogram $ABCD$, where $A = (1, 1)$, $B = (0, 1)$, $C = (-1, 0)$, $D = (0, 0)$. Set $g : \mathbb{R}^2 \rightarrow \mathbb{R}$, $g(x, t) = x + t$, $h = 0$ and $\Gamma = \{(x, t) : (x, 1), 0 \leq x \leq 1\} \cup \{(0, t) : 0 \leq t \leq 1\}$ (Γ is the union of two segments AB and BD). The optimal cost function $u(x, t)$ is defined on the closed set Ω (here $\Omega = \Omega_0$); note that the adjective optimal is meaningless in this case because the controller has no choice in selecting control functions. It is obvious that

$$u(x, t) = \begin{cases} t - x & (x, t) \in BCD \\ 2 + x - t, & (t, x) \in ABD, \end{cases}$$

where BCD (resp. ABD) stands for the triangle BCD (resp. the triangle ABD without the side BD). Therefore, $u(x, t)$ is C^∞ at each point of $\Omega = ABCD$ except for the points $(0, t)$, $0 \leq t < 1$, where $u(x, t)$ is discontinuous with a "jump" $2 - 2t$. It is easy to see that $0 \leq u(x, t) \leq 2$ on Ω ; also, assumptions (5)–(9) and (13), with $H(t, x, p) = p$, are satisfied. By Theorem 3.1, $u(x, t)$ is a solution of the Bellman equation

$$w_t(t, x) + \frac{\partial w}{\partial x}(t, x) = 0 \quad (t, x) \in \Omega \setminus \Gamma$$

with $w(t, x) = x + t$ on Γ . According to Definition 2.1, if we choose $\varphi(t, x) = u(x, t)$ then we shall obtain $u_t(x, t) + H^*(t, x, u_x(x, t)) = u_t(x, t) + u_x(x, t) = 0$ for $(x, t) \in \Omega \setminus \Gamma$; in this particular case, the Bellman equation is also satisfied on the side AB (which belongs to Γ). Summing up, although the optimal cost function is discontinuous on a part of Γ , it satisfies the Bellman equation in the classical sense at each other point of Ω .

Comment 3.1. We could, of course, remove the side $[AD]$ from $\Omega = ABCD$ and next repeat the whole reasoning, obtaining the same solution $u(t, x)$ on $ABCD$, despite the fact that the original domain was not closed. This circumstance, however, does not cause any trouble because $u(t, x)$ has usc and lsc envelopes (as a bounded function); as a matter of fact, $u^*(x, t) = u_*(x, t) = 2$ for $(x, t) \in [AD]$.

Now, we shall give a uniqueness result. In applications, it practically requires the continuity and boundedness of a viscosity solution to be unique.

THEOREM 3.2. Assume $\Omega \subset Q = [0, T] \times \mathbb{R}^n$ and $v(\cdot)$, $w(\cdot)$ satisfy, respectively, the inequalities $v_t(t, x) + H(t, x, v_x(t, x)) \geq 0$ in $\bar{\Omega} \setminus \Gamma$ and $w_t(t, x) + H(t, x, w_x(t, x)) \leq 0$ in $\bar{\Omega} \setminus \Gamma$ in the viscosity solution sense (Definition 2.1). In addition, assume that conditions (9), (13) hold and there exists $\mu(\cdot) \in C[0, \infty)$, $\mu(0) = 0$, such that

$$v(t, x) - w(t, y) \leq \mu(\|x - y\|) \quad \text{if } (t, x) \in \bar{\Omega} \cap \Gamma \quad \text{or} \quad (t, y) \in \bar{\Omega} \cap \Gamma. \quad (17)$$

Finally, let $v(\cdot)$, $-w(\cdot)$ be bounded from above and upper semi-continuous on $\bar{\Omega}$. Then $v(t, x) \leq w(t, x)$ on $\bar{\Omega}$.

Proof. Suppose $\sup\{(v - w)(t, x) : (t, x) \in \bar{\Omega}\} > 0$. Let $\varepsilon, \alpha, \beta, \gamma \in (0, 1]$ and $\lambda > 0$ be given. Define

$$\bar{\Phi}(t, x, s, y) = v(t, x) - w(s, y) - \varepsilon(T + 1 - t) - \frac{e^{-\lambda t}}{\alpha} \|x - y\|^2 - \frac{1}{\beta}(t - s)^2 - \gamma \|x\|^2$$

and next fix s and γ so small that $\sup \bar{\Phi}(t, x, t, x) : (t, x) \in \bar{\Omega} \geq 0$ and consequently $\sup\{\bar{\Phi}(t, x, s, y) : (t, x) \in \bar{\Omega}, (s, y) \in \bar{\Omega}\} \geq 0$. Since $\bar{\Phi}$ is usc, there must exist a global maximum point of $\bar{\Phi}$ on $\bar{\Omega} \times \bar{\Omega}$. Let $(\bar{t}, \bar{x}, \bar{s}, \bar{y})$ be such a point. Using the boundedness of $v(t, x)$ and $-w(t, x)$, we have

$$0 \leq \bar{\Phi}(\bar{t}, \bar{x}, \bar{s}, \bar{y}) \leq C - \frac{e^{-\lambda \bar{t}}}{\alpha} \|\bar{x} - \bar{y}\|^2 - \frac{1}{\beta}(\bar{t} - \bar{s})^2 - \gamma \|\bar{x}\|^2 \quad (18)$$

for some $C > 0$ so that

$$\|\bar{x} - \bar{y}\| \leq (\alpha C e^{\lambda T})^{\frac{1}{2}}, \quad \|\bar{t} - \bar{s}\| \leq (\beta C)^{\frac{1}{2}}, \quad \gamma \|\bar{x}\| = \gamma^{\frac{1}{2}} (\gamma \|x\|^2)^{\frac{1}{2}} \leq (\gamma C)^{\frac{1}{2}}. \tag{19}$$

Now, let us fix $\lambda > 0$ so large that

$$m(2r) < \varepsilon + \lambda r \text{ for } 0 \leq r \leq C \tag{20}$$

and next $\alpha > 0$ so small that

$$\mu(s) < \varepsilon + \frac{e^{\lambda T}}{\alpha} s^2, \quad 0 \leq s \leq (C e^{\lambda T})^{\frac{1}{2}}. \tag{21}$$

In view of (19) we can choose a sequence of β converging to zero such that the corresponding coordinates \bar{x} , \bar{y} , \bar{t} and \bar{s} are convergent (because \bar{x} and \bar{y} lie in a compact set). Letting $\tilde{x} = \lim \bar{x}$, $\tilde{y} = \lim \bar{y}$ and $\tilde{t} = \lim \bar{t} = \lim \bar{s} = \tilde{s}$, we conclude, using the usc of $\bar{\Phi}$ that $\bar{\Phi}(\tilde{t}, \tilde{x}, \tilde{s}, \tilde{y}) \geq \lim\{\bar{\Phi}(\bar{t}, \bar{x}, \bar{s}, \bar{y}) \geq 0 : \beta \rightarrow 0\}$. It follows from (18), (19) that

$$e^{\frac{\lambda \tilde{t}}{\alpha}} \|\tilde{x} - \tilde{y}\|^2 \leq C, \quad \|\tilde{x} - \tilde{y}\| \leq (C e^{\lambda T})^{\frac{1}{2}}, \quad \gamma \|\tilde{x}\| \leq (\gamma C)^{\frac{1}{2}}.$$

If $(\tilde{t}, \tilde{x}) \in \bar{\Omega} \cap \Gamma$ or $(\tilde{t}, \tilde{y}) \in \bar{\Omega} \cap \Gamma$ then by assumption (17) and inequality (21) we have

$$v(\tilde{t}, \tilde{x}) - w(\tilde{t}, \tilde{y}) \leq \mu(\|\tilde{x} - \tilde{y}\|) < \varepsilon + \frac{e^{\lambda \tilde{t}}}{\alpha} \|\tilde{x} - \tilde{y}\|^2,$$

so that $\bar{\Phi}(\tilde{t}, \tilde{x}, \tilde{t}, \tilde{y}) < 0$, a contradiction with $\bar{\Phi}(\tilde{t}, \tilde{x}, \tilde{t}, \tilde{y}) \geq 0$. We thus see that both (\tilde{t}, \tilde{x}) and (\tilde{t}, \tilde{y}) are in $\bar{\Omega} \setminus \Gamma$. Therefore, almost all (\bar{t}, \bar{x}) and (\bar{s}, \bar{y}) are in $\bar{\Omega} \setminus \Gamma$ and, since $v(t, x)$ and $w(t, x)$ are, respectively, sub- and supersolutions, we have the following two inequalities:

$$-\varepsilon - \frac{\lambda e^{-\lambda \bar{t}}}{\alpha} \|\bar{x} - \bar{y}\|^2 + \frac{2}{\beta} (\bar{t} - \bar{s}) + H(\bar{t}, \bar{x}, \frac{2e^{\lambda \bar{t}}}{\alpha} (\bar{x} - \bar{y}) + 2\gamma \|\bar{x}\|) \geq 0$$

and

$$\frac{2}{\beta} (\bar{t} - \bar{s}) + H(\bar{s}, \bar{y}, \frac{2e^{\lambda \bar{t}}}{\alpha} (\bar{x} - \bar{y})) \geq 0.$$

Subtracting the former inequality from the latter, and next letting β go to zero (\bar{t} and \bar{s} will tend to the same limit), we arrive at

$$\varepsilon + \frac{\lambda e^{-\lambda \bar{t}}}{\alpha} \|\bar{x} - \bar{y}\|^2 \leq w_R(2\gamma \|\bar{x}\|) + m(\|\bar{x} - \bar{y}\| + \frac{2e^{\lambda \bar{t}}}{\alpha} \|\bar{x} - \bar{y}\|^2)$$

with $R = 2 \max(C, C^{\frac{1}{2}})$; clearly, we have made use here of assumption (13). Finally, sending γ to zero and next α to zero, we obtain $\varepsilon + \lambda r \leq m(2r)$ for some $r \in [0, C]$, a contradiction with (20), which completes the proof.

The first part of the theorem below is exactly the content of Theorem 3.1, while the second one follows from Theorem 3.2 by a contradiction argument.

THEOREM 3.3. If assumptions (5)–(8) hold then $u(t, x)$ is a solution of the Bellman equation

$$w_t(t, x) + \inf_{\lambda \in U(t)} \left\{ f(t, x, \lambda) \frac{\partial w}{\partial x}(t, x) + h(t, x, \lambda) \right\} = 0, \quad (t, x) \in \bar{\Omega} \setminus \Gamma \quad (22)$$

with the boundary conditions $w(t, x) = g(t, x)$ on Γ in the sense of Definitions 2.1, 2.2. If, in addition,

- (i) conditions (9), (13) are satisfied,
- (ii) $u(t, x)$ is bounded and continuous on Ω ,
- (iii) $g : \Gamma \rightarrow \mathbf{R}$ is uniformly continuous,

then $u(t, x)$ is the unique solution of the Bellman equation in the class of bounded continuous functions on $\bar{\Omega}$.

References

1. Aubin, J. and Cellina, A., *Differential Inclusions*. Springer-Verlag, Berlin, 1984.
2. Bardi, M., A boundary value problem for the minimum-time function, *SIAM J. Control* **26** (1989), p. 776–785.
3. Bardi, M. and Soravia, P., Hamilton–Jacobi equations with singular boundary conditions on a free boundary and applications to differential games. *Trans. Amer. Math. Soc.* (to appear).
4. Bardi, M. and Soravia, P., A PDE framework for games of pursuit-evasion type. In: *Differential Games and Applications*, ed. Basar, T. and Bernhard, P., Springer-Verlag (to appear).
5. Barron, E., Evans, L. C. and Jensen, R., Viscosity solutions of Isaacs' equations and differential games with Lipschitz controls. *Differential Equations* **53** (1984), p. 213–233.
6. Bellman, R., *Adaptive Control Processes: A Guide Tour*. Princeton University Press, Princeton, 1961.
7. Berkovitz, L. D., Characterizations of the values of differential games. *Appl. Math. Optim.* **17** (1988), pp. 177–183.
8. Crandall, M. G. and Lions, P. L., Viscosity solutions of Hamilton–Jacobi equations. *Trans. Amer. Math. Soc.* **277** (1983), pp. 1–42.
9. Crandall, M. G. and Lions, P. L., Hamilton–Jacobi equations in infinite dimensions, I. Uniqueness of viscosity solutions. *J. Funct. Anal.* **62** (1985), pp. 379–396.
10. Crandall, M. G. and Lions, P. L., Hamilton–Jacobi equations in infinite dimensions, II. Existence of viscosity solutions. *J. Funct. Anal.* **65** (1985), pp. 368–405.
11. Crandall, M. G., Evans, L. C. and Lions, P. L., Some properties of viscosity solutions of Hamilton–Jacobi equations. *Trans. Amer. Math. Soc.* **282** (1984), pp. 487–502.
12. Elliott, R. J. and Kalton, N. J., The Existence of Value in Differential Games, *Memoirs of the AMS, Amer. Math. Soc.*, Providence, R. I., 1972.
13. Elliott, R. J. and Kalton, N. J., Cauchy problems for certain Isaacs–Bellman equations and games of survival. *Trans. Amer. Math. Soc.* **198** (1974), pp. 45–72.

14. *Evans, L. C. and Ishii, H.*, Differential games and nonlinear first order PDE on bounded domains. *Manuscripta Math.* **49** (1984), pp. 109–139.
15. *Evans, L. C. and Souganidis, P. E.*: Differential games and representation formulas for solutions of Hamilton–Jacobi–Isaacs equations. *Indiana Univ. Math. J.* **33** (1984), pp. 773–797.
16. *Friedman, A.*, *Differential Games*. John Wiley, New York, 1971.
17. *Friedman, A.*, *Differential Games*. CBMS Regional Conference in Math. **18**, Amer. Math. Soc., Providence, R. I., 1974.
18. *Fleming, W. H.*, The Chauchy problem for degenerate parabolic equations. *J. Math. Mech.* **19** (1964), pp. 987–1008.
19. *Ishii, H.*, Remarks on existence of viscosity solutions of Hamilton–Jacobi equations. *Bull. Fac. Sci. Engrg. Chuo Univ.* **26** (1983), pp. 5–24.
20. *Ishii, H.*, Uniqueness of unbounded viscosity solution of Hamilton–Jacobi equations. *Indiana Univ. Math. J.* **33** (1984), pp. 721–748.
21. *Ishii, H.*, Hamilton–Jacobi equations with discontinuous Hamiltonians on arbitrary open sets. *Bull. Fac. Sci. Engrg. Chuo Univ.* **28** (1985), pp. 33–77.
22. *Ishii, H.*, Perron's method for Hamilton–Jacobi equations. *Duke Math. J.* **55** (1987), pp. 369–384.
23. *Ishii, H.*, A boundary value problem of the Dirichlet type for Hamilton–Jacobi equation (preprint).
24. *Krasovskii, N. N. and Subbotin, A. I.*, *Positional Differential Games*. Nauka, Moscow, 1974.
25. *Lions, P. L.*, *Generalized Solutions of Hamilton–Jacobi Equations*. Pitman, Boston, 1982.
26. *Lions, P. L.*, Existence results for first order Hamilton–Jacobi equations. *Richerche Math.* **32** (1983), pp. 1–23.
27. *Lions, P. L.*, Neuman type boundary conditions for Hamilton–Jacobi equations. *Duke Math. J.* **52** (1985), pp. 793–820.
28. *Lions, P. L. and Souganidis, P. E.*, Differential games, optimal and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations. *SIAM J. Control Optim.* **23** (1985), pp. 566–583.
29. *Olejnijk, O. A. and Krushkov, S. N.*, Quasi-linear second order parabolic equations with many independent variables. *Uspekhi Mat. Nauk* **16** (1961), pp. 115–155.
30. *Souganidis, P. E.*, Existence of viscosity solutions of Hamilton–Jacobi equations. *J. Differential Eqs.* **56** (1985) pp. 345–390.
31. *Souganidis, P. E.*, Max-min representations and product formulas for the viscosity solutions of Hamilton–Jacobi equations with applications to differential games. *Nonlinear Anal.* **9** (1985), pp. 217–257.
32. *Souganidis, P. E.*, A remark about viscosity solutions of Hamilton–Jacobi equations at the boundary. *Proc. Amer. Math. Soc.* **96** (1986), pp. 323–329.
33. *Staicu, V.*, Minimal time function and viscosity solutions. *J. Optim. Theory Appl.* **60** (1989), pp. 81–91.
34. *Subbotin, A. I.*, A generalization of the fundanental equation of the theory of differential games. *Dokl. Akad. Nauk SSSR* **254** (1980), pp. 293–297.
35. *Subbotin, A. I.*, A generalization of the main equation of differential game theory. *J. Optim. Theory Appl.* **43** (1984), pp. 103–133.
36. *Subbotin, A. I. and Taras'ev, A. M.*, Conjugate derivatives of the value function of a differential game. *Dokl. Akad. Nauk SSSR* **283** (1985), pp. 559–564.
37. *Subbotin, A. I. and Taras'ev, A. M.*, Stability properties of the value function of a differential game and viscosity solutions of Hamilton–Jacobi equations. *Problems Control and Inf. Theory* **15** (1986), pp. 451–463.

38. *Subbotin, A. I.*, Existence and uniqueness results for Hamilton–Jacobi equations. *Nonlinear Analysis* (to appear).
39. *Zaremba, L. S.*, Modification of the viscosity solution and its applications. *Proc. 6th Int. Conf. on Mathematical Modelling*. In: *Math. Comp. Modelling* 11 (1988), pp. 699–701.

Вязкие решение уравнений Беллмана на множестве достижимости

Г. ИШИИ, ДЖ.Л. МЕНАЛЬДИ, Л. ЗАРЕМБА

(Токио, Детройт, Ст. Джонс)

В работе рассматривается задача оптимального управления, в которой минимизируемый функционал и функция оптимального результата разрывны.

Показано, что функция оптимального результата совпадает с вязкими решениями соответствующего уравнения Беллмана. Вязкое решение определено на базе конструкции, введенной Крэндаллом и Лионсом, с использованием ее модификации, предложенной Ишии.

При дополнительных условиях, обеспечивающих непрерывные функции оптимального результата, доказано, что вязкое решение единственно.

Leszek Zaremba
Memorial University of Newfoundland
Department of Mathematics and Statistics
St. John's, Nfld, Canada A1C 5S7

ON DECENTRALIZED STABILIZATION OF LARGE-SCALE LINEAR DISCRETE SYSTEMS

D. ROSINOVÁ

(*Bratislava*)

(Received September 9, 1990)

In this paper the decentralized stabilization problem of large-scale linear discrete non-delayed systems is studied. We present the subsystems dominance approach which yields a simple control procedure employing local feedback controllers with minimal knowledge of interconnections. Sufficient conditions for stabilizability of a complex system are derived providing wider class of stabilizable systems than those mentioned in the literature.

1. Introduction

We consider the stabilization problem of discrete-time decentralized systems. Various decentralized control techniques have been developed so far, however, most of them for continuous systems and their extension for discrete-time requires if possible, nontrivial additional assumptions and modifications. Furthermore, there is a fundamental difference in stabilizability conditions between continuous and discrete-time systems [7], [5]. In continuous large-scale systems controllability of all single subsystems implies stabilizability of the global system while in discrete systems stabilizability of the global systems requires also some constraints on the interconnection magnitudes.

In practice the exact mathematical models of interconnections in large-scale systems are rare to know and often only constraints on their magnitudes are used in control algorithms. From practical point of view minimization of the necessary information about the interactions, required in control algorithm, is very useful.

In this note the subsystem dominance approach for discrete-time systems is presented which extends the subsystem quality control principle from [3] and [9]. This approach yields from Lyapunov theory and comparison principle for discrete-time systems. A sufficient condition for stabilizability of decentrally controlled system is derived which is less strict than those mentioned in the literature. Subsystems dominance approach provides a simple procedure for decentralized controller

design on subsystem level, which does not directly employ the interconnection model.

2. Problem Formulation and Preliminaries

Consider the large-scale discrete-time dynamical system given by:

$$\begin{aligned} x_i(k+1) &= A_i x_i(k) + \sum_{\substack{j=1 \\ j \neq i}}^N A_{ij} x_j(k) + B_i u_i(k) \\ y_i(k) &= C_i x_i(k) \quad i = 1, 2, \dots, N \end{aligned} \quad (1)$$

where N is the number of subsystems, $x_i(k) \in \mathbb{R}^{n_i}$, $u_i \in \mathbb{R}^{m_i}$, $y_i \in \mathbb{R}^{p_i}$ represent state, input, and output vector of the i -th subsystem, respectively. A_i , A_{ij} , B_i , C_i are constant real matrices of appropriate dimensions.

Remark 2.1. System (1) corresponds with a decentralized continuous system converted with small sampling period. When considering longer sampling periods $B_{ij} \neq 0$ for $i \neq j$ can occur and the matrix B in discrete version would not have block diagonal form.

Our aim is to stabilize the global system (1) by local controllers:

$$u_i(k) = K_i x_i(k) \quad i = 1, \dots, N \quad (2a)$$

$$\text{or: } u_i(k) = K_i' y_i(k) = K_i' C_i x_i(k) = K_i x_i(k) \quad (2b)$$

control law (2b) can be employed when all the subsystems are observable.

In further consideration we use this notation:

$\lambda_M(X)$, $\lambda_m(X)$: maximum and minimum eigenvalue of square matrix X , respectively,

$\|X\|$: Euclidean norm of vector X for matrix X : $\|X\| = \lambda_M^{1/2}(X^T X)$

$A = (a_{ij})_{n \times m}$: $n \times m$ matrix A with elements a_{ij}

$K = \text{diag}(K_i)$: block diagonal matrix K with matrices K_i as blocks on the diagonal of K .

We will consider the notion of stabilization as it was introduced in [5].

DEFINITION 2.1. The discrete system (1) is stabilizable by the local feedback control (2a) or (2b) if every solution $x(k)$ of the closed-loop discrete system:

$$x_i(k+1) = (A_i + B_i K_i) x_i(k) + \sum_{\substack{j=1 \\ j \neq i}}^N A_{ij} x_j(k) \quad (3a)$$

$$\text{or: } x_i(k+1) = (A_i + B_i K_i C_i) x_i(k) + \sum_{\substack{j=1 \\ j \neq i}}^N A_{ij} x_j(k) \quad (3b)$$

$$i = 1, \dots, N$$

starting from arbitrary initial $x_0(0)$ converges asymptotically to $x(k) = 0$ as $k \rightarrow \infty$.

DEFINITION 2.2. The discrete system (1) is said to have a degree of stability $\alpha > 1$ if there exists $\beta > 0$ such that the solution $x(k)$ of the closed-loop system (3a) or (3b) satisfies:

$$\|x(k_2)\| \leq \beta \|x(k_1)\| (1/\alpha)^{k_2-k_1}$$

for all $k_1, k_2 > 0, k_2 \geq k_1$.

Because of the linearity of the system (1), the asymptotic stability implies the global exponential stability.

To analyse the stability of the system (1), the second Lyapunov method can be employed, with the use of vector Lyapunov function and the discrete version of the comparison principle [11]. Vector Lyapunov function for the system (1) is:

$$v(k) = [v_1(x_1(k)), v_2(x_2(k)), \dots, v_N(x_N(k))]^T \tag{4}$$

where $v_i(x_i(k)) = v_i(k)$ is the Lyapunov function of the i -th subsystem. For a linear system appropriate subsystem Lyapunov function is:

$$v_i(k) = (x_i^T(k) P_i x_i(k))^{1/2} \quad i = 1, 2, \dots, N \tag{5}$$

where P_i is a symmetric positive definite matrix. For a symmetric positive definite matrix Y :

$$\lambda_m(Y) \|x\|^2 \leq x^T Y x \leq \lambda_M(Y) \|x\|^2$$

(6)

e.g. $\lambda_m^{1/2} \|x\| \leq (x^T Y x)^{1/2} \leq \lambda_M^{1/2} \|x\|$.

Square matrix A is called M -matrix iff: $a_{ij} \geq 0$ for all nondiagonal elements of A [10].

DEFINITION 2.3. Square matrix $A = (a_{ij})_{n \times n}$ is diagonally dominant if there exist $d_j > 0, j = 1, \dots, n$ such that:

$$d_i |a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^n d_j |a_{ij}| \quad i = 1, 2, \dots, n$$

or (7)

$$d_j |a_{jj}| > \sum_{\substack{i=1 \\ i \neq j}}^n d_i |a_{ij}| \quad j = 1, 2, \dots, n$$

Square matrix A is negative diagonally dominant if it is diagonally dominant and $a_{ii} < 0$ for all i .

Lemma 2.1. [10] M -matrix is stable iff it is negative diagonally dominant.

3. Subsystems Dominancy Approach, Stabilization Condition

Subsystems dominancy approach (SDA) is based on the fact, that in the decentralized control scheme (1)-(3) the subsystems play an active role in the control procedure while the interconnections make up a passive part of the global system as only the diagonal blocks in system matrix can be changed via decentralized control. The main idea of using this approach is to design local controllers on subsystem level without directly employing the interconnection model. SDA is approved by practical experience and can be derived using the second Lyapunov method. This approach yields the sufficient condition for stability of the system (3) and provides simple control algorithms. The second Lyapunov method in connection with vector Lyapunov function (4) and the discrete comparison principle enables us to determine the stability of large-scale system in terms of negative definiteness of a constant matrix (the so-called aggregation matrix). For the system (3) with vector Lyapunov function defined in (4) and (5), aggregation matrix W can be found in the following form [8]:

$$\Delta v(k) = v(k+1) - v(k) \leq W[\|x_1(k)\|, \dots, \|x_N(k)\|]^T \quad (8)$$

where W is an $N \times N$ constant real matrix with elements:

$$\begin{aligned} w_{ii} &= \lambda_M^{1/2} [(A_i + B_i K_i)^T P_i (A_i + B_i K_i)] - \lambda_m^{1/2}(P_i) \\ w_{ij} &= \lambda_M(P_i) / \lambda_m^{1/2}(P_i) \cdot \lambda_M^{1/2}(A_{ij} A_{ij}) \quad i \neq j, \quad i = 1, 2, \dots, N \end{aligned} \quad (9)$$

Substituting for $\|x_i(k)\|$ in (8) its upper boundary from (5) and (6) we obtain:

$$\Delta v(k) \leq W_0 v \quad \text{where} \quad W_0 = (w_{0ij})_{N \times N} \quad (10)$$

$$w_{0ii} = \frac{\lambda_M^{1/2}(S_i^T P_i S_i)}{\lambda_m^{1/2}(P_i)} - 1 \quad (11)$$

$$w_{0ij} = \frac{\lambda_M(P_i)}{\lambda_m^{1/2}(P_i) \lambda_m^{1/2}(P_j)} \lambda_M^{1/2}(A_{ij}^T A_{ij}) \quad i \neq j$$

From (10) and (11) the sufficient condition for stability of the system (3) is negative definiteness of W_0 [6]. Having in mind that W_0 is an M -matrix and Lemma 2.1 holds, obviously negative definiteness of W_0 is equivalent to negative definiteness of W in (9) and also to negative definiteness of $W' = (w'_{ij})_{N \times N}$:

$$w'_{ii} = \frac{\lambda_m^{1/2}(P_i)}{\lambda_M^{1/2}(P_i)} \lambda_M^{1/2}(S_i^T S_i) - \frac{\lambda_m(P_i)}{\lambda_M(P_i)} \quad (12a)$$

$$w'_{ij} = \lambda_M^{1/2}(A_{ij}^T A_{ij}). \quad (12b)$$

From Lemma 2.1 for stable W' must be $w'_{ii} < 0$, together with (12a) we obtain:

$$\lambda_M^{1/2}(S_i^T S_i) < 1 \quad \text{that implies : } S_i^T S_i - I_i = -Q_i \tag{13}$$

where I_i is an $n_i \times n_i$ identity matrix, Q_i is symmetric positive definite $n_i \times n_i$ matrix. Therefore, in (12a) $P_i = I_i$ can be considered to determine the stability condition and we reach the following theorem.

THEOREM 3.1. The system (3) is asymptotically stable if there exist real positive q_1, \dots, q_N such that:

$$q_i > \sum_{\substack{j=1 \\ j \neq i}}^N q_j \|A_{ij}\| + q_i \|A_i + B_i K_i\| \quad i = 1, \dots, N. \tag{14}$$

Proof. Evidently, W' given in (12a), (12b) is M -matrix. Considering $P_i = I_i$ in (12a) and applying (7) and Lemma 2.1 for W' we obtain (14).

Theorem 3.1 provides the sufficient condition for the stability of discrete large-scale decentralized system and also approves subsystem dominance approach as it is stated in the following corollaries.

Corollary 3.1. Let the matrix norm $\|A_i + B_i K_i\|$ can be made arbitrarily small by choosing appropriate K_i . If there exist $q_1, \dots, q_N > 0$ such that:

$$q_i > \sum_{\substack{j=1 \\ j \neq i}}^N q_j \|A_{ij}\| \quad \text{for } i = 1, \dots, N \tag{15}$$

(for special case, $q_i = 1$ for all $i : 1 > \sum_{\substack{j=1 \\ j \neq i}}^N \|A_{ij}\|$) then the system (1) is stabilizable via decentralized control and its stability is determined by local subsystems feedback matrices ($A_i + B_i K_i$) i.e. by w'_{ii} in aggregation matrix (12a), (12b).

Corollary 3.2. Let for the global system (1) the inequality (15) holds. Then local controllers can be designed like for isolated subsystems and the sufficient condition for keeping the stability of the global system is obtained from (14):

$$\|A_i + B_i K_i\| < 1 - \sum_{\substack{j=1 \\ j \neq i}}^N (q_j / q_i) \|A_{ij}\| \tag{16}$$

for some real positive $q_j, j = 1, \dots, N$;

for $q_i = 1$ for all i :

$$\|A_i + B_i K_i\| < 1 - \sum_{\substack{j=1 \\ j \neq i}}^N \|A_{ij}\|. \tag{17}$$

Condition (16) (or (17)) can be treated for each subsystem independently and it determines the class of stabilizing matrices $K = \text{diag}(K_i)$. Condition (16) (or (17)) guarantees the stability of the global system and indicates a procedure which stabilizes the system or shifts its stability degree. Such a procedure is given in the following corollary.

Corollary 3.3. Let the closed-loop system is given by (3a) or (3b), where $K_i(\alpha_i)$, $i = 1, \dots, N$ is a function of real parameter α_i . The global system is stabilizable by means of $K_i(\alpha_i)$ if there exist α_i^* , $i = 1, \dots, N$ for which (14) holds. If, furthermore,

$$(\partial w'_{ii} / \partial \alpha_i) < 0 \text{ for } \alpha_i \in (\alpha_{i0}, \alpha_i^*) \text{ for some } \alpha_{i0}, i = 1, \dots, N$$

where w'_{ii} is given in (12a); the stabilization procedure starts with $K_i(\alpha_{i0})$ and α_i is changed: $\alpha_i \leftarrow \alpha_i + \Delta \alpha_i$, $\alpha_i \in (\alpha_{i0}, \alpha_i^*)$ until the given system is stable (or (14) holds for some q_i , $i = 1, \dots, N$).

Remark 3.1. Because α_i does not influence w'_{ij} in (12b), for $w'_{ii} < 0$ and constant P_i : $(\partial w'_{ii} / \partial \alpha_i) < 0$ implies $\partial(\Delta v_i / v_i) / \partial \alpha_i > 0$, where $\Delta v_i / v_i$ is a measure of stability and its upper bound is given by the value of the i -th row of W' .

Procedure stated in Corollary 3.3 provides the systematic way how to "improve" the values of parametrized feedback matrices $K_i(\alpha_i)$ in the direction to system stability region. This property is important for practical applications of decentralized control, it enables the "tuning" of the system via appropriate changes of parameters α_i . (When prescribing the demanded system behavior, its limitations which follow from decentralized control structure as we mentioned above must be considered.)

Stabilizable class of decentralized systems stated in Theorem 3.1 and Corollaries 3.1 and 3.2 is wider than those presented in Lee, Radovic [4]. (Though stability conditions stated in [4] are obtained for a more general delay case, they claim to be less strict than ones developed so far for the nondelay case.) In [4] the authors require

$$\|A_i + B_i K_i\| + \left\| \sum_{\substack{j=1 \\ j \neq i}}^N N_j A_{ji}^T A_{ji} \right\|^{1/2} < 1 \text{ for each } i \quad (18)$$

where N_j is the cardinality of a set:

$$J_i = \{j | A_{ij} \neq 0, j = 1, \dots, N\}.$$

In [5] Lee and Radovic presented another result for CCM model where stability of the global system is guaranteed if there exist positive definite symmetric Q_i for all i , such that:

$$Q_i - \sum_{j=1}^N N_j L_{ji}^T (I_i + B_j^T P_j B_j) L_{ji} > 0$$

where P_i is the solution of the Riccati equation:

$$P_i = A_i^T P_i A_i - A_i^T P_i B_i (I_i + B_i^T P_i B_i)^{-1} B_i^T P_i A_i + C_i^T Q_i C_i$$

and stabilizing gains are:

$$K_i = -(I_i + B_i^T P_i B_i)^{-1} B_i^T P_i A_i, \quad i = 1, \dots, N. \tag{19}$$

However, the analysis and use of (19) for decentralized control design is not simple and in [5] the control algorithm is proposed only for the upper triangular A .

In (16) the multiplying of the interconnection matrices by N_j is not required and the introduction of q_i further extends the stabilizable class of systems in comparison to (18). For instance, following (16) systems with upper triangular matrices A are obviously stabilizable, although it is not the case of (18). In comparison to (19) our results given in Theorem 3.1 and its corollaries are relatively simple and easily applicable for control algorithm design. Furthermore, (16) enables us to avoid a direct employment of the interconnections in a design of feedback matrix K .

4. Decentralized Control Algorithms for Stabilization

Following Theorem 3.1 and its corollaries the reasonable way to stabilize the system seems to minimize matrix norms of subsystem matrices (minimization of w'_i in (12a)), which is close to minimization of stability degrees of subsystems, though there is a certain gap between these two procedures. Theorem 3.1 provides only the sufficient condition for stabilizability giving upper bounds on system behaviour. Therefore, a stability degree of the global system does not exactly follow stability degrees of isolated subsystems, though in bounds mentioned above both tendencies are connected. Because of these reasons we propose control algorithms with local parameters on subsystem level, which can be changed "to tune" the global system behaviour in the desired way. This "tuning" can be carried out according to a certain objective which can be tested in the system.

Minimization of w'_i (or $\|A_i + B_i K_i\|$) in (12b) w. r. t. K_i yields control law:

$$K_i = -(B_i^T B_i)^{-1} B_i^T A_i$$

or: $K_i = -(B_i^T B_i)^{-1} B_i^T A_i C_i^T (C_i C_i^T)^{-1} C_i$

for output feedback which can be parametrized (see Corollary 3.3):

$$K_i = -\alpha_i (B_i^T B_i)^{-1} B_i^T A_i \quad \alpha_i \simeq 1$$

or: $K_i = -\alpha_i (B_i^T B_i)^{-1} B_i^T A_i C_i^T (C_i C_i^T)^{-1} C_i$ (20)

$\alpha_i = 1$ brings the greatest possible stability degree for subsystems. However, in practice such a control law probably will not provide the best results for the global

system because of high feedback gains which can cause overshoot and nonlinear behaviour following from constraints on signals in the control loop (in such cases the solution would be different from that expected). We must consider also the difference between the local and global stability degree. The advantage of the use of Corollary 3.3 and parametrized control law is in the possibility to determine the appropriate α_i for the global system. We tested $\alpha_i \in (0.9 - 1.3)$ for (20) with good results for the global system stability degree.

Remark 4.1. For square invertible matrix B_i and $\alpha_i = 1$ (20) place all subsystem poles to zero.

The second control algorithm is based on subsystem controller design using a corresponding Riccati equation. Subsystem stability degree α_i is being increased until condition (17) holds.

$$\begin{aligned} K_i &= -\alpha_i^2 (I_i + \alpha_i^2 B_i^T P_i B_i)^{-1} B_i^T P_i A_i \\ \text{or: } K_i &= -\alpha_i^2 (I_i + \alpha_i^2 B_i^T P_i B_i)^{-1} B_i^T P_i A_i C_i^T (C_i C_i^T)^{-1} C_i \end{aligned} \quad (21)$$

for output feedback where

$$P_i = \alpha_i^2 A_i^T P_i A_i - \alpha_i^4 A_i^T P_i B_i (I_i + \alpha_i^2 B_i^T P_i B_i)^{-1} B_i^T P_i A_i + C_i^T C_i$$

Having in mind (13) the use of $P_i = I_i$ in computation of the aggregation matrix W in (12a) is recommended. It provides a less strict stability condition than the original P_i .

We can observe that for stabilizable systems (20) and (21) provide the way to the stabilization and enable to increase the stability degree of the global system (as far as it is possible) by means of changing parameters α_i .

5. Conclusions

A new method for the decentralized stabilizing problem is presented, based on subsystems dominance approach introduced here. Sufficient conditions for decentralized stabilization are derived in simple form. Two control algorithms are designed, where the interconnection model is not required directly in control law. The computation of constant feedback matrices is carried out on subsystem level. Parameters are introduced into control law in order to enable the shifting of the stability degree of the whole system. The designed algorithms provide the approach which is relatively simple to be implemented in practice. The results are illustrated on an example in the Appendix.

Appendix

The following example illustrates the results of the designed control algorithms. Improvement against similar results in literature are also mentioned.

Example. First subsystem S_1 :

$$x_1(k+1) = \begin{bmatrix} 0.85 & 0.3 \\ 0.2 & 0.5 \end{bmatrix} x_1(k) + \begin{bmatrix} -0.4 & -0.35 \\ 0.2 & 0.4 \end{bmatrix} x_2(k) + \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix} x_3(k) + \begin{bmatrix} 0.8 & 0.1 \\ 0.6 & 0.5 \end{bmatrix} u_1(k)$$

Second subsystem S_2 :

$$x_2(k+1) = \begin{bmatrix} -0.25 & -0.4 \\ 0.1 & -0.3 \end{bmatrix} x_1(k) + \begin{bmatrix} 0.6 & 0.3 \\ 0.1 & 0.6 \end{bmatrix} x_2(k) + \begin{bmatrix} -0.2 \\ 0.02 \end{bmatrix} x_3(k) + \begin{bmatrix} 0.6 \\ 0.5 \end{bmatrix} u_2(k)$$

Third subsystem S_3 :

$$x_3(k+1) = [0.05 \ 0.1]x_1(k) + [-0.05 \ 0.1]x_2(k) + 0.4x_3(k) + 0.4u_3(k)$$

The given system is unstable with eigenvalues:

$$1.155; 0.561 \pm 0.429i; 0.442; 0.252$$

The subsystems are stable with eigenvalues:

$$S_1 : 0.976; 0.374 \quad S_2 : 0.773; 0.427 \quad S_3 : 0.400$$

a) from (20) we got next results:

$$\alpha_i = 1 \quad \text{for } i = 1, 2, 3 :$$

$$K = \begin{bmatrix} -1.191 & -0.294 & 0 & 0 & 0 \\ 1.029 & -0.647 & 0 & 0 & 0 \\ 0 & 0 & -0.672 & -0.787 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

spectral radius of the controlled system (1/stability degree):

$$\text{global system: } 0.542; \quad S_1 : 0.000; \quad S_2 : 0.403; \quad S_3 : 0.000;$$

aggregation matrix:

$$W = \begin{bmatrix} -1.000 & 0.682 & 0.224 \\ 0.524 & -0.592 & 0.201 \\ 0.112 & 0.112 & -1.000 \end{bmatrix}.$$

Sufficient condition (14) is fulfilled e.g. for $q_1 = 0.8$, $q_2 = 1$, $q_3 = 0.3$. (Condition (18) from [4] does not hold.)

$$\alpha_i = 1.2 \quad \text{for } i = 1, 2, 3 :$$

$$K = \begin{bmatrix} -1.429 & -0.353 & 0 & 0 & 0 \\ 1.235 & -0.776 & 0 & 0 & 0 \\ 0 & 0 & -0.807 & -0.944 & 0 \\ 0 & 0 & 0 & 0 & -1.2 \end{bmatrix}$$

spectral radius of the controlled system:

global system: 0.505; S_1 : 0.195; S_2 : 0.406; S_3 : 0.800

$$W = \begin{bmatrix} -0.804 & 0.682 & 0.224 \\ 0.524 & -0.590 & 0.200 \\ 0.112 & 0.112 & -0.920 \end{bmatrix}$$

As in the previous case, (14) is fulfilled.

b) from (21) we obtain these results:

$\alpha_i = 2.4$ for $i = 1, 2, 3$:

$$K = \begin{bmatrix} 0.933 & 0.343 & 0 & 0 & 0 \\ -0.402 & 0.383 & 0 & 0 & 0 \\ 0 & 0 & 0.773 & 0.539 & 0 \\ 0 & 0 & 0 & 0 & 0.595 \end{bmatrix} \quad (22)$$

$$P = \begin{bmatrix} 2.525 & 0.0142 & 0 & 0 & 0 \\ 0.0142 & 1.348 & 0 & 0 & 0 \\ 0 & 0 & 5.429 & -2.813 & 0 \\ 0 & 0 & -2.813 & 4.202 & 0 \\ 0 & 0 & 0 & 0 & 1.595 \end{bmatrix}$$

spectral radius:

global system: 0.519; S_1 : 0.172; S_2 : 0.360; S_3 : 0.162

for P from (22) the respective aggregation matrix is not negative definite so it was computed for vector Lyapunov function with $P_i = I_i$ (see (13)):

$$W = \begin{bmatrix} -0.768 & 0.682 & 0.224 \\ 0.524 & -0.549 & 0.054 \\ 0.112 & 0.112 & -0.838 \end{bmatrix}$$

For this W (14) is fulfilled. If α_i are further increased, the global system remains stable, but its stability degree slightly decreases.

References

1. Chen, W. S., Desoer, C. A., Eigenvalue assignment and stabilization of interconnected systems using local feedback. *IEEE Trans. Autom. Control*, **AC-24** (1979), pp. 312-317.
2. Fiedler, M., Speciální matice a jejich použití v numerické matematice (Special Matrix and Their Use in Numerical Mathematics; in Czech) SNTL, Praha, 1981.
3. Veselý, V., Murgáš, J., Hejda, I., Decentralized Adaptive Control with Strictly Reduced Information Pattern. *IFAC/IFORS/IMACS Symposion Large Scale Systems: Theory and Applications*, Berlin, 1989.

4. Lee, T. N., Radovic, U. L., General Decentralized Stabilization of Large-Scale Linear Continuous and Discrete Time-Delay Systems. *Int. J. Control*, vol. **46** (1987), pp. 2127-2140.
5. Lee, Y. N., Radovic, U. L., Decentralized Stabilization of Linear Continuous and Discrete-Time Systems with Delays in Interconnection. *IEEE Trans. Autom. Control*, **AC-33** (1988), pp. 757-760.
6. Bitsoris, G., Burgat, C., Stability Analysis of Complex Discrete Systems with Locally and Globally Stable Subsystems. *Int. J. Control*, vol. **25** (1977), pp. 413-424.
7. Suh, I. H., Bien, Z., On Stabilization by Local State Feedback for Discrete-Time Large Scale Systems with Delays in Interconnections. *IEEE Trans. Autom. Control*, **AC-27** (1982), pp. 744-746.
8. Veselý, V., Murgaš, J., Makovická, A., Rosinová, D., Barč, V., Varga, J., On the Decentralized Control of Large-Scale Linear Dynamic Systems. *International Symposium Lignoautomatica '86*, Bratislava, November 17-21, 1986, preprints, pp. 143-147.
9. Veselý, V., Large-Scale Dynamic System Stabilization Using the Subsystems Dominant Principle Approach. (submitted to press.)
10. Voronov, A. A., *Vvedeniye v dinamiku slozhnykh upravlyajemykh sistem* (in Russian). Nauka, Moskva, 1985
11. Grujić, L. T., Siljak, D. D., On Stability of Discrete Composite systems. *IEEE Trans. Autom. Control*, October 1973, pp. 522-524.

Децентрализованная стабилизация сложных линейных дискретных систем

Д. РОСИНОВА

(Братислава)

Статья рассматривает стабилизацию сложных линейных дискретных систем при помощи принципа доминирующих подсистем. На этом подходе приведены простые законы управления, которые можно использовать с минимальным знанием связей между подсистемами. Приведено достаточное условие стабилизируемости системы менее строго, чем условия в литературе.

Danica Rosinová
Slovak Technical University
Faculty of Electrical Engineering
KASR TP
Ilkovičova 3
812 19 Bratislava, Czechoslovakia

NON-ASYMPTOTIC SOLUTION OF CONFIDENCE ESTIMATION PARAMETER TASK OF A NON-LINEAR REGRESSION BY MEANS OF SEQUENTIAL ANALYSIS

A. V. TIMOFEEV

(*Tomsk*)

(Received February 8, 1991)

This paper presents a sequential plan used for obtaining the confidence interval with necessary size for volume parameters of non-linear regression. The distributions of the observation noises are supposed to be unknown. The plan described in this paper enables us to obtain the confidence estimates of the unknown parameters in non-asymptotic state. Estimation for the mean observation time have been cited in the suggested sequential plan.

1. Statement of the problem

Consider the process $\{(y(x_t), x_t)\}_{t \in \mathbb{N}}$ is being observed. This process is described by the following equation

$$y(x_t) = f_t(x_t, \theta^*) + \xi(t), \quad t \in \mathbb{N} \quad (1)$$

where $\mathbb{N} = \{1, 2, \dots\}$ is an index set, $\{x_t\}_{t \in \mathbb{N}}$ is a set of input variables and

$$\forall t \in \mathbb{N} : x_t \in X = [a, b], \quad |a| < \infty, \quad |b| < \infty;$$

$\{\xi(t)\}_{t \in \mathbb{N}}$ are independent random variables (random measurement errors) and

$$\forall t \in \mathbb{N} : [E\xi(t) = 0, E\xi^4(t) < \sigma^4 < \infty].$$

Here E stands for expectation. The value σ , functions $\{f_t(\cdot)\}_{t \in \mathbb{N}}$ into which the estimated parameter $\theta^* \in \Theta$ enters non-linearly as well as a compact $\theta \supseteq \mathbb{N}$ are known to the statistician.

Observing the process $\{(y(x_t), x_t)\}_{t \in \mathbb{N}}$ it is necessary to build in θ such a closed interval $\Xi = [c, d]$ that

$$P_{\theta^*}(\theta^* \in \Xi) > P_c, \quad 0 < |c - d| < \delta < \infty, \quad P_c \in [0, 1]$$

Values P_c and δ are supposed to be prescribed.

2. Method of solution

Let function $(f_i(\cdot))_{i \in \mathbb{N}}$ are continuously differentiable function on the Θ and

$$\forall (n \in \mathbb{N}, \theta \in \Theta) : c_n(\theta) = \left[\sum_{i=1}^n 2 \left[\frac{\partial f_i(x_i, \theta)^2}{\partial \theta} \right]^2 \right]^{0.5} (n^2(1 - P_c))^{-0.5}.$$

Suppose the estimations θ_n^+ and θ_n^- are defined as follows

$$\theta_n^+ = \arg \inf_{\theta \in \Theta} \left| \frac{\partial I_n(\theta)}{\partial \theta} - 2 \cdot c_n(\theta) \right|$$

$$\theta_n^- = \arg \inf_{\theta \in \Theta} \left| \frac{\partial I_n(\theta)}{\partial \theta} + 2 \cdot c_n(\theta) \right|$$

for any $n > 1$.

Here

$$I_n(\theta) = \sum_{i=1}^n (y(x_i) - f_i(x_i, \theta))^2 / (n \cdot \sigma).$$

Values θ_n^+ and θ_n^- can be used as limits of confidence interval in time moment n .

The solution of the given task will be found by means of sequential analysis. For the sequential plan of confidence estimation of θ^* by observation (1) let us consider the form of the pair (d, τ) where the value $d_n = |\theta_n^+ - \theta_n^-|$ characterizes the achieved accuracy of the confidence estimation for time moment n , and τ is the moment of observation stop

$$\tau = \inf \{ n > 1 | d_n < \delta \}.$$

Thus, the required accuracy of confidence estimation will be achieved at the time moment τ .

Let us define the following value for some finite $A, B > 0$

$$\rho_\epsilon(A, B) = \inf \{ n \in \mathbb{N} | n > 2B^2((A^2\delta\sigma^{-1}2^{-1} - \epsilon)^2(1 - P_c))^{-1} \}.$$

Here $0 < \epsilon < A\delta/2$.

THEOREM 1. Let for some positive values l and L the recording is acceptable

$$\forall (t \in \mathbb{N}; x \in X; \theta \in \Theta) : \left[0 < l \leq \left| \frac{\partial f_i(x, \theta)}{\partial \theta} \right| \leq L < \infty \right]$$

Then the next statements are true.

- 1) $P_{\theta^*}(\tau < \alpha) = 1, \theta^* \in \Theta,$
- 2) $E_{\theta^*} \tau \leq \rho_\epsilon(l, L) + 2L^4 \epsilon^{-4} (\pi^2/2 - 3\rho_\epsilon(l, L)/(1 + \rho_\epsilon(l, L)) + 2[S_3 - 0.25 + 1/(2(\rho_\epsilon(l, L) + 1)(\rho_\epsilon(l, L) + 2))]), \theta^* \in \Theta,$

where

$$S_3 = \sum_{k=1}^{\infty} k^{-3} \sim 1.2020569.$$

- 3) $P_{\theta^*}(\theta^* \in [\theta_l(\tau), \theta_h(\tau)]) \geq P_c, \theta^* \in \Theta.$

Here E is the mean observation time in the sequential plan $(d_n, \tau), \theta_l(n) = \min(\theta_n^+, \theta_n^-), \theta_h(n) = \max(\theta_n^+, \theta_n^-).$

Proof of the Theorem. From the Theorem statement follows that

$$\forall(\theta_1, \theta_2 \in \Theta; x \in X; t \in \mathbb{N}) : |f_t(x, \theta_1) - f(x, \theta)| \leq L|\theta_1 - \theta_2|.$$

Let us consider the following representation for some θ

$$\begin{aligned} 0.5 \frac{\partial I_n(\theta)}{\partial \theta} &= \sum_{t=1}^n f'_t(x_t, \theta)(f_t(x_t, \theta^*) - f_t(x_t, \theta))(n\sigma)^{-1} + \sum_{t=1}^n f'_t(x_t, \theta)\xi(t)(n\sigma)^{-1} \\ &= R_n(\theta^*, \theta) + \chi_n(\theta). \end{aligned} \tag{2}$$

Here $f'_t(x_t, \theta) = \frac{\partial f_t(x_t, \theta)}{\partial \theta}.$

Using the finite increment formula and the condition of the Theorem we come to the conclusion that

$$\forall(x, y \in X; t \in \mathbb{N}; \theta_1, \theta_2 \in \Theta) : |f_t(x, \theta_1) - f_t(x, \theta_2)| \geq l|\theta_1 - \theta_2|.$$

That is why

$$\forall(n \in \mathbb{N}, \theta \in \Theta) : R_n(\theta^*, \theta) \geq l^2 \sigma^{-1} |\theta^* - \theta|. \tag{3}$$

It may be occur that

$$\Delta(\theta_n^+) = \left| \frac{\partial I_n(\theta_n)}{\partial \theta} - 2c_n(\theta_n^+) \right| \neq 0,$$

$$\Delta(\theta_n^-) = \left| \frac{\partial I_n(\theta_n^-)}{\partial \theta} + 2c_n(\theta_n^-) \right| \neq 0,$$

because of the Θ array restrictivity. For example, let us assume that

$$\begin{aligned} n = 1, f_1(x, \theta) = \theta, \theta^* = 1, \sigma^* = 1, \xi(1) = 0, \\ \hat{\theta} = \arg \inf_{\theta \in \Theta} (\theta) = 1 - (2/(1 - P_c)^{0.5}) + \phi \end{aligned}$$

with some $\phi > 0$.

In this case we have

$$\begin{aligned}\frac{\partial I_1(\theta)}{\partial \theta} &= 2(y(x_1) - \theta) = -2(\theta^* + \xi(1) - \theta) = 2(1 - \theta), \\ c_1(\theta) &= c_1 = (2/(1 - P_c))^{0.5},\end{aligned}$$

that is why

$$\theta_n^- = \hat{\theta} \text{ and } \Delta(\theta_n^-) = |2(1 - \theta + 2c_1)| = 2\phi > 0.$$

Consider the events:

$$\begin{aligned}\omega_n^+ &: \{R_n(\theta^*, \theta_n^+) = c_n(\theta_n^+) - \chi_n(\theta_n^+) - \Delta(\theta_n^+)/2 \leq l^2 \delta \sigma^{-1} 2^{-1}\}, \\ \omega_n^- &: \{R_n(\theta^*, \theta_n^-) = -c_n(\theta_n^-) - \chi_n(\theta_n^-) - \Delta(\theta_n^-)/2 \leq -l^2 \delta \sigma^{-1} 2^{-1}\}.\end{aligned}$$

Taking into consideration (3) it can be clearly seen that

$$\begin{aligned}\omega_n^+ &= \omega(n) : \{|\theta^* - \theta_n^+| < \delta/2\} \\ \omega_n^- &= \omega(n) : \{|\theta^* - \theta_n^-| < \delta/2\}\end{aligned}$$

Here the recording $\omega_1 = \omega_2$ denotes that event ω_2 is being conditioned by event ω_1 .

Further,

$$\begin{aligned}\omega_n^+ \cdot \omega_n^- &= \bar{\omega}_1(n) \cdot \bar{\omega}_2(n) = \bar{\omega}_3(n) : \{|\theta^* - \theta_n^+| + |\theta^* - \theta_n^-| < \delta\} \\ &= \bar{\omega}_4(n) : \{|\theta^+ - \theta_n^-| < \delta\} \\ &= \bar{\omega}_5(n) : \{\tau < \delta\}.\end{aligned}\tag{4}$$

It is evident that

$$\forall (n > \rho_\varepsilon(l, L), \theta \in \Theta) : \{c_n(\theta) < l^2 \delta \sigma^{-1} 2^{-1}\}.$$

So

$\forall (n > \rho_\varepsilon(l, L), \theta^* \in \Theta) :$

$$\begin{aligned}P_{\theta^*}(\omega_n^+ \omega_n^-) &> P_{\theta^*}(\{-\chi_n(\theta_n^+) < l^2 \delta \sigma^{-1} 2^{-1} - c_n(\theta_n^+)\}) \\ &\quad \times \{-\chi_n(\theta_n^-) < l^2 \delta \sigma^{-1} 2^{-1} - c_n(\theta_n^-)\}) \\ &> P_{\theta^*}(\{|\chi_n(\theta_n^-)| < l^2 \delta \sigma^{-1} 2^{-1} - c_n(\theta_n^-)\}) \\ &\quad \times \{|\chi_n(\theta_n^-)| < l^2 \delta \sigma^{-1} 2^{-1} - c_n(\theta_n^-)\}) \\ &\geq 1 - [P_{\theta^*}(|\chi_n(\theta_n^+)| > l^2 \delta \sigma^{-1} 2^{-1} - 2^{0.5} \cdot L((1 - P_c)n)^{-0.5}) \\ &\quad + P_{\theta^*}(|\chi_n(\theta_n^-)| > l^2 \delta \sigma^{-1} 2^{-1} - 2^{0.5} \cdot L((1 - P_c)n)^{-0.5})] \\ &= 1 - [P_{\theta^*}(|\chi_n(\theta_n^+)| > \varepsilon) + P_{\theta^*}(|\chi_n(\theta_n^-)| > \varepsilon)].\end{aligned}\tag{5}$$

It is easy to see that

$$\forall (n \in \mathbb{N}; \theta^*, \theta \in \Theta) : \quad (6)$$

$$E_{\theta^*}(\chi_n(\theta))^4 < n^{-3}L^4 + 3n(n-1)L^4n^{-4} = L^4(3n^{-2} - 2n^{-3}).$$

Using the Chebychev's inequality we have

$$\forall (\theta \in \Theta, n \in \mathbb{N}; \theta, \theta^* \in \Theta) : \quad (7)$$

$$P_{\theta^*}(|\chi_n(\theta)| > \varepsilon) < E_{\theta^*}(\chi_n(\theta))^4 \varepsilon^{-4} < L^4 \varepsilon^{-4} (3n^{-2} - 2n^{-3})$$

From (4) follows that for any $\alpha \in]0, 1[$ true implication

$$[P_{\theta^*}(\omega_n^+ \omega_n^-) > \alpha] \Rightarrow [P_{\theta^*}(\tau < n) > \alpha] \Rightarrow [P_{\theta^*}(\tau > n) < 1 - \alpha].$$

From (4) and (7) we get

$$P_{\theta^*}(\tau > n) < P_{\theta^*}(|\chi_n(\theta_n^+)| > \varepsilon) + P_{\theta^*}(|\chi_n(\theta_n^-)| > \varepsilon) < 2L^4 \varepsilon^{-4} (3n^{-2} - 2n^{-3}).$$

Further,

$$\begin{aligned} E_{\theta^*} \tau &= \sum_{n=1}^{\infty} P_{\theta^*}(\tau > n) < \rho_\varepsilon(l, L) + \sum_{n=\rho_\varepsilon(l, L)+1}^{\infty} (3n^{-2} - 2n^{-3}) 2L^4 \varepsilon^{-4} \\ &= \rho_\varepsilon(l, L) + 2L^4 \varepsilon^{-4} \left(\left[\pi^2/6 - \sum_{n=1}^{\rho_\varepsilon(l, L)} n^{-2} \right] + 2 \left[S_3 - \sum_{n=1}^{\rho_\varepsilon(l, L)} n^{-3} \right] \right) \quad (8) \\ &< \infty. \end{aligned}$$

There has been taken into account that $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$. From (8) and the Borel-Cantelli lemma we obtain that $P_{\theta^*}(\tau < \infty) = 1$. Thus, the first statement of the Theorem has been proved.

Taking into consideration (8) and the fact that

$$\forall \rho > 1 : \left\{ \begin{aligned} \sum_{n=1}^{\rho} n^{-3} &> \sum_{n=1}^{\rho} (n(n+1)(n+2))^{-1} = 0.25 - 1/(2(\rho+1)(\rho+2)), \\ \sum_{n=1}^{\rho} n^{-2} &> \sum_{n=1}^{\rho} (n(n+1))^{-1} = \rho/(\rho+1) \end{aligned} \right\},$$

we come to the conclusion that the second statement of Theorem is also true. From the condition of the Theorem we get that functions $\{f_t(\cdot)\}_{t \in \mathbb{N}}$ are strictly monotonous in θ . Let us consider the function

$$U_t(\theta^*, \theta) = (f_t(x_t, \theta^*) - f_t(x_t, \theta))^2/2, \quad \theta \in \Theta, \quad t \in \mathbb{N}.$$

From strictly monotonicity of the functions $\{f_t(\cdot)\}_{t \in \mathbb{N}}$ in θ follows that

$$\forall(\theta \in \Theta, x \in X, t \in \mathbb{N}) : \frac{\partial U_t(\theta^*, \theta)}{\partial \theta} = \frac{\partial f_t(x_t, \theta)}{\partial \theta} (f_t(x_t, \theta^*) - f_t(x_t, \theta)) \geq 0$$

with fixed $\theta^* \in \Theta$. Using the statement B.3.6 [2, p. 450] now we conclude that $U_t(\theta^*, \theta)$ is a convex function in θ . Function

$$\Phi_n(\theta^*, \theta) = h_n(u_1(\theta^*, \theta), u_2(\theta^*, \theta) \dots u_n(\theta^*, \theta)) = \sum_{t=1}^n u_t(\theta^*, \theta) n^{-1} \sigma^{-1}$$

is a linear combination of convex functions and $h_n(\cdot)$ is not decreasing with every of the function-independent variables. It can be deduced that $\Phi_n(\theta^*, \theta)$ is a convex function in θ if we use the statement of Theorem B.7.a [2, p. 470] in this case. From here we easily come to the conclusion that

$$R_n(\theta^*, \theta) = \frac{\partial \Phi_n(\theta^*, \theta)}{\partial \theta}, \quad n \in \mathbb{N}$$

is a monotonic function in θ with fixed $\theta^* \in \Theta$. So, taking into account the representation (2) we have

$$\begin{aligned} \tilde{\omega}_0(n) &: (\{|\chi_n(\theta_n^+)| < c_n(\theta_n^+)\} \cdot \{|\chi_n(\theta_n^-)| < c_n(\theta_n^-)\}) \\ \supset \tilde{\omega}_1(n) &: \{\text{sign}[R_n(\theta^*, \theta_n^+)] = -\text{sign}[R_n(\theta^*, \theta_n^-)]\} \\ \supset \tilde{\omega}_2(n) &: \{\theta^* \in [\theta_1(n), \theta_h(n)]\} \end{aligned} \tag{9}$$

Now, using the Lyapunov inequality [1] it can be written

$$\forall t \in \mathbb{N} : E\xi^2(t) < (E\xi^4(t))^{0.5} < \sigma^2.$$

From here we obtain

$$\forall(n \in \mathbb{N}; \theta^*, \theta \in \Theta) : \left\{ E_{\theta^*} \chi_n(\theta) = 0, E_{\theta^*} \chi_n^2(\theta) < \sum_{t=1}^n \left[\frac{\partial f_t(x_t, \theta)}{\partial \theta} \right]^2 n^{-2} \right\}. \tag{10}$$

An application of Chebychev's inequality yields

$$\forall(n \in \mathbb{N}; \theta^*, \theta \in \Theta) : P_{\theta^*}(|\chi_n(\theta)| > c_n(\theta)) < (1 - P_c)/2.$$

Using the Boolean inequality it is easy to see that

$$\begin{aligned} P_{\theta^*}(\tilde{\omega}_0(n)) &> 1 - [P_{\theta^*}(|\chi_n(\theta_n^+)| > c_n(\theta_n^+)) + P_{\theta^*}(|\chi_n(\theta_n^-)| > c_n(\theta_n^-))] \\ &> P_c. \end{aligned} \tag{11}$$

From (11) and (9) we obtain that

$$\forall (n \in \mathbb{N}; \theta^*, \theta \in \Theta) : P_{\theta^*}(\tilde{\omega}_2(n)) > P_c.$$

As $P_{\theta^*}(\tau < \infty) = 1$, $\theta^* \in \Theta$ and $\tau \in \mathbb{N}$ we finally conclude that

$$P_{\theta^*}(\tilde{\omega}_2(\tau)) > P_c.$$

Thus, the Theorem has been completely proved.

3. Linear regression

Let in (1)

$$\forall n \in \mathbb{N} : f_t(x_t, \theta^*) = x_t \theta^*$$

and $\theta^* \in \mathbb{R}^1$. In this case we have

$$\theta_n^+ = \arg \inf_{\theta \in \mathbb{R}^1} \left| \frac{\partial I_n(\theta)}{\partial \theta} - 2c_n(\theta) \right|.$$

It is clearly seen that we come to the following conclusions

$$\begin{aligned} \theta_n^+ &= \sum_{i=1}^n x_i y(x_i) \left(\sum_{i=1}^n x_i^2 \right)^{-1} + \left(\sum_{i=1}^n x_i^2 (1 - P_c) \right)^{-0.5} \sigma \sqrt{2}, \\ \theta_n^- &= \sum_{i=1}^n x_i y(x_i) \left(\sum_{i=1}^n x_i^2 \right)^{-1} - \left(\sum_{i=1}^n x_i^2 (1 - P_c) \right)^{-0.5} \sigma \sqrt{2}. \end{aligned}$$

We define

$$J_n = \sum_{i=1}^n x_i^2.$$

It is easily seen that

$$\begin{aligned} \theta_n^+ &= \left[\theta_{n-1}^+ - \left(J_{n-1} (1 - P_c) \right)^{-0.5} \sigma \sqrt{2} \right] J_{n-1} J_n^{-1} + x_n y(x_n) J_n^{-1} \\ &\quad + \sigma \left(J_n (1 - P_c) \right)^{-0.5} \sigma \sqrt{2} \end{aligned} \tag{12}$$

$$\theta_n^- = \theta_n^+ - 2\sigma \left(J_n (1 - P_c) \right)^{-0.5} \sigma \sqrt{2} \tag{13}$$

$$J_n = J_{n-1} + x_n^2, \quad J_0 = 0, \quad n \geq 1. \tag{14}$$

Equations (11)–(14) describe the recurrent algorithm intended for the calculation of values θ_n^+ , θ_n^- on each of the sequential plan (τ, d_n) steps. From the essence of a set problem we have:

$$\forall n \in \mathbb{N} : \min(|a|, |b|) \leq |x_n| \leq \max(|a|, |b|).$$

From here follows that all the statements of Theorem 1 are true for the algorithm (12)–(14) with

$$L = \max(|a|, |b|) \quad \text{and} \quad l = \min(|a|, |b|).$$

Now, let us assume that

$$\forall n \in \mathbb{N} : E\xi(t) = \sigma^2 \tag{15}$$

and $\{\xi_t(t)\}_{t \in \mathbb{N}}$ is a Gaussian sequence. In this case [3]

$$P(\theta^* \in [\beta_n - \gamma_n, \beta_n + \gamma_n]) \geq P_c$$

where $n \in \mathbb{N}$

$$\beta_n = \sum_{i=1}^n x_i y(x_i) \left(\sum_{i=1}^n x_i^2 \right)^{-1}, \quad \gamma_n = \sigma U(P_c) \left(\sum_{i=1}^n x_i^2 \right)^{-1},$$

and for a value $U(P_c)$ the following notation is possible

$$2\Phi(U(P_c)) - 1 = P_c$$

where $\Phi(\cdot)$ is the standard distribution function.

The confidence interval size for value θ^* is equal to $2\gamma_n$ in this case. For concrete values $n \in \mathbb{N}$, $P_c \in]0, 1[$, $(x_1, x_2 \dots x_n)$ we have the relatively large value

$$\varepsilon_n = ||\theta_n^+ - \theta_n^-| - 2\gamma_n|. \tag{16}$$

This is the peculiar charge for the security of a successful plan (d_n, τ) working with large noise distribution class.

If we can cancel the non-parametricity in the set problem, the value ε_n could be notably decreased. The distribution of the observation noise are supposed to be known in this case. Let us assume that variables $\{\xi(t)|t \in \mathbb{N}\}$ have Gaussian distributions and equation (15) is true. It is easy to see that almost all information about noise distributions is taken into account related to the choice of $\{c_n(\cdot)\}_{n \in \mathbb{N}}$. Substituting $(2/(1 - P_c))^{-0.5}$ into the other coefficient, a more complete accounting for the Gaussian nature of the noise distribution into the sequence $\{c_n(\cdot)\}_{n \in \mathbb{N}}$, it is possible to increase the effectiveness of the method.

From the proof of Theorem 1 follows that inequality (11) should be true for the sequence $\{c_n(\cdot)\}_{n \in \mathbb{N}}$. In the Gaussian case it is possible that

$$\forall c_n(\theta) = \left(\sum_{i=1}^n 2 \left[\frac{\partial f_i(x_i, \theta)}{\partial \theta} \right]^2 \right)^{0.5} U((1 + P_c)/2).$$

Indeed, it is easy to see that

$$\begin{aligned} \forall (\theta \in \Theta, n \in \mathbb{N}) : P(|\chi_n(\theta)| > c_n(\theta)) &= 2\Phi \left(\frac{c_n(\theta)}{\sqrt{E_{\theta^*} \chi_n^2(\theta)}} \right) - 1 \\ &= 2\Phi(U((1 + P_c)/2)) - 1 \\ &= (1 + P_c)/2, \end{aligned}$$

accounting for both the Gaussian nature of the value $\chi_n(\theta)$, $\theta \in \Theta$ and the truth of the following statement

$$\forall \theta \in \Theta : E_{\theta^*} \chi_n^2(\theta) = 2 \sum_{i=1}^n \left[\frac{\partial f_i(x_i, \theta)}{\partial \theta} \right]^2 n^{-2}.$$

Using the Boolean inequality we obtain (11).

In the linear regression case it can be written:

$$\varepsilon^* = \sigma (U((1 + P_c)/2) - U(P_c)) \left(\sum_{i=1}^n x_i^2 \right)^{-0.5}.$$

Value ε_n^* is approximately only one eighteenth of ε_n in (16) with $P_c = 0.95$. It is an excellent advantage!

4. A practical example

Let us check the Theorem statement for a case of log-linear model regression function widely used in economics. Thus, we consider that

$$y(x_t) = A \cdot \exp(\theta^* x_t) + \xi(t), \quad t \in \mathbb{N},$$

where $\theta^* \in [\alpha, \beta]$, $\alpha > 0$, $\beta < \infty$, $0 < A < \infty$, $\mathbb{N} = \{1, 2, \dots\}$ is an index set,

$$\forall t \in \mathbb{N} : x_t \in [C_1, C_2], \quad C_2 < \infty, \quad C_1 > 0$$

$$\forall t \in \mathbb{N} : [E\xi(t) = 0, E\xi^4(t) < \sigma^4 < \infty]$$

In this case it is easy to see that

$\forall (t \in \mathbb{N}; \theta \in [\alpha, \beta], x \in [C_1, C_2]) :$

$$C_1 \exp(\alpha C_1) < \left| \frac{\partial}{\partial \theta} (A \cdot \exp(\theta^* x)) \right| < C_2 \exp(\beta C_2) < \infty.$$

Thus, all the statements of the Theorem have been proved for the considered case.

4. Conclusions

The suggested algorithm enables us not only to build a confidence interval for the estimated parameter of the nonlinear regression but also to control the quality estimation in arbitrary observation time moments. These results are important for practice where the sample volume obtained by statisticians are always upper bounded. The solution has been obtained with nonparametric a priori uncertainty relative to the noise distributions with a limited fourth moment. So, the suggested plan of estimation is of working capacity in the case of arbitrary noise distribution with limited fourth moment.

References

1. Shiryayev, A. N., Probability. Moscow, Nauka, 1989
2. Marshall, A. W., Olkin, I., Inequalities: Theory and Its Applications, Academic Press, New York, 1979.
3. Kendall, M. G., Stuart, A., The advanced Theory of Statistics, vol. 2, Charles Griffing & Company Limited, London, 1966.

**Неасимптотическое решение
задачи доверительного оценивания параметра
нелинейной регрессии
с позиций последовательного анализа**

А. В. ТИМОФЕЕВ

(Томск)

В статье предлагается последовательный план, позволяющий за конечное число наблюдений в условиях априорной определенности относительно распределения

шумов построить доверительный интервал фиксированного размера для параметра, нелинейно входящего в уравнение регрессии. Границы доверительного интервала в каждый момент времени наблюдения определяются из решения двух оптимизационных задач на минимум некоторых функционалов специально выбранной структуры. Приводятся оценки сверху для среднего времени наблюдения в предложенном последовательном плане.

А. В. Тимофеев
НИИ автоматики и электромеханики,
Отдел оптимальных и адаптивных систем,
СССР, 634004, Томск, ул. Белинского, 53

ANALYSIS OF ADMISSIBLE PERTURBATIONS AND STABILIZATION OF UNCERTAIN DISCRETE-TIME PLANTS

S. V. EMELYANOV, P. V. ZHIVOGLYADOV, S. K. KOROVIN

(*Moscow*)

(Received June 12, 1990)

For stabilization of stationary and nonstationary discrete-time plants under compact uncertainty, solvability criteria are proved that are necessary or necessary and sufficient stabilizability conditions. For plants under stationary uncertainty a stabilization method is proposed which reduces design of feedback to finite choice of parameters through proper parametrization and quantification of uncertainty factors. Examples are provided.

1. Introduction

Robustness of dynamic systems in the face of variations of its operator has been moving to the top of research agenda. Unlike the problem of robust stability for a stationary linear system where the results are nearly exhaustive [1, 2, 6], stabilization remains an underexplored field. The conditions for solvability of this problem is the subject of this article.

Various aspects of control of uncertain discrete-time plants have been analyzed [3-5, 7, 8], but stabilization of a discrete-time dynamic plant under a compact, in particular, interval uncertainty has to be studied more thoroughly, especially as far as nonstationary plants are concerned.

The specific of the plant and assumptions on the features of uncertainties largely dictate the choice of the technique for design of feedback. Thus min-max control combined with recurrent estimation of unknown parameters has been used [7] for the stabilization of an uncertain stationary plant; also, a dividing manifold can serve the purpose [4]. Various stabilization techniques which call for accumulation of information have been proposed [3].

This article will address both stationary and nonstationary discrete-time dynamic plants. Perturbations must have certain asymptotic properties if an uncertain nonstationary plant is to be stabilized. Stabilizability criteria will be formulated and proved for various ways to describe the uncertainty. A localization method will be proposed for stabilization of plants under stationary uncertainty. Prop-

er parametrization and quantification of uncertainty factors reduces the design of feedback to a finite choice of parameters for which numerous efficient localization methods will be proposed.

Section 2 will state in formal terms the control problem. Stabilizability criteria for uncertain nonstationary plants are formulated in Section 3. Section 4 will describe the localization method and various ways to obtain feedbacks. Examples of applying the findings are provided in Section 5. The Appendix will contain proofs of Theorems and comments.

2. Problem statement

The discrete-time plants of this article are described by the formula

$$\begin{aligned} \Sigma_t(M, B, L) : x_{t+1} &= A(t)x_t + Bu_t + \xi_t, \\ A(t) \in M \subset \mathbf{R}^{n \times n}, \quad \xi_t \in L \subset \mathbf{R}^n, \quad t \in \mathbf{N}, \end{aligned} \quad (1)$$

where $x \in \mathbf{R}^n$; \mathbf{R} , \mathbf{N} are sets of real and natural numbers, respectively; $A(t)$ is the matrix of plant parameters; $u \in \mathbf{R}^m$ is the control vector; ξ_t is an unobservable disturbance; the pair $\{A, B\}$ is stabilizable for any matrix A from M ; the compact sets M and L reflect the uncertainty of the plant. The plant will be denoted as $\Sigma_t(M, B)$, if $\xi_t \equiv 0$ and $\Sigma(M, B, L)$ if $A(t) = A = \text{const}$.

It is required to:

1. Asymptotically stabilize the process in zero in the absence of an exogenous disturbance ($\xi_t = 0$), or to obtain feedback such that the origin of coordinates of \mathbf{R}^n is the globally asymptotically stable equilibrium position of a closed-loop system;
2. ε -stabilize the plant in the presence of an exogenous disturbance when feedback is such that in the vicinity of the origin of coordinates of \mathbf{R}^n there exists a global stable attractor whose size is dependent on the three-tuple (M, B, L) .

The stabilization problem is stated in the narrow sense in that the stabilizing feedback is to be linear and stationary.

3. Stabilization of nonstationary plants

It is required to stabilize the nonstationary plant $\Sigma_t(M, B, L)$. For convenience, equation (1) will be used also in its equivalent form

$$x_{t+1} = (A + \Delta A(t))x_t + Bu_t + \xi_t, \quad \xi_t \in L \subset \mathbf{R}^n,$$

where $A \in \mathbb{R}^{n \times n}$ is a known (nominal) matrix, $\Delta A(t) \in M \subset \mathbb{R}^{n \times n}$, $t \in \mathbb{N}$; $\Delta A(t)$ is the matrix of parametric perturbations.

The definitions below will specify the geometrical and asymptotic properties of the perturbations.

DEFINITION 1. The compact set $L \subset \mathbb{R}^n$, $n \in \mathbb{N}$ is symmetrizable if there exists a symmetry center of the convex hull $L^* \in \text{conv } L$, or for any $L^1 = (L^* + \Delta L) \in \text{conv } L$ the inclusion holds

$$L^2 = (L^* - \Delta L) \in \text{conv } L.$$

DEFINITION 2. If it is symmetrizable, the compact set $L \subset \mathbb{R}^n$, $n \in \mathbb{N}$ is even and for any $L^1 = (L^* + \Delta L) \in \text{conv } L$ the inclusions holds

$$L_i = (L^* + P_i \Delta L) \in \text{conv } L; \quad i = 1, \dots, 2^n$$

where $P_i \in \mathbb{R}^{n \times n}$ are matrices of the form

$$\text{diag}(\pm 1, \dots, \pm 1).$$

Because in addition to the geometrical properties of the set M the asymptotic behaviour of sequences made of elements of M is also important for this problem, the following definitions will also be useful.

DEFINITION 3. The sequence of matrices $S = \{S(t_0), S(t_0 + 1), \dots, S(t_0 + i), \dots\}$, $S(i) \in \mathbb{R}^{n \times n}$ is recurrently stable if the discrete-time process

$$z_{t+1} = S(t)z_t, \quad z \in \mathbb{R}^n; \quad t = t_0, t_0 + 1, \dots$$

is uniformly asymptotically stable, otherwise S is a recurrently unstable sequence.

The set of all sequences made of elements of the compact $M \subset \mathbb{R}^{n \times n}$ will be denoted as S_M .

DEFINITION 4. If any sequence $S \in S_M$ is recurrently stable, then so is M .

DEFINITION 5. The uncertain plant $\Sigma_t(M, B, L)$ is M -stabilizable (asymptotically M -stabilizable) if with any sequence $\Delta A(t) \in M$; $t = t_0, t_0 + 1, \dots$; $t_0 \geq 0$; the problem of its ε -stabilization (asymptotic stabilization) is solvable.

Definitions such as 1 and 2 are also applicable to the matrix set $M \subset \mathbb{R}^{n \times n}$. The most important stabilizability condition will be shown to be "proper" asymptotic behaviour of the parametric perturbations $\Delta A(t) \in M$; therefore without loss of generality, assume that $\xi_t \equiv 0$.

The following result is the most important for linear systems.

THEOREM 1 (RECURRENT STABILITY OF THE PERTURBATIONS). If for some even set $M \subset \mathbb{R}^{n \times n}$, $M^* = 0$ the uncertain plant

$$\Sigma_t(M, B) : x_{t+1} = (A + \Delta A(t))x_t + Bu_t, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}^m$$

is asymptotically M -stabilizable, then M is a recurrently stable set. The inverse is also true, if M is a recurrently unstable even set ($M^* = 0$), then the plant $\Sigma_t(M, B)$ is asymptotically M -nonstabilizable.

This Theorem is proved in the Appendix.

Theorem 1 confines parametric uncertainty to recurrently stable sets. If the compact set M is assumed to be only symmetrizable, a similar result is true for plants whose feedback is stationary and linear $u_t = Kx_t$.

To improve the usefulness of the recurrent stability condition, let us specify the form of a priori estimates of uncertainty.

3a. Interval estimates

For $\Sigma_t(M, B)$ under interval uncertainty

$$A(t) \in A_I \in I(\mathbb{R}^{n \times n}), \quad A_I = [A^-, A^+], \quad m(A) = \frac{1}{2}(A^+ + A^-),$$

$$W(A) = (A^+ - A^-), \quad w(a_{ij}) = w(a_{ij}^+) - w(a_{ij}^-) \geq 0, \quad W^* = \frac{1}{2}W(A),$$

where $m(A) \in \mathbb{R}^{n \times n}$ and $W(A) \in \mathbb{R}^{n \times n}$ are matrices of mean values and width of the interval matrix A_I and $I(\mathbb{R}^{n \times n})$ is the set of real interval $n \times n$ matrices holds

THEOREM 2. Stability of the matrix W^* is the necessary condition for the asymptotic A_I -stabilizability of $\Sigma_t(A_I, B)$.

Using $m(A)$ as the matrix of nominal values and denoting $\Delta A(t) = A(t) - m(A)$ it follows from the inequality $|A(t) - m(A)| \leq W^*$ which holds at any $t \in \mathbb{N}$ that any sequence of the perturbations is recurrently stable.

3b. Metric estimates

If uncertainty of $\Sigma_t(M, B)$ is specified in the form

$$\Delta A(t) \in Q = \{\Delta A \in \mathbb{R}^{n \times n} : \|\Delta A\| \leq \alpha, \alpha > 0\}, \quad t \in \mathbb{N}$$

where $\|\Delta A\| = \max_{\|y\|=1} \|\Delta A y\|$ is the operator norm, then the necessary stabilizability condition is proved by

THEOREM 3. The condition $0 < \alpha < 1$ is necessary for asymptotic Q -stabilizability of $\Sigma_t(Q, B)$.

It follows from Theorem 3 that with $\alpha \geq 1$ for any stabilization control $u(x_t)$ at least one strategy of the perturbations behaviour can be obtained, in particular

$$\Delta A(t) = \begin{cases} \alpha I & \text{if } (A + \alpha I)x_t + Bu_t \notin \beta M(x_t), \\ -\alpha I & \text{if } (A - \alpha I)x_t + Bu_t \notin \beta M(x_t), \\ -\alpha I, \alpha I & \text{in the remaining cases,} \end{cases}$$

then the above conditions of recurrent stability of the perturbations become necessary and sufficient stabilizability conditions. For convenience, let us use an equivalent presentation of equation (2) as an autoregression equation

$$x_{t+1} = A_t X_t + u_t; \quad t = t_0, \quad t_0 + 1, \dots,$$

where $x \in \mathbb{R}$, $u \in \mathbb{R}$, $X_t = (x_{t-n+1}, \dots, x_t)^T$ is a vector from \mathbb{R}^n ; $A_t = (a_{n-1}(t), \dots, a_0(t))$ is a vector of plant parameters; $A_t^T \in C \subset \mathbb{R}^n$, $t \in \mathbb{N}$ and C is a compact symmetrizable set which reflects uncertainty of the plant description.

Denote

$$C_1^\wedge = \{C^\wedge \in \mathbb{R}^n; C^\wedge = A^T - C^*, A^T \in C\}, \quad M = \left| \begin{array}{c} 0 \quad \vdots \quad I_{n-1} \\ \hline (C_1^\wedge)^T \end{array} \right|,$$

where C^* is the symmetry center of $\text{conv } C$.

The solvability condition of the C -stabilization problem is provided by

THEOREM 4. With $C \subset \mathbb{R}^n$ being a compact even set, the uncertain plant $\Sigma_i^c(C, b)$ is asymptotically C -stabilizable iff M is a recurrently stable set.

Recurrent stability of the set M is equivalent to asymptotic stability of the plant $\Sigma_i^c(C, b)$ having linear feedback

$$u^*(X_t) = -(C^*)^T X_t \tag{3}$$

which is the best in a certain sense. In particular, nonstabilizability of the plant by $u^*(X_t)$ suggests essential C -nonstabilizability.

In the case of independent (interval) perturbations of parameters of $\Sigma_i^c(C, b)$ use the following notation

$$\begin{aligned} A^T \in C &= C_I \in I(\mathbb{R}^n), \\ a_i(t) &\in \bar{c}_i = [c_i^-, c_i^+], \\ m(c_i) &= \frac{1}{2}(c_i^+ + c_i^-), \\ w(c_i) &= (c_i^+ - c_i^-) \geq 0, \quad i = 0, 1, \dots, n-1 \end{aligned}$$

$m(c_i)$, $w(c_i)$ being the mean and width, respectively, of the interval number \bar{c}_i . Let us compile a polynomial of the form

$$G(z) = z^n + \beta_0 z^{n-1} + \dots + \beta_{n-1}, \quad \beta_i = -\frac{1}{2}w(c_i); \quad i = 0, 1, \dots, n-1.$$

The following Theorem is true.

THEOREM 5. Asymptotic stability of the polynomial $G(z)$ is the necessary and sufficient condition of asymptotic C_I -stabilizability of $\Sigma_i^c(C_I, b)$.

Because the parallelepiped representing interval uncertainty is an even set in the parameter space, Theorem 5 is proved by noting that the polynomial $G(z)$ defines the comparison system for $\Sigma_i^{\xi}(C_I, b)$ whose feedback is

$$u^*(X_t) = -(C^*)^T X_t = - \sum_{i=0}^{n-1} m(c_i) x_{t-i}.$$

This Theorem provides an exhaustive solution of the stabilization problem for a discrete interval plant having a canonical form. In obtaining $u^*(X_t)$, for an arbitrary $Y \in \mathbb{R}$, $Y \neq 0$ the optimization problems are solved

$$\mathbf{A}Y \rightarrow \min_{\mathbf{A}}, \quad \mathbf{A}Y \rightarrow \max_{\mathbf{A}}, \quad \mathbf{A}^T \in C.$$

The minimum and maximum of the function $f(\mathbf{A}) = \mathbf{A}Y$ are obtained with certain values of $(\mathbf{A}^-)^T \in C$ and $(\mathbf{A}^+)^T \in C$. The set C is even and its center of symmetry is provided by the equality $C^* = \frac{1}{2}(\mathbf{A}^+ + \mathbf{A}^-)^T$ while the stabilizing control is

$$u^*(X_t) = -\frac{1}{2}(\mathbf{A}^+ + \mathbf{A}^-)X_t.$$

These results are extended to the case of $\xi_t \neq 0$ while asymptotic stability of the closed-loop system is replaced by dissipativity.

4. Stabilization of stationary plants

In the problem of stabilizing an uncertain stationary plant $\Sigma(M, B, L)$ the perturbations must satisfy less restrictive constraints. The conditions below make it possible to obtain asymptotic stabilization or ϵ -stabilization of $\Sigma(M, B, L)$ with any intensity of exogenous disturbances by using a localization method.

For convenience, the plants will be scalar

$$\Sigma(M, b, L) : x_{t+1} = Ax_t + bu_t + \xi_t, \quad A \in M \subset \mathbb{R}^{n \times n}, \quad \xi_t \in L \subset \mathbb{R}^n, \quad t \in \mathbb{N},$$

where $x \in \mathbb{R}^n$; $u \in \mathbb{R}$ is scalar control; and ξ is an unobservable disturbance.

4a. The approach

On M define a finite decomposition into possibly intersecting sets $M_i \subset \mathbb{R}^{n \times n}$, or $M = \bigcup_{i=0}^S M_i$, and assume a parametric family of feedback functions $u(K, x)$, $K \in \Omega$. Under certain conditions M and its decomposition induce on the set Ω a

subset $\Omega^M \subset \Omega$ and its decomposition $\Omega^M = \bigcup_{i=0}^S \Omega_i$ so that for any pairs $\{K, A\}$, $K \in \Omega_i, A \in M_i$ the closed-loop system

$$x_{t+1} = Ax_t + bu(K, x_t) + \xi_t, \quad \xi_t \in L$$

has a globally stable attractor in the vicinity of the origin of coordinates of \mathbb{R}^n . In this procedure the parametrization and quantification reduce the feedback design to a finite problem of choice of parameters. $M_r = \bigcup_i M_i$ will denote the union of sets such that $A \in M_i$ and, consequently, $\Omega_r = \bigcup_i \Omega_i$. The plant can be stabilized if a finite number of measurements of the state path suggests some element $K \in \Omega_r$. This technique will be referred to as localization whose efficiency depends on “proper” decomposition of M , choice of the parametric set $u(K, x)$, and the localization procedure, or the measurement technique and ways to use information for finding elements of Ω_r .

4b. Localization methods

Let us have a look at localization methods in which the quadratic and linear forms of the state vector are measured and a comparison system is used.

DEFINITION 6. For positive definite matrix $H^T = H$ and number $\varepsilon > 0$ the plant $\Sigma(M, b, L)$ is (H, ε) -quadratically stabilizable by feedback $u(x)$ with any $A \in M$ and $\xi_t \in L$ if

- 1) there exists a globally stable attractor $\mathbf{A}_\varepsilon \in B_\varepsilon(0)$;
- 2) for every x_0 there exists a time $t_\varepsilon(x_0)$ such that for $x_t(x_0) \notin \mathbf{A}_\varepsilon$ at $t > t_\varepsilon(x_0)$ the inequality holds

$$\langle (Ax_t + bu(x_t) + \xi_t), H(Ax_t + bu(x_t) + \xi_t) \rangle - \langle x_t, Hx_t \rangle < 0.$$

Here $B_\varepsilon(0) = \{x \in \mathbb{R}^n, \langle x, Hx \rangle \leq \varepsilon\}$.

Let $\|A\|_H = \|S^{-1}AS\|$, $S^T S = H$ is the H -norm of matrix A and ∂M is the boundary of M . The solvability condition of the (H, ε) -quadratic stabilization problem is provided by

THEOREM 6. The plant $\Sigma(M, b, L)$ is (H, ε) -quadratically stabilizable with some $\varepsilon > 0$ iff

$$\max_{A \in \partial M} \|(I - (b^T H b)^{-1} b b^T H)A\|_H \leq 1. \tag{4}$$

Without loss of generality it will be assumed hereafter that $M = \text{conv}\{A_i\}_1^m$, $A_i \in \mathbb{R}^{n \times n}$; $i = 1, \dots, m$ are known matrices. Because feedback $u_t = -(b^T H b)^{-1} b^T H A x_t$

is optimal in terms of decrease of the quadratic form $\langle x, Hx \rangle$ on the plant paths $x_{t+1} = Ax_t + bu_t$ it follows from the inequalities

$$\|(I - (b^T Hb)^{-1}bb^T H) \text{conv}\{A_i\}_1^m\|_H \leq \max_{1 \leq i \leq m} \|(I - (b^T Hb)^{-1}bb^T H)A_i\|_H < 1$$

that linear feedbacks

$$u(K, x) = Kx, \quad K \in \Omega^M = \text{conv}\{K_i\}_1^m, \quad K_i = -(b^T Hb)^{-1}b^T HA_i; \quad i = 1, \dots, m$$

would be sufficient with subsequent localization of the vector K from Ω^M .

M is decomposed as follows

$$M = \bigcup_{i=0}^s M_i, \quad M_i = \left\{ \begin{array}{l} A \in \mathbf{R}^{n \times n}; \quad A = A^i + \Delta A, \quad A^i = \sum_{j=1}^m \mu^{ij} A_j, \\ \Delta A = \sum_{j=1}^m \Delta \mu_j A_j, \quad |\Delta \mu_j| \in [-r, r], \quad j = 1, \dots, m \end{array} \right\} \quad (5)$$

where $r > 0$ is such that for any $|\Delta \mu_j| \leq r, j = 1, \dots, m$:

$$(\tilde{A}_i + \sum_{j=1}^m \Delta \mu_j A_j)^T H (\tilde{A}_i + \sum_{j=1}^m \Delta \mu_j A_j) - H \quad i = 1, \dots, m$$

where $\tilde{A}_i = (I - (b^T Hb)^{-1}bb^T H)A_i, 0 < \beta^* < 1 - \max_{1 \leq i \leq m} \|\tilde{A}_i\|_H$ and the basic vectors of the decomposition $\mu^i = (\mu^{i1}, \dots, \mu^{im})^T$ can be computed recurrently.

Assume for simplicity that $1/r$ is an integer; then the basic vectors $\mu^i = (\mu^{i1}, \dots, \mu^{im})^T, i = 0, \dots, s$ can be obtained by the recurrent procedure

$$\begin{aligned} \mu^0 &= \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{vmatrix}, \quad \mu^1 = \mu^0 + \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ r \\ -r \end{vmatrix}, \dots, \mu^{1/r} = \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 1 \\ 0 \end{vmatrix}, \quad \mu^{1/r+1} = \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ r \\ 0 \\ 1-r \end{vmatrix}, \\ \mu^{1/r+2} &= \mu^{1/r+1} \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ r \\ -r \end{vmatrix}, \dots, \mu^{2/r+1} = \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ r \\ 1-r \\ 0 \end{vmatrix}, \quad \mu^{2/r+2} = \begin{vmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ r+1 \\ 0 \\ -r \end{vmatrix}, \dots, \end{aligned}$$

$$\mu^{s-1} = \begin{vmatrix} r \\ 1-r \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{vmatrix}, \quad \mu^s = \mu^{s-1} + \begin{vmatrix} r \\ -r \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{vmatrix} = \begin{vmatrix} 1 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{vmatrix}.$$

Let $L = \{\xi \in \mathbb{R}^n; \langle \xi, H\xi \rangle \leq \alpha, \alpha > 0\}$. For localization on a finite set

$$\mathbf{K} = \{\mathbf{K}_i\}_0^s, \quad \mathbf{K}_i = -(b^T Hb)^{-1} b^T H A^i; \quad i = 0, \dots, s$$

use the function

$$\varphi_t(K) = \langle (A + bK)x_t + \xi_t, H((A + bK)x_t + \xi_t) \rangle - (1 - \beta^*) \langle x_t, Hx_t \rangle - \alpha$$

which is equal to the difference between the decrease of the quadratic form $\langle x, Hx \rangle$ on the paths of $\Sigma(M, b, L)$ with feedback $u_t = Kx_t$ and an estimate of minimal decrease of this form among all systems of the form

$$x_{t+1} = (\tilde{A}^i + \Delta A)x_t + \xi_t, \quad \xi_t \in L, \quad \Delta A = \sum_{j=1}^m \Delta \mu_j A_j, \quad |\Delta \mu_j| \leq r, \quad j = 0, \dots, s.$$

By $u_t^-, u_t^+, (u_t^- < u_t^+)$ let us denote real (by virtue of (4)) zeros of the function $\varphi_t(K)$. Then holds

THEOREM 7. When $\Sigma(M, b, L)$ satisfies the following inequality condition $\max_{A \in \partial M} \|(I - (b^T Hb)^{-1} b b^T H A)\|_H < 1$ with some matrix $H = H^T > 0$, there are positive numbers β^*, α, r and s such that feedback $u(x)$ from the set

$$H(u) = \begin{cases} u_t = \mathbf{K}^t x_t, \\ \mathbf{K}^t \in \Omega_t(K) = \Omega_{t-1}(K) \cap \tilde{\Omega}_t(K), \quad \Omega_0(K) = \mathbf{K}, \\ \tilde{\Omega}_t(K) = \{K^T \in \mathbb{R}^n; u_{t-1}^- \leq Kx_{t-1} \leq u_{t-1}^+\}, \\ u_{t-1}^\pm = u_{t-1} - (b^T Hb)^{-1} b^T Hx_t \pm ((b^T Hb)^{-2} (b^T Hx_t)^2 - \\ \quad - (b^T Hb)^{-1} (\langle x_t, Hx_t \rangle - (1 - \beta^*) \langle x_{t-1}, Hx_{t-1} \rangle - \alpha))^{\frac{1}{2}} \end{cases} \quad (6)$$

(H, ε) -stabilizes the plant $\Sigma(M, b, L)$ and for arbitrary initial conditions the inequality

$$\langle x_{t+1}, Hx_{t+1} \rangle - \langle x_t, Hx_t \rangle \leq -\beta^* \langle x_t, Hx_t \rangle + \alpha$$

is disturbed a maximum of a finite number of times.

Dissipativity of a closed-loop system follows from the fact that the sequence $\Omega_t(K); t = 0, 1, \dots$ in (6) has the following properties:

- 1) $\Omega_t(K) \subseteq \Omega_{t-1}(K) \subseteq \dots \subseteq \Omega_0(K)$,
- 2) $\Omega_{t+1}(K) \subset \Omega_t(K)$ if $\varphi_t(\mathbf{K}^t) > 0$, $\mathbf{K}^t \in \Omega_t(K)$, $u_t = \mathbf{K}^t x_t$.

Let us consider a localization method in which the linear form $\sigma = cx$ is measured. Let

$$G = \{x \in \mathbb{R}^n; \sigma = cx = 0\},$$

$$G_\Delta = \{x_t \in \mathbb{R}^n; |\sigma_t| \leq \Delta \|x_{t-1}\| + c_0, \Delta > 0, c_0 > 0\}.$$

DEFINITION 7. The set G is the stabilizing set of $\Sigma(M, b)$ if for any matrix $A \in M$ there exists feedback $u(x)$ such that the equation $x_{t+1} = Ax_t + bu(x_t)$ is asymptotically stable in zero with $x_t \in G$ at every $t > 0$

DEFINITION 8. The set G_Δ is the stabilizing set of $\Sigma(M, b, L)$ if for any matrix $A \in M$ there are numbers $\Delta > 0$, $c_0 \geq 0$ and $\varepsilon(x_0) > 0$ and feedback $u(x)$ such that every solution $x_t(x_0)$ of the equation $x_{t+1} = Ax_t + bu(x_t) + \xi_t$, $\xi_t \in L$ belongs to the sphere $B_{\varepsilon(x_0)}(0) = \{x \in \mathbb{R}^n; \langle x, x \rangle \leq \varepsilon(x_0)\}$ with $x_t \in G_\Delta$ at every $t > 0$.

DEFINITION 9. $\Sigma(M, b)$ is globally G -stabilizable by feedback $u(x)$ with any matrix $A \in M$ if the set G is

- 1) the stabilizing set of $\Sigma(M, b)$ and
- 2) finitely attracting set of the closed-loop control system.

DEFINITION 10. $\Sigma(M, b, L)$ is globally G_Δ -stabilizable by $u(x)$ with any $A \in M$ and $\xi_t \in L$ if there are numbers $\Delta > 0$ and $c_0 \geq 0$ such that the set G_Δ is (1) the stabilizing set of $\Sigma(M, b, L)$ and (2) the finitely attracting set of the closed-loop control system.

Let A_μ be an arbitrary matrix of the parametric family $M = \{A \in \mathbb{R}^{n \times n}; A = A_\mu = \sum_{i=1}^m \mu_i A_i, \sum_{i=1}^m \mu_i = 1, 0 \leq \mu_i \leq 1\}$. When the feedback is a member in the parametric family $u_\lambda = -(cb)^{-1}cA_\lambda x$ the closed-loop system equation with $\xi_t = 0$ takes the form

$$x_{t+1} = PA_\mu x_t - (cb)^{-1}bc \sum_{i=1}^m (\lambda_i - \mu_i) A_i x_t, \quad P = (I - (cb)^{-1}bc). \quad (7)$$

It follows from the equality $\sigma_{t+1} = -\sum_{i=1}^m (\lambda_i - \mu_i) cA_i x_t$ that the linear form $\sigma = cx$ contains information on the unknown vector $\mu = (\mu_1, \dots, \mu_m)^T$; besides, if $|\lambda_{\max}(PA\mu)| < 1$, then there is a number $\Delta > 0$ such that with $|\sigma_{t+1}| \leq \Delta \|x_t\|$ equation (7) is asymptotically stable. In the subsequent discussion $\Sigma(M, b, L)$ are such that $cb \neq 0$, $\max_{A \in M} |\lambda_{\max}(PA)| < 1$.

Because the properties of the linear system are continuously dependent on the parameters, there are positive numbers $\Delta, \alpha_1, \alpha_2, r, c_0 \geq \max_{\xi \in L} |c\xi|$ such that

for any $A_\mu \in M$, $|\lambda_i - \mu_i| = |\Delta\mu_i| \leq r$; $i = 1, \dots, m$ every solution of the equation $x_{t+1} = (PA_\mu + \sum_{i=1}^m \Delta\mu_i A_i)x_t + \xi_t$, $\xi_t \in L$ at $t > 0$ belongs to the stabilizing set G_Δ for which the inequality holds $\langle (PA_\mu + \sum_{i=1}^m \Delta\mu_i A_i)x_t + \xi_t, U((PA_\mu + \sum_{i=1}^m \Delta\mu_i A_i)x_t + \xi_t) \rangle - \langle x_t, Ux_t \rangle \leq -\alpha_1 \langle x_t, x_t \rangle + \alpha_2$ and $U = U^T > 0$ is the solution of the Lyapunov equation $A_\mu^T P^T U P A_\mu - U = -\alpha_1 I$.

Note that if $\Sigma(M, b, L)$ is (H, ε) -quadratically stabilizable with some matrix $H = H^T > 0$, then it is sufficient that the vector $c^T \in \mathbb{R}^n$ in linear form $\sigma = cx$ be chosen in the form $c = b^T H$ and the condition $\max_{A \in M} |\lambda_{\max}(PA)| < 1$ holds by virtue of (4).

For instance, if $A = A_0 + b\Delta A$, $\Delta A^T \in Q \subset \mathbb{R}^n$ where Q is an arbitrary compact set and A_0 is a known matrix, there is certainly a matrix $H = H^T > 0$ such that the pair $\{A_0, b\}$ is H -quadratically stabilizable and from the equality $P(A_0 + b\Delta A) = PA_0$ which holds for any $\Delta A^T \in Q$ follows the truth of the condition $\max_{A \in M} |\lambda_{\max}(PA)| < 1$ with $c = b^T H$. Consequently, M can be decomposed as in equation (5). When the localizing function

$$\varphi_t(K) = |c(A + bK)x_t + c\xi_t| - \Delta \|x_t\| - c_0$$

is used, holds

THEOREM 8. If for $\Sigma(M, b, L)$ and some vector $c^T \in \mathbb{R}^n$, ($cb \neq 0$) the inequality $\max_{A \in M} |\lambda_{\max}(PA)| < 1$ holds where $P = (I - (cb)^{-1}bc)$, positive numbers Δ , c_0 , r , α_1 , α_2 , s exist such that feedback $u(x)$ from the set

$$\zeta_\Delta(u) = \begin{cases} u_t = K^t x_t, \\ K^t \in \Omega_t(K) = \Omega_{t-1}(K) \cap \tilde{\Omega}_t(K), \quad \Omega_0(K) = K, \\ \tilde{\Omega}_t(K) = \{K^T \in \mathbb{R}^n; u_{t-1}^- \leq Kx_{t-1} \leq u_{t-1}^+\}, \\ u_{t-1}^\pm = u_{t-1} - (cb)^{-1}\sigma_t \pm |(cb)^{-1}|(\Delta \|x_{t-1}\| + c_0), \end{cases}$$

globally G_Δ -stabilizes $\Sigma(M, b, L)$ and there exists a matrix $U = U^T > 0$ such that the inequality

$$\langle x_{t+1}, Ux_{t+1} \rangle - \langle x_t, Ux_t \rangle \leq -\alpha_1 \langle x_t, x_t \rangle + \alpha_2$$

is disturbed on solutions of a closed-loop system a maximum of a finite number of times.

For G -stabilization of $\Sigma(M, b)$ which is a limit case of G_Δ -stabilization, there exists a unique feedback $u_t = -(cb)^{-1}cAx_t$ which ensures motion in the manifold

$\sigma = cx = 0$. Let $\zeta(u)$ denote the set of feedbacks which result from $\zeta_\Delta(u)$ as a consequence of a limit transition as $\Delta \rightarrow 0$ and $c_0 \rightarrow 0$:

$$\zeta(u) = \begin{cases} u_t = K^t x_t, \\ K^t \in \Omega_t(K) = \Omega_{t-1}(K) \cap \tilde{\Omega}_t(K), \\ \Omega_0(K) = \{K^T \in \mathbb{R}^n; K = -(cb)^{-1}cA, A \in M\}, \\ \tilde{\Omega}_t(K) = \{K^T \in \mathbb{R}^n; Kx_{t-1} = u_{t-1} - (cb)^{-1}\sigma_t\}, \end{cases}$$

Solvability of the G -stabilization problem is defined by

THEOREM 9. If for some vector $c^T \in \mathbb{R}^n$, ($cb \neq 0$) the inequality condition $\max_{A \in M} |\lambda_{\max}(PA)| < 1$ holds, then $\Sigma(M, b)$ with $u_t = u(x_t) \in \zeta(u)$ is asymptotically stable and any motion path of the closed-loop system belongs to the hyperplane $\sigma = 0$ with a possible exception of a finite number of points $q \leq n$ where n is the system dimension.

Localization with the use of a comparison system is a convenient tool if the parameters and exogenous disturbances of $\Sigma(M, b, L)$ satisfy the matching conditions, namely,

$$A = A_0 + b\Delta A, \xi_t = b\check{\xi}_t, \Delta A^T \in Q = \text{conv}\{Q_i\}_1^m, |\check{\xi}_t| \leq \xi, \xi \geq 0,$$

where $\{A_0, b\}$ is the controllable pair and $Q_i \in \mathbb{R}^n; i = 1, \dots, m$ are known vectors. Such a plant is described by a scalar n -th order difference equation

$$y_{t+1} = A_\mu Y_t + u_t + \check{\xi}_t,$$

where $y \in \mathbb{R}, Y_t = (y_t, \dots, y_{t-n+1})^T$ is a vector from $\mathbb{R}^n, A_\mu = \sum_{i=1}^m \mu_i A_i, \sum_{i=1}^m \mu_i = 1, 0 \leq \mu_i \leq 1, A_i^T \in \mathbb{R}^n$; a convex combination of vectors $A_i; i = 1, \dots, m$ is in one-to-one correspondence with a convex combination of matrices $A_0 + bQ_i^T; i = 1, \dots, m$. Introduce the following exponentially stable comparison system

$$z_{t+1} = BZ_t + \Psi_t,$$

where $z \in \mathbb{R}, Z_t = (z_t, \dots, z_{t-n+1})^T, B = (\beta_0, \dots, \beta_{n-1})$ is a vector of positive parameters, and $|\check{\xi}_t| \leq \Psi_t < \infty$ is the majorant of disturbance. The stabilization problem is understood in the following sense.

DEFINITION 11. The plant

$$y_{t+1} = A_\mu Y_t + u_t + \check{\xi}_t, A_\mu \in \text{conv}\{A_i\}_1^m, |\check{\xi}_t| \leq \xi$$

is globally B -stabilizable by $u(x)$ with any $A_\mu \in \text{conv}\{A_i\}_1^m, \check{\xi}_t \in [-\xi, \xi]$ if

- 1) there exists a stable attractor whose size is dictated by the pair $\{B, \xi\}$; and

- 2) for every Y_0 there is a time $t(Y_0)$ such that at every $t \geq t(Y_0)$ the magnitude of the solution y_t is majored by the associated solution of the comparison system, or $|y_t| \leq z_t$ with $Z_{t(Y_0)} = |Y_{t(Y_0)}|$.

When the control is $u_\lambda = -A_\lambda Y$ the closed-loop system equation is

$$y_{t+1} = \sum_{i=1}^m (\mu_i - \lambda_i) A_i Y_t + \tilde{\xi}_t.$$

Assume that the set $\text{conv}\{A_i\}_1^m = \bigcup_{i=0}^S J_i$ is decomposed as follows $J_i = \{A^T \in \mathbb{R}^n; A = A^i + \Delta A, \Delta A = \sum_{i=1}^m \Delta\mu_i A_i, \Delta\mu_i \in [-r_i, r_i]\}$, where the parameters $r_i > 0$ in the inequalities $|\Delta\mu_i| = |\mu_i - \lambda_i| \leq r_i; i = 1, \dots, m$ are such that the inequality holds $|\sum_{i=1}^m \Delta\mu_i A_i| \leq B$. For localization on the set $\Omega_0(K) = \{-A^i\}_0^s$ it would be sufficient to use the localizing function $\varphi_t(K) = |(A_\mu - K)Y_t + \tilde{\xi}_t| - B|Y_t| - \Psi_t$. Then the set of feedback takes the form

$$B(u) = \begin{cases} u_t = K^t Y_t, \\ K^t \in \Omega_t(K) = \Omega_{t-1}(K) \cap \tilde{\Omega}_t(K), \Omega_0(K) = \{-A^i\}_0^s, \\ \tilde{\Omega}_t(K) = \{K^T \in \mathbb{R}^n; u_{t-1}^- \leq K Y_{t-1} \leq u_{t-1}^+\}, \\ u_{t-1}^\pm = u_{t-1} - y_t \pm (B|Y_t| + \Psi_t). \end{cases}$$

5. Examples

Example 1. For an uncertain plant

$$\begin{vmatrix} x_{t+1}^1 \\ x_{t+1}^2 \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ a_1(t) & a_0(t) \end{vmatrix} \begin{vmatrix} x_t^1 \\ x_t^2 \end{vmatrix} + \begin{vmatrix} 0 \\ 1 \end{vmatrix} u_t$$

where $a_1(t) \in [-\varepsilon, \varepsilon], a_0(t) \in [0.2, 1.4], t \in \mathbb{N}, \varepsilon \geq 0$ the polynomial $G(z) = z^2 - \frac{1}{2}\omega(a_0)z - \frac{1}{2}\omega(a_1) = z^2 - 0.6z - \varepsilon$ has roots $z_1 = 0.3 + \sqrt{0.09 + \varepsilon}, z_2 = 0.3 - \sqrt{0.09 - \varepsilon}$ and by Theorem 5 the plant is stabilizable iff $0 \leq \varepsilon < 0.4$.

Example 2. In developing a localization algorithm for an (H, ε) -quadratically stabilizable plant

$$x_{t+1} = A_\mu x_t + b u_t + \xi_t, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R}$$

$$A_\mu \in M = \text{conv}\{A_1, A_0\}, \quad \|\xi\|_H \leq \alpha, \quad \alpha > 0,$$

because $A_\mu = A_0 + \mu(A_1 - A_0)$, $\|\tilde{A}_\mu\|_H < 1$ for any $\mu \in [0, 1]$ let us obtain estimates of the parameters δ_0 and δ_1 in the inequalities $|\Delta\mu_i| \leq \delta_i$, $i = 0, 1$, with which

$$\begin{aligned} (\tilde{A}_i + \Delta\mu_i(A_1 - A_0))^T H(\tilde{A}_i + \Delta\mu_i(A_1 - A_0)) - H &\leq -\beta_i H, \\ 0 < \beta_i < 1 - \|\tilde{A}_i\|_H; \quad i = 0, 1. \end{aligned}$$

Specify a decomposition

$$M = \bigcup_{i=0}^{s+1} M_i, \quad M_i = \{A; A = A_\mu, \mu^i \leq \mu \leq \mu^{i+1}, \mu^i = \text{sat}(i\delta)\}$$

where $s = [1/\delta]$, $\delta = \min\{\delta_0, \delta_1\}$, $\text{sat}(\cdot)$ is the saturation function. With a localizing function

$$\varphi_t(\lambda) = \langle (A_\mu + bK_\lambda)x_t + \xi_t, H((A_\mu + bK_\lambda)x_t + \xi_t) \rangle - (1 - \beta^*)\langle x_t, Hx_t \rangle - \alpha$$

where $\beta^* = \min\{\beta_0, \beta_1\}$, $K_\lambda = -(b^T H b)^{-1} b^T H A_\lambda$, the localization algorithm has the form

$$\begin{cases} u_t = -(b^T H b)^{-1} b^T H A_\lambda, \\ \lambda^t \in \Omega_t(\mu) = \Omega_{t-1}(\mu) \cap [\lambda_t^1, \lambda_t^2], \quad \Omega_0(\mu) = \{\mu^i\}_0^{s+1}, \end{cases} \quad (8)$$

where λ_t^1, λ_t^2 are zeros of the function $\varphi_t(\lambda)$.

Example 3. Let $x_{t+1} = (A_0 + b\Delta A)x_t + bu_t + \xi_t$, $\Delta A \in \text{conv}\{Q_1, Q_0\}$ and A_0 be a known matrix. For an arbitrary matrix $H = H^T > 0$ such that the pair $\{A_0, b\}$ is H -quadratically stabilizable choose a vector c in a linear form $\sigma = cx$ so that $c = b^T H$. Let $H = I$, $\|\xi_t\| \leq \alpha$, $\alpha > 0$ and $P = (I - (cb)^{-1}bc)$. Estimate the parameter Δ in the inequality $|\sigma_{t+1}| \leq \Delta\|x_t\|$ so that every solution of the equation $x_{t+1} = PA_0x_t + (cb)^{-1}b\sigma_{t+1}$ satisfies the inequality

$$\langle x_{t+1}, x_{t+1} \rangle - \langle x_t, x_t \rangle \leq -\beta\langle x_t, x_t \rangle, \quad 0 < \beta < 1 - \|PA_0\|.$$

To do this, it is sufficient that Δ is chosen from the equality $0 < \Delta < \beta\|b\|$. Because $A_\mu = (A_0 + bQ_0) + \mu b(Q_1 - Q_0)$, by selecting the decomposition step δ so that $0 < \delta \leq \Delta\|cb(Q_0 - Q_1)\|^{-1}$ the set $\Omega_0(\mu) = \{\mu^i\}_0^{s+1}, \mu^i = \text{sat}(i\delta); i = 0, \dots, s + 1$ is obtained.

If $\varphi_t = |c(A_\mu + bK_\lambda)x_t + c\xi_t| - \Delta\|x_t\| - c_0$, where

$$K_\lambda = -(cb)^{-1}cA_\lambda, \quad c_0 \geq \max_\xi |c\xi|$$

then the relation (8) where λ_t^1 and λ_t^2 are replaced by zeros of the function $\varphi_t(\lambda)$ may be used for localization.

6. Conclusion

In the problem of robust stabilization for a nonstationary uncertain discrete-time plant the admissible parametric perturbations have to satisfy requirements, much more restrictive than in continuous stabilization. This is true, in particular, of the asymptotic and geometric features. In these terms the necessary and sufficient stabilizability conditions are formulated.

The problem of stabilizing a stationary discrete-time process is solved by a newly proposed and refined localization method which reduces the design of feedback to a finite problem of parameters choice by proper parametrization and quantification of uncertainty factors.

Appendix

Proof of Theorem 1. Assume that there exists a recurrently unstable sequence of the perturbations $S = \{S(t_0), S(t_0 + 1), \dots, S(t_0 + i), \dots\}$, $S \in S_M$. Introduce a majoring sequence $S^* = \{|S(t_0)|, |S(t_0 + 1)|, \dots, |S(t_0 + i)|, \dots\}$ which is recurrently unstable and, because the set $M \in \mathbb{R}^{n \times n}$ is even, is nonempty, or $S^* \in S_M$. The key point of the proof is that by virtue of nonstationarity of the plant and evenness of M for any $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $t \in \mathbb{N}$, $t \geq t_0$ there is a matrix $\Delta A(t) \in M$ such that

$$|x_{t+1}| = |\Delta A(t)| \cdot |x_t| + |Ax_t + Bu_t|.$$

The operation $|\cdot|$ is understood component-wise. Without loss of generality assume that $t_0 = 0$. In a recurrent form the sequence

$$\begin{aligned} |x_1| &= |\Delta A(0)||x_0| + |Ax_0 + Bu_0|, \\ |x_2| &= |\Delta A(1)| \cdot |\Delta A(0)| \cdot |x_0| + |\Delta A(1)| \cdot |Ax_0 + Bu_0| + |Ax_1 + Bu_1|, \\ &\vdots \\ |x_N| &= \prod_{i=0}^{N-1} |\Delta A(i)||x_0| + \prod_{i=1}^{N-1} |\Delta A(i)||Ax_0 + Bu_0| + \dots + |Ax_{N-1} + Bu_{N-1}|. \end{aligned}$$

As $N \rightarrow \infty$ going to the limit we have

$$\begin{aligned} \lim_{N \rightarrow \infty} |x_N| &= \lim_{N \rightarrow \infty} \prod_{i=1}^{N-1} |\Delta A(i)||x_0| + \lim_{N \rightarrow \infty} \Delta_N, \\ \Delta_N &= \prod_{i=1}^{N-1} |\Delta A(i)||Ax_0 + Bu_0| + \prod_{i=2}^{N-1} |\Delta A(i)||Ax_1 + Bu_1| + \dots + |Ax_{N-1} + Bu_{N-1}|. \end{aligned}$$

Because $\Delta_N \geq 0$, $\lim_{N \rightarrow \infty} \Delta_N \geq 0$ we have

$$\lim_{N \rightarrow \infty} |x_N| \geq \lim_{N \rightarrow \infty} \prod_{i=0}^{N-1} |\Delta A(i)| |x_0| > 0 \quad x_0 \neq 0, \tag{9}$$

if a matrix sequence $\Delta A^* = \{|\Delta A(0)|, |\Delta A(1)|, \dots, |\Delta(i)|, \dots\}$ is recurrently unstable. From (9) it follows that the plant is M -nonstabilizable with any feedback $u(x_t)$.

Proof of Theorem 2 proceeds as above. Indeed, representing the plant equation in an equivalent form of a difference inclusion

$$x_{t+1} \in A_I x_t + B u_t = [-W^*, W^*] x_t + m(A) x_t + B u_t$$

and denoting $x_{t+1}^* = \max_{A(t) \in A_I} |x_{t+1}|$ with fixed (x_t, u_t) we have

$$x_N^* = (W^*)^N |x_0| + (W^*)^{N-1} |m(A)x_0 + B u_0| + \dots + |m(A)x_{N-1} + B u_{N-1}|,$$

$$\lim_{N \rightarrow \infty} x_N^* \geq \lim_{N \rightarrow \infty} (W^*)^N |x_0| > 0, \quad x_0 \neq 0$$

if the condition of asymptotic stability is not met for the matrix W^* .

To prove Theorem 3 it is sufficient to derive at least one behavioural strategy of the perturbations $\Delta A = \{\Delta A(0), \Delta A(1), \dots, \Delta A(i), \dots\}$ such that $\|x_{t+1}\| \geq \alpha \|x_t\|$, $t \in \mathbb{N}$ with $\alpha \geq 1$ and with any feedback $u(x_t)$. One of such strategies is given in the body of the next.

Proof of Theorem 4 is facilitated by using an equivalent representation of the equation describing $\Sigma_t^c(C, b)$ as an autoregression

$$x_{t+1} = A_t X_t + u_t, \quad t = 0, 1, \dots,$$

where $x \in \mathbb{R}$, $u \in \mathbb{R}$, $X_t = (x_{t-n+1}, \dots, x_t)^T$, $A_t = (a_{n-1}(t), \dots, a_0(t))$, $A_t^T \in C \subset \mathbb{R}^n$.

Assume that there exists stabilizing feedback $u(X_t)$. Represent $u(X_t)$ as a sum

$$u(X_t) = u^*(X_t) + \Delta u(X_t), \quad u^*(X_t) = \arg \min_{u_t} \max_{A_t} |X_{t+1}|,$$

$$\Delta u(X_t) = u(X_t) - u^*(X_t).$$

Introduce a non-negative function $f_0 : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ in the form

$$f_0(Y) = \max_{\mathbf{A}} \mathbf{A} Y - \min_{\mathbf{A}} \mathbf{A} Y, \quad \mathbf{A}^T \in C$$

where $Y = (y_1, \dots, y_n)^T$. From analysis of the function $f_0(Y)$ with the set C assumed even it follows that

$$y_i \partial f_0(Y) / \partial y_i \geq 0, \quad f_0(Y) = f_0(|Y|), \quad Y \neq 0, \quad i = 1, \dots, m. \tag{10}$$

Because the plant is nonstationary, there exists a vector $A_t^T \in \partial C$ such that

$$|x_{t+1}(u(X_t))| = \frac{1}{2} \left(\max_{\mathbf{A}} \mathbf{A} X_t - \min_{\mathbf{A}} \mathbf{A} X_t \right) + |\Delta u(X_t)|.$$

From the definition of the min-max control it follows that

$$\begin{aligned} |x_{t+1}(u(X_t))| &\geq |x_{t+1}(u^*(X_t))|, \\ |x_{t+2}(u(X_{t+1}), u(X_t))| &\geq |x_{t+2}(u^*(X_{t+1}), u(X_t))|. \end{aligned}$$

By virtue of (10) the equalities hold

$$\begin{aligned} |x_{t+2}(u^*(X_{t+1}), u(X_t))| &\geq |x_{t+2}(u^*(X_{t+1}), u^*(X_t))|, \\ |x_{t+2}(u(X_{t+1}), u(X_t))| &\geq |x_{t+2}(u^*(X_{t+1}), u^*(X_t))|. \end{aligned}$$

Extending this reasoning to arbitrary times we arrive at a conclusion that there always exists a strategy $\{A_0, \dots, A_N\}$ such that on the paths of the closed-loop system

$$|x_N(u(X_{N-1}), \dots, u(X_0))| \geq |x_N(u^*(X_{N-1}), \dots, u^*(X_0))|. \tag{11}$$

The majoring condition (11) holds for any feedback $u(X_t)$. If $u^*(X_t)$ is not a stabilizing control, then

$$\lim_{N \rightarrow \infty} |x_N(u(X_{N-1}), \dots, u(X_0))| \geq \lim_{N \rightarrow \infty} |x_N(u^*(X_{N-1}), \dots, u^*(X_0))| > 0$$

and feedback $u(X_t)$ does not stabilize $\Sigma_i^c(C, b)$ either. Assuming that the set C is even the control $u^*(X_t) = -(C^*)^T X_t$ is linear, whence immediately follows the requirement that M be recurrently stable.

Proof of Theorem 5 can be obtained in two ways. First, directly from Theorem 4, because $u^*(X_t) = -\sum_{i=0}^{n-1} m(c_i)x_{t-1}$ and the polynomial $G(z)$ dictates the comparison model for the closed-loop system and, second, as a particular case in the proof of Theorem 2.

Proof of Theorem 6. Necessity of condition (4) follows from optimality, in terms of reduction of the quadratic form $\langle x, Hx \rangle$ on the paths of the plant $x_{t+1} = Ax_t + bu_t$ of feedback $u_t = -\langle b, Hb \rangle^{-1} b^T H A x_t$. To prove sufficiency, feedback $u(x) \in H(u)$ is used which satisfies the conditions of (H, ε) -quadratic stabilizability.

To prove Theorems 7 and 8 it is sufficient to note that by virtue of the decomposition the sequence of the set $\Omega_t(K)$; $t = 0, 1, \dots$ has properties 1 and 2, whence it follows that with $u(x) \in H(u)$ and $u(x) \in \zeta_\Delta(u)$ for g , the number of times when the inequality $\varphi_t \leq 0$ does not hold, the estimate $q \leq s$ is true.

Proof of Theorem 9. With t_1, t_2, \dots, t_m being arbitrary time, $t_i \geq 0$; $i = 1, \dots, n$, compose an equation $KX = B_0$, where

$$X = |x_{t_1} x_{t_2} \dots x_{t_n}|, \quad B_0^T = \begin{vmatrix} u(x_{t_1}) - (cb)^{-1} \sigma_{t_1+1} \\ \vdots \\ u(x_{t_n}) - (cb)^{-1} \sigma_{t_n+1} \end{vmatrix}.$$

If $\det X \neq 0$, then the desired vector is $K = B_0 X^{-1} = -(cb)^{-1} cA$. To prove the Theorem, it is sufficient to show that if on the paths of the system $x_{t+1} = Ax_t + bu_t$, $A \in M$ whose feedback is $u(x) \in \zeta(u)$ the inequalities hold $|\sigma_{t_i+1}| > 0$; $i = 1, \dots, n$ with some $t_i > 0$; $i = 1, \dots, n$, then the vectors are linearly independent.

References

1. Barmish, B. R., Generalization of Kharitonov's four-polynomial concept for robust stability problems with linearly dependent coefficient perturbations. *IEEE Trans. Aut. Contr.* **34** (1989), 2, pp. 157-165.
2. Bartlett, A. C., A necessary and sufficient condition for Schur invariance and generalized stability of polytopes of polynomials. *IEEE Trans. Aut. Contr.* **33** (1988), 6, pp. 575-578.
3. Fomin, V. N., Fradkov, A. L., Yakubovich, V. A., Adaptive control of dynamic plants. Nauka, Moscow, 1981 (in Russian).
4. Furuta, K., Sliding mode control of a discrete system. *Systems & Control Letters* **14** (1990), pp. 145-152.
5. Isermann, R., Digital control systems. Springer-Verlag, Berlin-Heidelberg-New York, 1981.
6. Kharitonov, V. L., Asymptotic stability of an equilibrium position of a family of systems of linear differential equations. *Differentsialnye Uravneniya* **14** (1978), 11, pp. 1483-1485.
7. Kuntsevich, V. M., Lychak, M. M., Design of optimal and adaptive control systems. A game approach. Naukova dumka, Kiev, 1985 (in Russian).
8. Mahmoud, M. S., Bahnasawi, A. A., Asymptotic stability for a class of linear discrete systems with bounded uncertainties. *IEEE Trans. Aut. Contr.* **33** (1988), 6, pp. 378-383.
9. Yedavalli, R. K., Perturbation bounds for robust stability in linear state space models. *Int. J. Contr.* **42** (1985) 6, pp. 1507-1517.

**Анализ допустимых возмущений и стабилизация
неопределенных дискретных объектов**

С. В. ЕМЕЛЬЯНОВ, П. В. ЖИВОГЛЯДОВ, С. К. КОРОВИН

(Москва)

Рассматривается задача стабилизации стационарных и нестационарных дискретных объектов с компактной неопределенностью. В терминах асимптотических свойств допустимых возмущений сформулированы критерии стабилизируемости нестационарного дискретного объекта. Для стабилизации объектов со стационарной неопределенностью предложен метод локализации, сводящий задачу синтеза обратной связи к задаче конечного выбора параметров. Приведены примеры.

С. В. Емельянов
Всесоюзный научно-исследовательский
институт системных исследований
СССР, 117312, Москва, В-312,
Проспект 60-летия Октября, 9

Typesetting by TYPOT_EX Kft, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

**MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA**

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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 V. STREJČ
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

or to L. GYÖRFI
Technical University of Budapest
H-1111 Budapest, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

The papers, preferably in English or Russian, should be typed double spaced on one side of good-quality paper with wide margins (4–5 cm). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as $\text{diag} (d_1, d_2, \dots, d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

Объём статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском языке в трех экземплярах, на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Smagina, Ye. M.</i> : A method of designing of observable output ensuring given zeros location (<i>Смагина Е. М.</i> , Метод проектирования наблюдаемого выхода обеспечивающего заданные нули)	299-
<i>Shiryayev, V. I.</i> : Minimax filtering in real time of multistage systems (<i>Ширяев В. И.</i> , Минимаксная фильтрация в реальном времени многошаговых систем)	309
<i>Ishii, H., Menaldi, J-L., Zaremba, L.</i> : Viscosity solutions of the Bellman equation on an attainable set (<i>Ишии Г., Менальди Дж-Л., Заремба Л.</i> , Вязкие решение уравнений Беллмана на множестве достижимости)	317
<i>Rosinová, D.</i> : On decentralized stabilization of large-scale linear discrete systems (<i>Росинова Д.</i> Децентрализованная стабилизация сложных линейных дискретных систем)	329
<i>Timofeev, A. V.</i> : Non-asymptotic solution of confidence estimation parameter task of a non-linear regression by means of sequential analysis (<i>Тимофеев А. В.</i> , Неасимптотическое решение задачи доверительного оценивания параметра нелинейной регрессии с позиций последовательного анализа)	341
<i>Emelyanov, S. V., Zhivoglyadov, P. V., Korovin, S. K.</i> : Analysis of admissible perturbations and stabilization of uncertain discrete-time plants (<i>Емельянов С. В., Живоглядов П. В., Коровин С. К.</i> , Анализ допустимых возмущений и стабилизация неопределенных дискретных объектов)	353

316920

VOL. 20 • NUMBER 6
TOM HOMEP

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

PROBLEMS OF
CONTROL AND
INFORMATION
THEORY

13
10

ПРОБЛЕМЫ
УПРАВЛЕНИЯ И
ТЕОРИИ
ИНФОРМАЦИИ

КАДЕМИЯ НАУК СССР 1991
HUNGARIAN ACADEMY OF SCIENCES
CZECHOSLOVAK ACADEMY OF SCIENCES

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

PROBLEMS OF CONTROL AND INFORMATION THEORY

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

- theory of control processes
- theory of adaptive systems
- theory of estimation and identification
- theory of controlling robot-technologic and flexible manufacturing systems
- information theory
- information-theoretic aspects of multiple access networks.

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

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

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

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

- теория процессов управления;
- теория адаптивных систем;
- теория оценивания и идентификации;
- теория управления робототехническими и гибкими производственными системами;
- теория информации;
- теория информации в области сетей с множественным доступом.

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

Distributors

For the Soviet Union:

SOYUZPECHATY, Moscow 123 308 USSR

For Albania, Bulgaria, China, Cuba, Czech and Slovak Federal Republic, Korean People's Republic, Mongolia, Poland, Rumania, Vietnam and Yugoslavia:

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

For all other countries:

PERGAMON PRESS PLC Headington Hill Hall, Oxford OX3 0BW, England

or

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

1991 Subscription Rate DM 627,— per annum including postage and insurance.

PROBLEMS OF CONTROL AND INFORMATION THEORY

ПРОБЛЕМЫ УПРАВЛЕНИЯ И ТЕОРИИ ИНФОРМАЦИИ

EDITOR

N. N. KRASOVSKII (USSR)

COORDINATING EDITORS

USSR

S. V. EMELYANOV

E. P. POPOV

V. S. PUGACHEV

V. I. SIFOROV

K. V. FROLOV

A. B. KURZHANSKI

I. A. OVSEEVICH

E. D. TERYAEV

R. Z. KHASHMINSKI

HUNGARY

T. VÁMOS

A. PRÉKOPA

S. CSIBI

I. CSISZÁR

L. KEVICZKY

L. GYÖRFI

J. KOCSIS

CZECHOSLOVAKIA

J. BENEŠ

V. STREJC

I. VAJDA

РЕДАКТОР ЖУРНАЛА

Н. Н. КРАСОВСКИЙ (СССР)

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ

СССР

С. В. ЕМЕЛЬЯНОВ

Е. П. ПОПОВ

В. С. ПУГАЧЕВ

В. И. СИФОРОВ

К. В. ФРОЛОВ

А. Б. КУРЖАНСКИЙ

И. А. ОВСЕЕВИЧ

Е. Д. ТЕРЯЕВ

Р. З. ХАСЬМИНСКИЙ

ВНР

Т. ВАМОШ

А. ПРЕКОПА

Ш. ЧИБИ

И. ЧИСАР

Л. КЕВИЦКИ

Л. ДЪЕРФИ

Я. КОЧИШ

ЧССР

Й. БЕНЕШ

В. СТРЕЙЦ

И. ВАЙДА



AKADÉMIAI KIADÓ

PUBLISHING HOUSE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST

Typesetting by TYPOT_EX Kft, Budapest
PRINTED IN HUNGARY
Akadémiai Kiadó és Nyomda Vállalat, Budapest

MAGYAR
TUDOMÁNYOS AKADÉMIA
KÖNYVTÁRA

Taking Farewell

With the appearance of this last, sixth issue of its twentieth volume the Editors announce the termination of the journal *Problems of Control and Information Theory*.

While such a step is necessarily made with uneasy feelings, it is still firmly hoped no additional undue inconvenience is being caused by this neither to the readers/subscribers nor to the potential authors. Care has been taken throughout 1991 for making no further announcement concerning the Journal, and of course no further call was made for future subscriptions for 1992. Manuscripts, submitted next to the deadline of this present sixth issue, have been promptly returned to their authors, informing these about the termination with apologies.

The Hungarian Academy of Sciences branch of the Editorial Board, directly responsible for the technical process, came up with the idea of stopping the Journal by the end of 1991; as it became obvious already in 1990 that even for keeping the outputs of the consecutive issues strictly according to schedule, unconventional extra care and interventions were needed, because of the devaluation and decay of the financial backgrounds. In addition it has been realized by all of us that there is no more real need, for operating an additional anonymously reviewed inter-academic publication channel in English, in addition to those classically existing at each place, specifically for those working in Control and Information Theory at the sponsoring Academies; and for those either visiting at or co-working with these communities from abroad. This is no more indispensable, as all renowned journals in the field became globally accessible in the meantime by authors from any part of the world, at least by those with something really significant to say. Such a journal as the *Problems of Control and Information Theory* can hardly make a sense in the future within such an active arena without a truly global publicity, and a truly global editorial activity.

While even several members of the present Editorial Board are willing to start further activities towards these ambitious objectives, under a more constrained profile within Control Theory, an entirely new endeavor appears as most appropriate even for this purpose; and not just some slight furtherings along the existing lines. This is while the Editorial Board unanimously proposed the sponsoring Academies to stop, by the end of 1991, the operations of the present Journal; even though all of its members still so much enjoy of acting together. A steady academic give-and-take is existing within this community apart from any editorial activity anyhow.

This final Editorial is, also to acknowledge that long list of soul-seeking contributions included in the past twenty volumes, due not only to the schools at the sponsoring Academies, but also from many other centers, nearby and overseas. We

particularly recollect those ones that actually provoked fair further reflections from their field. We also wish to thank for the competent technical support, regularly received from the Publishing House and the Press of the Hungarian Academy of Sciences, and particularly for the support due to the desc-top publishing group we were associated with.

Finally let us express in this final Editorial our deep feelings and appreciation towards those with whom we were honored to co-work in the present capacity for long whiles, who however passed by in the meantime. More distinctly, our commemoration is of late F. Csáki (Editor), B. N. Petrov (Editor), G. Bognár (Editor), L. Kalmár, M. A. Gavrilov and A. M. Letov, all forming personalities of their times.

The Editorial Board

A DECENTRALIZED CONTROL SCHEME FOR CONTINUOUS-TIME SYSTEMS THROUGH PARTIAL AGGREGATION

V. VESELÝ, V. BARČ, K. S. HINDI

(Bratislava, Manchester)

(Received February 21, 1991)

A new method for designing a control law for a subsystem connected to a large-scale system such that the interaction is minimized, is proposed. The method yields a robust design while ensuring the best possible conditions for the stability of the overall system. The conditions for a successful design are not too strong and can be easily met. The calculation of the control law requires only modest computation.

1. Introduction

In many cases, when the subsystem is to be connected to an already existing large-scale system, the objective is to design the control of the subsystem to optimize a given objective function while ensuring that the overall system remains stable. The design control of a new power station to be integrated within an existing power system is one example. The second example we can take from the control of industrial processes when the subsystems can be jointed or disconnected in a prescribed way. In this case

- the first/last subsystem stability,
 - the prescribed way jointed/disconnected subsystem stability and
 - the complex system stability
- must be ensured by control system.

Two approaches have so far been adopted. The first is to design a decentralized controller for the subsystem concerned, checking overall stability a posteriori. If stability is not attained, the design must be repeated, which may require a number of design iterations [1, 2]. The second approach, i.e. the centralized approach, is to make a design taking the whole system into consideration, utilizing only the output of the subsystem. Thus, this approach necessitates working with a high-order mathematical model of the overall system.

The approach described in this paper is based on considering a detailed model of the subsystem and an aggregate model of the large-scale system. The interaction

between the two is minimized, while augmenting the subsystem to cater for the dynamics of the large-scale system.

2. Problem statement

Suppose that a large-scale system can be divided into two subsystems, the first of which is to be controlled while the second is already controlled. Let us consider a linear time-invariant system S_1

$$\begin{aligned} S_1 : \quad \dot{x}_1 &= A_{11}x_1 + A_{12}x_2 + B_1u \\ \dot{x}_2 &= A_{21}x_1 + A_{22}x_2 \\ y_1 &= C_1x_1 \end{aligned} \quad (2.1)$$

where

$x_1 \in \mathbb{R}^{n_1}$, $x_2 \in \mathbb{R}^{n_2}$ are the state vectors of the two subsystems, respectively; $u \in \mathbb{R}^{m_1}$ is the control vector of the first subsystem, A_{11} , A_{12} , A_{21} , A_{22} , B_1 are constant matrices having appropriate dimensions.

We assume that $n_1 \ll n_2$, matrix A_{22} is stable, and the triplet (A_{11}, B_1, C_1) is controllable/observable, and

$$u = K C_1 x_1 \quad (2.2)$$

It is required that u is determined such that an objective function for the first subsystem is optimized, while:

- (a) minimizing interaction with the second subsystem, and
- (b) ensuring the best possible conditions from the viewpoint of the first subsystem, for the stability of the overall system. Assuming this conditions, the first subsystem's feedback matrix K must be chosen so that the contribution of the first subsystem to the stability of the whole dynamical system will be positive.

The objective function is:

$$J = \int_{t=t_0}^{\infty} (x_1^T Q_1 x_1 + u^T R u) dt \quad (2.3)$$

where Q_1 and R are positive definite matrices.

3. The main result

Let us define a new state vector $v_1(t)$ as follows

$$v_1 = L x_2 \quad (3.1)$$

where $v_1 \in \mathbb{R}^{n_1}$. Matrix L will be referred to as the aggregation matrix. The dimension of matrix L and the magnitude of this elements are determined by the requirements following from the aggregation objective. In our case, this objective is to retain those properties of the aggregated system portion, which are essential when considering the stability of the whole system [6]. Letting

$$L = A_{12},$$

the most simple but not optimal results are obtained.

Multiplying equation (2.1) by A_{12} ,

$$\dot{v}_1 = A_{12}A_{21}x_1 + A_{12}A_{22}x_2$$

Let

$$A_{12}A_{22} = [MA_{12} + E]$$

It is now possible to minimize an interaction between the augmented system and the second subsystem by

$$\min_M \|E\| = \|A_{12}A_{22} - MA_{12}\|$$

from which

$$\begin{aligned} M &= A_{12}A_{22}A_{12}^+ \\ E &= A_{12}A_{22}[I - A_{12}^+A_{12}] \end{aligned} \quad (3.2)$$

where A_{12}^+ is the pseudo-inverse of A_{12} . Now, (2.1) can be written as:

$$\begin{aligned} S_2 : \quad \dot{x}_1 &= A_{11}x_1 + v_1 + B_1u \\ \dot{v}_1 &= A_{12}A_{21}x_1 + Mv_1 + Ex_2 \end{aligned} \quad (3.3a)$$

$$\dot{x}_2 = A_{21}x_1 + A_{22}x_2 \quad (3.3b)$$

Let $z_1 = \begin{bmatrix} x_1 \\ v_1 \end{bmatrix}$ and $z_2 = x_2$, then

$$S_2 : \quad \dot{z}_1 = D_1z_1 + D_2z_2 + Bu_1 \quad (3.4a)$$

$$\dot{z}_2 = D_3z_1 + D_4z_2 \quad (3.4b)$$

$$y_1 = Cz_1$$

where

$$D_1 = \begin{bmatrix} A_{11} & I \\ A_{12}A_{21} & M \end{bmatrix} \quad D_2 = \begin{bmatrix} 0 \\ E \end{bmatrix} \quad D_3 = [A_{21} \quad 0]$$

$$D_4 = A_{22} \quad B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix} \quad C = [C_1 \quad 0]$$

Since D_2 is small, by virtue of minimizing $\|E\|$, it is now possible to find the condition for tearing the two subsystems apart, while maintaining the stability of the overall system. Let us consider that the conditions for tearing the two subsystems hold (see the next Section). For system S_3

$$S_3 : \quad \dot{z}_1 = D_1 z_1 + Bu \quad (3.4)$$

we can find the control law (2.2), which will minimize the next augmented objective function

$$J_1 = \int_{t=t_0}^{\infty} (z_1^T Q z_1 + u^T R u) dt \quad (3.5)$$

where

$$Q = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix}$$

Matrix Q_2 is a positive definite one associated with the extra state vector $v_1(t)$. The state vector $v_1(t)$ is regarded as the interaction between the original first subsystem and the other part of the system. An appropriate choice of Q_2 leads to a minimization of this interaction.

Let us consider a Lyapunov function of the system (3.4) with control law (2.2) in the form

$$v_1 = z_1^T P z_1 \quad (3.6)$$

For Bellman–Lyapunov equation we have

$$\begin{aligned} B(x) &= (2Pz_1)^T (D_1 z_1 + Bu) + z_1^T Q z_1 + u^T R u = \\ &= z_1 [(D_1 + BKC)^T P + P(D_1 + BKC) + Q + C^T K^T RKC] z_1 \end{aligned} \quad (3.7)$$

To minimize (3.5) we obtain

$$\min J_1 = V_1(t_0) = z_1^T(t_0) P z_1(t_0) \leq \text{Tr}(P) \|z_1(t_0)\|^2 \quad (3.8)$$

under the condition $B(x) = 0$.

To minimize $\text{Tr}\|P\|$, we write the following Lagrangian function:

$$\begin{aligned} L = \min_{K,P} \max_W \{ &\text{Tr}[P + W((D_1 + BKC)^T P + P(D_1 + BKC) + \\ &+ Q + C^T K^T RKC)] \} \end{aligned} \quad (3.9)$$

where $\text{Tr}[\cdot] = \text{trace}[\cdot]$ and W is a matrix of Lagrangian multipliers. The necessary conditions for optimality are:

$$\nabla_P L = I + W(D_1 + BKC)^T + (D_1 + BKC)W = 0 \quad (3.10a)$$

$$\nabla_W L = P(D_1 + BKC) + (D_1 + BKC)^T P + Q + C^T K^T RKC = 0 \quad (3.10b)$$

$$\nabla_K L = RKCWC^T + B^T PWC^T = 0, \quad (3.10c)$$

from which

$$K = -R^{-1}B^T[PWC^T(CWC^T)^{-1}] \quad (3.11)$$

which, when C is an identity matrix, reduces to the well-known equation for the linear quadratic problem

$$K = -R^{-1}B^T P. \quad (3.12)$$

The three nonlinear matrix equations (3.10) can be solved by a two-level iteration process, similar to that employed by Xinogalis [3] to solve a set of similar equations. The solution steps are as follows:

1. Choose K^1 such that $(D_1 + BKC)$ is stable.
2. Substitute this value of K^1 in (3.10a) and (3.10b) to calculate W^1 and P^1 .
3. Substitute it in (3.10c) to find K^2 .
4. If $\|K^1 - K^2\| \leq \varepsilon$, where ε is a small positive number, then stop; else set $K^1 = K^2$ and go to Step 2.

For the industrial process mentioned above, which can be jointed/disconnected a control law will be determined by the next way.

1. For the first isolated subsystem a control law can be determined such that the corresponding objective function is optimized by centralized approaches.

2. The second subsystem is jointed to the first one. The mathematical model of two subsystems is given by (2.1). The first subsystem model and its control law are switch on the matrix A_{22} . The controller for the first subsystem can be determined by Eqs (3.10) and (3.11) through minimizing the corresponding objective function. So, we have got a new subsystem and a new control law calculation for the next jointed subsystem can be repeated.

4. Stability analysis

In order to check sufficient stability conditions for the interconnected system, the Lyapunov function for subsystems are constructed which are followed by check up that some linear combination of them is the Lyapunov function for the global system.

$$V = a_1 z_1^T H_1 z_1 + a_2 z_2^T H_2 z_2 \quad (4.1)$$

where a_1, a_2 are positive constants and

$$(D_1 + BKC)^T P_1 + P_1 (D_1 + BKC) = -H_1 \quad (4.2a)$$

$$D_4^T P_2 + P_2 D_4 = -H_2 \quad (4.2b)$$

Lemma. If $D_{11} = D_1 + BKC$ and D_4 are stable, then the condition

$$\|D_2\| \leq \frac{\lambda_{\min}(H_1)\lambda_{\min}(H_2)}{4\|P_1\| \cdot \|D_3^T P_2\|} \quad (4.3)$$

ensures that the global system is stable, too. It makes the disconnection of the two subsystems (3.4) possible. The way of the proof of this Lemma is similar to that in [5].

In order to maximize the right-side of inequality (4.3) the optimization technique can be used for finding

$$\max \left[\frac{\lambda_{\min}(H_i)}{\lambda_M(P_i)} \right] \quad i = 1, 2 \tag{4.4}$$

under the conditions (4.2).

Let us consider the non-singular constant matrices T_j , $j = 1$ and 4 , and let

$$\begin{aligned} z_1 &= T_1 \tilde{z}_1 \\ z_2 &= T_4 \tilde{z}_2 \end{aligned} \tag{4.5}$$

For the disconnected two subsystems (3.4) with controller (2.2) using (4.3) we can get

$$\begin{aligned} \dot{\tilde{z}}_1 &= L_1 \tilde{z}_1 \\ \dot{\tilde{z}}_2 &= L_4 \tilde{z}_2 \end{aligned} \tag{4.6}$$

where

$$\begin{aligned} L_1 &= T_1^{-1}(D_1 + BKC)T_1 \\ L_4 &= T_4^{-1}D_4T_4 \end{aligned}$$

and

$$L = \text{diag} \left\{ \left[\begin{array}{cc} \sigma_1^i & \omega_1^i \\ -\omega_1^i & \sigma_1^i \end{array} \right], \dots, \left[\begin{array}{cc} \sigma_p^i & \omega_p^i \\ -\omega_p^i & \sigma_p^i \end{array} \right], \sigma_{p+1}^i, \dots, \sigma_{n_i-p}^i \right\} \tag{4.7}$$

where $\sigma_p^i, \sigma_q^i \pm j\omega_q^i$ are eigenvalues of matrix L_i with $\sigma_q^i < 0$ for $q = 1, 2, \dots, n_i - p$ and $0 \leq p \leq n_i/2$. For Lyapunov matrix equation (4.2), we obtain

$$L_i^T \tilde{P}_i + \tilde{P}_i L_i = -\tilde{H}_i \tag{4.8}$$

where

$$\begin{aligned} \tilde{P}_i &= T_i^T P_i T_i \\ \tilde{H}_i &= T_i^T H_i T_i \end{aligned} \quad i = 1, 4$$

Lemma [6]. Let us consider that $\tilde{H}_i = c_i I_i$ then from Eq. (4.8) for \tilde{P}_i we obtain

$$\tilde{P}_i = \frac{1}{2} c_i \left[\text{diag} \left\{ -\sigma_1^i, -\sigma_1^i, \dots, -\sigma_p^i, -\sigma_p^i, -\sigma_{p+1}^i, \dots, -\sigma_{n_i-p}^i \right\} \right]^{-1} \tag{4.9}$$

where $c_i > 0$ is an arbitrary constant, and

$$\max \left[\frac{\lambda_{\min}(\tilde{H}_i)}{\lambda_M(\tilde{P}_i)} \right] = -2\sigma_M^i \tag{4.10}$$

where σ_M^i is the maximal eigenvalue of matrix L_i . Applying Eq. (4.10) to Eq. (4.3), we obtain

$$\|D_2\| \leq \frac{|\sigma_M^1| \cdot |\sigma_M^4|}{\|D_3\|} \tag{4.11}$$

where σ_M^1, σ_M^4 are the maximal eigenvalues of matrix $(D_1 + BKC)$ and matrix D_4 , respectively.

Condition (4.11) gives a more agreeable result than the one given by (4.3) (see example).

5. A practical example

Let us consider the linear invariant dynamical system (2.1) with

$$A = \begin{bmatrix} -0.9 & 0.3 & 0.7 & 0.12 & 0.2 & 0.31 & 0 \\ 0.1 & -1.7 & -0.1 & 0 & 0 & 0.22 & 0.2 \\ 0.6 & 0 & -0.4 & 0.19 & 0.05 & -0.05 & 0 \\ 0.6 & 1.0 & 0 & -0.36 & 0.4 & 0.11 & -0.01 \\ -0.7 & 0.28 & -0.3 & 0.22 & -0.1 & 0.2 & 0 \\ 0.3 & 0 & 0.6 & -0.01 & 0.42 & -1.2 & 0 \\ 0.82 & -0.15 & -0.025 & 0 & 0.1 & 0.6 & -3.6 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 1.2 & 0 \\ 0.1 & 0 \\ 0 & 0.7 \end{bmatrix} \quad B_2 = \begin{bmatrix} 1.2 & 0 \\ 0 & 0.3 \end{bmatrix} \quad B_3 = \begin{bmatrix} 0.4 \\ 0.85 \end{bmatrix}$$

The matrices C, Q, R are identities.

1. For the first isolated subsystem the feedback matrix K_1 can be determined by the centralized approach. We have got

$$K_1 = \begin{bmatrix} -0.6526 & -0.0956 & -0.4596 \\ 0.2663 & -0.0211 & -0.7099 \end{bmatrix}$$

2. The first and second subsystems are jointed now. The feedback matrix K_2 can be determined by (3.10). We have got

$$K_2 = \begin{bmatrix} -0.8716 & -0.4595 \\ -0.0969 & -0.6275 \end{bmatrix}$$

In order to check the two jointed subsystems stability we use (4.3). We obtain

$$\|D_2\| = 0.3452 \leq \frac{\lambda_{\min}(H_1)\lambda_{\min}(H_2)}{4\|P_1\| \cdot \|D_3^T P_2\|} = 0.8759$$

or for Eq. (4.11)

$$0.3452 \leq 1.346$$

The two jointed subsystems are stable. Now, we connect to the jointed subsystems the third one.

3. For the feedback matrix K_3 from (3.10), we obtain

$$K_3 = [-0.2574 \quad -0.1308]$$

In order to check the stability of the overall system we use Eqs (4.3) and (4.11). We obtain these inequalities

$$0.5104 \leq 0.7716 \quad \text{Eq. (4.3)}$$

$$0.5104 \leq 0.927 \quad \text{Eq. (4.11)}$$

The three jointed subsystems are stable. As we can see from this practical example condition (4.11) gives a more agreeable result than condition (4.3).

5. Conclusions

A new method for designing a control law for a subsystem connected to a large-scale system such that the interaction is minimized, has been proposed. The method yields a robust design which optimizes an objective function for the subsystem while ensuring the best possible conditions for the stability of the overall system. The conditions for a successful design are not too strong and can be easily met (i.e. that A_{ii} , B_i , C_i must be stabilizable and A_{22} must be stable). In addition, the calculation of the control law requires only modest computation.

The suggested approach is not contingent upon deriving a control law through optimizing an objective function, and, therefore, can also be used if a different control design method, say a frequency response method is desired.

References

1. Singh, M. G., Decentralized Control. North-Holland, 1981.
2. Siljak, D. D., Overlapping decentralized control. In: M. G. Singh and A. Titli (eds), Handbook of Large Scale Systems Engineering Applications. North-Holland, 1979.
3. Xinogalis, T. C., Hierarchical Optimization Structures. Ph.D. Thesis, UMIST 1982.
4. Gantmacher, F. R., The Theory of Matrices. Chelsea Publishing Company. New York 1959.
5. Vesely, V., Decentralized Control of Linear Dynamic systems with partial aggregation. Kybernetika, Vol. 25 (1989), No. 5, pp. 408-418.
6. Barc, V., Decentralized Control of Linear Dynamic Control. Ph.D. Thesis, EF SVST Bratislava 1988.

7. Aoki, M., Control of Large-Scale dynamic systems by aggregation. IEEE Trans. Aut. Control AC-13, 1968, pp. 246-253.
8. Murgas, J., Hejda, I., Decentralized Adaptive Stabilization with State Regulators. Kybernetika, Vol. 26 (1990), No. 6, pp. 496-504.

Децентрализованный алгоритм управления многосвязными системами

В. Веселы, В. Барч и К. С. Хинди

(Братислава, Манчестер)

Предложен новый метод синтеза локальных алгоритмов управления для многосвязных систем, обеспечивающий минимизацию взаимосвязей между различными подсистемами.

Метод обеспечивает робастность замкнутой системы в целом. Получены конструктивные условия синтеза алгоритмов управления.

Vojtech Veselý, Vladimír Barč
Slovak Technical University
Faculty of Electrical Engineering
KASR TP
Ilkovičova 3
812 19 Bratislava
Czechoslovakia

Khali S. Hindi
Control Systems Centre
UMIST, P.O. Box 88
Manchester M60 1QD
U.K.

NEAREST NEIGHBOR SEARCH AND CLASSIFICATION IN $O(1)$ TIME

A. FARAGÓ, T. LINDER, G. LUGOSI

(*Budapest*)

(Received February 21, 1991)

A method of finding the nearest neighbor is presented. The effectiveness of the algorithm has been shown in computer simulations. This paper gives a *probabilistic* analysis of the performance. The algorithm is shown to have $O(1)$ expected asymptotic complexity, measured in the number of distance calculations for n sample points. A reduced complexity classification rule is derived which has the same error probability as that of the nearest neighbor discrimination rule.

1. Introduction

Similar versions of a Nearest Neighbor algorithm have been presented independently by several authors (Vidal [1], Motoishi and Uemura [2], Faragó et al. [3]) which finds the nearest neighbor considerably faster than the exhaustive search at the price of $O(n^2)$ memory requirement. We will refer to this algorithm as *Geometric Search (GS)*. The effectiveness of the algorithm is proved in Vidal [1] via extensive computer simulations. However, no exact analysis of the performance of this type of algorithms have been published. This paper is devoted to filling this gap. In Section 2 we present *GS* and mention that, in a given probabilistic model, the number of necessary distance calculations tends to zero compared to the number of sample points. In Section 3 we propose a modified version of *GS*, called *Modified Geometric Search (MGS)*, which requires $O(n)$ storage capacity, and its average complexity (measured in distance calculations) is $O(1)$, that is, asymptotically constant. In Section 4 we introduce a nonparametric classification rule derived from *MGS*, which has the same asymptotic error probability as that of the Nearest Neighbor classification rule and requires no more than $d + 1$ distance calculations in a d -dimensional Euclidean space.

2. Geometric Search

Let X, X_1, X_2, \dots, X_n be i.i.d. random variables taking their values from \mathbb{R}^d . Assume that the X_i have a common density f of compact support (all random variables are defined on the same probability space $(\Omega, \mathcal{A}, \mathcal{P})$). Let ρ be a metric on \mathbb{R}^d . The task is to determine the nearest neighbor of the observation point X among the sample points X_1, X_2, \dots, X_n . Denote the nearest neighbor by X_n^{NN} . The algorithm “Geometric Search” uses the distances between the sample points X_1, X_2, \dots, X_n , thus they must be calculated and stored in a preprocessing stage.

Algorithm 2.1 (Geometric Search)

Initialization. Set $T \leftarrow X_1, T^{NN} \leftarrow X_1, R_{\min} \leftarrow \infty, \mathcal{T} \leftarrow \{X_1, \dots, X_n\}$.

Step 1. Calculate $\rho(X, T)$.

Step 2. If $\rho(X, T) < R_{\min}$ then $R_{\min} \leftarrow \rho(X, T)$ and $T^{NN} \leftarrow T$.

Step 3. Update \mathcal{T} in the following way: Exclude all sample points $T^* \in \mathcal{T}$ from \mathcal{T} , for which

$$\rho(X, T) + R_{\min} < \rho(T, T^*)$$

or

$$\rho(X, T) - R_{\min} > \rho(T, T^*)$$

holds. Delete T from \mathcal{T} , as well.

Step 4. If \mathcal{T} is empty then STOP, the last value of T^{NN} is the nearest neighbor and its distance from X is R_{\min} . If \mathcal{T} is not empty then go to Step 5.

Step 5. If the distances $\rho(X, X_{i_1}), \rho(X, X_{i_2}), \dots, \rho(X, X_{i_m})$ have already been evaluated, then

$$T \leftarrow \arg \min_{U \in \mathcal{T}} \gamma_m(U),$$

where

$$\gamma_m(U) = \max_{j \leq m} |\rho(U, X_{i_j}) - \rho(X, X_{i_j})|.$$

Go to Step 1.

It follows from elementary geometric arguments, using the triangle inequality, that Step 3 never excludes the true nearest neighbor. Since Step 5 effects only the order of distance calculations, the reader can immediately check that the algorithm ends with the correct nearest neighbor. The idea of the exclusion in Step 3 also appears in other nearest neighbor algorithms which use branch and bound techniques to reduce complexity ([4], [5]).

Although Algorithm 2.1 finds the nearest neighbor for any metric on \mathbb{R}^d , for the analysis we need to assume that the metric is of the form

$$\rho(x, y) = (x - y, x - y)^{1/2} = \|x - y\|,$$

where (\cdot, \cdot) is an arbitrary inner product in \mathbb{R}^d and $\|\cdot\|$ is the corresponding norm. It is readily verified that every inner product in \mathbb{R}^d can be written in the form: $(x, y) = x^T R y$, where R is a positive definite symmetric matrix.

THEOREM 1. Denote by N_n the number of distance calculations performed by *GS* (Algorithm 2.1). Then

$$\lim_{n \rightarrow \infty} \frac{N_n}{n} = 0$$

with probability 1.

We omit the proof of Theorem 1 for the ideas in it are similar to those in the proof of Theorem 3.

As a measure of complexity we have considered only the number of distance computations. We find it a reasonable assumption for this is the most time consuming part of the computation, especially when the distance measure is a complicated one. All the simulation results show that the processing time is essentially determined by the number of distance computations.

3. Finding the Nearest Neighbor in $O(1)$ Time

A serious drawback of Algorithm 2.1 is its $O(n^2)$ memory requirement. In the sequel we present a modified version of *GS* with memory requirement *linear* in the number of points. Now, instead of computing the distances between the points X_1, X_2, \dots, X_n , the preprocessing is the following:

Let p_1, \dots, p_{d+1} be $d+1$ points in \mathbb{R}^d of pairwise equal distances whose convex hull \mathcal{P} contains the support of the density of X . (Therefore, $X_i \in \mathcal{P}$ with probability 1.) Denote by s_1, s_2, \dots, s_{d+1} the hyperplanes whose finite segments are the faces of the regular simplex \mathcal{P} . Calculate and store the $n(d+1)$ distances $\rho(s_i, X_j)$, $1 \leq i \leq d+1; 1 \leq j \leq n$, i.e. the distances of the points X_j from all the hyperplanes. Having this done, we propose the following algorithm:

Algorithm 3.1

Initialization. $\mathcal{T} \leftarrow \{X_1, X_2, \dots, X_n\}$.

Step 1. Calculate the distances $\rho(X, s_i)$, $i = 1, \dots, d+1$.

Step 2. Calculate the values

$$h_X(X_i) = \max_{j \leq d+1} |\rho(X_i, s_j) - \rho(X, s_j)|, \tag{1}$$

determine their minimum and put

$$X_n^* \leftarrow \arg \min_{U \in \mathcal{T}} h_X(U). \tag{2}$$

Step 3. $\mathcal{T} \leftarrow \mathcal{T} - \mathcal{U}$, where \mathcal{U} consists of the X_i for which

$$h_X(X_i) > \frac{2}{\sqrt{2}} h_X(X_n^*). \tag{3}$$

Step 4. Calculate the distances of the elements of \mathcal{T} from X , and put

$$X_n^{NN} \leftarrow \arg \min_{U \in \mathcal{T}} \rho(X, U).$$

Before analysing the complexity, we verify that the algorithm works properly.

THEOREM 2. Algorithm 3.1. always finds the nearest neighbor of X .

The *proof* of the theorem uses the following lemma:

Lemma 1. For each point y inside \mathcal{P}

$$\frac{\sqrt{2}}{2} \rho(X, y) \leq h_X(y) \leq \rho(X, y),$$

where h_X is defined as in (1).

Proof. Denote by $s_j(X)$, $j = 1, \dots, d+1$ the hyperplane that contains X and is parallel to s_j . Then clearly,

$$|\rho(y, s_j) - \rho(X, s_j)| = \rho(y, s_j(X)),$$

therefore,

$$h_X(y) = \max_{j \leq d+1} \rho(y, s_j(X)).$$

Since $X \in s_j(X)$ for all j thus

$$\rho(y, s_j(X)) \leq \rho(y, X)$$

for all j which proves the second inequality.

To prove the first inequality let $s_{j_y}(X) = \arg \max_j \rho(s_j(X), y)$, that is $h_X(y) = \rho(s_{j_y}(X), y)$. Then it is easy to check that y falls in a right cone centered at X with base parallel to $s_{j_y}(X)$ and angle which is twice the angle α_d between an edge of \mathcal{P} and the corresponding height of \mathcal{P} . (Every distance and angle is understood in the given metric and inner product.) We will prove that angle α_d is not greater than $\pi/4$. Then it follows that denoting the projection of y onto $s_{j_y}(X)$ by v the angle Xyv is not greater than $\pi/4$, thus we have

$$\rho(y, v) = \rho(y, X) \cos(Xyv) \geq \rho(y, X) \cos(\pi/4) = \frac{\sqrt{2}}{2} \rho(X, y),$$

and the statement is proved.

All we have to prove is $\alpha_d \leq \pi/4$. Let m_k be the altitude and r_k be the circumradius of the k -dimensional regular simplex with edge of length 1. Thus α_k , the angle of interest, is just the angle between an edge and the corresponding height. Using elementary plane geometry we can write the following two equalities:

$$\begin{aligned} \cos^2(\alpha_k) + r_k^2 \cos^2(\pi/2 - 2\alpha_k) &= 1 \\ \sin(\alpha_{k+1}) &= r_k \end{aligned}$$

After elementary steps we get

$$\cos^2(\alpha_{k+1}) = 1 - \frac{1}{\cos^2(\alpha_k)}.$$

Now, it is obvious that $\cos(\alpha_k)$ monotonically decreases to $1/2$, thus $\cos^2(\alpha_k) \geq 1/2$ for all k and the lemma is proved. \square

Proof of Theorem 2. We need to check that Algorithm 3.1 does not exclude the nearest neighbor in Step 3., that is,

$$\frac{\sqrt{2}}{2} h_X(X_n^{NN}) \leq h_X(X_n^*).$$

Applying Lemma 1

$$h_X(X_n^{NN}) \leq \rho(X_n^{NN}, X) \leq \rho(X_n^*, X) \leq \frac{\sqrt{2}}{2} h_X(X_n^*)$$

thus, the algorithm never excludes the nearest neighbor. \square

Now, we can turn to the complexity analysis of Algorithm 3.1. As the measure of complexity again, the number of distance calculations are considered. The following theorem states that on the average, the algorithm executes no more than a constant number of distance calculations.

Introduce the following notations: Let $S_{x,r}$ be the ball of radius r centered at x , and let λ be the Lebesgue measure on \mathbb{R}^d and let P_X denote the probability measure on \mathcal{R}^d induced by X . In addition to the compact support condition we need a regularity condition imposed on the density. We assume that there exist functions $c_1(x), c_2(x) \geq 0$ and constant $r_0 > 0$ such that for almost all x (mod P_X) and all $r \leq r_0$

$$c_1(x)f(x) \leq \frac{1}{\lambda(S_{x,r})} \int_{S_{x,r}} f(y) dy \leq c_2(x)f(x) \tag{4}$$

and

$$\int_{\mathbb{R}^d} \frac{c_1(x)}{c_2(x)} f(x) dx < \infty. \tag{5}$$

Notice that this condition holds if e.g. the support of the density is a convex set and $a \leq f(x) \leq b$ for almost all x (mod P_X) for some $a, b > 0$.

THEOREM 3. Denoting by M_n the number of distance calculations performed by Algorithm 3.1, we have

$$\lim_{n \rightarrow \infty} E(M_n) \leq c_d,$$

where $c_d = d + 2^d + 1$ is a constant depending on the dimension d only.

The following lemmas show that having Step 3 of Algorithm 3.1 executed, the number of the remaining points, on the average, is not greater than 2^d .

Lemma 2. If $y \in \mathcal{P}$ and $\frac{\sqrt{2}}{2}h_X(y) \leq h_X(X_n^*)$, then

$$\rho(y, X) \leq 2\rho(X_n^{NN}, X).$$

Proof. Applying Lemma 1, the condition and the minimality of $h_X(X_n^*)$ we can write

$$\frac{\sqrt{2}}{2}\rho(y, X) \leq h_X(y) \leq \frac{2}{\sqrt{2}}h_X(X_n^*) \leq \frac{2}{\sqrt{2}}h_X(X_n^{NN}) \leq \frac{2}{\sqrt{2}}\rho(X_n^{NN}, X),$$

which completes the proof. □

Lemma 3. Let Z be a random variable taking its values in \mathbb{R}^d . Assume there exists a compact set $A \in \mathbb{R}^d$ with $P_Z(A) = 1$. Then for any $y > 0$ there is an $\epsilon = \epsilon(y) > 0$ such that

$$\Pr\{P_Z(S_{Z,y}) \geq \epsilon\} = 1.$$

Proof. For any set $H \in \mathbb{R}^d$ define the quantity

$$\mu(H) = \inf_{x \in H} P_Z(S_{x,y}).$$

If $\mu(A) > 0$, then $\epsilon = \mu(A)$ will clearly satisfy the requirement, so assume that $\mu(A) = 0$. Then there are sequences $x_n \in A$, $\epsilon_n > 0$ with

$$\lim_{n \rightarrow \infty} \epsilon_n = 0 \quad \text{and} \quad P_Z(S_{x_n,y}) < \epsilon_n.$$

Now, cover A by a finite number of open balls of radius $y/2$, all centered in A . Then at least one of these balls contains infinitely many of the points x_n . Let B_0 be such a ball. As B_0 has radius $y/2$

$$B_0 \in S_{x_{n'},y}$$

holds for some infinite subsequence $\{x_{n'}\}$ of $\{x_n\}$ which implies

$$P_Z(B_0) \leq P_Z(S_{x_{n'},y}) < \epsilon_{n'},$$

that is, $P_Z(B_0) = 0$ must hold.

Now, replace A by $A_1 = A - B_0$ and repeat the whole procedure. If $\mu(A_1) > 0$, then the proof is finished, otherwise we can delete a ball B_1 of radius $y/2$ centered in A_1 with $P_Z(B_1) = 0$. Continuing in this manner we must get stuck after a finite number of such deletions by the boundedness of A , and the procedure yields a set $A_l \in A$ with $P_Z(A_l) = 1$ and $\mu(A_l) > 0$, which completes the proof. \square

Let $X_n^{NN}(x)$ be the nearest neighbor of the point x , and let the random variable $R_n(x) = \rho(x, X_n^{NN}(x))$ be its distance to x . The next lemma states that the average number of points not excluded in Step 3 of Algorithm 3.1 asymptotically does not increase with n .

Lemma 4. For all $c > 1$

$$\lim_{n \rightarrow \infty} E \left(\sum_{i=1}^n I_{\{X_i \in S_{X, cR_n(X)}\}} \right) = c^d.$$

Proof.

$$\begin{aligned} E \left(\sum_{i=1}^n I_{\{X_i \in S_{X, cR_n(X)}\}} \right) &= n E \left(I_{\{X_n \in S_{X, cR_n(X)}\}} \right) \\ &= n \Pr\{X_n \in S_{X, cR_n(X)}, X_n = X_n^{NN}\} \\ &\quad + n \Pr\{X_n \in S_{X, cR_n(X)}, X_n \neq X_n^{NN}\} \end{aligned}$$

The i.i.d. property of the X_i implies that the first term in the brackets is $1/n$, while the second term is the following:

$$\begin{aligned} &\Pr\{X_n \in S_{X, cR_n(X)}, X_n \neq X_n^{NN}\} \\ &= E \left(I_{\{X_n \in S_{X, cR_{n-1}(X)}\}} \right) - E \left(I_{\{X_n \in S_{X, R_{n-1}(X)}\}} \right) \\ &= E \left(I_{\{X_n \in S_{X, cR_{n-1}(X)}\}} \right) - \frac{1}{n}, \end{aligned}$$

thus, we have

$$E \left(\sum_{i=1}^n I_{\{X_i \in S_{X, cR_n(X)}\}} \right) = n \Pr\{X_n \in S_{X, cR_{n-1}(X)}\}.$$

Now, for arbitrary $y > 0$

$$\begin{aligned} n \Pr\{X_n \in S_{X, cR_{n-1}(X)}\} &\leq n \Pr\{X_n \in S_{X, cR_{n-1}(X)}, R_{n-1}(X) \leq y\} \\ &\quad + n \Pr\{R_{n-1}(X) > y\}. \end{aligned} \tag{6}$$

First, we treat the second term of (6). By Lemma 3 there exists an $\epsilon > 0$ (independent of n) with

$$\Pr\{R_{n-1}(X) > y\} = E(1 - P_X(S_{X,y}))^{n-1} \leq (1 - \epsilon)^{n-1},$$

which implies

$$\lim_{n \rightarrow \infty} n \Pr\{R_{n-1}(X) > y\} = 0.$$

The first term of (6) can be written as

$$\int_{\mathbb{R}^d} n \Pr\{X_n \in S_{x,cR_{n-1}(x)}, R_{n-1}(x) \leq y\} f(x) dx \tag{7}$$

We will show that $n \Pr\{X_n \in S_{x,cR_{n-1}(x)}, R_{n-1}(x) \leq y\} \rightarrow c^d$ as $n \rightarrow \infty$ for almost all x (mod P_X), and use Lebesgue's dominated convergence theorem to finish the proof. We can write

$$\begin{aligned} & \lim_{n \rightarrow \infty} n \Pr\{X_n \in S_{x,cR_{n-1}(x)}, R_{n-1}(x) \leq y\} \\ &= \lim_{n \rightarrow \infty} \frac{\Pr\{X_n \in S_{x,cR_{n-1}(x)}, R_{n-1}(x) \leq y\}}{\Pr\{X_n \in S_{x,R_{n-1}(x)}\}} \\ &= \lim_{n \rightarrow \infty} \frac{\Pr\{X_n \in S_{x,cR_{n-1}(x)}, R_{n-1}(x) \leq y\}}{\Pr\{X_n \in S_{x,R_{n-1}(x)}, R_{n-1}(x) \leq y\}}. \end{aligned} \tag{8}$$

The last equality is true because $\Pr\{R_{n-1}(x) > y\}$ tends to zero exponentially fast for almost all x (mod P_X) (see Cover, Hart [6]). Introducing the notation $F_{n-1}(y) = \Pr\{R_{n-1}(x) \leq y\}$, Lebesgue's density theorem implies (Wheeden, Zygmund [7]) that for almost all x (mod λ)

$$\begin{aligned} & \frac{\Pr\{X_n \in S_{x,cR_{n-1}(x)}, R_{n-1}(x) \leq y\}}{\Pr\{X_n \in S_{x,R_{n-1}(x)}, R_{n-1}(x) \leq y\}} \\ &= \frac{\int_0^y \int_{S_{x,ct}} f(z) dz dF_{n-1}(t)}{\int_0^y \int_{S_{x,t}} f(z) dz dF_{n-1}(t)} \end{aligned} \tag{9}$$

$$= c^d \frac{\int_0^y (f(x) + h'_x(t)) \lambda(S_{x,t}) dF_{n-1}(t)}{\int_0^y (f(x) + h_x(t)) \lambda(S_{x,t}) dF_{n-1}(t)} \tag{10}$$

where $h_x(t), h'_x(t) \rightarrow 0$ as $t \rightarrow \infty$. Obviously, (10) is arbitrarily close to c^d if y is small enough and $f(x) > 0$. Therefore, the limit in (8) is c^d for almost all x (mod P_X). On the other hand, using condition (4) we can upper bound (9) as follows:

$$\frac{\int_0^y \int_{S_{x,ct}} f(z) dz dF_{n-1}(t)}{\int_0^y \int_{S_{x,t}} f(z) dz dF_{n-1}(t)} \leq c^d \frac{c_2(x)}{c_1(x)},$$

which is integrable by condition (5), therefore, the dominated convergence theorem can be applied to complete the proof. \square

Proof of Theorem 3. Since the number of distance calculations is $d + 1$ (in Step 1) plus the number of points not excluded in Step 3, Lemmas 2 and 4 readily imply the theorem. \square

4. A fast nonparametric classification algorithm

We have seen that Algorithm 3.1 finds nearest neighbor after $d+1+2^d$ distance calculations, on the average. In many cases the nearest neighbor is used to classify the input vector X into one of M categories, that is, the task is to estimate the value of the random variable $Y \in \{0, 1, \dots, M - 1\}$ through X , using independent copies of $(X, Y) : \xi_n = (X_1, Y_1), \dots, (X_n, Y_n)$. Cover and Hart's [6] well known result states that if the estimation is Y_n^{NN} , the label of X_n^{NN} , then the asymptotic error probability of the nearest neighbor classification rule is

$$\lim_{n \rightarrow \infty} \Pr\{Y_n^{NN} \neq Y\} = E \left(1 - \sum_{i=0}^{M-1} p_i^2(X) \right), \tag{11}$$

if the $p_i(x) = \Pr\{Y = i|X = x\}$ a posteriori probabilities are continuous. In this Section we propose a classification rule which provides the same asymptotic error probability as that of the nearest neighbor classification and calculates only (deterministically) $d+1$ distances, and we also drop the continuity condition of the $p_i(x)$. This classification rule is simply the truncated version of Algorithm 3.1:

Classification rule 4.1. Estimate Y by Y_n^* , where Y_n^* is the label of $X_n^* = \arg \min_{U \in \mathcal{T}} h_X(U)$, the point obtained after executing the first two steps of Algorithm 3.1.

Clearly, $d + 1$ distance calculations are necessary to obtain Y_n^* . We have the following theorem for the probability of misclassification:

THEOREM 4.

$$\lim_{n \rightarrow \infty} \Pr\{Y_n^* \neq Y\} = E \left(1 - \sum_{i=0}^{M-1} p_i^2(X) \right). \tag{12}$$

First, we prove the theorem for continuous a posteriori probabilities.

Lemma 5. If the $p_i(x)$ a posteriori probability functions are continuous, then (12) holds.

Proof. In the proof of (11) the only property of the nearest neighbor Cover and Hart used was that $\rho(X, X_n^{NN}) \rightarrow 0$ as $n \rightarrow \infty$ with probability 1. But, by Lemma 2 $\rho(X, X_n^*) \leq 2\rho(X, X_n^{NN})$, thus $\rho(X, X_n^*) \rightarrow 0$ as $n \rightarrow \infty$ with probability 1, therefore, the statement can be proved in the same way. \square

Lemma 6. If f is a non-negative measurable function on \mathbf{R}^d , then

$$E[f(X_n^*)] \leq 2^{d+1} E[f(X)].$$

Proof. We use Stone’s technique [8] to prove the statement. Put

$$\omega_{in}(x, x_1, \dots, x_n) = \begin{cases} 1, & \text{if } h_x(x_i) \leq h_x(x_j) \text{ for all } i \neq j \\ 0, & \text{otherwise} \end{cases}$$

then exploit the i.i.d. property of the X_i .

$$\begin{aligned} E[f(X_n^*)] &= E\left(\sum_{i=1}^n f(X_i)\omega_{in}(X, X_1, \dots, X_n)\right) \\ &= \sum_{i=1}^n E(f(X)\omega_{in}(X_i, X_1, \dots, X, \dots, X_n)) \\ &= E\left(f(X)\sum_{i=1}^n \omega_{in}(X_i, X_1, \dots, X, \dots, X_n)\right) \end{aligned}$$

Therefore, it is enough to prove that

$$\sum_{i=1}^n \omega_{in}(x_i, x_1, \dots, x, \dots, x_n) \leq 2^{d+1}.$$

Here $\sum_{i=1}^n \omega_{in}(x_i, x_1, \dots, x, \dots, x_n)$ is just the number of the x_i for which $h_x(x) \leq h_{x_j}(x)$ for all $j \neq i$. We can upper bound this number as follows. Let u and u_i be the following $d + 1$ dimensional vectors:

$$u = (u^{(1)}, u^{(2)}, \dots, u^{(d+1)}) = (\rho(x, s_1), \rho(x, s_2), \dots, \rho(x, s_{d+1})),$$

and

$$u_i = (u_i^{(1)}, u_i^{(2)}, \dots, u_i^{(d+1)}) = (\rho(x_i, s_1), \rho(x_i, s_2), \dots, \rho(x_i, s_{d+1})),$$

for $i = 1, 2, \dots, n$. Observe that

$$\begin{aligned} h_x(x_i) &= \max_{j \leq d+1} |\rho(x_i, s_j) - \rho(x, s_j)| \\ &= \max_{j \leq d+1} |u_i^{(j)} - u^{(j)}| = \|u - u_i\|_\infty. \end{aligned}$$

Thus, it is obvious that $x_n^* = x_i$ iff u_i is the nearest neighbor of u (with respect to the maximum norm) among the points u_1, u_2, \dots, u_n . Therefore, $\sum_{i=1}^n \omega_{in}(x_i, x_1, \dots, x_n)$ is just the number of u_i for which u is the nearest neighbor of u_i among $u_1, \dots, u_{i-1}, u, u_{i+1}, \dots, u_n$. An elementary argument shows that this is not greater than the number of orthants of \mathbb{R}^{d+1} , that is, 2^{d+1} . \square

Lemma 7. Let Z, Z_1, \dots, Z_n be random variables taking their values from the set $\{0, 1, \dots, M-1\}$ such that $(X, Y, Z), (X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)$ form an i.i.d. sequence. Put $q_i(x) = \Pr\{Z = i | X = x\}$, $i = 0, \dots, M-1$. Denoting by Z_n^* the label of X_n^* we have

$$\limsup_{n \rightarrow \infty} |\Pr\{Y_n^* \neq Y\} - \Pr\{Z_n^* \neq Y\}| \leq 2^{d+1} \sum_{i=0}^{M-1} E|p_i(X) - q_i(X)|.$$

Proof.

$$\begin{aligned} |\Pr\{Y_n^* \neq Y\} - \Pr\{Z_n^* \neq Y\}| &= |E(\Pr\{Y_n^* \neq Y | X, X_n^*\} - \Pr\{Z_n^* \neq Y | X, X_n^*\})| \\ &= \left| E \left(\sum_{i=0}^{M-1} \Pr\{Y = i | X\} (p_i(X_n^*) - q_i(X_n^*)) \right) \right| \\ &\leq E \left(\sum_{i=0}^{M-1} |p_i(X_n^*) - q_i(X_n^*)| \right). \end{aligned}$$

Applying Lemma 6 the proof is completed. \square

Proof of Theorem 4. Since the set of continuous functions is dense in $L_1(P_X)$, it is easy to see that for every $\epsilon > 0$ there exists non-negative continuous functions

$q_i(x)$ with $\sum_{i=0}^{M-1} q_i(x) = 1$ such that

$$\sum_{i=0}^{M-1} E|p_i(X) - q_i(X)| < \frac{\epsilon}{2 + 2^{d+1}}.$$

Now, the random variables Z, Z_1, \dots, Z_n can be defined such that $\Pr\{Z = i | X = x\} = q_i(x)$ and $(X, Y, Z), (X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)$ form an i.i.d. sequence. Then from Lemma 7

$$\limsup_{n \rightarrow \infty} |\Pr\{Y_n^* \neq Y\} - \Pr\{Z_n^* \neq Y\}| < \frac{2^{d+1}\epsilon}{2 + 2^{d+1}}.$$

On the other hand,

$$\begin{aligned} \left| E \sum_{i=0}^{M-1} (1 - p_i^2(X))(1 - q_i^2(X)) \right| &< 2 \sum_{i=0}^{M-1} E |p_i(X) - q_i(X)| \\ &< \frac{2\epsilon}{2 + 2^{d+1}}. \end{aligned}$$

Therefore, by the triangle inequality, using Lemma 5 and the continuity of the $q_i(x)$, we have

$$\limsup_{n \rightarrow \infty} \left| \Pr\{Y_n^* \neq Y\} - E \sum_{i=0}^{M-1} (1 - p_i^2(X)) \right| < \frac{2\epsilon}{2 + 2^{d+1}} + \frac{2^{d+1}}{2 + 2^{d+1}} = \epsilon.$$

Since ϵ is arbitrary, the proof is completed. \square

Acknowledgement

The authors wish to express their thanks to Laci Györfi for his valuable help and encouragement.

References

1. Vidal, E., An algorithm for finding nearest neighbours in (approximately) constant average time. *Pattern Recognition Letters*, **4** (1986), pp. 145-157.
2. Motoishi, K. and Uemura, N., On a fast vector quantization algorithm. The VIIth Symposium on Information Theory and Its Applications, Kinugava, Japan, in Japanese, 1984.
3. Faragó, A., Linder, T., Lugosi, G. and Pikler, T., On the algorithmic problems of the nearest neighbor method (in Hungarian). *Híradástechnika (Telecommunication)*, Vol. **XXXIX** (1986), No. 8.
4. Fukunaga, K. and Narendra, P. M., A branch and bound algorithm for finding k -nearest neighbors. *IEEE Trans. Comput.*, Vol. **24** (1975), pp. 750-753.
5. Kamgar-Parsi, B., Kanal, L. N., An improved branch and bound algorithm for computing k -nearest neighbors. *Pattern Recognition Letters*, **3** (1985), pp. 7-12.
6. Cover, T. M., Hart, P. E., Nearest neighbor pattern classification. *IEEE Trans. on Information Theory*, Vol. **IT-13** (1967), pp. 21-27.
7. Wheeden, R. L., Zygmund, A., *Measure and integral*. Marcel Dekker, New York, 1977.
8. Stone, C. J., Consistent nonparametric regression. *Annals of Statistics*, Vol. **8** (1977), pp. 1348-1360.
9. Linder, T., Lugosi, G., Classification with a low complexity nearest neighbor algorithm. *IEEE International Symposium on Information Theory*, San Diego, CA, 1990.

Поиск ближайшего соседа и классификация за время $O(1)$

А. ФАРАГО, Т. ЛИНДЕР и Г. ЛУГОШИ

(Будапешт)

Представлен метод поиска ближайшего «соседа». Эффективность алгоритма показана с помощью компьютерных вычислений. В статье дается вероятностный анализ его результативности. Показано, что алгоритм имеет асимптотическую сложность $O(1)$, которая измеряется по числу вычисляемых расстояний до n выбранных точек.

Выведенное правило классификации пониженной сложности имеет ту же вероятность ошибки, что и правило отбора ближайшего соседа.

A. Faragó, T. Linder, G. Lugosi
Technical University of Budapest
H-1521 Budapest
Stoczek u. 2.
Hungary

UNIVERSAL ALMOST SURE DATA COMPRESSION FOR ABSTRACT ALPHABETS AND ARBITRARY FIDELITY CRITERIONS

EN-HUI YANG

(Tianjin)

(Received December 30, 1990)

The problem of universal almost sure data compression for abstract source alphabets is considered. Under certain mild conditions on fidelity criterions, universal almost sure data compression theorems are established for abstract source alphabets and reproducing alphabets. The methods are distortion program-size complexity oriented, and the constructions of the universal sequence of codes used are based on distortion Chaitin complexity. These results are the generalization of the universal almost sure data compression theorem of Ornstein-Shields for finite alphabets and the fidelity criterion of Hamming distance to abstract alphabets and arbitrary fidelity criterions and, this author believe, should have a profound impact on the development of theory of distortion program-size complexity.

Keywords and phrases: Bounded distortion variable rate code, Chaitin complexity, distortion Chaitin complexity, operational rate distortion function, universal data compression.

1. Introduction

Let A and \hat{A} be two abstract alphabets, henceforth called the source alphabet and the reproducing alphabet, respectively. Let \mathcal{A} and $\hat{\mathcal{A}}$ be σ -fields of subsets of A and \hat{A} , respectively. We denote by $(A^\infty, \mathcal{A}^\infty)$ the infinite Cartesian product $\prod_{k=1}^{\infty} (A_k, \mathcal{A}_k)$ and by (A^n, \mathcal{A}^n) for each positive integer n , the n -fold Cartesian product $\prod_{k=1}^n (A_k, \mathcal{A}_k)$, where $(A_k, \mathcal{A}_k) = (A, \mathcal{A})$ for each positive integer k . $(\hat{A}^\infty, \hat{\mathcal{A}}^\infty)$ and $(\hat{A}^n, \hat{\mathcal{A}}^n)$ are defined similarly. If $x = (x_i)$ is a finite or infinite sequence from A or \hat{A} , let $x_m^n = (x_m, x_{m+1}, \dots, x_n)$ and, for simplicity, write x_1^n as x^n . For our purposes, a source μ is a stationary, ergodic process $\{X_n\}$ taking values in the source alphabet A .

Let $\rho : A \times \hat{A} \rightarrow [0, \infty)$ be a single-letter distortion measure for which there exists a finite subset $\hat{A} \subset \hat{A}$ such that

$$\sup_{x \in A} \inf_{y \in \hat{A}} \rho(x, y) \leq D, \quad (1)$$

where $D \geq 0$ is a fixed real number, and let $F_\rho = \{\rho(x^n, y^n) | \rho(x^n, y^n) = (1/n) \times \sum_{i=1}^n \rho(x_i, y_i)\}$ be the single-letter fidelity criterion generated by ρ . A D -bounded distortion variable rate code \mathcal{C}_n of order n is a quadruple (ϕ, \bar{A}, n, τ) , where \bar{A} is a finite subset of \hat{A}^n , ϕ is a measurable mapping from A^n to \bar{A} such that, for every $x^n \in A^n$, $\rho(x^n, \phi(x^n)) \leq D$, and τ is a length function from \bar{A} to $\{1, 2, \dots\}$ which satisfies the Kraft inequality, that is,

$$\sum_{y^n \in \bar{A}} 2^{-\tau(y^n)} \leq 1.$$

Following [1], we refer to $l(\mathcal{C}_n, x^n) = \tau(\phi(x^n))$ as the length function and $r(\mathcal{C}_n, x^n) = \tau(\phi(x^n))/n$ as its associated compression factor. The expected compression factor, $R(\mathcal{C}_n) = E_\mu(r(\mathcal{C}_n, x^n))$, is called the rate of the code \mathcal{C}_n .

The operational rate-distortion function $R(\mu, D)$ is defined as follows. Let

$$R_n(\mu, D) = \inf\{R(\mathcal{C}_n) | \mathcal{C}_n \text{ is a } D\text{-bounded distortion variable rate code of order } n\}.$$

This is well-defined since (1) guarantees that there exists for each positive integer n at least one D -bounded distortion variable rate code of order n . The operational rate-distortion function is then defined by

$$R(\mu, D) = \lim_{n \rightarrow \infty} R_n(\mu, D).$$

Standard subadditivity arguments can be used to show that the above limit does exist and $R(\mu, D) = \inf\{R_n(\mu, D) | n \text{ is a positive integer}\}$. If there exists a letter $b^* \in \hat{A}$ for which

$$E_\mu \rho(X, b^*) < \infty,$$

and if (1) is a strict inequality, then it follows from [2] that our definition is equivalent to the usual mutual information definition of the rate-distortion function.

When A and \hat{A} are the same finite alphabet, and the single-letter distortion measure ρ is the Hamming distance on A , that is

$$\rho(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{otherwise,} \end{cases}$$

Ornstein and Shield [1] and Shield [3] proved the following theorem.

THEOREM 1. Let A and \hat{A} be the same finite alphabet, and let the single-letter distortion measure ρ be the Hamming distance on A . Then for any $D \geq 0$ there exists a sequence $\{\mathcal{C}_n\}$ of D -bounded distortion variable rate codes such that for any ergodic source μ , the sample compression factor $r(\mathcal{C}_n, x)$ converges almost surely to $R(\mu, D)$.

Note that in the case of Theorem 1, our definition of D -bounded distortion variable rate codes is equivalent to the Ornstein-Shields's definition of D -semifaithful codes in [1].

Theorem 1 is the first general result of the theory of universal almost sure data compression. In this paper, we generalize Theorem 1 to the very general case of abstract alphabets and arbitrary distortion measures. Specifically, we prove the following theorems.

THEOREM 2. Let the source alphabet A , the reproducing alphabet \hat{A} , and the single-letter distortion measure ρ satisfy the assumption (1). If \hat{A} is finite, then there exists a sequence $\{C_n\}$ of D -bounded distortion variable rate codes such that for any ergodic source μ with the source alphabet A , the sample compression factor $r(C_n, x^n)$ converges almost surely to $R(\mu, D)$.

THEOREM 3. Suppose the assumption (1) holds. If \hat{A} is countably infinite, then there exists a sequence $\{C_n\}$ of D -bounded distortion variable rate codes such that for any ergodic source μ , the sample compression factor $r(C_n, x^n)$ converges almost surely to $R(\mu, D)$.

THEOREM 4. Suppose the assumption (1) holds. In addition, suppose there exists a denumerable subset $\bar{A} \subset \hat{A}$ such that for any ergodic source μ , $R(\mu, D, \bar{A}) = R(\mu, D)$ where $R(\mu, D, \bar{A})$ is the operational rate-distortion function for the reproducing alphabet \bar{A} . Then there is a sequence $\{C_n\}$ of D -bounded distortion variable rate codes such that for any ergodic source μ , the sample compression factor $r(C_n, x^n)$ converges almost surely to $R(\mu, D)$.

The reasons why we state separately the Theorems 2 and 3 can be seen in the following Sections. Theorem 4 follows directly from Theorems 2 and 3. Since the reproducing alphabet \hat{A} in Theorem 4 is abstract, so far Theorem 4 can be considered as the most general result of the theory of universal almost sure data compression. From [2], the following two examples satisfy the conditions of Theorem 4, and hence the corresponding universal almost sure data compression theorems hold.

Example 1. \hat{A} is a separable metric space, ρ is bounded, and $\rho(x, \cdot)$ is continuous for each $x \in A$. In addition, there exists a finite subset $\bar{A} \subset \hat{A}$ such that

$$\sup_{x \in A} \inf_{y \in \bar{A}} \rho(x, y) < D.$$

Example 2. \hat{A} is a totally bounded metric space, A is a Borel subset of \hat{A} , and ρ is the metric on \hat{A} .

The proofs of Theorems 2 and 3 are given in Section 2 and 3, respectively. The methods used in the proofs are distortion program-size complexity oriented, and the constructions of codes described are based on distortion Chaitin complexity. The concept of distortion Chaitin complexity was first proposed in [4], and for some basic properties of this concept the reader is suggested to refer to [4].

2. Proof of Theorem 2

Throughout this Section the reproducing alphabet \hat{A} is assumed to be finite. We begin this Section by reviewing some basic properties used in this paper of distortion Chaitin complexity.

For each positive integer n and each $x^n \in A^n$, the D -distortion Chaitin complexity $C_D(x^n)$ of x^n is defined by

$$C_D(x^n) = \min\{C(y^n) | y^n \in \hat{A}^n \text{ and } \rho(x^n, y^n) \leq D\}$$

and the D -distortion conditional Chaitin complexity $C_D(x^n | n)$ of x^n given the length n is defined by

$$C_D(x^n | n) = \min\{C(y^n | n) | y^n \in \hat{A}^n \text{ and } \rho(x^n, y^n) \leq D\},$$

where $C(y^n)$ and $C(y^n | n)$ are the Chaitin complexity (see [5], p. 331) of $y^n \in \hat{A}^n$ and the conditional Chaitin complexity (in old fashion, see the appendix of [5], p. 338) of $y^n \in \hat{A}^n$ given the length n , respectively. (For some properties of Chaitin complexity and conditional Chaitin complexity, please refer to [5]. While only binary alphabet was dealt with in [5], most of the results in [5], including the definitions of Chaitin complexity and conditional Chaitin complexity, can be easily extended to the case of any finite alphabet.) It is easy to see that both $C_D(x^n)$ and $C_D(x^n | n)$ are \mathcal{A}^n -measurable. The following properties were proved in [4].

Property 1. For any $x \in A^\infty$,

$$\limsup_{n \rightarrow \infty} \frac{1}{n} C_D(x^n) = \limsup_{n \rightarrow \infty} \frac{1}{n} C_D(x^n | n),$$

and

$$\liminf_{n \rightarrow \infty} \frac{1}{n} C_D(x^n) = \liminf_{n \rightarrow \infty} \frac{1}{n} C_D(x^n | n).$$

Property 2. $\forall x, y \in A^*$, $C_D(x * y) \leq C_D(x) + C_D(y) + O(1)$, where A^* is the set of all finite sequences from A , and the symbol “*” denotes the concatenation between finite sequences.

Property 3. Let m be a fixed integer. For any $x \in A^\infty$ and $n \geq m$,

$$C_D(x^n | n) \leq \frac{1}{m} \sum_{i=1}^n C_D(x_i^{i+m-1} | m) + O(1).$$

From Property 2, it follows that for each positive integer n and each $x^n \in A^n$,

$$C_D(x^n) \leq C_D(x_2^n) + O(1),$$

and hence for any $x \in A^\infty$

$$\liminf_{n \rightarrow \infty} \frac{1}{n} C_D(x^n | n) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} C_D(x_2^{n+1} | n). \tag{2}$$

Based on distortion Chaitin complexity, next we describe for each positive integer n a D -bounded distortion variable rate code \mathcal{C}_n of order n . Toward this end, let $\phi_n : A^n \rightarrow \hat{A}^n$ be a measurable mapping such that for each $x^n \in A^n$, $\rho(x^n, \phi_n(x^n)) \leq D$ and $C(\phi_n(x^n) | n) = C_D(x^n | n)$. The code \mathcal{C}_n is now furnished by $\mathcal{C}_n = (\phi_n, \hat{A}^n, n, C(\cdot | n))$, where $C(\cdot | n)$ is the measure of conditional Chaitin complexity. Henceforth, the code $\mathcal{C}_n = (\phi_n, \hat{A}^n, n, C(\cdot | n))$ is said to be generated by the measure $C_D(\cdot | n)$ of D -distortion conditional Chaitin complexity. It is easy to see that the compression factor $r(\mathcal{C}_n, x^n)$ equals $C_D(x^n | n)/n$. To complete the proof of Theorem 2, it is enough to show that

$$\lim_{n \rightarrow \infty} \frac{1}{n} C_D(x^n | n) = R(\mu, D) \text{ a.s.} \tag{3}$$

We first prove that

$$\liminf_{n \rightarrow \infty} \frac{1}{n} C_D(x^n | n) \geq R(\mu, D) \text{ a.s.} \tag{4}$$

To this end, we make use of the sample path covering argument originated by Ornstein and Weiss [6] and modified by Shields [7]. For any $x \in A^\infty$, let $f(x) = \liminf_{n \rightarrow \infty} (1/n) C_D(x^n | n)$. By stationarity, it follows from (2) that $f(x)$ is invariant almost surely. The ergodicity of μ then guarantees that $f(x)$ is constant almost surely. Let H denote this constant so that $H = \liminf_{n \rightarrow \infty} (1/n) C_D(x^n | n)$ with probability one. Inequality (4) is valid once we show that $H \geq R(\mu, D)$. Let ϵ be any positive number, then for almost every $x \in A^\infty$, $C_D(x^n | n) < n(H + \epsilon)$ for infinitely many indices n . As in [7], let us fix a positive number δ , choose $M \geq 3/\delta$, and define $M_n(x) = m(n, x) - n + 1$, where $m(n, x)$ is the least integer $m \geq n$ such that

$$C_D(x_n^m | m - n + 1) < (m - n + 1)(H + \epsilon) \text{ and } m - n + 1 \geq M.$$

In what follows, when the infinite sequence $x \in A^\infty$ is clear from the context, we shall write $M_n(x)$ and $m(n, x)$ for convenience as M_n and $m(n)$, respectively. It is easy to see that M_n is almost surely finite and that $\{M_n\}$ is stationary and ergodic. Thus, there exists an integer L such that $Pr\{M_n > L\} < \delta^2/3$. We define g_n to be the characteristic function of the event $\{M_n > L\}$. that is, g_n equals 1 if $M_n > L$ and 0 otherwise, so that for any K ,

$$E \left(\frac{1}{K} \sum_{i=1}^K g_i \right) < \delta^2/3.$$

For $K > L$, let G_K be the event

$$\frac{1}{K - L + 1} \sum_{i=1}^{K-L+1} g_i < \delta/3.$$

From Markov inequality it follows that $Pr\{G_K\} \geq 1 - \delta$. As in [7], we can identify G_K with a measurable subset of A^K . For each $x^K = (x_1, x_2, \dots, x_K) \in G_K$ we define a sequence of nonoverlapping intervals $[n_i, m_i]$ inductively, letting

$$n_1 = \min\{n | g_n(x^K) = 0\}, \quad m_1 = m(n_1)$$

and

$$n_i = \min\{n | n > m_{i-1} + 1, g_n(x^K) = 0\}, \quad m_i = m(n_i).$$

The construction stops the first time n_i or m_i exceeds $K - L + 1$. A similar argument to [7] can be used to prove that there are at most δK indices n in $[1, K]$ but not in $\cup_i [n_i, m_i]$ in case $K \geq 3L/\delta$. We define ψ_K to be a measurable mapping from A^K to \hat{A}^K such that

- (i) if $x^K = (x_1, x_2, \dots, x_K) \notin G_K$, then $(\psi_K(x^K))_i = t(x_i)$ for $1 \leq i \leq K$;
- (ii) if $x^K \in G_K$, then

$$\begin{aligned} (\psi_K(x^K))_{n_i}^{m_i} &= \phi_{m_i - n_i + 1}(x_{n_i}^{m_i}), \\ (\psi_K(x^K))_j &= t(x_j), \quad j \notin \cup_i [n_i, m_i], \end{aligned}$$

where $t : A \rightarrow \hat{A}$ is a measurable mapping such that for any $x \in A$, $\rho(x, t(x)) = \min\{\rho(x, y) | y \in \hat{A}\}$, and $\phi_n : A^n \rightarrow A^n$ is the measurable mapping of the code $\mathcal{C}_n = (\phi_n, \hat{A}^n, n, \mathcal{C}(\cdot | n))$. Let $F_K = \{\psi_K(x^K) | x^K \in G_K\}$. The following lemma gives an upper bound to the cardinality $|F_K|$ of F_K . (If S is a finite set, $|S|$ denotes the cardinality of S ; all logarithms are to base 2.)

Lemma 1. If δ is small enough then $|F_K| < 2^{K(H+2\epsilon)}$.

Proof. For each $x^K \in G_K$ we call $\{[n_i, m_i]\}$ as in [7] the block decompositions of $[1, K]$ associated with x^K . The number of possible block decomposition is upper bounded by $2^{KH(\delta)}$, where $H(\delta) = -\delta \log \delta - (1 - \delta) \log(1 - \delta)$. Since $m_i - n_i + 1 \geq M$ and $M \geq 3/\delta$, there are at most $2^{(m_i - n_i + 1)(H + \epsilon + \delta/3)}$ sequences $y_{n_i}^{m_i}$ from \hat{A} such that

$$\mathcal{C}(y_{n_i}^{m_i} | m_i - n_i + 1) < (m_i - n_i + 1)(H + \epsilon).$$

Therefore, corresponding to any given block decomposition $\{[n_i, m_i]\}$, if we let

$$\begin{aligned} F_K(\{[n_i, m_i]\}) &= \\ &= \{\psi_K(x^K) | x^K \in G_K \text{ has the given block decomposition } \{[n_i, m_i]\}\}, \end{aligned}$$

then

$$|F_K(\{[n_i, m_i]\})| \leq |\hat{A}|^{\delta K} \prod 2^{(m_i - n_i + 1)(H + \epsilon + \delta/3)} \leq 2^{K(H + \epsilon + \delta/3 + \delta \log |\hat{A}|)}$$

This implies

$$|F_K| \leq \sum |F_K(\{[n_i, m_i]\})| \leq 2^{H(\delta)K} 2^{K(H + \epsilon + \delta/3 + \delta \log |\hat{A}|)},$$

where the summation is over all possible block decompositions. Letting δ small enough yields the lemma.

To sum up, we have obtained for each positive integer $K \geq 3L/\delta$ a measurable mapping ψ_K from $A^K \rightarrow \hat{A}^K$, a measurable subset $G_K \subset \hat{A}^K$ with probability $Pr\{G_K\} \geq 1 - \delta$, and a subset $F_K \subset \hat{A}^K$ of cardinality at most $2^{K(H + 2\epsilon)}$ such that

- (i) for any $x^K \in A^K$, $\rho(x^K, \psi_K(x^K)) \leq D$;
- (ii) for any $x^K \in G_K$, $\psi_K(x^K) \in F_K$.

For each $K \geq 3L/\delta$ we define a length function $\tau_K : \hat{A}^K \rightarrow \{1, 2, \dots\}$ so that

$$\begin{aligned} \tau_K(y^K) &\leq K \log |\hat{A}| + 2, \text{ if } y^K \in \hat{A}^K \\ \tau_K(y^K) &\leq K(H + 2\epsilon) + 2, \text{ if } y^K \in F_K. \end{aligned}$$

Therefore, we obtain for each $K \geq 3L/\delta$ a D -bounded distortion variable rate code $C'_K = (\psi_K, \hat{A}^K, K, \tau_K)$ of order K with expected compression factor

$$\begin{aligned} R(C'_K) &= \int_{G_K} \frac{1}{K} \tau_K(\psi_K(x^K)) d\mu + \int_{A^K - G_K} \frac{1}{K} \tau_K(\psi_K(x^K)) d\mu \\ &\leq \frac{1}{K} (K(H + 2\epsilon) + 2) Pr\{G_K\} + \frac{1}{K} (K \log |\hat{A}| + 2) \delta \\ &\leq H + 2\epsilon + \delta \log |\hat{A}| + 4/K. \end{aligned}$$

From the definition of $R(\mu, D)$ it follows that

$$R(\mu, D) \leq H + 2\epsilon + \delta \log |\hat{A}| + 4/K.$$

Letting $K \rightarrow \infty$ and then letting ϵ and $\delta \rightarrow 0$ yield

$$R(\mu, D) \leq H$$

which completes the proof of (4).

We next turn to prove that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} C_D(x^n | n) \leq R(\mu, D) \text{ a.s..} \tag{5}$$

Toward this end, we take arbitrarily a D -bounded distortion variable rate code $\mathcal{C}'_m = (\psi, \hat{A}^n, m, \tau)$ of order m . A similar argument to the proof of the Theorem 3 of [4] can be used to prove that for any $x \in A^\infty$ and $n \geq 2m$,

$$C_D(x^n|n) \leq \frac{1}{m} \sum_{i=1}^{n-m+1} \tau(\psi(x_i^{i+m-1})) + O(1).$$

Hence,

$$\frac{1}{n} C_D(x^n|n) \leq \frac{1}{n} \sum_{i=1}^n r(\mathcal{C}'_m, x_i^{i+m-1}) + O(1)/n.$$

From the ergodic theorem it follows that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} C_D(x^n|n) \leq R(\mathcal{C}'_m) \text{ a.s..}$$

Since \mathcal{C}'_m is taken arbitrarily, from the above inequality (5) follows. The proof of (3), hence the proof of Theorem 2, is now complete.

Note that when A and \hat{A} are the same finite alphabet, and when ρ is the Hamming distance on A and $D = 0$, the identity (3), hence Theorem 2, was established in [8].

3. Proof of Theorem 3

Throughout this Section we assume that the reproducing alphabet \hat{A} is countably infinite. Let \tilde{A} be a finite subset of \hat{A} satisfying

$$\sup_{x \in A} \inf_{y \in \tilde{A}} \rho(x, y) \leq D. \tag{6}$$

For convenience we write \hat{A} as $\tilde{A} \cup \{a_1, a_2, \dots\}$ and denote by A_n the finite alphabet $\tilde{A} \cup \{a_1, a_2, \dots, a_n\}$. Let $R(\mu, D, A_i)$ denote the operational rate-distortion function for the reproducing alphabet A_i and let $C^i_D(\cdot)$ denote the measure of distortion Chaitin complexity for the reproducing alphabet A_i , that is, for each positive integer n and each $x^n \in A^n$,

$$C^i_D(x^n) = \min\{C^i(y^n) | y^n \in A_i^n, \rho(x^n, y^n) \leq D\},$$

where $C^i(\cdot)$ is measure of Chaitin complexity defined on the set of all finite sequences from A_i . From Property 2, there exists a constant d_i depending only on A_i such that for each n and each $x^n \in A^n$,

$$C^i_D(x^n) \leq C^i_D(x^n_2) + d_i. \tag{7}$$

We next construct for each n a D -bounded distortion variable rate code \mathcal{C}_n of order n . Let $1 = n_1 < n_2 < \dots < n_i < \dots$ be an unbounded sequence of integers such that

$$\max_{j \leq i} |d_j|/n_i \rightarrow_{i \rightarrow \infty} 0,$$

and let $i(n) = \max\{i | n \geq n_i\}$. As in the above Section, we define ϕ_n^i to be a measurable mapping from $A^n \rightarrow (A_i)^n$ so that

$$\rho(x^n, \phi_n^i(x^n)) \leq D \text{ and } \mathbf{C}^i(\phi_n^i(x^n)) = \mathbf{C}_D^i(x^n).$$

Let $\psi_n : A^n \rightarrow (A_{i(n)})^n$ be a measurable mapping so that for any $x^n \in A^n$,

$$\psi_n(x^n) = \phi_n^{j(x^n)}(x^n),$$

where

$$j(x^n) = \min\{j | 1 \leq j \leq i(n), \mathbf{C}_D^j(x^n) = \min\{\mathbf{C}_D^k(x^n) | 1 \leq k \leq i(n)\}\}.$$

We define τ_n to be a length function from $(A_{i(n)})^n \rightarrow \{1, 2, \dots\}$ so that

$$\tau_n(y^n) = \min\{\mathbf{C}^j(y^n) | k(y^n) \leq j \leq i(n)\} + \lceil \log i(n) \rceil,$$

where $k(y^n) = \min\{j | y^n \in (A_j)^n\}$. Here $\lceil r \rceil$ denotes the least integer m such that $m \geq r$. The code \mathcal{C}_n is now defined to be $(\psi_n, (A_{i(n)})^n, n, \tau_n)$. It is not hard to see that for each $x^n \in A^n$,

$$\tau_n(\psi_n(x^n)) = \min_{j \leq i(n)} \mathbf{C}_D^j(x^n) + \lceil \log i(n) \rceil.$$

Let $S(x^n) = \min_{j \leq i(n)} \mathbf{C}_D^j(x^n)$. The compression factor then is given by

$$r(\mathcal{C}_n, x^n) = \frac{1}{n}(S(x^n) + \lceil \log i(n) \rceil).$$

To complete the proof of Theorem 3, it is now enough to prove that

$$\lim_{n \rightarrow \infty} r(\mathcal{C}_n, x^n) = R(\mu, D) \text{ a. s.} \tag{8}$$

We first prove that

$$\limsup_{n \rightarrow \infty} r(\mathcal{C}_n, x^n) \leq R(\mu, D) \text{ a. s.} \tag{9}$$

For each j , it is easy to see that

$$\limsup_{n \rightarrow \infty} r(\mathcal{C}_n, x^n) \leq \limsup_{n \rightarrow \infty} \frac{1}{n} \mathbf{C}_D^j(x^n).$$

From (3) and Property 1,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{C}_D^j(x^n) = R(\mu, D, A_j) \text{ a. s.}$$

and hence

$$\limsup_{n \rightarrow \infty} r(\mathcal{C}_n, x^n) \leq R(\mu, D, A_j) \text{ a. s.}$$

Letting $j \rightarrow \infty$ yields (9), since

$$R(\mu, D, A_j) \rightarrow_{j \rightarrow \infty} R(\mu, D).$$

We next need to show that

$$\liminf_{n \rightarrow \infty} r(\mathcal{C}_n, x^n) \geq R(\mu, D) \text{ a. s.} \tag{10}$$

To this end, as in the previous Section we make use of the sample path covering arguments. From (7),

$$S(x^n) \leq S(x_2^n) + \max_{j \leq i(n)} |d_j|$$

and hence

$$\liminf_{n \rightarrow \infty} \frac{1}{n} S(x^n) \leq \liminf_{n \rightarrow \infty} \frac{1}{n} S(x_2^n)$$

so that

$$\liminf_{n \rightarrow \infty} r(\mathcal{C}_n, x^n) \leq \liminf_{n \rightarrow \infty} r(\mathcal{C}_{n-1}, x_2^n),$$

from which it follows that $\liminf_{n \rightarrow \infty} r(\mathcal{C}_n, x^n)$ is invariant with probability one. The ergodicity of μ then tells us that $\liminf_{n \rightarrow \infty} r(\mathcal{C}_n, x^n)$ is a constant almost surely. Let H denote this constant so that $H = \liminf_{n \rightarrow \infty} r(\mathcal{C}_n, x^n)$ with probability one. Thus, (10) is established once we show that $H \geq R(\mu, D)$. Let ϵ be any positive number, then for almost all $x \in A^\infty$, $r(\mathcal{C}_n, x^n) < H + \epsilon$ for infinitely many indices n . For each n we define $m(n)$ to be the least integer $m \geq n$ such that

$$r(\mathcal{C}_{m-n+1}, x_m^n) < H + \epsilon \text{ and } m - n + 1 \geq M,$$

where $M \geq 3/\delta$ and δ is a positive number to be specified later. The number L , the random variables M_n and g_n , the measurable subset $G_K \subset A^K$, and, for each $x^K = (x_1, x_2, \dots, x_K) \in G_K$, the intervals $[n_i, m_i]$ are then defined similarly as in the previous Section, replacing $\mathbf{C}_D(x^n|n)$ by $nr(\mathcal{C}_n, x^n)$. Recall that $\mathcal{C}_n = (\psi_n, (A_{i(n)})^n, n, \tau_n)$. We then define $\hat{\psi}_K$ to be a measurable mapping from $A^K \rightarrow (A_{i(L)})^K$ so that

- (i) for $x^K = (x_1, x_2, \dots, x_k) \notin G_K$, $(\hat{\psi}_K(x^K))_i = t(x_i)$ for $1 \leq i \leq K$;

(ii) for $x^K = (x_1, x_2, \dots, x_K) \in G_K$,

$$\begin{aligned} \left(\hat{\psi}_K(x^K)\right)_{n_i}^{m_i} &= \psi_{m_i, -n_i+1}(x_{n_i}^{m_i}), \\ \left(\hat{\psi}_K(x^K)\right)_j &= t(x_j), \quad j \notin \cup_i [n_i, m_i], \end{aligned}$$

where $t : A \rightarrow \tilde{A}$ is a measurable mapping such that $\rho(x, t(x)) = \min\{\rho(x, y) \mid y \in \tilde{A}\}$. Let $F_K = \{\hat{\psi}_K(x^K) \mid x^K \in G_K\}$. A similar argument to the proof of Lemma 1 can lead to

$$|F_K| \leq 2^{H(\delta)K} |\tilde{A}|^{\delta K} 2^{K(H+\epsilon+\delta/3)}.$$

If δ is small enough then $|F_K| < 2^{K(H+2\epsilon)}$. Let $\sigma_K : (\tilde{A})^K \cup F_K \rightarrow \{1, 2, \dots\}$ be a length function so that

$$\begin{aligned} \sigma_K(y^K) &\leq \lceil K \log |\tilde{A}| \rceil + 1, & y^K \notin F_K, \\ \sigma_K(y^K) &\leq \lceil K(H + 2\epsilon) \rceil + 1, & y^K \in F_K. \end{aligned}$$

Then we obtain a D -bounded distortion variable rate code $C'_K = (\hat{\psi}_K, (\tilde{A})^K \cup F_K, K, \sigma_K)$ of order K with expected compression factor

$$\begin{aligned} R(C'_K) &= \int_{G_K} \frac{1}{K} \sigma_K(\hat{\psi}_K(x^K)) d\mu + \int_{A^K - G_K} \frac{1}{K} \sigma_K(\hat{\psi}_K(x^K)) d\mu \\ &\leq H + 2\epsilon + 2/K + \delta(\log |\tilde{A}| + 2/K) \\ &\leq H + 2\epsilon + 4/K + \delta \log |\tilde{A}|. \end{aligned}$$

From the definition of $R(\mu, D)$ it follows that

$$R(\mu, D) \leq H + 2\epsilon + \delta \log |\tilde{A}| + 4/K.$$

Letting $K \rightarrow 0$ and then letting ϵ and $\delta \rightarrow 0$ yield

$$R(\mu, D) \leq H$$

which completes the proof of (10). The proof of (8), hence the proof of Theorem 3, is now complete.

References

1. Ornstein, D. and Shields, P., Universal almost sure data compression. *Annals of Prob.*, 18 (1990), pp. 441-452.

2. Yang En-hui and Shen Shi-yi, Bounded distortion variable rate source coding. Submitted for publication
3. Shields, P., Universal almost sure data compression using Markov types. *Probl. Control Inform. Theory*, Vol. 19 (1990), pp. 269–277.
4. Yang En-hui and Shen Shi-yi, Distortion program-size complexity with respect to a fidelity criterion and rate distortion function. *IEEE Trans. Inform. Theory*, to appear.
5. Chaitin, G. J., A theory of program-size formally identical to information theory. *J. ACM*, Vol. 22 (1975), pp. 329–340.
6. Ornstein, D. and Weiss, B., The Shannon–McMillan–Breiman theorem for a class of amenable groups. *Israel J. Math.*, Vol. 44 (1983), pp. 56–60.
7. Shields, P., The ergodic and entropy theorems revisited. *IEEE Trans. Inform. Theory*, IT-33 (1987), pp. 263–266.
8. Yang En-hui, The proof of Levin's conjecture. *Chinese Science Bulletin*, Vol. 34 (1989), pp. 1761–1765.

Универсальное почти всюду сжатие данных для абстрактных алфавитов и произвольных критериев верности

ЕН-ХУ-ЯНГ

(Тяньян)

В работе рассмотрено обобщение результата у Орнштейна и Шильдса [1] о сходимости почти всюду для некоторой последовательности кодов показателя сжатия к rate-distortion функции для конечного алфавита. В качестве критерия верности и эргодического источника на случай абстрактного алфавита и произвольного аддитивного по размерности критерия верности выбрано расстояние Хэмминга с дополнительным свойством $\sup_{x \in A} \inf_{y \in \tilde{A}} f(x, y) \leq 1$ для алфавитов A, \tilde{A} . Кроме того, работа представляет развитие результатов автора [4] о свойствах сложности Чайтина [5].

En-hui Yang
 Department of Mathematics
 Nankai University
 Tianjin 300071
 P. R. China

SYNTHESIS OF OPTIMAL CONTROLS ON NONEXACT MEASUREMENTS OF OUTPUT SIGNALS

R. GABASOV, F. M. KIRILLOVA, P. V. GAISHUN, S. V. PRISCHEPOVA

(Minsk)

(Received May 27, 1991)

A method of synthesis of discrete estimator and regulator optimizing the behaviour of dynamic systems under incomplete and inexact observations on the control process is proposed.

1. Introduction

Feedback controls are convenient for parrying of unexpected perturbations arising in control processes of dynamic systems. In the first classical statements of the synthesis problem the nature of perturbations was not described and it was supposed that exact and complete state measurements are possible [1]. Taking into account the random nature of active perturbations the synthesis problem is investigated in the theory of stochastic control [2, 3]. Another approach to the perturbation registration in the control process is developed on the base of the guaranteed control theory [4].

Perturbations created by the interested participants of the optimization process are considered in the game approaches [5, 6].

In this paper the problem of optimal control synthesis is studied on the base of extremal problem solution suggested by the authors [7, 8].

2. Statement of the problem

Consider a discrete linear system the behaviour of which on the discrete interval $T(t_*) = \{t_*, t_* + h, \dots, t^* - h\}$ is described by the equation

$$x(t+h) = A(t, h)x(t) + b(t, h)u(t) \quad (2.1)$$

Here $x(t)$ is the state n -vector of the discrete system (2.1) at the moment t , $u(t)$ is the value of the one-dimensional controlling influence; symbols $A(t, h)$, $b(t, h)$

denote the parameters of the optimization object and the input device, respectively (matrix $A(t, h)$, $t \in T(t_*)$, is supposed nondegenerate).

Suppose that the initial state of system (2.1) is known inexactly. The a priori information about it has the form

$$\begin{aligned} x(t_*) = z \in \check{X}_* = \{z \in \mathbb{R}^n : Gz = f, d_* \leq z \leq d^*\}, \\ (f \in \mathbb{R}^n, \text{rank } G = r \leq n). \end{aligned} \quad (2.2)$$

To each control $u(t)$, $t \in T(t_*)$, limited by the restrictions

$$u_*(t) \leq u(t) \leq u^*(t), \quad t \in T(t_*), \quad (2.3)$$

it corresponds the totality of trajectories of system (1)

$$\check{X}(t|u(\cdot)) = \{x(t|z, u(\cdot)), z \in \check{X}_*\}, \quad t \in T^*(t_*) = T(t_*) \cup t^*.$$

Let in the space of states of discrete systems the terminal set

$$X^* = \{x \in \mathbb{R}^n : h'_i x \geq g_i, i = 1, m\}. \quad (2.4)$$

Following the principle of getting the guaranteed result, let us call the control $u(\cdot) = (u(t), t \in T(t_*))$, admissible if the corresponding movement $\check{X}(t|u(\cdot))$, $t \in T(t_*)$, satisfies the terminal inclusion

$$\check{X}(t^*|u(\cdot)) \subset X^*. \quad (2.5)$$

In the frames of the accepted approach the value of quality criterion on the admissible control is called the number

$$J(u(\cdot)) = \min_{x \in \check{X}_*} h'_0 x(t^*|z, u(\cdot)) \quad (2.6)$$

The admissible control $u^0(t)$, $t \in T(t_*)$, having the property

$$J(u^0(\cdot)) = \max_{u(\cdot)} J(u(\cdot)) \quad (2.7)$$

we shall call optimal.

Because of the indefiniteness (2.2) mentioned above the problem (2.1)–(2.7) not always has the solution as it is often impossible to supply demand (2.5). On the other hand if the admissible controls exists then, for the same reasons, the efficiency (2.7) of optimal control may be low.

With the purpose of increasing the control efficiency the procedure of discrete system optimization is supplemented by the measuring device described by the equation

$$y(t) = c'(t)x(t) + \xi(t), \quad (y \in \mathbb{R}^1). \quad (2.8)$$

Suppose the measurement errors $\xi(t), t \in T(t_*)$ satisfy the restrictions

$$\xi_*(t) \leq \xi(t) \leq \xi^*(t), t \in T(t_*). \tag{2.9}$$

Let the measuring device (2.8), (2.9) recorded the signal $y_\tau(\cdot) = (y(t), t = t_*, t_* + h, \dots, \tau), \tau$ - the given time moment from $T(t_*) \cup t^*$, corresponding to the chosen control $u(t), t \in T(t_*)$. Verify with it the a priori distribution \check{X}_* of the initial states.

Call the set $\hat{X}_*^\tau = \check{X}_*(y_\tau(\cdot))$ the a posteriori distribution of initial states corresponding to the observation process till the moment τ if it consists of those and only those initial states $x(t_*) \in X_*$ which can generate observed signal $y_\tau(\cdot)$ together with some measurement errors $\xi(t), t \geq t_*$, and the used control $u(\cdot)$.

In itself set \hat{X}_*^τ is not necessary for solution of the synthesis problem. We need its following numerical characteristics (estimates) connected with the terminal discrete system states

$$\hat{\alpha}_i^\tau(t^*) = \hat{\alpha}_i^\tau(t^*|u(\cdot)) = \min_{z \in \hat{X}_*^\tau} h'_i x(t^*|z, u(\cdot)), \quad i = \overline{0, m} \tag{2.10}$$

The calculation of the estimates $\hat{\alpha}_i^\tau(t^*), i = \overline{0, m}$, we shall call τ -observation problems accompanying the original problem (2.1)-(2.7).

The control $\hat{u}(\cdot) = (\hat{u}(t), t \in T(t_*))$, with the known starting part $u(t), t_* \leq t \leq \tau - h$, is called τ -a posteriori admissible if

$$\hat{\alpha}_i^\tau(t^*) \geq g_i, \quad i = \overline{1, m} \tag{2.11}$$

define the τ -a posteriori optimal control by the equality

$$\hat{\alpha}_0^\tau(t^*|\hat{u}^0(\cdot)) = \max_{\hat{u}(\cdot)} \hat{\alpha}_0^\tau(t^*|\hat{u}(\cdot)) \tag{2.12}$$

We shall call the search of controls $\hat{u}^0(t), t = \tau, \tau + h, \dots, t^* - h$, as τ -problem of optimal control accompanying the problem (2.1)-(2.7).

As a whole we shall call the problem (2.1)-(2.12) by the problem of optimal control on incomplete and inexact measurements of systems states.

In this paper solutions of two types are given: program solution for any fixed $\tau \in T(t_*)$ and feedback optimal solution consisting of the optimal estimator and the optimal regulator.

3. Program solution of τ -observation problem

Let except of the mathematical model (2.1)-(2.7) and the control $u(t)$ used on the interval $\{t_*, t_* - h, \dots, \tau - h\}$, the signal $y(t), t_* \leq t \leq \tau$, written with the device (2.8), (2.9), is known.

Denote through $F(t, \tau)$, $t, \tau \in T(t_*)$, the fundamental matrix of system (1) solutions

$$\begin{aligned} F(t+h, \tau) &= A(t, h)F(t, \tau), & F(\tau+h, \tau) &= E, \\ F(t, \tau-h) &= F(t, \tau)A(\tau, h), & F(t, t-h) &= E. \end{aligned} \quad (3.1)$$

(E is a unit diagonal $n \times n$ matrix).

Let $x_u(t)$, $t_* \leq t \leq \tau$, be a control system trajectory

$$x_u(t+h) = A(t, h)x_u(t) + b(t, h)u(t), \quad x(t_*) = 0, \quad (3.2)$$

and let

$$y_0(t) = y(t) - c'x_u(t), \quad t_* \leq t \leq \tau. \quad (3.3)$$

Since

$$\hat{\alpha}_i^T(t^*) = \hat{\alpha}_i^T(t^* | u(\cdot)) = \min_{z \in \bar{X}_i^T} h_i' F(t^*, t_* - h)z + h_i' x_u(t^*),$$

the problem of τ -observation (2.10) is reduced to the following extremal problems

$$\begin{aligned} \tilde{\gamma}_i^T(t^*) &= \min_z h_i' F(t^*, t_* - h)z, \\ \xi_*(t) &\leq y_0(t) - c'(t)F(t, t_* - h)z \leq \xi^*(t), \quad t_* \leq t \leq \tau, \\ Gz &= f, \quad d_* \leq z \leq d^*, \quad i = \overline{0, m} \end{aligned} \quad (3.4)$$

At the same time

$$\hat{\alpha}_i^T(t^*) = h_i' x_u(t^*) + \hat{\gamma}_i^T(t^*), \quad i = \overline{0, m}. \quad (3.5)$$

Denote

$$\begin{aligned} a'(t) &= (a_1(t), a_2(t), \dots, a_n(t))' = -c'(t)F(t, t_* - h) \\ \eta_i &= -h_i' F(t^*, t_* - h), \quad i = \overline{0, m}. \end{aligned} \quad (3.6)$$

Then problem (3.4) will be written in the form

$$\begin{aligned} \hat{\gamma}_i^T &= \max_z \eta_i' z, \\ \xi_*(t) &\leq y_0(t) + a'(t)z \leq \xi^*(t), \quad t_* \leq t \leq \tau, \\ Gz &= f, \quad d_* \leq z \leq d^*, \quad i = \overline{0, m}. \end{aligned} \quad (3.7)$$

By virtue of uniformity of the problem (3.7), in the sequel index i will be omitted and we shall speak of arbitrary problem family (3.7).

Solve the problem (3.7) with methods of the linear programming [8]. Let $\{z(\tau), S_{\text{sup}}(\tau)\}$ be an optimal feasible solution of problem (3.7). According to [9] the optimal support $S_{\text{sup}}(\tau) = \{J_{\text{sup}}(\tau), T_{\text{sup}}(\tau)\}$ is a totality from the set $J_{\text{sup}}(\tau) \subset J = \{1, 2, \dots, n\}$ of supporting indices of the feasible solution $z(\tau)$ and the set $T_{\text{sup}}(\tau) \subset T^\tau = \{t : t_* \leq t \leq \tau\}$ of supporting moments $t_* \leq \theta_1 = \theta_1(\tau) <$

... < $\theta_l = \theta_l(\tau) \leq \tau$. At the same time the relations $r + |T_{\text{sup}}(\tau)| = |J_{\text{sup}}(\tau)|$, $\det P \neq 0$, $0 \leq l \leq n - r$,

$$P = P(\tau) = P(\{T_{\text{sup}}(\tau), M\}, J_{\text{sup}}(\tau)) = \begin{bmatrix} a_j(t) : j \in J_{\text{sup}}(\tau) \\ t \in T_{\text{sup}}(\tau) \\ G(M, J_{\text{sup}}(\tau)) \end{bmatrix}$$

are carried out.

Introduce the notations

$$Q = Q(\tau) = Q(J_{\text{sup}}(\tau), \{T_{\text{sup}}(\tau), M\}) = P^{-1}(\tau) = \begin{bmatrix} ((q_j(t) : t \in T_{\text{sup}}(\tau)), (q_{ji} : i \in M))' \\ j \in J_{\text{sup}}(\tau) \end{bmatrix}.$$

Construct the sets $T_N = T_N(\tau) = T^r \setminus T_{\text{sup}}(\tau)$; $J_N = J_N(\tau) = J \setminus J_{\text{sup}}(\tau)$. To every moments of time $t \in T^r$ and indices $j \in J$, $i \in M$ add the numbers

$$\begin{aligned} \nu(t) &= \nu(t|\tau), \Delta_j = \Delta_j(\tau), \mu_i = \mu_i(\tau); \\ \nu(t) &= 0, t \in T_N(\tau); \Delta_j(\tau) = 0, j \in J_{\text{sup}}(\tau); \\ \mu &= \mu(\tau) = (\mu_i(\tau), i \in M); \nu_{\text{sup}} = (\nu(T_{\text{sup}}(\tau)) = (\nu(\theta_1(\tau)), \nu(\theta_2(\tau)), \dots, \nu(\theta_l(\tau))); \\ \nu_N &= \nu(T_N(\tau)) = (\nu(t), t \in T_N(\tau)); \lambda_{\text{sup}} = (\nu(T_{\text{sup}}(\tau)), \mu(\tau)) = (\nu_{\text{sup}}, \mu); \\ \lambda_{\text{sup}} &= \eta'_{\text{sup}} Q(\tau), \eta_{\text{sup}} = (\eta_j, j \in J_{\text{sup}}(\tau)); \\ \lambda_N &= \lambda(T_N(\tau)) = (\nu(T_N(\tau)), \mu(\tau)); \\ \lambda &= \lambda(T^r) = (\lambda(t), t \in T^r) = (\lambda(T_{\text{sup}}(\tau)), \lambda(T_N(\tau))) = (\lambda_{\text{sup}}, \lambda_N); \\ \Delta'(J_N) &= \Delta'(\tau|J_N(\tau)) = (\Delta_j(\tau), j \in J_N(\tau))' \\ &= \nu'_{\text{sup}} A(T_{\text{sup}}(\tau), J_N(\tau)) + \mu' G(M, J_N(\tau)) - \eta'_N, \\ \eta_N &= \eta_N(\tau) = (\eta_j, j \in J_N(\tau)). \end{aligned}$$

Feasible solution $z(\tau)$ is optimal iff there exists a support $S_{\text{sup}}(\tau)$ when

$$\begin{aligned} \Delta_j(\tau) &\leq 0 \text{ if } z_j(\tau) = d_j^*; \\ \Delta_j(\tau) &\geq 0 \text{ if } z_j(\tau) = d_{*j}; \quad j \in J_N(\tau), \\ \Delta_j(\tau) &= 0 \text{ if } d_{*j} < z_j(\tau) < d_j^*; \\ \nu(\theta_K(\tau)) &\geq 0 \text{ if } y_0(\theta_K(\tau)) + a'(\theta_K(\tau))z(\tau) = \xi^*(\theta_K(\tau)); \\ \nu(\theta_K(\tau)) &\leq 0 \text{ if } y_0(\theta_K(\tau)) + a'(\theta_K(\tau))z(\tau) = \xi_*(\theta_K(\tau)); \quad K = \overline{1, l}. \\ \nu(\theta_K(\tau)) &= 0 \text{ if } \xi_*(\theta_K(\tau)) < y_0(\theta_K(\tau)) + a'(\theta_K(\tau))z(\tau) < \xi^*(\theta_K(\tau)); \end{aligned}$$

4. Synthesis of optimal estimator

Let on the measuring results of output signals $y(t)$, $t_* \leq t \leq \tau - h$, and also on the values of controlling influence $u(t)$, $t_* \leq t \leq \tau - 2h$, produced with the regulator (see below). It is solved the problem (20) of finding the estimators, i.e. the problem

$$\begin{aligned} \eta'z \rightarrow \max, \quad Gz = f, \quad d_* \leq z \leq d^*, \\ \xi_*(t) \leq y_0(t) + a'(t)z \leq \xi^*(t), \quad t_* \leq t \leq \tau - h, \end{aligned} \quad (4.1)$$

and let $\{z(\tau - h), S_{\text{sup}}(\tau - h)\}$ be the optimal support feasible solution of the problem (4.1).

Give the estimates found from (4.1) (at $\eta = \eta_i$, $i = \overline{0, m}$) to the regulator. Denote controlling influence for the moment of time $\tau - h$ worked out with regulator through $u(\tau - h)$. Write the signal $y(\tau)$ of the measuring device (2.8), (2.9) at the moment τ . Proceeding from this information we find the optimal feasible solution $\{z(\tau), S_{\text{sup}}(\tau)\}$ of the problem

$$\begin{aligned} \eta'z \rightarrow \max, \quad Gz = f, \quad d_* \leq z \leq d^*, \\ \xi_*(t) \leq y_0(t) + a'(t)z \leq \xi^*(t), \quad t_* \leq t \leq \tau, \end{aligned} \quad (4.2)$$

where $y_0(\tau) = y(\tau) - c'(\tau)x_u(\tau)$; $x_u(\tau) = A(\tau - h, h)x_u(\tau - h) + b(\tau - h, h)u(\tau - h)$.

Call the construction of the optimal feasible solution $\{z(\tau), S_{\text{sup}}(\tau)\}$ of the problem (4.2) for any $y(\tau)$ proceeding from the optimal feasible solution $\{z(\tau - h), S_{\text{sup}}(\tau - h)\}$ of the problem (4.1) as an optimal estimator synthesis of discrete control system (2.1) at the moment τ with measuring device (2.8), (2.9).

Begin with the solution of the synthesis problem. According to the information available to the moment $\tau - h$ calculate the value

$$w(\tau - h) = y_0(\tau) + a'(\tau)z(\tau - h). \quad (4.3)$$

If $\xi_*(\tau) \leq w(\tau - h) \leq \xi^*(\tau)$ then $\{z(\tau), S_{\text{sup}}(\tau)\} = \{z(\tau - h), S_{\text{sup}}(\tau - h)\}$. Therefore, the problem of optimal estimator synthesis at the moment τ does not occur (or in another words, it is solved trivially). It occurs at $w(\tau - h) \notin [\xi_*(\tau), \xi^*(\tau)]$. Let, for definiteness $w(\tau - h) > \xi^*(\tau)$.

Imbed the problem (4.2) in the family of extremal problems depending on the parameter ρ

$$\begin{aligned} \eta'z \rightarrow \max, \quad Gz = f, \quad d_* \leq z \leq d^*, \\ \xi_*(t) \leq y_0(t) + a'(t)z \leq \xi^*(t), \quad t_* \leq t \leq \tau - h, \\ \xi_*(\tau) \leq y(\tau) + a'(\tau)z \leq \rho. \end{aligned} \quad (4.4)$$

The problem (4.4) at $\rho = w(\tau - h)$ has the solution $\{z(\tau - h), S_{\text{sup}}(\tau - h)\}$. To find $\{z(\tau), S_{\text{sup}}(\tau)\}$ we shall iteratively decrease the parameter ρ : $w(\tau - h) = \rho_0 >$

$\rho_1 > \dots > \rho_p = \xi^*(\tau)$, constructing simultaneously the solutions $\{z_K, S_{\text{sup}}^K\} = \{z(\tau - h|\rho_K), S_{\text{sup}}(\tau - h|\rho_K)\}$ of the problem (4.4). Then let $\{z(\tau), S_{\text{sup}}(\tau)\} = \{z_p, S_{\text{sup}}^p\}$.

Proceeding to the description of the optimal estimator algorithm denote through $T_{\text{sup}}^K, J_{\text{sup}}^K$ the sets of supporting moments of time and indices from J on the k -th iteration step of the algorithm and let

$$T_N^K = [(\{t_*\} \cup \{t_* + h\} \cup \{t \pm h, t \in T_{\text{sup}}^K\}) \cap T^\tau],$$

$$L_{\text{sup}}^K = \{T_{\text{sup}}^K, M\}.$$

We call the totality

$$C^K(\tau - h) = \{z^K; S_{\text{sup}}^K; T_N^K; y(T_N^K); u(T_{\text{sup}}^K); x_u(T_{\text{sup}}^K); F(T_{\text{sup}}^K, t_* - h);$$

$$Q^K = Q(J_{\text{sup}}^K, L_{\text{sup}}^K); \lambda^K = (\nu^K(T_{\text{sup}}^K), \mu^K(M)); \Delta^K(J_N^K); \rho_K\}$$

as the state of the algorithm on the k -th iteration step at the moment $\tau - h$.

Compose from the components

$$z^0 = z(\tau - h); S_{\text{sup}}^0 = S_{\text{sup}}(\tau - h),$$

$$T_N^0 = [(\{t_*\} \cup \{t_* + h\} \cup \{t \pm h, t \in T_{\text{sup}}(\tau - h)\}) \cap T^\tau];$$

$$y(T_N^0); u(T_{\text{sup}}(\tau - h)); x_u(T_{\text{sup}}(\tau - h)); F(T_{\text{sup}}(\tau - h), t_* - h);$$

$$Q^0 = Q(\tau - h); \lambda^0 = \lambda(\tau - h); \Delta^0(J_N^0) = \Delta(\tau - h)J_N(\tau - h); \rho_0 = w(\tau - h),$$

the zero state of the algorithm.

Iteration of the algorithm $C^K(\tau - h) \rightarrow C^{K+1}(\tau - h)$ consists of the following steps.

Step 1. Verify the condition $\tau \in T_{\text{sup}}^K$. If it is fulfilled we proceed to the step 2. Otherwise we proceed to step 5.

Step 2. Let $q^K(\tau) = Q^K(J_{\text{sup}}^K, \tau) = q^K(\theta_l(\tau - h)) = q_l^K = (q_{jl}^K, j \in J_{\text{sup}}^K)$. Calculate

$$\beta_j^K = \begin{cases} (z_j^K - d_j^*)/q_{jl}^K & \text{at } q_{jl}^K < 0, \\ (z_j^K - d_{*j})/q_{jl}^K & \text{at } q_{jl}^K > 0, \\ \infty & \text{at } q_{jl}^K = 0, \end{cases} \quad j \in J_{\text{sup}}^K \quad (4.5)$$

Estimate

$$x_u(t) = \begin{cases} A(t - h, h)x_u(t - h) + b(t - h, h)u(t - h), & \text{when } t - h = \theta \in T_{\text{sup}}^K, \\ A^{-1}(t, h)x(t + h) - b(t, h)u(t), & \text{when } t + h = \theta \in T_{\text{sup}}^*, \end{cases}$$

$$(t \in T_N^K)$$

$$x_u(t_*) = 0; x_u(t_* + h) = b(t_*, h)u(t_*), \quad (4.6)$$

$$y_0(t) = y(t) - c'(t)x_u(t),$$

$$(t \in T_N^K)$$

$$a'(t) = \begin{cases} -c'(t)A(t - h, h)F(t - h, t_* - h), & \text{when } t - h = \theta \in T_{\text{sup}}^K, \\ -c'(t)A^{-1}(t, h)F(t + h, t_* - h), & \text{when } t + h = \theta \in T_{\text{sup}}^K, \end{cases}$$

$$(t \in T_N^K).$$

Construct

$$\beta^K(t) = \begin{cases} [y_0(t) + a'(t)z^K - \xi^*(t)]/a'_{\text{sup}}(t)g_l^K, & \text{when } a'_{\text{sup}}(t)q_l^K < 0; \\ [y_0(t) + a'(t)z^K - \xi_*(t)]/a'_{\text{sup}}(t)g_l^K, & \text{when } a'_{\text{sup}}(t)q_l^K > 0; \\ \infty, & \text{when } a'_{\text{sup}}(t)q_l^K = 0, \end{cases} \quad (4.7)$$

$$(t \in T_N^K)$$

$$\beta^K(\tau) = \rho_K - \xi^*(\tau), \quad (4.8)$$

$$\beta_{j_0}^K = \min \beta_j^K, \quad j \in J_{\text{sup}}^K;$$

$$\beta^K(t^0) = \min \beta^K(t), \quad t \in T_N^K; \quad (4.9)$$

$$\beta_0^K = \min\{\beta_{j_0}^K, \beta^K(t^0), \beta^K(\tau)\}$$

Let

$$z_{\text{sup}}^{K+1} = (z_j^{K+1}, j \in J_{\text{sup}}^K) = z_{\text{sup}}^K - \beta_0^K q_{\text{sup}l}^K, \quad \rho_{K+1} = \rho_K - \beta_0^K$$

Here

$$q_{\text{sup}l}^K = (q_{jl}^K, j \in J_{\text{sup}}^K); \quad z_{\text{sup}}^K = (z_j^K, j \in J_{\text{sup}}^K);$$

$$z^{K+1} = (z_{\text{sup}}^{K+1}, z_N^K); \quad z_N^K = (z_j^K, j \in J_N^K).$$

The following cases are possible

$$\text{a) } \beta_0^K = \beta_{j_0}^K; \quad \text{b) } \beta_0^K = \beta^K(t^0); \quad \text{c) } \beta_0^K = \beta^K(\tau).$$

If the case a) is realized we proceed to step 3. In case b) we proceed to step 4. In case c) we proceed correlation (4.27) of step 6.

Step 3. Estimate

$$\begin{aligned} \Delta\lambda^{K'} &= (\Delta\nu^K, \Delta\mu^K)' = (\Delta\nu^K(T_{\text{sup}}^K), (\Delta\mu_j^K, j \in M))' = \\ &= e'_{j_0} Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) \text{sign } q_{j_0}^K, \end{aligned} \quad (4.10)$$

$$e_{j_0} = (e_j : e_j = 0, j \neq j_0, e_{j_0} = 1, j \in J_{\text{sup}}^K),$$

$$\Delta\delta^{K'} = \Delta\delta^{K'}(J) = \Delta\lambda^{K'} \rho^K(L_{\text{sup}}^K, J);$$

$$\sigma^K(t) = \begin{cases} -\nu^K(t)|\Delta\nu^K(t), & \text{when } \nu^K(t)\Delta\nu^K(t) < 0, \\ \infty, & \text{when } \nu^K(t)\Delta\nu^K(t) \geq 0, \end{cases} \quad (4.11)$$

$$(t \in T_{\text{sup}}^K);$$

$$\sigma_j^K = \begin{cases} -\Delta_j/\Delta\delta_j^K, & \text{when } \Delta_j\Delta\delta_j^K < 0, \\ \infty, & \text{when } \Delta_j\Delta\delta_j^K \geq 0, \end{cases} \quad j \in J_N^K;$$

$$\sigma^K(t^0) = \min \sigma^K(t), \quad t \in T_{\text{sup}}^K;$$

$$\sigma_{j_*}^K = \min \sigma_j^K, \quad j \in J_N^K; \quad (4.12)$$

$$\sigma_0^K = \min\{\sigma^K(t^0), \sigma_{j_*}^K\}$$

Let

$$\begin{aligned} S_{\text{sup}}^{K+1} &= \{T_{\text{sup}}^{K+1}, J_{\text{sup}}^{K+1}\}; \\ T_{\text{sup}}^{K+1} &= T_{\text{sup}}^K \setminus t^0; \quad \text{when } \sigma_0^K = \sigma^K(t^0), \end{aligned} \tag{4.13}$$

$$\begin{aligned} J_{\text{sup}}^{K+1} &= J_{\text{sup}}^K \setminus j^0; \\ T_{\text{sup}}^{K+1} &= T_{\text{sup}}^K; \\ J_{\text{sup}}^{K+1} &= J_{\text{sup}}^K \cup j_*, \quad \text{when } \sigma_0^K = \sigma_{j_*}^K \\ L_{\text{sup}}^{K+1} &= \{T_{\text{sup}}^{K+1}, M\}. \end{aligned} \tag{4.14}$$

Let the situation (4.13) be realized. Then

$$\begin{aligned} Q^{K+1} &= Q^{K+1}(J_{\text{sup}}^K, L_{\text{sup}}^{K+1}) = Q^K(J_{\text{sup}}^K \setminus j_0, L_{\text{sup}}^K \setminus t^0) - \\ &\quad - Q^K(J_{\text{sup}}^K \setminus j_0, t^0)Q^K(j_0, L_{\text{sup}}^K \setminus t^0)/q_{j_0 t^0}^K, \\ q_{j_0 t^0}^K &= Q^K(j_0, t^0), \end{aligned} \tag{4.15}$$

$$\begin{cases} \nu^{K+1}(T_{\text{sup}}^{K+1}) = \nu^K(T_{\text{sup}}^K \setminus t^0) + \sigma_0^K \Delta \nu^K(T_{\text{sup}}^K \setminus t^0), \\ \mu^{K+1}(M) = \mu^K + \sigma_0^K \Delta \mu^K, \\ \Delta^{K+1}(J_N^{K+1} \setminus j_0) = \Delta^K(J_N^K) + \sigma_0^K \Delta \delta^K(J_N^K), \end{cases} \tag{4.16}$$

If situation (4.14) is realized then

$$\begin{aligned} Q^{K+1} &= Q^{K+1}(J_{\text{sup}}^{K+1}, L_{\text{sup}}^{K+1}) \\ &= Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) - Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K)[P^K(L_{\text{sup}}^K, j_*) - P^K(L_{\text{sup}}^K, j_0)] \times \\ &\quad \times Q^K(j_0, L_{\text{sup}}^K)/[Q^K(j_0, L_{\text{sup}}^K)P^K(L_{\text{sup}}^K, j_*)]; \end{aligned} \tag{4.17}$$

$$\begin{cases} \nu^{K+1}(T_{\text{sup}}^{K+1}) = \nu^K(T_{\text{sup}}^K) + \sigma_0^K \Delta \nu^K(T_{\text{sup}}^K), \\ \mu^{K+1}(M) = \mu^K + \sigma_0^K \Delta \mu^K, \\ \Delta^{K+1}(J_N^{K+1} \setminus j_0) = \Delta^K(J_N^K \setminus j_*) + \sigma_0^K \Delta \delta^K(J_N^K \setminus j_*), \\ \Delta_{j_0}^{K+1} = \sigma_0^K \text{sign } q_{j_0 t^0}^K. \end{cases} \tag{4.18}$$

Proceed to step 6.

Step 4. Having estimated according to (4.5) and preserved $a(t^0)$, $x_u(t^0)$, construct

$$\begin{aligned} \Delta \lambda^{K'} &= (\Delta \nu^K, \Delta \mu^K)' = (\Delta \nu^K(T_{\text{sup}}^K), (\Delta \mu_j^K, j \in M))' \\ &= a'_{\text{sup}}(t^0)Q^K \text{sign}(a'_{\text{sup}}(t^0)q_l^K), \\ \Delta \delta^{K'} &= \Delta \delta^K(J) = \Delta \lambda^{K'} P^K(L_{\text{sup}}^K, J) - a'(t^0) \text{sign}(a'_{\text{sup}}(t^0)q_l^K) \end{aligned} \tag{4.19}$$

Following (4.10)–(4.12), (4.19), find σ_0^K .

Change the support $S_{\text{sup}}^K \rightarrow S_{\text{sup}}^{K+1}$:

$$T_{\text{sup}}^{K+1} = (T_{\text{sup}}^K \setminus t^*) \cup t^0; \quad J_{\text{sup}}^{K+1} = J_{\text{sup}}^K; \quad \sigma_0^K = \sigma^K(t^*) \quad (4.20)$$

$$T_{\text{sup}}^{K+1} = T_{\text{sup}}^K \cup t^0; \quad J_{\text{sup}}^{K+1} = J_{\text{sup}}^K \cup j_*; \quad \sigma_0^K = \sigma_j^K, L_{\text{sup}}^{K+1} = \{T_{\text{sup}}^{K+1}, M\}. \quad (4.21)$$

If the situation (4.20) is realized then having put $P^K(t^0, J_{\text{sup}}^K) = (a_j(t^0), j \in J_{\text{sup}}^K)$ we get

$$\begin{aligned} Q^{K+1} &= Q^{K+1}(J_{\text{sup}}^{K+1}, L_{\text{sup}}^{K+1}) \\ &= Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) - Q^K(J_{\text{sup}}^K, t^*) [P^K(t^*, J_{\text{sup}}^K) - P^K(t^0, J_{\text{sup}}^K)] \times \\ &\quad \times Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) / [-P^K(t^0, J_{\text{sup}}^K) Q^K(J_{\text{sup}}^K, t^*)], \end{aligned} \quad (4.22)$$

$$\begin{cases} \nu^{K+1}(T_{\text{sup}}^{K+1} \setminus t^0) = \nu^K(T_{\text{sup}}^K \setminus t^*) + \sigma_0^K \Delta \nu^K(T_{\text{sup}}^K \setminus t^*), \\ \nu^{K+1}(t^0) = -\sigma_0^K \text{sign}(a'_{\text{sup}}(t^0) q_i^K), \\ \mu^{K+1}(M) = \mu^K + \sigma_0^K \Delta \mu^K, \\ \Delta^{K+1}(J_N^{K+1}) = \Delta^K(J_N^K) + \sigma_0^K \Delta \delta^K(J_N^K). \end{cases} \quad (4.23)$$

Let the situation (4.21) is realized. Then

$$\begin{aligned} Q^{K+1} &= Q^{K+1}(J_{\text{sup}}^{K+1}, L_{\text{sup}}^{K+1}) = \\ &= \left[Q^{K+1}(J_{\text{sup}}^K, L_{\text{sup}}^K) + Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) P^K(L_{\text{sup}}^K, j_*) P^K(t^0, J_{\text{sup}}^K) Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) / W, \right. \\ &\quad \left. - P^K(t^0, J_{\text{sup}}^K) Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) / W, \right. \\ &\quad \left. - Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) P^K(L_{\text{sup}}^K, j_*) / W \right], \quad P^K(t^0, J_{\text{sup}}^K) = (a_j(t^0), j \in J_{\text{sup}}^K); \end{aligned} \quad (4.24)$$

$$\begin{aligned} P^K(t^0, j_*) &= a_{j_*}(t^0); \\ W &= P^K(t^0, j_*) - P^K(t^0, J_{\text{sup}}^K) Q^K(J_{\text{sup}}^K, L_{\text{sup}}^K) P(L_{\text{sup}}^K, j_*); \\ \nu^{K+1}(T_{\text{sup}}^{K+1} \setminus t^0) &= \nu^K(T_{\text{sup}}^K) + \sigma_0^K \Delta \nu^K(T_{\text{sup}}^K), \\ \nu^{K+1}(t^0) &= -\sigma_0^K \text{sign}(a'_{\text{sup}}(t^0) q_i^K), \\ \mu^{K+1}(M) &= \mu^K + \sigma_0^K \Delta \mu^K, \\ \Delta^{K+1}(J_N^{K+1}) &= \Delta^K(J_N^K \setminus j_*) + \sigma_0^K \Delta \delta^K(J_N^K \setminus j_*). \end{aligned} \quad (4.25)$$

Proceed to step 6.

Step 5. Introduce moment τ in the support. For this case estimate

$$\begin{aligned} \Delta \lambda^{K'} &= (\Delta \nu^K, \Delta \mu^K)' = (\Delta \nu^K(T_{\text{sup}}^K), (\Delta \mu_j, j \in M))' \\ &= a'_{\text{sup}}(\tau) Q^K \text{sign}(\xi^*(\tau) - w(\tau - h)), \\ \Delta \delta^{K'} &= \Delta \delta^{K'}(J) = \Delta \lambda^{K'} P^K(L_{\text{sup}}^K, J) - a'(\tau) \text{sign}(\xi^*(\tau) - w(\tau - h)). \end{aligned} \quad (4.26)$$

Calculate according (4.10)–(4.12), (4.19) the value σ_0^K . Change the support $S_{\text{sup}}^K \rightarrow S_{\text{sup}}^{K+1}$ according to (4.20), (4.21). Following (4.22)–(4.27) construct Q^{K+1} , ν^{K+1} , μ^{K+1} , Δ_N^{K+1} . Let $S_{\text{sup}}^K = S_{\text{sup}}^{K+1}$; $Q^{K+1} = Q^K$; $P^K = P^{K+1}$; $\nu^K = \nu^{K+1}$; $\mu^K = \mu^{K+1}$; $\Delta_N^K = \Delta_N^{K+1}$ and proceed to step 2.

Step 6. If $\rho_{K+1} > \xi^*(\tau)$ then K -th iteration of the algorithm $C^K(\tau - h) \rightarrow C^{K+1}(\tau - h)$ at the moment $\tau - h$ is completed.

At

$$\rho_{K+1} \leq \xi^*(\tau), \tag{4.27}$$

the work of the optimal estimator at the moment $\tau - h$ is completed ($K + 1 = p$). Zero state of the algorithm at moment τ :

$$C^0(\tau) = C^{K+1}(\tau - h) \setminus \rho_{K+1} \cup (\rho_0 = w(\tau)).$$

The work of the algorithm for the case $w(\tau - h) > \xi^*(\tau)$ is described completely. The case $w(\tau - h) < \xi_*(\tau)$ is analysed similarly.

Remark. While realizing step 5 in the recount formulae of the potentials and estimations (4.23), (4.25) we suppose

$$\text{sign}(a'_{\text{sup}}(\tau)q_i^K) = \begin{cases} \text{sign}(\xi^*(\tau) - w(\tau - h)), & \text{when } w(\tau - h) > \xi^*(\tau); \\ \text{sign}(\xi_*(\tau) - w(\tau - h)), & \text{when } w(\tau - h) < \xi_*(\tau). \end{cases}$$

5. Program solution of the τ -control problem

According to (2.12) the τ -a posteriori optimal control $\hat{u}^0(t), t = \tau, \tau + h, \dots, t^* - h$, is the solution of the following extremal problem

$$\begin{aligned} h'_0 x(t^*) \rightarrow \max, \quad & x(t + h) = A(t, h)x(t) + b(t, h)u(t); \\ x(\tau) = 0; \quad & h'_i x(t^*) \geq \hat{g}_i^\tau, \quad i = \overline{1, m} \\ u^*(t) \leq u(t) \leq u^*(t), \quad & \tau \leq t \leq t^* - h, \end{aligned} \tag{5.1}$$

where $\hat{g}_i^\tau = g_i - \hat{\gamma}_i^{\tau-h} - \sum_{t=0}^{\tau-h} h'_i x_u(t), \hat{\gamma}_i^{\tau-h}$ is the estimator value of i -th observation problem (3.7), $i = \overline{1, m}$.

The solution of the terminal control problem (5.1) [7,8] is the totality of $\{\hat{u}^0(\cdot|\tau), S_{\text{sup}}(\tau)\}$ where $S_{\text{sup}}(\tau) = \{I_{\text{sup}}(\tau), T_{\text{sup}}(\tau)\}$, $I_{\text{sup}}(\tau) \subset I = \{1, 2, \dots, m\}$, $T_{\text{sup}}(\tau) = \{\tau_1, \dots, \tau_l\}, \tau \leq \tau_1(\tau) < \dots < \tau_l(\tau) \leq t^* - h$. Along with this the relations

$$\det P(\tau) \neq 0, \quad P(\tau) = \begin{bmatrix} h'_i F(t^*, t)b(t, h), & t \in T_{\text{sup}}(\tau) \\ & i \in I_{\text{sup}}(\tau) \end{bmatrix}$$

are fulfilled.

The vector of the potentials

$$\begin{aligned} \nu' &= \nu'(\tau) = c_{\text{sup}} Q(\tau), \quad C_{\text{sup}} = (c(t); t \in T_{\text{sup}}(\tau)), \\ c(t) &= h'_0 F(t^*, t) b(t, h), \quad \tau \leq t \leq t^* - h, \quad Q(\tau) = P^{-1}(\tau), \end{aligned}$$

corresponds to the support $S_{\text{sup}}(\tau)$. The co-trajectory $\psi(t) = \psi(t|\tau)$, $\tau \leq t \leq t^* - h$, accompanying the support $S_{\text{sup}}(\tau)$ as the solution of the conjugated system

$$\begin{aligned} \psi'(t-h) &= \psi'(t) A(t, h), \quad \psi'(t^* - h) = h'_0 - \nu'(I_{\text{sup}}) H(I_{\text{sup}} J), \\ H(I_{\text{sup}}, J) &= \begin{bmatrix} h'_i(J) \\ i \in I_{\text{sup}}(\tau) \end{bmatrix} \end{aligned}$$

is constructed with the help of the vector $\nu(\tau)$.

The co-trajectory generates the co-control

$$\Delta(t) = \Delta(t|\tau), \quad \tau \leq t \leq t^* - h : \Delta(t) = -\psi'(t) b(t, h) \quad (5.2)$$

According to the construction at the supporting moments co-control (5.2) is equal to zero

$$\Delta(t) = 0, \quad t \in T_{\text{sup}}(\tau).$$

The optimal control $\hat{u}^0(t)$, $t \in T_N(\tau) = T(\tau) \setminus T_{\text{sup}}(\tau)$ at nonsupporting moments of time has the form

$$\hat{u}^0(t) \begin{cases} = u_*(t), & \text{when } \Delta(t) > 0; \\ = u^*(t), & \text{when } \Delta(t) < 0; \\ \in [u_*(t), u^*(t)], & \text{when } \Delta(t) = 0, t \in T_N(\tau). \end{cases}$$

Without loss of generality we can consider that the equations $h'_i x^0(t^*) = \hat{g}_i^T$, $i \in I_{\text{sup}}(\tau)$ ($\nu(i) \leq 0$, $i \in I_{\text{sup}}(\tau)$) are fulfilled for the trajectory $x^0(t)$, $t \in T(\tau)$ generated with the control $\hat{u}^0(t)$, $t \in T(\tau)$, of the problem (5.1).

The totality of the optimal control values at the support moment $\hat{u}^0_{\text{sup}} = (\hat{u}^0(t), t \in T_{\text{sup}}(\tau))$ is calculated according to the formula

$$\begin{aligned} \hat{u}^0_{\text{sup}} &= Q(\tau) g(\tau), \quad g(\tau) = \begin{bmatrix} g_i(\tau) \\ i \in I_{\text{sup}}(\tau) \end{bmatrix}, \\ g_i(\tau) &= \hat{g}_i^T - \sum_{t \in T_N(\tau)} H(I_{\text{sup}}, J) x_{\hat{u}^0}(t) \end{aligned}$$

Later on it will be necessary the additional information about the optimal support control $\{\hat{u}^0(\cdot|\tau), S_{\text{sup}}(\tau)\}$:

$$\begin{aligned} T_{N+}(\tau) &= \{t \in T_N(\tau) : \Delta(t) > 0, \Delta(t)\Delta(t-h) < 0\} \cup \\ &\quad \cup \{t \in T_N(\tau) : \Delta(t) > 0, (t-h) \in T_{\text{sup}}(\tau)\}, \\ T_{N-}(\tau) &= \{t \in T_N(\tau) : \Delta(t) < 0, \Delta(t)\Delta(t-h) < 0\} \cup \\ &\quad \cup \{t \in T_N(\tau) : \Delta(t) < 0, (t-h) \in T_{\text{sup}}(\tau)\}. \end{aligned}$$

Limit oneself to the case when the number of elements $|T_{N+}(\tau)| + |T_{N-}(\tau)|$ exceeds $|T_{\text{sup}}(\tau)|$ by not more than four units ($\tau_N \in T_{N+}(\tau) \cup T_{N-}(\tau)$, $\tau_N - h \notin T_{\text{sup}}(\tau)$).

Suppose that the masses of information $F(t^*, t)$, $t \in T_{\text{sup}}(\tau) \cup \tau_N \cup (t^* - h)$ are known.

6. Optimal regulator synthesis

Proceed to the description of acting the optimal regulator algorithm.

We call the totality $C^K(\tau) = \{u^{(K)}(t), t \in T(\tau + h); W^K; S_{\text{sup}}^K = \{I_{\text{sup}}^K, T_{\text{sup}}^K\}, T_{N+}^K; T_{N-}^K; \Delta g^K; Q^K(t), t \in T_{\text{sup}}^K \cup \tau_N \cup (t^* - h); \psi^K(t), t \in T_{N+}^K \cup T_{N-}^K \cup (t^* - h); \nu^K; Q^K\}$ as the state of the algorithm on k -th iteration at the moment τ . As the initial state $C^0(\tau)$ at the moment τ choose the totality with the following components:

$$\begin{aligned} u^{(0)}(t) &= \dot{u}^0(t|\tau), t \in T(\tau + h); W^0 = Hx^0(t^*) - \dot{g}^\tau; \\ S_{\text{sup}}^0 &= S_{\text{sup}}(\tau) = \{I_{\text{sup}}(\tau), T_{\text{sup}}(\tau)\}; T_{N+}^0 = T_{N+}(\tau); \\ T_{N-}^0 &= T_{N-}(\tau); \Delta g^0 = \tilde{\gamma}^{\tau-2h} - \tilde{\gamma}^{\tau-h}; \\ \phi^0(t) &= F(t^*, t), t \in T_{\text{sup}}(\tau) \cup \tau_N \cup (t^* - h); \\ \psi^0(t) &= \psi(t|\tau), t \in T_{N+}(\tau) \cup T_{N-}(\tau) \cup (t^* - h); \\ \nu^0 &= \nu(\tau); Q^0 = Q(\tau). \end{aligned}$$

Iteration of the algorithm $C^K(\tau) \rightarrow C^{K+1}(\tau)(C^K(\tau) \rightarrow C^0(\tau + h))$ consists of the following steps.

Step 1. If $l = 0$ proceed to step 2. Let compare τ with τ_1 . If $\tau < \tau_1$ proceed to step 8.

Step 2. Calculate the vectors

$$\begin{aligned} \Delta u^K(T_{\text{sup}}^K) &= Q^K \Delta g^K(I_{\text{sup}}^K); \Delta u^K(T_N^K) = 0, T_N^K = T(\tau + h) \setminus T_{\text{sup}}^K; \\ \Delta W^K(I_N^K) &= \sum_{t \in T_{\text{sup}}^K} H(I_N^K, J) \phi^K(t) b(t, h) \Delta u^K(T), \\ \Delta W^K(I_{\text{sup}}^K) &= 0, I_N^K = I \setminus I_{\text{sup}}^K. \end{aligned}$$

Step 3. Calculate the numbers $\alpha^K, \beta^K, \theta^K$:

$$\alpha^K = \alpha(\tau_s) = \min \alpha(t), t \in T_{\text{sup}}^K :$$

$$\alpha(t) = \begin{cases} \frac{u_*(t) - u^{(K)}(t)}{\Delta u^K(t)}, & \text{when } \Delta u^K(t) < 0; \quad t \in T_{\text{sup}}^K, \\ \frac{u^*(t) - u^{(K)}(t)}{\Delta u^K(t)}, & \text{when } \Delta u^K(t) > 0; \\ \infty, & \text{when } \Delta u^K(t) = 0; \end{cases}$$

$$\beta^K = \beta(i_0) = \min \beta(i), i \in I_N^K :$$

$$\beta(i) = \begin{cases} \frac{-W_i^K}{\Delta W_i^K - \Delta g_i^K}, & \text{when } \Delta W_i^K - \Delta g_i^K < 0; \\ \infty, & \text{when } \Delta W_i^K - \Delta g_i^K \geq 0. \end{cases}$$

Let $\theta^K = \min\{1, \alpha^K, \beta^K\}$. If $\theta^K = 1$ proceed to Step 4. At $\theta^K < 1$ proceed to step 5.

Step 4. At $\theta^K < 1$ proceed to step 5.

Step 4. Let $\hat{u}^0(\tau + h|\tau + h) = u^{(K)}(\tau + h) + \Delta u^K(\tau + h)$. If $\tau + h = t^* - lh$ the algorithm completes the work: $\hat{u}^0(\tau + ih|\tau + ih) = u^{(K)}(\tau + ih) + \Delta u^K(\tau + ih)$, $i = \overline{1, l}$.

At $\tau + h < t^* - lh$ construct the initial state $C^0(\tau + h)$ for the moment $\tau + h$ with the following components:

$$u^{(0)}(t) = u^{(K)}(t) + \Delta u^K(t), t \in T(\tau + 2h); W^0 = W^K + \Delta W^K;$$

$$S_{\text{sup}}^0 = S_{\text{sup}}^K; T_{N+}^0 = T_{N+}^K; T_{N-}^0 = T_{N-}^K;$$

$$\Delta g_0 = \hat{\gamma}^{\tau-h} - h\gamma^\tau; \phi^0(t) = \phi^K(t), t \in T_{\text{sup}}^0 \cup T_N^0 \cup (t^* - h);$$

$$\psi^0(t) = \psi^K(t), t \in T_{N+}^0 \cup T_{N-}^0 \cup (t^* - h); \nu^0 = \nu^K; Q^0 = Q^K.$$

Step 5. Calculate $\Delta^K(t) = -\psi^K(t)'b(t, h)$, $t \in T_{N+}^K \cup T_{N-}^K \cup (t^* - h)$. In case $\theta^K = \beta^K = \beta(i_0)$ let

$$\mu^K(i) = [h'_{i_0} \phi^K(t)b(t, h), t \in T_{\text{sup}}^K] Q^K(T_{\text{sup}}^K, i), i \in I_{\text{sup}}^K;$$

$$\xi^K(t) = [h'_{i_0} - \mu^K(I_{\text{sup}}^K)'H(I_{\text{sup}}^K, J)] Q^K(t), t \in T_{N+}^K \cup T_{N-}^K \cup (t^* - h),$$

$$\phi^K(t) = A^{-1}(\theta, h)\phi^K(\theta), t = \theta + h, \theta \in T_{\text{sup}}^K.$$

In case $\theta^K = \alpha^K = \alpha(\tau_s)$ let

$$\mu^K(I_{\text{sup}}^K) = \rho g(\tau_s),$$

$$\xi^K(t) = \rho g(\tau_s)H(I_{\text{sup}}^K, J)\phi^K(t), t \in T_{N+}^K \cup T_{N-}^K \cup (t^* - h),$$

$$\phi^K(t) = A^{-1}(\theta, h)\phi^K(\theta), t = \theta + h, \theta \in T_{\text{sup}}^K, \rho = -\text{sign } \Delta u^K(\tau_s).$$

Calculate $\delta^K(t) = \xi^K(t)'b(t, h)$, $t \in T_{N+}^K \cup T_{N-}^K \cup (t^* - h)$. Transition to step 6.

Step 6. Calculate $\sigma^K = \min\{\sigma(t), t \in T_{N+}^K \cup T_{N-}^K \cup (t^* - h); \omega(i), i \in I_{\text{sup}}^K\}$;
 $s(t), t \in T_{N+}^K \cup T_{N-}^K$:

$$\sigma(t) = -\frac{\Delta^K(t)}{\delta^K(t)}, s(t) = 0 \text{ either } t \in T_{N+}^K, \delta^K(t) < 0 \text{ or } t \in T_{N-}^K, \delta^K(t) > 0;$$

$$\sigma(t) = -\frac{\Delta^K(t-h)}{\delta^K(t-h)} = -\frac{\psi^K(t)'A(t, h)b(t-h, h)}{\xi^K(t)'A(t, h)b(t-h, h)} s(t) = h,$$

either $t \in T_{N+}^K, \delta^K(t-h) > 0$ or $t \in T_{N+}^K, \delta^K(t-h) > 0$

or $t \in T_{N-}^K, \delta^K(t-h) < 0, (t-h) \notin T_{\text{sup}}^K$;

$$\sigma(t) = -\frac{\Delta^K(t-2h)}{\delta^K(t-2h)} = -\frac{\psi^K(t)'A(t, h)A(t-h, h)b(t-2h, h)}{\xi^K(t)'A(t, h)A(t-h, h)b(t-2h, h)}, s(t) = 2h,$$

either $t \in T_{N+}^K, \delta^K(t-2h) > 0$ or $t \in T_{N-}^K$,

$\delta^K(t-2h) < 0, t-2h \notin T_{\text{sup}}^K, t-h \in T_{\text{sup}}^K$;

$$\sigma(t^* - h) = -\frac{\Delta^K(t^* - h)}{\delta^K(t^* - h)}, \text{ when } \Delta^K(t^* - h)\delta^K(t^* - h) < 0, t^* - h \notin T_{\text{sup}}^K,$$

$\sigma(t) = \infty$ in other cases;

$$\omega(i) = -\frac{\nu^K(i)}{\mu^K(i)}, \text{ either } \nu^K(i)\mu^K(i) < 0 \text{ or } \nu^K(i) = 0, \mu^K(i) > 0;$$

$\omega(i) = \infty$ in other cases.

Proceed to step 7.

Step 7. Transform the set $S_{\text{sup}}^K = \{I_{\text{sup}}^K, T_{\text{sup}}^K\}$ and the matrix Q^K .

1) Let $Q^K = \beta(i_0) < 1, \sigma^K = \omega(i_*)$. Then

$$I_{\text{sup}}^{K+1} = (I_{\text{sup}}^K \setminus i_*) \cup i_0, T_{\text{sup}}^{K+1} = T_{\text{sup}}^K,$$

$$Q^{K+1}(\tau_j, i) = Q^K(\tau_j, i) + Q^K(\tau_j, i_*)r^K(i)|r^K(i_*), i \neq i_*;$$

$$Q^{K+1}(\tau_j, i_*) = Q^K(\tau_j, i_*)|r^K(i_*),$$

where $r^K = (r^K(i), i \in I_{\text{sup}}^K) = [h'_{i_0} \phi^K(t)b(t, h), t \in T_{\text{sup}}^K]Q^K(T_{\text{sup}}^K, I_{\text{sup}}^K)$.

2) Let $\theta^K = \beta(i_0) < 1, \sigma^K = \sigma(t_g)$. Then

$$I_{\text{sup}}^{K+1} = I_{\text{sup}}^K \cup i_0, T_{\text{sup}}^{K+1} = T_{\text{sup}}^K \cup (tg - s(tg)),$$

$$Q^{K+1}(\tau_j, i) = Q^K(\tau_j, i) + r_1^K(\tau_j)r_2^K(i)|\rho\delta(tg - s(tg)),$$

$$Q^{K+1}(tg - s(tg), i) = -r_2^K(i)|\rho\delta(tg - s(tg)),$$

$$Q^{K+1}(\tau_j, i_0) = -r_1^K(\tau_j)|\rho\delta(tg - s(tg)), Q^{K+1}(tg - s(tg), i_0) = 1|\rho\delta(tg - s(tg)),$$

$$r_1^K = (r_1^K(\tau_j), j = \overline{1, l}) = Q^K H(I_{\text{sup}}^K, J)\phi^K(tg - s(tg))b(tg - s(tg), h),$$

$$r_2^K = (r_2^K(i), i = \overline{1, l}) = [h'_{i_0} \phi^K(t)b(t, h), t \in T_{\text{sup}}^K]Q^K.$$

3) Let $\theta^K = \alpha^K = \alpha(\tau_s) < 1$, $\sigma = \omega(i_*)$. Then

$$\begin{aligned} I_{\text{sup}}^{K+1} &= I_{\text{sup}}^K \setminus i_*, \quad T_{\text{sup}}^{K+1} = T_{\text{sup}}^K \setminus T_s, \\ Q^{K+1}(T_{\text{sup}}^K \setminus \tau_s, I_{\text{sup}}^K \setminus i_*) &= \\ Q^K(T_{\text{sup}}^K \setminus \tau_s, I_{\text{sup}}^K \setminus i_*) - Q^K(T_{\text{sup}}^K \setminus \tau_s, i_*)Q^K(\tau_s, I_{\text{sup}}^K \setminus i_*) \setminus \rho\mu(i_*). \end{aligned}$$

4) Let $\theta^K = \alpha^K = \alpha(\tau_s) < 1$, $\delta = \delta(tg)$. Then

$$\begin{aligned} I_{\text{sup}}^{K+1} &= I_{\text{sup}}^K, \quad T_{\text{sup}}^K = (T_{\text{sup}}^K \setminus \tau_s) \cup (tg - s(tg)), \\ Q^{K+1}(\tau_j, i) &= Q^K(\tau_j, i) - Q^K(\tau_s, i)r^K(\tau_j)|r^K(\tau_s), \quad i \neq s, \\ Q^{K+1}(\tau_s, i) &= Q^K(\tau_s, i)|r^K(\tau_s), \\ r^K &= (r^K(\tau_j), j = \overline{1, l}) = Q^K H(I_{\text{sup}}^K, J)\phi^K(tg - s(tg))b(tg - s(tg), h). \end{aligned}$$

Matrix $\phi^K(tg - s(tg))$ is calculated in a standard way:

$$\begin{aligned} \phi^K(tg - s(tg)) &= \phi^K(tg) = A^{-1}(tg)\phi^K(tg - h), \quad \text{when } tg - h \in T_{\text{sup}}^K, \quad s(tg) = 0; \\ \phi^K(tg - s(tg)) &= \phi^K(\tau_N), \quad \text{when } tg - s(tg) = \tau_N, \\ \phi^K(tg - s(tg)) &= \phi^K(tg)A(tg), \quad \text{when } tg = \tau_N, \quad s(tg) = h, \\ \phi^K(tg - s(tg)) &= \phi^K(tg - h)A(tg - h), \quad \text{when } tg - h \in T_{\text{sup}}^K, \quad s(tg) = 2h. \end{aligned}$$

Let

$$\begin{aligned} U^{K+1}(t) &= u^{(K)}(t) + \theta^K \Delta u^K(t), \quad t \in T(\tau + h); \quad \Delta g^{K+1} = (1 - \theta^K)\Delta g^K; \\ \psi^{K+1}(t) &= \psi^K(t) + \sigma^K \xi^K(t), \quad t \in T_{N+}^K \cup T_{N-}^K \cup (t^* - h); \\ \nu^{K+1} &= \nu^K + \sigma^K \mu^K; \quad W^{K+1} = W^K + \theta^K \Delta W^K \end{aligned}$$

In cases 1), 3): $(T_{N+}^{K+1} \cup T_{N-}^{K+1}) = (T_{N+}^K \cup T_{N-}^K)$,
 in case 2) $(T_{N+}^{K+1} \cup T_{N-}^{K+1}) = (T_{N+}^K \cup T_{N-}^K) \cup (tg - s(tg) + h)$,
 on case 4) $(T_{N+}^{K+1} \cup T_{N-}^{K+1}) = (T_{N+}^K \cup T_{N-}^K) \setminus (\tau_s + h) \cup (tg - s(tg) + h)$.
 Proceed to step 2.

Step 8. Let $s = 1$, $\theta^0 = \alpha^0 = \alpha(\tau_1) = 0$, $\Delta u^{(0)}(\tau_1) = g(\tau_1)\Delta g^0(I_{\text{sup}}^0)$ and proceed to step 5.

7. Example

Illustrate the results by an example of optimization of mechanical movement.

It is necessary at a given moment to transpose a material point which begins its movement along a rectilinear path from some neighbourhood of the given point to a certain region and provide at the moment the velocity the guaranteed value of which is maximal. Besides it is necessary to take into account that all the information about the control result comes from a device which is able to measure the summarized value of the position and velocity from the point with limited exactness.

The mathematical model of the problem has the form

$$\begin{aligned} x_2(3) &\rightarrow \max, & x_1(t+h) &= x_1(t) + hx_2(t), \\ x_2(t+h) &= x_2(t) + hu(t), & |x_1(0)| &\leq 1, \quad x_2(0) = 0, \\ x_1(3) &\leq 1, & 0 \leq u(t) &\leq 1, \quad t = 0, h, 2h, \dots, 3; \\ h &= 0.5; & y &= x_1 + x_2 + \xi, \quad 0 \leq \xi \leq 1. \end{aligned}$$

Present the results of the optimal estimator and the regulator work for the case when the point actually began movement from the point $x_1(0) = 0$ and the following measuring errors were realized

$$\xi(0) = 1/2, \quad \xi(0.5) = 1/4, \quad \xi(1) = 1/2, \quad \xi(1.5) = 1/5, \quad \xi(2) = 1/4,$$

but this information is not known by either the estimator or the regulator.

The a priori optimal control $\check{u}^0(\cdot)$ constructed at the moment $t = 0$ without the results of observation has the form represented at Fig. 1.

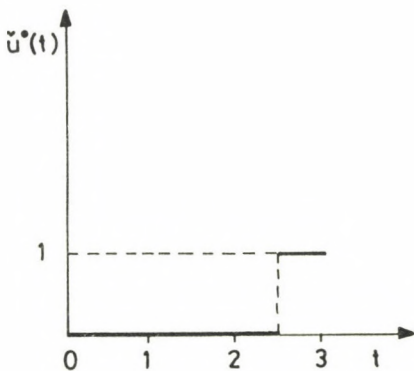


Fig. 1

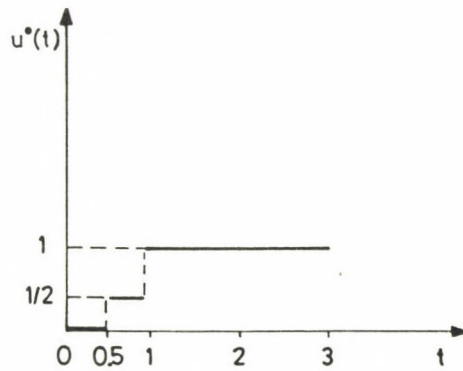


Fig. 2

The guaranteed value of the quality criterion is equal to $J(\check{u}^0(\cdot)) = 1/2$.

If the initial state $x_1(0) = 0$ would be known for the regulator at the moment $t = 0$ then the optimal control $u^0(\cdot)$ has the form shown at Fig. 2. The value of the quality criterion would reach $J(u^0(\cdot)) = 9/4$.

After processing the signal $y(0) = 1/2$ with the estimator the regulator, acting according to the algorithm described above, produced the control $\hat{u}_1^0(\cdot)$, presented at Fig. 3 ($J(\hat{u}_1^0(\cdot)) = 3/4$).

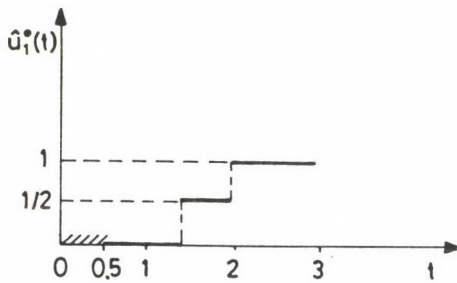


Fig. 3

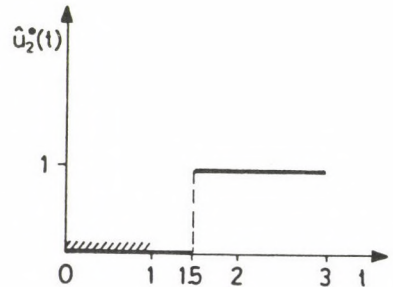


Fig. 4

Making an analogous signal processing $y(0.5) = 1/4$, $y(1) = 1/2$, $y(1.5) = 1/5$, $y(2) = 1/4$, the regulator constructed a priori optimal controls presented at Fig. 4 ($\hat{u}^0(\cdot) = \hat{u}_2^0(\cdot) = \hat{u}_3^0(\cdot) = \hat{u}_4^0(\cdot)$). It is clear that the processing of measurements $y(t)$, $t \geq 1$, does not influence the form of the synthesized control.

The value of the quality criterion on the constructed control is equal to $J(\hat{u}^0(\cdot)) = 3/2$.

The value $J(\hat{u}^0(\cdot)) - J(\tilde{u}^0(\cdot)) = 1$ characterizes the increase of control efficiency at the expense of measuring device.

The loss of efficiency because of the errors of the measuring device equals to $J(u^0(\cdot)) - J(\hat{u}^0(\cdot)) = 3/4$.

References

1. Feldbaum, A. A., Principles of the theory of optimal automatic systems. Moscow, GIFMP, 1963.
2. Germeier, Yu. B., Games with nonopposite interesess. Moscow, Nauka, 1976.
3. Chernous'ko, F. A., Kolmanovskiy, V. B., Optimal control under accidental perturbations. Moscow, Nauka, 1978.
4. Krasovskiy N. N., Theory of control with movement. Moscow, Nauka, 1968.
5. Krasovskiy, N. N., Subbotin, A. I., Position differential games. Moscow, Nauka, 1974.
6. Kurzansky, A. B., Control and observation in indefiniteness conditions. Moscow, Nauka, 1977.
7. Gabasov, R., Kirillova, F. M., Constructive methods of optimization. Part 2. Minsk, Universitetskoe, 1984.
8. Gabasov, R., Kirillova, F. M., Methods of linear programming. Part 3. Minsk, BGU, 1980.

Синтез оптимальных управлений по неточным измерениям выходных сигналов

Р. ГАБАСОВ, Ф. М. КИРИЛЛОВА, П. В. ГАЙШУН, С. В. ПРИЩЕПОВА

(Минск)

В статье предложен метод синтеза дискретных эстиматора и регулятора, оптимизирующих поведение динамической системы в условиях неполного и неточного наблюдения за процессом управления. Синтез осуществляется в режиме реального времени по мере поступления очередного сигнала от измерительного устройства. В основе работы алгоритма лежит процесс коррекции очередного опорного решения задачи.

Р. Габасов, Ф. М. Кириллова
П. В. Гайшун, С. В. Прищепова
Институт математики АН БССР
220604 Минск, Сурганова 11

PARAMETER ESTIMATION FOR NEAREST NEIGHBOR GAUSSIAN RANDOM FIELDS IN THE PLANE

ANTONÍN OTÁHAL

(Prague)

(Received December 30, 1990)

A parameter-estimation method for both scalar and vector-valued Gauss-Markov random fields is presented.

Introduction

Stationary Gauss-Markov random fields represent a rather nice model of spatial randomness as it is possible to get some, more or less explicit, results in their statistical analysis. Künsch (1981) developed ideas of Dobrushin (1980) and applied them to asymptotic statistical analysis or, as he chose to call it, thermodynamics of stationary Gaussian fields. The last term indicates that the Gaussian (and the Gauss-Markov) fields can be viewed in frame of statistical physics generalizations, a Gaussian field being considered as a Gibbs field corresponding to given potentials (interactions).

Both Künsch (1981) and Janžura (1988), who gave some deeper asymptotic results for the Gauss-Markov fields, considered what we might call "fitting-of-moments" estimates of parameters. Namely, the lags corresponding to non-vanishing interactions were taken, the corresponding sample covariances were calculated and the interactions estimates were fitted to these sample covariances. This is quite a natural way of doing it and both the mentioned authors proved good asymptotic properties (such as consistency, asymptotic normality etc.) of the estimates.

What we concentrate on in the present paper is how to actually calculate these estimates. We confine our effort to the nearest neighbor model in the plane that is not only the simplest non-trivial case but also the most interesting one from the practical point of view. We consider both the scalar and the vector-valued cases.

First, the general set-up is formulated, then the estimation procedures are derived. The first one concerns the scalar case estimates. The vector-valued field estimating procedure is based on that the problem is, in a sense, transformed to a set of scalar ones.

1. Problem Formulation

A system $X = (X(t): t \in T)$ of n -dimensional (column) random vectors indexed by two-component integer vectors is a *vector-valued random field in the plane*. That is, $T = Z^2$ where Z is the set of all integers and, for every $t \in T$, it is $X(t) = (X_1(t), \dots, X_n(t))^*$ where asterisk means matrix transposition. The field X is *Gaussian* if all finite-dimensional distributions of all variables in X are Gaussian. It is *stationary* if the mean value is constant, say $EX(t) = 0$ for every $t \in T$, and the second moments are shift-invariant, in other words, the covariances depend on the lags only: $EX(t) \cdot X(s)^* = R(t - s)$ where R is an $n \times n$ matrix *covariance function*. If a matrix function f exists such that

$$R(t) = (2\pi)^{-2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cos(t_1 x + t_2 y) \cdot f(x, y) dx dy$$

holds for every $t \in T$, we call f a *spectral density* of the field X .

Let us denote $\mathcal{U} \in \mathcal{U}$ the set of all $U = (U_0, U_1, U_2)$ such that U_0, U_1, U_2 are symmetric real $n \times n$ -matrices and $U_0 \pm U_1 \pm U_2$ is a positive definite matrix for every choice of signs. For $U \in \mathcal{U}$ we define

$$f_U(x, y) = (U_0 + U_1 \cos x + U_2 \cos y)^{-1}$$

where $^{-1}$ denotes the matrix inversion. From the properties of the set \mathcal{U} it follows that the matrix function f_U is positive definite for every $U \in \mathcal{U}$, and every $x, y \in [-\pi, \pi)$.

We say that a Gaussian stationary vector-valued random field is a *nearest neighbor* one if it has a spectral density which is, for some $U \in \mathcal{U}$, expressed in the form f_U . In other words, the considered field is nearest neighbor if its covariance function is given

$$R^U(t) = (2\pi)^{-2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \cos(t_1 x + t_2 y) \cdot f_U(x, y) dx dy$$

for every $t \in T$.

We denote $A = \{(0, 0), (0, 1)(1, 0)\}$. Let us consider a finite large subset M of the index set T ; M is typically a rectangle and "large" means that $|M_t|/|M| \approx 1$ for $t \in A$ where $M_t = \{s \in M: s + t \in M\}$ and $|\cdot|$ denotes the cardinality of a set. We define *sample covariances*

$$\hat{R}(t) = (2|M_t|)^{-1} \sum_{s \in M_t} (X(s+t)X(s)^* + X(s)X(s+t)^*)$$

for $t \in A$. An estimate \hat{U} of $U \in \mathcal{U}$ is then defined by equations

$$R^{\hat{U}}(t) = \hat{R}(t), \quad t \in A.$$

The estimate \hat{U} is defined with a probability near to 1 for large M . This and some good asymptotic properties of U were proved in a scalar case by Janžura (1988). It is not difficult to generalize those results to the vector case.

2. Scalar Estimate

The above considerations simplify for the scalar case $n = 1$; namely, scalar values take places of $n \times n$ -matrices. Let us write r_0, r_1, r_2 for $\hat{R}(0, 0), \hat{R}(1, 0), \hat{R}(0, 1)$, respectively. The estimate \hat{U} we represent as a triplet (u_0, u_1, u_2) and we put $\rho = r_1/r_2, \sigma = r_2/r_0, \alpha = u_1/u_0, \beta = u_2/u_0$. The condition $U \in \mathcal{U}$ takes the form $|\alpha| + |\beta| < 1$.

If we denote

$$I_0(\alpha, \beta) = (2\pi)^{-2} \int_{-\pi}^{\pi} \frac{dx dy}{1 + \alpha \cos x + \beta \cos y}$$

$$I_1(\alpha, \beta) = (2\pi)^{-2} \int_{-\pi}^{\pi} \frac{\cos x dx dy}{1 + \alpha \cos x + \beta \cos y}$$

$$I_2(\alpha, \beta) = (2\pi)^{-2} \int_{-\pi}^{\pi} \frac{\cos y dx dy}{1 + \alpha \cos x + \beta \cos y}$$

then u_0, α, β are given by the following basic system of equations

$$r_0 u_0 = I_0(\alpha, \beta)$$

$$\rho r_0 u_0 = I_1(\alpha, \beta)$$

$$\sigma r_0 u_0 = I_2(\alpha, \beta)$$

and the original parameters u_1, u_2 are expressed as $u_1 = \alpha \cdot u_0, u_2 = \beta \cdot u_0$.

Generally, the integrals I_0, I_2 are not integrable explicitly. Let us denote [cf. e.g. Bateman (1953), Chap. XIII]

$$F(k) = (2/n) \int_0^{\pi/2} (1 - k^2 \sin^2 \varphi)^{\frac{1}{2}} d\varphi$$

and

$$H(k, p) = (2/\pi) \int_0^{\pi/2} (1 - (p-1) \sin^2 \varphi)(1 - k^2 \sin^2 \varphi)^{\frac{1}{2}} d\varphi$$

the "normed" complete *elliptic integrals* of the first and third kinds, respectively. After some integration and substitution we get, for $\alpha \neq 0$, $\beta \neq 0$,

$$\begin{aligned} I_0(\alpha, \beta) &= m \cdot F(k) \\ I_1(\alpha, \beta) &= \alpha^{-1} m ((\beta - 1)F(k) + (1 - \alpha - \beta)H(k, p)) \\ I_2(\alpha, \beta) &= \beta^{-1} m ((\alpha - 1)F(k) + (1 - \alpha - \beta)H(k, q)) \end{aligned}$$

where

$$\begin{aligned} m &= (1 - (\alpha - \beta)^2)^{\frac{1}{2}}, \\ k^2 &= 4|\alpha\beta|m^2, \\ p &= (1 - \alpha - \beta)/(1 + \alpha - \beta), \\ q &= (1 - \alpha - \beta)/(1 - \alpha + \beta). \end{aligned}$$

The cases $\alpha = 0$, $\beta = 0$ are trivial: suppose that one of the correlations ρ , σ vanishes, say $\alpha = 0$. Then β has to vanish, both I_0 and I_1 are explicitly integrable and the corresponding solution of the basic system is

$$\alpha = -2p/(1 - \rho^2), \quad \beta = 0, \quad u_0 = (1 + \rho^2)/[r_0(1 - \rho^2)];$$

and symmetrically for $\rho = 0$.

If none of the correlations vanishes we may consider the case $\alpha > 0$, $\beta > 0$ only. In fact, any other case is reduced to this one by taking into account the following rules (that are easy to derive from the expressions of the integrals): $I_0(\pm\alpha, \pm\beta) = I_0(\alpha, \beta)$, $I_1(\alpha \pm \beta) = I_1(\alpha, \beta)$, $I_2(\pm\alpha, \beta) = I_2(\alpha, \beta)$, for every choice of signs and $I_1(-\alpha, \beta) = -I_1(\alpha, \beta)$, $I_2(\alpha, -\beta) = I_2(\alpha, \beta)$.

At this moment we have to stop thinking about an explicit solution of the basic system. Instead we describe a numerical procedure which converges to the solution.

The procedure needs evaluations of complete elliptic integrals. Though general numerical integration could be used, it is much better to apply specialized procedures of Bullirsch (1965) which are very fast and accurate.

The solution (for the mentioned case $\rho > 0$, $\sigma > 0$) is found in two main steps.

2.1. Isotropic Approximation

We put $r = (\rho + \sigma)/2$ and solve an equation in ξ

$$F(\xi) \cdot (1 + r \cdot \xi) = 1$$

to which the basic system reduces for $\rho = \sigma = \tau$, u_0 being separated as $r_0 u_0 (1 + r\xi) = 1$. The equation is solved iteratively

$$\xi_{n+1} = F^{-1}(1/(1 + \tau \cdot \xi_n))$$

with an initial value $\xi_0 = -2\tau/(1 + \tau^2)$ which would correspond to a vanishing correlation. The inversion F^{-1} is again computed iteratively, by the regula falsi method. The initial pair of values $\tau_1 = (1 - 4 \cdot \exp\{-\pi\tau\})^{\frac{1}{2}}$, $\tau_2 = (1 - \exp\{\pi(1 - \tau)\})^{\frac{1}{2}}$, follows from Tricomi's inequality

$$2 \cdot \ln 4 \leq \pi \cdot F(k) + \ln(1 - k^2) \leq \pi;$$

cf. Bateman (1953), 13.8.9.

2.2. General Case

Again we separate u_0 as

$$1/u_0 = r_0 \cdot (1 + \rho \cdot \alpha + \sigma \cdot \beta)$$

using an obvious identity $I_0 + \alpha \cdot I_1 + \beta \cdot I_2 \equiv 1$. Thus, the basic system reduces to

$$\begin{aligned} f(\alpha, \beta) &:= I_1(\alpha, \beta) - \rho/(1 + \rho\alpha + \sigma\beta) = 0 \\ g(\alpha, \beta) &:= I_2(\alpha, \beta) - \sigma/(1 + \rho\alpha + \sigma\beta) = 0. \end{aligned}$$

Now, take initial values $\alpha_0 = \beta_0 = \xi/2$ which correspond to the isotropic approximation solution and we iterate "Newton-like"

$$\begin{pmatrix} \alpha_{n+1} \\ \beta_{n+1} \end{pmatrix} = \begin{pmatrix} \alpha_n \\ \beta_n \end{pmatrix} - c_n \cdot J^{-1}(\alpha_n, \beta_n, h) \cdot \begin{pmatrix} f(\alpha_n, \beta_n) \\ g(\alpha_n, \beta_n) \end{pmatrix}.$$

Here $J(\alpha, \beta, h)$ is a "difference Jacobi matrix",

$$J(\alpha, \beta, h) = (1/h) \begin{pmatrix} f(\alpha + h, \beta) & f(\alpha, \beta + h) \\ g(\alpha + h, \beta) & g(\alpha, \beta + h) \end{pmatrix} - (1/h) \begin{pmatrix} f(\alpha, \beta) & f(\alpha, \beta) \\ g(\alpha, \beta) & g(\alpha, \beta) \end{pmatrix}$$

and $c_n = 2^{-j}$ where j is the minimal non-negative integer for which the corresponding new values α_{n+1} , β_{n+1} fulfil the regularity condition $|\alpha_{n+1}| + |\beta_{n+1}| < 1$.

2.3. Remarks

1. Theoretically, the conditions $|\rho| < 1, |\sigma| < 1$ are necessary and sufficient for the feasibility of computations. Practically, the nearer the values $|\rho|, |\sigma|$ are to the bound 1 the higher accuracy of computations is needed. For example FORTRAN double precision (i.e. 8 hexadecimal digits accuracy) is still able to manage the situation when both the correlations are (in absolute value) about 0.9.

2. All the considered iterations terminate when the consecutive values differ less than some $\epsilon > 0$. In the mentioned (i.e. FORTRAN double precision) implementation, $\epsilon = 10^{-8}$ and $h = 10^{-6}$ have proved to be useful.

3. Vector Estimate

We want to solve the general system of equations

$$R^{\hat{U}}(t) = \hat{R}(t), \quad t \in A$$

with $n \times n$ matrices on both sides. This is a difficult task even for moderate values of n . That is why we consider a more restricted problem: we suppose the *separability of channels* of the given vector-valued field. Namely, we suppose that there exists a regular $n \times n$ matrix L for which all the matrices $L \cdot R(t) \cdot L^*$ are diagonal for every $t \in T$. In other words, we suppose an existence of a regular linear transformation which transforms the original field onto a new one and this new one is an n -tuple of mutually (stochastically) independent scalar-valued random fields (*channels*).

A matrix L separates channels of a nearest neighbor field if and only if its spectral density f_U corresponds to such $U = (U_0, U_1, U_2, \dots) \in \mathcal{U}$ that the matrices $L \cdot U_j \cdot L^*$, $j = 0, 1, 2, \dots$ are diagonal. In fact, diagonal U_j 's obviously lead to a covariance function which has diagonal $R(t)$'s; on the other hand, every separated channel is a nearest neighbor scalar field. Hence, our statement follows from that the correspondence between spectral densities and covariance functions is one-to-one, cf. Stein, Weiss (1971), Thm. VII.1.7.

Let us define a "correlation" function on T by means of $\rho(t) = R(0)^{-\frac{1}{2}} \cdot R(t) \cdot R(0)^{-\frac{1}{2}}$. It is easy to see that a random field has separable channels if and only if there exists an orthogonal matrix Q such that matrices $Q \cdot \rho(t) \cdot Q^*$ are diagonal for all $t \in T$. For a nearest neighbor field it is the same as that the matrices $\rho(0, 1)$ and $\rho(1, 0)$ form a *reducible pair* what means there exists an orthogonal matrix Q for which both $Q \cdot \rho(0, 1) \cdot Q^*$ and $Q \cdot \rho(1, 0) \cdot Q^*$ are diagonal.

Sample correlation function $\hat{\rho}$ is defined in a natural way on the base of the above defined sample covariance function.

$$\hat{\rho}(t) = \hat{R}(t) \hat{R}(0)^{-\frac{1}{2}};$$

this is a reasonable definition because $\hat{R}(0)$ is a positive definite matrix with probability 1. Clearly, $\hat{\rho}(0, 1)$ and $\hat{\rho}(1, 0)$ need not form a reducible pair. So as to meet the separability-of-channels assumption we should find a new pair of matrices, say $\tilde{\rho}(0, 1)$ and $\tilde{\rho}(1, 0)$, which would be, in some sense, the best approximation of the original one in the class of reducible pairs.

3.1. Reducible Approximation Problem

For an $n \times n$ matrix $V = (v_{jk})$ we define a diagonal matrix $\Delta(V)$ whose main diagonal is the same as that of V i.e. $\Delta(V)_{jk} = \delta_{jk}v_{jk}$ where δ is the usual Kronecker symbol. By a norm of V we understand $\|V\| = \Sigma_j \Sigma_k v_{jk}^2)^{\frac{1}{2}}$.

Let there be given two symmetric $n \times n$ matrices E, F . For an orthogonal $n \times n$ matrix Q we denote $E_Q = QEQ^*, F_Q = QFQ^*$ and put

$$c(E, F, Q) = \|E_Q - \Delta(E_Q)\|^2 + \|F_Q - \Delta(F_Q)\|^2.$$

The problem of *reducible approximation* consists in:

- (RA) find an orthogonal $n \times n$ matrix Q^0 such that $c(E, F, Q^0) = \min_Q c(E, F, Q)$ where the minimum is taken over all orthogonal matrices Q .

Before going into solution of (RA) we introduce some symbolics. Let us take two integers l, m such that $1 \leq l < m \leq n$.

If $V = (v_{jk})$ is an $n \times n$ matrix then $P_{lm}(V)$ denotes the corresponding 2×2 submatrix: writing $W = P_{lm}(V)$ we put $w_{11} = v_{ll}, w_{12} = v_{lm}, w_{21} = v_{ml}, w_{22} = v_{mm}$.

The other way round, if W is a 2×2 matrix then $V = P_{lm}^{-1}(W)$ denotes such an $n \times n$ matrix for which $v_{ll} = w_{11}, v_{lm} = w_{12}, v_{ml} = w_{21}, v_{mm} = w_{22}$, and $v_{jk} = \delta_{jk}$ for $(j, k) \notin \{(l, l), (l, m), (m, l), (m, m)\}$.

For a real number z we denote $H(z)$ the corresponding *planar rotation matrix*

$$H(z) = \begin{pmatrix} \cos z & \sin z \\ -\sin z & \cos z \end{pmatrix}$$

and, on the base of this, we define a so-called *Givens matrix* $G(l, m, z) = P_{lm}^{-1}(H(z))$.

The problem (RA) will be solved iteratively: in every step a Givens matrix will be determined and the desired orthogonal matrix Q^0 will be approximated by a product of thus obtained sequence of Givens matrices. The method generalizes the known *Jacobi method* for diagonalization of a real symmetric matrix, cf. Wilkinson (1965), Chap. 5.

3.2. Small Reducible Approximation

As a starting point, the problem (RA) will be solved in case E, F are symmetric 2×2 matrices. In this case it is possible to solve (RA) explicitly; we may restrict the minimization onto the above defined planar rotation matrices $H(z)$, so we look for a real z_0 which minimizes the criterion $c(E, F, H(z))$.

Let us denote

$$\begin{aligned} p &= (e_{11} - e_{22})/2 \\ q &= (f_{11} - f_{22})/2 \\ x &= p^2 + q^2 - e_{12}^2 - f_{12}^2 \\ y &= 2(p \cdot e_{12} + q \cdot f_{12}). \end{aligned}$$

If we express the criterion c in terms of x, y, z it is not difficult to derive

1. if $x = y = 0$ then the criterion value is independent of z and we may put $z_0 = 0$,
2. if $x^2 + y^2 > 0$ then the criterion is minimized by that real z_0 for which

$$\begin{aligned} \cos z_0 &= (h + h(h + x \cdot h/(x^2 + y^2)^h)^h)^h \\ \sin z_0 &= (h - h(h + x \cdot h/(x^2 + y^2)^h)^h)^h \cdot (-\operatorname{sgn} y) \end{aligned}$$

where $h = \frac{1}{2}$ and sgn is the usual sign function. These expressions are not the most elegant ones; but viewed as a recipe for computation those very expressions are the most stable ones from the numerical point of view when being repeatedly used in the later solution of the general problem.

The decrement of the criterion value which corresponds to the optimal z_0 is

$$c(E, F, H(0)) - c(E, F, H(z_0)) = (x^2 + y^2)^{\frac{1}{2}} - x.$$

3.3 Reducible Approximation Procedure

For $n > 2$, (RA) is solved iteratively.

As an initialization, we put $E^{(1)} = E$, $F^{(1)} = F$, $Q^{(1)} = I$ where I is an identity $n \times n$ matrix.

The N -th iteration step consists of several substeps. Let us for a while denote $D = \{(l, m): 1 \leq l < m \leq n, l, m \text{ integers}\}$.

1. *Choice of 2×2 submatrices.* For $(l, m) \in D$ we take the corresponding 2×2 submatrices $P_{lm}(E^{(N)})$, $P_{lm}(F^{(N)})$. For these 2×2 matrices we determine p, q, x, y as in 3.2. Further, we denote Δ_{lm} the corresponding criterion decrement, $\Delta_{lm} = (x^2 + y^2)^{\frac{1}{2}} - x$. Now, we take (j, k) such that $\Delta_{jk} = \max\{\Delta_{lm}: (l, m) \in D\}$.
2. *Small solution.* For the chosen (j, k) we solve the small reducible approximation problem on the matrices $P_{jk}(E^{(N)})$, $P_{jk}(F^{(N)})$ what provides a planar rotation matrix $H(z_0)$.
3. *Innovation.* We take a Givens matrix $G = P_{jk}^{-1}(H(z_0))$ and put

$$\begin{aligned} E^{(N+1)} &= G \cdot E^{(N)} \cdot G, \\ F^{(N+1)} &= G \cdot F^{(N)} \cdot G, \\ Q^{(N+1)} &= G \cdot Q^{(N)}. \end{aligned}$$

Termination. If, in the 1-st substep of the N -th iteration step, the maximal possible criterion decrement Δ_{jk} is “almost zero”, i.e. $\Delta_{jk} < \epsilon$ where ϵ is some pre-defined small positive number, the iterations terminate and the matrix $Q^{(N)}$ is taken for a solution of the problem (RA).

3.4. Convergence of Iterations

It is easy to see that the criterion $c(E, F, Q)$ decreases during the iterations; in fact, the actual decrement is equal to Δ_{jk} in every iteration step. From this the convergence of iterations follows. As for the convergence rate, any theoretical bounds would be extremely difficult to derive. An implementation shows that the convergence is very quick even for every small values of ϵ (it was put $\epsilon = 10^{-8} \cdot (\|E\|^2 + \|F\|^2)$).

Another open (and difficult) question consists in that it is not a priori known whether the iterations terminate in the very minimum of the criterion.

3.5. Separable-channels Estimate

Let us put $\hat{\rho}(0, 1) = E$ and $\hat{\rho}(1, 0) = F$ in the problem (RA). We denote $r(0, 1) = \Delta(E^{(N)})$, $r(1, 0) = \Delta(F^{(N)})$, and $Q = Q^{(N)}$ where the superscript (N) denotes the matrices given by the terminal iteration step of the above described procedure. The matrices $\tilde{\rho}(0, 1) = Q^* \cdot r(0, 1) \cdot Q$ and $\tilde{\rho}(1, 0) = Q^* \cdot r(1, 0) \cdot Q$ represent the reducible pair that is the best approximation of $\hat{\rho}(0, 1)$ and $\hat{\rho}(1, 0)$.

Now, it is obvious how to use the scalar estimation procedure in estimating parameters of a separable-channels nearest neighbor random field. In fact, for $j = 1, \dots, n$ we put $\rho = r(0, 1)_{jj}$, $\sigma = r(1, 0)_{jj}$, $r_0 = 1$ and apply the scalar estimation procedure, getting values u_0^j , u_1^j , u_2^j . Depending on further analysis purposes either these values may be taken as the final result or, if we tend to get an estimate independent of the channels separation, we can go back putting, for $k = 0, 1, 2$,

$$U_k = M \cdot \begin{pmatrix} u_k^1 & & 0 \\ & \dots & \\ 0 & & u_k^n \end{pmatrix} \cdot M^*$$

where $M = \hat{R}(0)^{\frac{1}{2}} \cdot Q^*$.

3.6. Remark

Absolute values of eigenvalues of $\hat{\rho}(0, 1)$, $\hat{\rho}(1, 0)$ have to be less than 1. This implies that the terminal criterion value in reducible approximation of these matri-

ces can not be too large and, consequently, the pair $\tilde{\rho}(0, 1)$, $\tilde{\rho}(1, 0)$ can not depart too far from $\hat{\rho}(0, 1)$, $\hat{\rho}(1, 0)$. In other words, the reducible approximation could be taken as a means of an approximate estimation procedure for vector-valued nearest neighbor random fields, without even mentioning the separability-of-channels assumption.

4. Application Range

From the estimation problem formulation it follows that the considered type of random fields is a suitable model in digital image processing. Here the interactions could represent a good characterization e.g. of textures. More generally, the interactions could serve as some global characteristics of any planar data, providing the stationarity assumption is acceptable.

The results concerning the vector-valued fields would be easily generalized for Gauss–Markov fields, for which the interactions are not restricted to the nearest neighbors, as well as for d -dimensional index set Z^d . Only a generalization of the reducible approximation procedure to more than two matrices would be needed and that could be easily done. Of course, the more interactions are taken into consideration the more restrictive it is to suppose that the channels are separable.

References

1. *Bateman, H.*, Higher Transcendental Functions, Vol. II. McGraw-Hill, New York–Toronto–London, 1953.
2. *Bullirsch, R.*, Numerical calculation of elliptic integrals and elliptic functions. Handbook Series Special Functions, Numerische Mathematik 7 (1965), pp. 78–90.
3. *Dobrushin, R. L.*, Gaussian random fields — Gibbsian point of view. In: Multicomponent Random Systems, eds. R. L. Dobrushin, Ya. G. Sinai. M. Dekker, New York, 1980.
4. *Janžura, M.*, Asymptotic theory of parameter estimation for Gauss–Markov random fields, Kybernetika 24 (1988), No. 3, pp. 161–176.
5. *Künsch, H.*, Thermodynamics and statistical analysis of Gaussian random fields. Z. Wahrscheinlichkeitstheorie und verw. Gebiete 55 (1981), pp. 407–421.
6. *Stein, E. M., Weiss, G.*, Introduction to Fourier Analysis on Euclidean Spaces. Princeton Univ. Press, 1971.
7. *Wilkinson, J. H.*, The Algebraic Eigenvalue Problem. Clarendon Press, Oxford, 1965.

**Оценка параметров марковских гауссовских случайных полей в
плоскости**

А. ОТАХАЛ

(Прага)

В статье дается метод оценивания параметров для марковских гауссовских случайных полей и как для скалярной, так и для векторной области значений данного поля.

RNDr. Antonín Otáhal, CSc.,
Institute of Information Theory and Automation
of the Czechoslovak Academy of Sciences,
Pod vodárenskou věží 4,
18208 Praha 8,
Czechoslovakia

NONPARAMETRIC ENTROPY ESTIMATION BASED ON RANDOMLY CENSORED DATA

A. CARBONEZ, L. GYÖRFI, E. C. VAN DER MEULEN

(*Leuven, Budapest*)

(Received October 1, 1991)

The Shannon entropy of a random variable X with density function $f(x)$ is defined as $H(f) = - \int f(x) \log f(x) dx$.

Based on randomly censored observations a nonparametric estimator for $H(f)$ is proposed if $H(f)$ is finite and is nonnegative. This entropy estimator is histogram-based in the sense that it involves a histogram-based density estimator \hat{f}_n constructed from the censored data. We prove the a. s. consistency of this estimator.

1. Introduction

The entropy of a probability density function $f(x)$ of a nonnegative random variable X is defined by

$$H(f) = - \int_0^{+\infty} f(x) \log f(x) dx. \quad (1)$$

In the literature, several estimators of entropy have been proposed for non-censored observations. Typically, these estimators are based on obtaining first (cf. Györfi and van der Meulen [5]) a suitable density estimate $f_n(x)$ for $f(x)$ and then substituting f_n for f in an entropy-like functional. In the random censorship model, one observes random variables $Z_i = \min(X_i, Y_i)$ and indicator variables $\delta_i = I[X_i \leq Y_i]$, $i = 1, \dots, n$. The random variables X_i of interest, and the censoring variables Y_i , are i.i.d. and nonnegative. Moreover X_i and Y_i are independent for all i . We assume $F(x) = P[X > x]$, $G(x) = P[Y > x]$ and $K(x) = P[Z > x] = F(x)G(x)$ are continuous. The product-limit estimator \hat{F}_n (Kaplan and Meier [7]), which is based on the observations (Z_i, δ_i) , $i = 1, \dots, n$, is often used to estimate $F(x)$. Here the nonnegative random variable X is supposed to have a density function f with probability measure μ defined on the Borel sets of \mathbb{R} .

The notion of *fair censoring* was introduced in the paper by Carbonez, Györfi and van der Meulen [1]): censoring is called *fair* if for any T for which $F(T) > 0$, it holds that $G(T) > 0$.

Define $T^* = \sup\{T : 0 < F(T)G(T) < 1\}$. Let $\mathcal{P}_n = \{I_{n1}, I_{n2}, \dots\}$ be a partition of the real line $n \geq 1$, with $I_{nj} = [t_{j-1,n}; t_{j,n})$.

Remark. In the uncensored case, the a. s. L_1 rate of convergence of a density estimate implies the a. s. convergence of the corresponding entropy estimate (see Theorem 2, Györfi and van der Meulen [6]).

Given the partition \mathcal{P}_n , the intervals I_{nj} have length h_n , $\lambda(I_{nj}) = h_n$, $0 < h_n < 1$, where $\lambda(I_{nj})$ denotes the Lebesgue measure of I_{nj} and with Kaplan–Meier measure $\hat{\mu}_n(I_{nj}) = \hat{F}_n(t_{j-1,n}) - \hat{F}_n(t_{j,n})$.

The histogram estimate of $f(x)$ in this censored situation then has the following form:

$$\hat{f}_n(x) = \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})}, \quad x \in I_{nj}. \quad (2)$$

Now choose $0 < a_n < 1$, $0 < b_n < 1/2$ and introduce the notations

$$\mathcal{F}_n = \{j : \hat{\mu}_n(I_{nj}) \geq a_n h_n\}$$

$$\hat{T}_n = K_n^{-1}(b_n)$$

$$\hat{\mathcal{G}}_n = \{j : I_{nj} \subset [0, \hat{T}_n]\}$$

where

$$K_n(x) = \frac{1}{n} \sum_{i=1}^n I_{\{Z_i \geq x\}}$$

then our estimate of $H(f)$ is defined by

$$H_n = - \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \hat{\mu}_n(I_{nj}) \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})}. \quad (3)$$

2. Main result

We then have the following Theorem which states the a. s. consistency of the randomly censored version of the histogram-based entropy estimate.

THEOREM. Assume that the censoring is fair, and if \mathcal{P}_0 denotes the partition of \mathbf{R} by unit intervals that

$$- \sum_{A \in \mathcal{P}_0} \mu(A) \log \mu(A) < \infty, \quad (4)$$

and

$$0 < a_n < 1, \quad \lim_{n \rightarrow \infty} a_n = 0, \tag{5}$$

$$0 < b_n < 1/2, \quad \lim_{n \rightarrow \infty} b_n = 0, \tag{6}$$

and that

$$\sum_{n=1}^{\infty} \frac{1}{a_n^2 h_n^2 b_n} \exp(-C_1 n b_n^8 a_n^2 h_n^2) < \infty \tag{7}$$

for all $C_1 > 0$ and h_n^{-1} is an integer such that

$$\lim_{n \rightarrow \infty} h_n = 0. \tag{8}$$

Then, if

$$H(f) = - \int_0^{T^*} f(x) \log f(x) dx \text{ is finite,} \tag{9}$$

$$\lim_{n \rightarrow \infty} H_n = H(f) \text{ a. s.}$$

In order to prove this Theorem, we need the following lemma's.

Lemma 1 (Rényi [8], Csiszár [2, 3]). If $H(f)$ is finite, then under conditions (4) and (8) we have

$$\lim_{n \rightarrow \infty} - \sum_j \mu(I_{nj}) \log \frac{\mu(I_{nj})}{\lambda(I_{nj})} = H(f). \tag{10}$$

Lemma 2 (extension of Lemma 4 of Györfi and van der Meulen [5]). For each $\epsilon > 0$ and every interval $I \subset [0, T]$, $T < T^*$, $K(T) < 1/2$,

$$P \left(\left| \log \frac{\mu(I)}{\hat{\mu}_n(I)} \right| > \epsilon \right) \leq 2 \exp \left\{ - \frac{nK(T)}{16} (\mu(I)\delta_\epsilon)^2 \right\} + \frac{11520}{K(T)\mu(I)\delta_\epsilon} \exp \left\{ -n \frac{K(T)^8}{288} (\mu(I)\delta_\epsilon)^2 \right\} \tag{11}$$

where $\delta_\epsilon = 1 - 2^{-\epsilon}$.

Proof. First observe that as in Györfi and van der Meulen [5],

$$\left\{ \left| \log \frac{\mu(I)}{\hat{\mu}_n(I)} \right| > \epsilon \right\} \subset \{ |\hat{\mu}_n(I) - \mu(I)| > \mu(I)(1 - 2^{-\epsilon}) \}.$$

Lemma 1 from Carbonez, Györfi and van der Meulen [1], states that

$$P[|\hat{\mu}_n(I) - \mu(I)| > \epsilon] \leq 2 \exp(-n\epsilon^2 G(T)/16) + \frac{11520}{G(T)\epsilon} \exp \left\{ -n\rho^6 \frac{G(T)^2 \epsilon^2}{288} \right\}, \tag{12}$$

where $\rho = \min\{K(T), 1 - K(T)\} > 0$, $K(T) < 1/2$ therefore $\rho = K(T)$. Moreover $G(T) \geq K(T)$ so (12) implies (11).

Lemma 3. If conditions (5), (6), (7) and (8) hold for all $C_1 > 0$ then

$$\sum_{n=1}^{\infty} \frac{c_n}{b_n} \exp\left(-C_1 \frac{nb_n^8}{c_n^2}\right) < \infty \tag{13}$$

where $c_n = \log \frac{1}{h_n} + \log \frac{1}{a_n}$.

Lemma 4 (Theorem 1 of Carbonez, Györfi and van der Meulen [1]). Assume that the censoring is fair. If

$$\lim_{n \rightarrow \infty} h_n = 0 \tag{14}$$

and

$$\lim_{n \rightarrow \infty} nh_n = \infty \tag{15}$$

then

$$J_n = \int_0^{+\infty} |f_n - f| \rightarrow 0 \quad \text{a. s. as } n \rightarrow \infty \tag{16}$$

for all f .

3. Proof of the theorem

Since

$$H_n = - \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \hat{\mu}_n(I_{nj}) \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})}, \tag{17}$$

we can write

$$H_n - H(f) = U_n + V_n + W_n + Z_n. \tag{18}$$

Hereby, we put

$$U_n = \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} (-\hat{\mu}_n(I_{nj}) + \mu(I_{nj})) \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})}, \tag{19}$$

$$V_n = - \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{\hat{\mu}_n(I_{nj})}{\mu(I_{nj})}, \tag{20}$$

$$W_n = \sum_{j \notin \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{\mu(I_{nj})}{\lambda(I_{nj})}, \tag{21}$$

$$Z_n = - \sum_j \mu(I_{nj}) \log \frac{\mu(I_{nj})}{\lambda(I_{nj})} - H(f), \tag{22}$$

and thus

$$|H_n - H(f)| \leq |U_n| + |V_n| + |W_n| + |Z_n|. \tag{23}$$

(I) Observe that $|U_n| = U_n^+ + U_n^- = U_n^+ + (-U_n)^+$.

Introduce the notations

$$T_n = K^{-1}(b_n/2),$$

$$\mathcal{G}_n = \{j : I_{nj} \subset [0, T_n]\}.$$

Obviously

$$P(|U_n| > \epsilon) \leq P(|U_n| > \epsilon, \hat{T}_n < T_n) + P(\hat{T}_n \geq T_n).$$

So for $\hat{T}_n < T_n$ we have $\hat{\mathcal{G}}_n \subset \mathcal{G}_n$, therefore for $\hat{T}_n < T_n$

$$|U_n| \leq \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\mu(I_{nj}) - \hat{\mu}_n(I_{nj}))^+ \left\{ \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})} \right\}^+$$

$$+ \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\mu(I_{nj}) - \hat{\mu}_n(I_{nj}))^- \left\{ \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})} \right\}^-$$

$$+ \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\hat{\mu}_n(I_{nj}) - \mu(I_{nj}))^+ \left\{ \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})} \right\}^+$$

$$+ \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\hat{\mu}_n(I_{nj}) - \mu(I_{nj}))^- \left\{ \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})} \right\}^-.$$

For $j \in \mathcal{F}_n$:

$$\left\{ \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})} \right\}^+ \leq \log \frac{1}{\lambda(I_{nj})} = \log \frac{1}{h_n}$$

and

$$\left\{ \log \frac{\hat{\mu}_n(I_{nj})}{\lambda(I_{nj})} \right\}^- = \left\{ \log \frac{\lambda(I_{nj})}{\hat{\mu}_n(I_{nj})} \right\}^+$$

$$\leq \log \frac{\lambda(I_{nj})}{a_n h_n} = \log \frac{1}{a_n}.$$

Therefore, with the notations

$$A_n = \bigcup \{I_{nj} : j \in \mathcal{F}_n \cap \mathcal{G}_n, \mu(I_{nj}) \geq \hat{\mu}_n(I_{nj})\}$$

$$B_n = \bigcup \{I_{nj} : j \in \mathcal{F}_n \cap \mathcal{G}_n, \mu(I_{nj}) < \hat{\mu}_n(I_{nj})\}$$

$$c_n = \log \frac{1}{h_n} + \log \frac{1}{a_n}$$

we get that for $\hat{T}_n < T_n$

$$\begin{aligned}
 |U_n| &\leq \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\mu(I_{nj}) - \hat{\mu}_n(I_{nj}))^+ \log \frac{1}{h_n} \\
 &\quad + \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\mu(I_{nj}) - \hat{\mu}_n(I_{nj}))^- \log \frac{1}{a_n} \\
 &\quad + \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\hat{\mu}_n(I_{nj}) - \mu(I_{nj}))^+ \log \frac{1}{h_n} \\
 &\quad + \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\hat{\mu}_n(I_{nj}) - \mu(I_{nj}))^- \log \frac{1}{a_n} \\
 &= c_n \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\mu(I_{nj}) - \hat{\mu}_n(I_{nj}))^+ \\
 &\quad + c_n \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n} (\hat{\mu}_n(I_{nj}) - \mu(I_{nj}))^+ \\
 &= c_n \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n, \mu(I_{nj}) \geq \hat{\mu}_n(I_{nj})} (\mu(I_{nj}) - \hat{\mu}_n(I_{nj})) \\
 &\quad + c_n \sum_{j \in \mathcal{F}_n \cap \mathcal{G}_n, \hat{\mu}_n(I_{nj}) > \mu(I_{nj})} (\hat{\mu}_n(I_{nj}) - \mu(I_{nj})) \\
 &= c_n(\mu(A_n) - \hat{\mu}_n(A_n)) + c_n(\hat{\mu}_n(B_n) - \mu(B_n)).
 \end{aligned}$$

Thus

$$\begin{aligned}
 P(|U_n| > \epsilon, \hat{T}_n < T_n) &\leq P\left(|\mu(A_n) - \hat{\mu}_n(A_n)| > \frac{\epsilon}{2c_n}\right) + P\left(|\mu(B_n) - \hat{\mu}_n(B_n)| > \frac{\epsilon}{2c_n}\right) \\
 &\leq 2 \sup_A P\left(|\mu(A) - \hat{\mu}_n(A)| > \frac{\epsilon}{2c_n}\right) \tag{24} \\
 &\leq 4 \exp\left\{-\frac{nK(T_n)}{16} \left(\frac{\epsilon}{2c_n}\right)^2\right\} + 2 \frac{11520}{K(T_n)} \frac{\epsilon}{2c_n} \exp\left\{-n \frac{K(T_n)^8}{288} \left(\frac{\epsilon}{2c_n}\right)^2\right\} \\
 &= 4 \exp\left\{-\frac{nb_n \epsilon^2}{128c_n^2}\right\} + \frac{92160c_n}{b_n \epsilon} \exp\left\{-n \frac{b_n^8 \epsilon^2}{294912c_n^2}\right\}.
 \end{aligned}$$

On the other hand

$$\begin{aligned}
 P(\hat{T}_n \geq T_n) &= P(K_n^{-1}(b_n) \geq K^{-1}(b_n/2)) \\
 &= P(b_n \leq K_n(K^{-1}(b_n/2)))
 \end{aligned}$$

$$\begin{aligned}
 &= P(b_n - b_n/2 \leq K_n(K^{-1}(b_n/2)) - K(K^{-1}(b_n/2))) \tag{25} \\
 &\leq \exp\{-2n(b_n - b_n/2)^2\} \\
 &= \exp(-nb_n^2/2)
 \end{aligned}$$

where the first inequality is Hoeffding’s inequality. (7), (24), (25) and Lemma 3 imply that

$$\sum_n P(|U_n| > \epsilon) < \infty. \tag{26}$$

(II) Let $\mathcal{L}_n = \{j : \mu(I_{nj}) \geq a_n h_n\}$. Then

$$\begin{aligned}
 V_n &= - \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{\hat{\mu}_n(I_{nj})}{\mu(I_{nj})} \\
 &= \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{\mu(I_{nj})}{\hat{\mu}_n(I_{nj})} \tag{27}
 \end{aligned}$$

$$\begin{aligned}
 &\leq \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n \cap \mathcal{L}_n} \mu(I_{nj}) \log \frac{\mu(I_{nj})}{\hat{\mu}_n(I_{nj})} \\
 &\leq \sum_{j \in \mathcal{L}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \left| \log \frac{\mu(I_{nj})}{\hat{\mu}_n(I_{nj})} \right|. \tag{28}
 \end{aligned}$$

Therefore, proceeding in the same way as in (I), applying Lemma 2 to (28) yields

$$\begin{aligned}
 P(V_n > \epsilon) &\leq P(V_n > \epsilon, \hat{T}_n < T_n) + P(\hat{T}_n \geq T_n) \\
 &\leq \sum_{j \in \mathcal{L}_n \cap \mathcal{G}_n} P\left(\left| \log \frac{\mu(I_{nj})}{\hat{\mu}_n(I_{nj})} \right| > \epsilon\right) + P(\hat{T}_n \geq T_n) \\
 &\leq \sum_{j \in \mathcal{L}_n \cap \mathcal{G}_n} \left\{ 2 \exp\left[-n \frac{K(T_n)}{16} (\mu(I_{nj})(1 - 2^{-\epsilon}))^2\right] \right. \\
 &\quad \left. + \frac{11520}{K(T_n)\mu(I_{nj})(1 - 2^{-\epsilon})} \exp\left[-n \frac{K(T_n)^8}{288} (\mu(I_{nj})(1 - 2^{-\epsilon}))^2\right] \right\} \\
 &\quad + \exp[-nb_n^2/2] \\
 &\leq \frac{1}{a_n h_n} \left[2 \exp\left\{-n \frac{b_n}{32} (a_n h_n (1 - 2^{-\epsilon}))^2\right\} \right. \\
 &\quad \left. + \frac{23040}{b_n a_n h_n (1 - 2^{-\epsilon})} \exp\left\{-n \frac{b_n^8}{73728} (a_n h_n (1 - 2^{-\epsilon}))^2\right\} \right] \\
 &\quad + \exp(-nb_n^2/2)
 \end{aligned}$$

where in the last inequality we used that for $j \in \mathcal{L}_n \cap \mathcal{G}_n, \mu(I_{nj}) \geq a_n h_n$ and $K(T_n) = b_n/2$.

It follows from (7) and the Borel–Cantelli lemma that

$$\lim_{n \rightarrow \infty} V_n^+ = 0 \quad \text{a. s.} \tag{29}$$

On the other hand, by (27),

$$\begin{aligned} -V_n &= - \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{\mu(I_{nj})}{\hat{\mu}_n(I_{nj})} \\ &\leq \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{\sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \hat{\mu}_n(I_{nj})}{\sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj})}, \end{aligned} \tag{30}$$

where the latter inequality follows from the so-called Log-Sum Inequality (Csiszár and Körner [4], p. 48). Continuing (30) we have that

$$\begin{aligned} \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{\sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \hat{\mu}_n(I_{nj})}{\sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj})} &\leq \sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \log \frac{1}{\sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj})} \\ &\leq \log \frac{1}{\sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj})}. \end{aligned} \tag{31}$$

Now, we can show that (cf. (39) below)

$$\sum_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} \mu(I_{nj}) \rightarrow 1 \quad \text{a. s., as } n \rightarrow \infty,$$

hence (30) and (31) imply that

$$I_{\{-V_n > \epsilon\}} \leq I_{\{1 \geq 2^\epsilon \mu(\cup_{j \in \mathcal{F}_n \cap \hat{\mathcal{G}}_n} I_{nj})\}} \rightarrow 0 \quad \text{a. s.}$$

Combining this last result with (29) yields the proof of the fact that

$$\lim_{n \rightarrow \infty} V_n = 0 \quad \text{a. s.} \tag{32}$$

(III) We also have

$$\lim_{n \rightarrow \infty} Z_n = 0 \tag{33}$$

by Lemma 1 and (8).

(IV) Finally introduce the notations

$$A_n = \bigcup_{j \notin \mathcal{F}_n \cap \hat{G}_n} I_{nj}.$$

and

$$g_n(x) = \frac{\mu(I_{nj})}{\lambda(I_{nj})} \quad \text{if } x \in I_{nj}$$

then

$$\begin{aligned} W_n &= \sum_{j \notin \mathcal{F}_n \cap \hat{G}_n} \mu(I_{nj}) \log \frac{\mu(I_{nj})}{\lambda(I_{nj})} \\ &= \int_{A_n} f(x) \log g_n(x) \, dx \\ &= \int_{A_n} f(x) \log f(x) \, dx - \int_{A_n} f(x) \log \frac{f(x)}{g_n(x)} \, dx. \end{aligned} \tag{34}$$

$\int_{A_n} f = \int_{A_n} g_n$ implies that

$$\int_{A_n} f(x) \log \frac{f(x)}{g_n(x)} \, dx \geq 0 \tag{35}$$

and in the same way

$$\int_{A_n^c} f(x) \log \frac{f(x)}{g_n(x)} \, dx \geq 0. \tag{36}$$

Thus (34), (35) and (36) imply that

$$\begin{aligned} |W_n| &\leq \int_{A_n} f(x) |\log f(x)| \, dx + \int_{A_n} f(x) \log \frac{f(x)}{g_n(x)} \, dx \\ &= \int_{A_n} |\log f(x)| \mu(dx) + Z_n. \end{aligned} \tag{37}$$

Since $H(f)$ is finite,

$$\nu(A) = \int_A |\log f(x)| \mu(dx)$$

is absolutely continuous with respect to μ .

Moreover, we observe that

$$\begin{aligned} \sum_{j \notin \mathcal{F}_n} \mu(I_{nj}) &= \mu\left(\bigcup_{j \notin \mathcal{F}_n} I_{nj}\right) = \mu(\{x : \hat{f}_n(x) \leq a_n\}) \\ &= \int_{\hat{f}_n(x) \leq a_n} f(x) dx \\ &\leq \int_{\frac{1}{2}f(x) \leq a_n} f(x) dx + \int_{\frac{1}{2}f(x) \geq a_n \geq \hat{f}_n(x)} f(x) dx \\ &\leq \int_{\frac{1}{2}f(x) \leq a_n} f(x) dx + 2 \int |f(x) - \hat{f}_n(x)| dx. \end{aligned}$$

Therefore, by Lemma 4 and (5) we have that

$$\sum_{j \notin \mathcal{F}_n} \mu(I_{nj}) \rightarrow 0 \quad \text{a. s., } n \rightarrow \infty. \tag{38}$$

On the other hand,

$$\sum_{j \notin \hat{\mathcal{G}}_n} \mu(I_{nj}) \leq \mu([\hat{T}_n, T^*]) \rightarrow 0$$

a. s. since $b_n \rightarrow 0$, therefore by (38)

$$\mu(A_n) \rightarrow 0 \quad \text{a. s. } n \rightarrow \infty. \tag{39}$$

Hence, (33), (37) and (39) imply that

$$\lim_{n \rightarrow \infty} |W_n| = 0 \quad \text{a. s.} \tag{40}$$

Now, from (26), (32), (33) and (40) it follows that

$$\lim_{n \rightarrow \infty} H_n = H(f) \quad \text{a. s.} \tag{41}$$

and thus the proof of the Theorem is complete.

References

1. Carbonez, A., Györfi, L. and van der Meulen, E. C., L_1 -consistency of randomly censored version of histogram estimate (1991). Submitted for publication.

2. Csiszár, I., On generalized entropy. *Studia Sci. Math. Hung.*, **4** (1969), pp. 401–419.
3. Csiszár, I. Generalized entropy and quantization problems. *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, Academia, Prague, 1973.*
4. Csiszár, I. and Körner, J., *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Akadémiai Kiadó, Budapest, and Academic Press, New York, 1981.
5. Györfi, L. and van der Meulen, E. C., Density-free convergence properties of various estimators of entropy. *Comput. Statist. Data Anal.*, **5** (1987), pp. 425–436.
6. Györfi, L. and van der Meulen, E. C., Entropy estimation based on L_1 -consistent density estimates. Preprint series Dept. Mathematics K. U. Leuven, **2** (1990), pp. 75–105.
7. Kaplan, E. L. and Meier, P., Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53** (1958), pp. 457–481.
8. Rényi, A. On the dimension and entropy of probability distributions. *Acta Math. Acad. Sci. Hung.*, **10** (1959), pp. 193–215.

Непараметрическая оценка энтропии на основе случайно цензурированных данных

А. КАРБОНЕЗ, Л. ДЬЁРФИ и Э. К. ван дер МАЙЛЕН

(Лувен, Будапешт, Лувен)

Шэннонская энтропия случайной величины X с функцией плотности $f(x)$ определяется как

$$H(f) = - \int f(x) \log f(x) dx.$$

На основе цензурированных наблюдений предложен непараметрический оценитель для $H(f)$, если $H(f)$ является конечной.

A. Carbonez
E. C. van der Meulen
Department of Mathematics
Kath. Univ. Leuven
Celestijnenlaan 200B
B-3001 Heverlee Belgium

L. Györfi
Technical University of Budapest
H-1521 Budapest
Stoczek u. 2.
Hungary

EMPIRICAL LOG-OPTIMAL PORTFOLIO SELECTION

GUSZTÁV MORVAI¹

(Budapest)

(Received October 1, 1991)

We show that the empirical log-optimal portfolio performs asymptotically under certain conditions as well as the optimal one.

1. Introduction

Let $\mathbf{X} \in \mathbb{R}^m$ denote a random stock market return vector, where X_i is the value of one unit investment in stock i at the end of the trading day. We require that $X_i \geq 0$ for $i = 1, 2, \dots, m$, that is, an investor can not loose more than the invested capital. Let \mathbf{b} , $b_i \geq 0$, $\sum_{i=1}^m b_i = 1$, denote a portfolio, that is, an allocation of investor's capital across the investment alternatives. Let B denote the set of such portfolios. Thus b_i is the proportion of current capital invested in stock i . The resulting wealth is $S = \sum_{i=1}^m b_i X_i = \mathbf{bX}$. This is the wealth resulting from a unit investment allocated to the m stocks according to portfolio \mathbf{b} . If the current capital is reallocated according to portfolio \mathbf{b}_i at time i in repeated investments against stock vectors $\mathbf{X}_1, \mathbf{X}_2, \dots$ then the wealth S_n at time n is given by

$$S_n = \prod_{i=1}^n \mathbf{b}_i \mathbf{X}_i.$$

Suppose the stock market process $\mathbf{X}_1, \mathbf{X}_2, \dots$ is independent and identically distributed. A portfolio \mathbf{b}^* is called log-optimal if $E \ln \mathbf{b}^* \mathbf{X} = \sup_{\mathbf{b} \in B} E \ln \mathbf{bX}$. Let B^* denote the set of log-optimal portfolios. It can be shown that $\limsup_{n \rightarrow \infty} \frac{1}{n} \ln S_n \leq$

¹ This paper was prepared under the auspices of E. C. TEMPUS Office grant IMG-HVS-0062-90.

$\lim_{n \rightarrow \infty} \frac{1}{n} \ln S_n^* = E \ln \mathbf{b}^* \mathbf{X}$ a. s., where S_n, S_n^* denote capitals achieved by an arbitrary and the log-optimal portfolio in n repeated games, respectively. For more about the log-optimal portfolio see [1]–[8].

If the probability distribution of the stocks is not known in advance, consider as a goal to find a portfolio selector $\hat{\mathbf{b}}(\cdot)$ which achieves the same asymptotic capital growth rate as the log-optimal portfolio does, that is,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \hat{S}_n = E \ln \mathbf{b}^* \mathbf{X} \quad \text{a. s.},$$

where $\hat{S}_n = \prod_{i=1}^n \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i$.

2. The empirical log-optimal portfolio

We suppose that the sequence of random stock market variables $\mathbf{X}_1, \mathbf{X}_2, \dots$ is stationary and ergodic. We examine the performance of the following portfolio selector:

$$\begin{aligned} \hat{\mathbf{b}}(\cdot) &= (1/m, 1/m, \dots, 1/m) && \text{for } n = 0 \\ \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) &= \arg \max_{\mathbf{b} \in B} \frac{1}{n} \sum_{i=1}^n \ln \mathbf{b} \mathbf{X}_i = \arg \max_{\mathbf{b} \in B} \int \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) && \text{for } n \geq 1 \end{aligned}$$

where

$$\hat{\mu}_n(A) = \frac{1}{n} \sum_{i=1}^n I_{\{\mathbf{X}_i \in A\}}$$

and

$$I_{\{\mathbf{X}_i \in A\}} = \begin{cases} 1 & \text{if } \mathbf{X}_i \in A \\ 0 & \text{if } \mathbf{X}_i \notin A \end{cases}.$$

In other words, we choose the log-optimal portfolio according to the empirical distribution of the past.

The following theorem implies that the asymptotically optimal growth rate is achieved by the proposed portfolio selector if the sequence of random stock vectors is independent and identically distributed rather than merely ergodic. The portfolio selector proposed in Cover [9] achieves this goal but our selector is much simpler.

THEOREM 1. Suppose the sequence of random stock market variables $\mathbf{X}_1, \mathbf{X}_2, \dots$, is stationary, ergodic, and $E|\ln X_j| < \infty$ for $j = 1, 2, \dots, m$. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \ln \hat{S}_n = E \ln \mathbf{b}^* \mathbf{X} \quad \text{a. s.},$$

where $\hat{S}_n = \prod_{i=1}^n \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})\mathbf{X}_i$ and $E \ln \mathbf{b}^*\mathbf{X} = \sup_{\mathbf{b} \in B} E \ln \mathbf{b}\mathbf{X}$.

Let μ denote the distribution of the random stock vector \mathbf{X} .

Lemma 1. Suppose $-\infty < \sup_{\mathbf{b} \in B} E \ln \mathbf{b}\mathbf{X} < \infty$. Let $\{\mathbf{b}_n\}$ be a fixed sequence of portfolios. If $\lim_{n \rightarrow \infty} \int \ln \mathbf{b}_n \mathbf{x} \mu(dx) = E \ln \mathbf{b}^*\mathbf{X}$ then the accumulation points of $\{\mathbf{b}_n\}$ are log-optimal according to the true distribution μ .

Proof. Suppose \mathbf{b}' is an accumulation point which is not log-optimal.

Let $\{\mathbf{b}_{n_i}\}$ be a subsequence of $\{\mathbf{b}_n\}$ converging to \mathbf{b}' . Since the function $E \ln \mathbf{b}\mathbf{X}$ is continuous in \mathbf{b} ,

$$E \ln \mathbf{b}'\mathbf{X} = \int \ln \lim_{i \rightarrow \infty} \mathbf{b}_{n_i} \mathbf{x} \mu(dx) = \lim_{i \rightarrow \infty} \int \ln \mathbf{b}_{n_i} \mathbf{x} \mu(dx).$$

Since \mathbf{b}' is not log-optimal, $E \ln \mathbf{b}'\mathbf{X} < E \ln \mathbf{b}^*\mathbf{X}$. Thus

$$\lim_{i \rightarrow \infty} \int \ln \mathbf{b}_{n_i} \mathbf{x} \mu(dx) < E \ln \mathbf{b}^*\mathbf{X}.$$

But this contradicts the assumption $\lim_{n \rightarrow \infty} \int \ln \mathbf{b}_n \mathbf{x} \mu(dx) = E \ln \mathbf{b}^*\mathbf{X}$.

Lemma 2 (Cover [8]). Suppose $-\infty < \sup_{\mathbf{b} \in B} E \ln \mathbf{b}\mathbf{X} < \infty$. Let L be the subspace of \mathbb{R}^m of least dimension satisfying $P(\mathbf{X} \in L) = 1$. Each log-optimal portfolio $\mathbf{b}^* \in B^*$ has the same orthogonal projection \mathbf{b}_L onto L .

Lemma 3. Suppose $-\infty < \sup_{\mathbf{b} \in B} E \ln \mathbf{b}\mathbf{X} < \infty$. Let the process $\mathbf{X}_1, \mathbf{X}_2, \dots$, be stationary and ergodic. Consider any function $\mathbf{b}^*(\cdot)$ such that $\mathbf{b}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_i) \in B^*$ for all i . Let \mathbf{b}^* be a log-optimal portfolio. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \mathbf{b}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})\mathbf{X}_i = E \ln \mathbf{b}^*\mathbf{X} \quad \text{a. s.}$$

Proof.

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \mathbf{b}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})\mathbf{X}_i &= \\ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln(\mathbf{b}_L \mathbf{X}_i + (\mathbf{b}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) - \mathbf{b}_L)\mathbf{X}_i) &= \\ \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \mathbf{b}_L \mathbf{X}_i &= \\ E \ln \mathbf{b}_L \mathbf{X} &= \\ E \ln \mathbf{b}^*\mathbf{X} \quad \text{a. s.,} \end{aligned}$$

since $(\mathbf{b}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) - \mathbf{b}_L)\mathbf{x} = 0$ for $\mathbf{x} \in L$ and $P(\mathbf{X} \in L) = 1$ by Lemma 2, where \mathbf{b}_L is the unique projection of the log-optimal portfolios.

Lemma 4. Suppose $-\infty < \sup_{\mathbf{b} \in B} E \ln \mathbf{b}\mathbf{X} < \infty$. Then the set of log-optimal portfolios B^* is closed.

Proof. Suppose \mathbf{b}' is from the boundary of B^* but $\mathbf{b}' \notin B^*$. Let \mathbf{b}_n^* be a sequence converging to \mathbf{b}' . By the continuity of the function $E \ln \mathbf{b}\mathbf{X}$, $E \ln \mathbf{b}'\mathbf{X} = E \ln \lim_{n \rightarrow \infty} \mathbf{b}_n^*\mathbf{X} = \lim_{n \rightarrow \infty} E \ln \mathbf{b}_n^*\mathbf{X} = E \ln \mathbf{b}^*\mathbf{X}$. Thus \mathbf{b}' is log-optimal. But this contradicts the assumption $\mathbf{b}' \notin B^*$.

Lemma 5.

$$I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{X} - I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}'\mathbf{X} \leq \frac{m}{\epsilon^2} \|\mathbf{b} - \mathbf{b}'\|,$$

where $0 < \epsilon < 1$ and

$$I_{\{\mathbf{x} \in A_\epsilon\}} = \begin{cases} 1 & \text{if } \mathbf{X} \in [\epsilon, 1/\epsilon]^m \\ 0 & \text{if } \mathbf{X} \notin [\epsilon, 1/\epsilon]^m \end{cases}.$$

Proof.

$$\begin{aligned} I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{X} - I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}'\mathbf{X} &= \\ I_{\{\mathbf{x} \in A_\epsilon\}} \ln \frac{\mathbf{b}\mathbf{X}}{\mathbf{b}'\mathbf{X}} &= I_{\{\mathbf{x} \in A_\epsilon\}} \ln \left(1 + \frac{(\mathbf{b} - \mathbf{b}')\mathbf{X}}{\mathbf{b}'\mathbf{X}} \right) \leq \\ I_{\{\mathbf{x} \in A_\epsilon\}} \ln \left(1 + \frac{\sum_{i=1}^m |b_i - b'_i| \frac{1}{\epsilon}}{\sum_{i=1}^m b'_i \epsilon} \right) &\leq I_{\{\mathbf{x} \in A_\epsilon\}} \ln \left(1 + \frac{\sum_{i=1}^m |b_i - b'_i|}{\epsilon^2} \right) \leq \\ I_{\{\mathbf{x} \in A_\epsilon\}} \frac{\sum_{i=1}^m |b_i - b'_i|}{\epsilon^2} &\leq \frac{m}{\epsilon^2} \|\mathbf{b} - \mathbf{b}'\|. \end{aligned}$$

Lemma 6. For $0 < \epsilon < 1$

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) \leq 0 \quad \text{a. s.}$$

Proof. We cover the simplex $B = \left\{ \mathbf{b} : \sum_{i=1}^m b_i = 1, b_i \geq 0 \text{ for } i = 1, 2, \dots, m \right\}$ by regions D_j with diameter Δ , where $j = 1, 2, \dots, r(\Delta)$. Let \mathbf{b}_j denote a portfolio from the region D_j .

$$\sup_{\mathbf{b} \in B} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) =$$

$$\begin{aligned} & \max_j \sup_{\mathbf{b} \in D_j} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) \leq \\ & \max_j \sup_{\mathbf{b} \in D_j} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \hat{\mu}_n(d\mathbf{x}) + \\ & \max_j \sup_{\mathbf{b} \in D_j} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \mu(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) + \\ & \max_j \sup_{\mathbf{b} \in D_j} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \mu(d\mathbf{x}). \end{aligned}$$

From Lemma 5,

$$\sup_{\mathbf{b} \in D_j} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \hat{\mu}_n(d\mathbf{x}) \leq \frac{m}{\epsilon^2} \Delta.$$

Similarly,

$$\sup_{\mathbf{b} \in D_j} \int \ln \mathbf{b}_j\mathbf{x} I_{\{\mathbf{x} \in A_\epsilon\}} \mu(d\mathbf{x}) - \int \ln \mathbf{b}\mathbf{x} I_{\{\mathbf{x} \in A_\epsilon\}} \mu(d\mathbf{x}) \leq \frac{m}{\epsilon^2} \Delta.$$

Thus

$$\begin{aligned} & \sup_{\mathbf{b} \in B} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) \leq \\ & \frac{2m\Delta}{\epsilon^2} + \max_j \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \mu(d\mathbf{x}). \end{aligned}$$

By the strong law of large numbers for ergodic sequence,

$$\lim_{n \rightarrow \infty} \max_j \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}_j\mathbf{x} \mu(d\mathbf{x}) = 0 \quad \text{a. s.},$$

hence

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) \leq \frac{2m\Delta}{\epsilon^2} \quad \text{a. s.}$$

Since Δ was arbitrary,

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int I_{\{\mathbf{x} \in A_\epsilon\}} \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) \leq 0 \quad \text{a. s.}$$

Lemma 7. Under the conditions of Theorem 1,

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int \ln \mathbf{b}\mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int \ln \mathbf{b}\mathbf{x} \mu(d\mathbf{x}) \leq 0 \quad \text{a. s.}$$

Proof.

$$\begin{aligned} & \sup_{\mathbf{b} \in B} \int \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) \leq \\ & \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \in A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int_{I_{\{\mathbf{x} \in A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) + \\ & \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}). \end{aligned}$$

From Lemma 6, for arbitrary $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \in A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int_{I_{\{\mathbf{x} \in A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) \leq 0 \quad \text{a. s.}$$

Furthermore,

$$\begin{aligned} & \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) \leq \\ & \sup_{\mathbf{b} \in B} \left| \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) \right| + \sup_{\mathbf{b} \in B} \left| \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) \right| \leq \\ & \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \left| \ln \sum_{i=1}^m b_i x_i \right| \hat{\mu}_n(d\mathbf{x}) + \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \left| \ln \sum_{i=1}^m b_i x_i \right| \mu(d\mathbf{x}) \leq \\ & \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \max \left\{ \max_{i=1,2,\dots,m} \ln x_i, - \min_{i=1,2,\dots,m} \ln x_i \right\} \hat{\mu}_n(d\mathbf{x}) + \\ & \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \max \left\{ \max_{i=1,2,\dots,m} \ln x_i, - \min_{i=1,2,\dots,m} \ln x_i \right\} \mu(d\mathbf{x}) \leq \\ & \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \sum_{i=1}^m |\ln x_i| \hat{\mu}_n(d\mathbf{x}) + \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \sum_{i=1}^m |\ln x_i| \mu(d\mathbf{x}). \end{aligned}$$

It follows that

$$\begin{aligned} \limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) & \leq \\ & 2 \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \sum_{i=1}^m |\ln x_i| \mu(d\mathbf{x}) \quad \text{a. s.,} \end{aligned}$$

since

$$\lim_{n \rightarrow \infty} \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \sum_{i=1}^m |\ln x_i| \hat{\mu}_n(d\mathbf{x}) = \int_{I_{\{\mathbf{x} \notin A_\epsilon\}}} \sum_{i=1}^m |\ln x_i| \mu(d\mathbf{x}) \quad \text{a. s.}$$

Thus for arbitrary $\epsilon > 0$,

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) \leq 2 \int I_{\{\mathbf{x} \notin A_\epsilon\}} \sum_{i=1}^m |\ln x_i| \mu(d\mathbf{x}) \quad \text{a. s.}$$

Since $\epsilon > 0$ was arbitrary and by assumption $E|\ln X_i| < \infty$ for $i = 1, 2, \dots, m$,

$\lim_{\epsilon \rightarrow 0} 2 \int I_{\{\mathbf{x} \notin A_\epsilon\}} \sum_{i=1}^m |\ln x_i| \mu(d\mathbf{x}) = 0$ by the Lebesgue dominated convergence theorem. Thus

$$\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) \leq 0 \quad \text{a. s.}$$

Lemma 8. Under the conditions of Theorem 1, the accumulation points of $\hat{\mathbf{b}}(\cdot)$ are log-optimal with probability one.

Proof. By the definition of log-optimality,

$$\begin{aligned} 0 &\leq \int \ln \mathbf{b}^* \mathbf{x} \mu(d\mathbf{x}) - \int \ln \hat{\mathbf{b}}(\mathbf{X}_1 \mathbf{X}_2 \dots, \mathbf{X}_n) \mathbf{x} \mu(d\mathbf{x}) = \\ &\int \ln \mathbf{b}^* \mathbf{x} \mu(d\mathbf{x}) - \int \ln \mathbf{b}^* \mathbf{x} \hat{\mu}_n(d\mathbf{x}) + \int \ln \mathbf{b}^* \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \\ &\int \ln \hat{\mathbf{b}}(\mathbf{X}_1 \mathbf{X}_2 \dots, \mathbf{X}_n) \mathbf{x} \hat{\mu}_n(d\mathbf{x}) + \int \ln \hat{\mathbf{b}}(\mathbf{X}_1 \mathbf{X}_2 \dots, \mathbf{X}_n) \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \\ &\int \ln \hat{\mathbf{b}}(\mathbf{X}_1 \mathbf{X}_2 \dots, \mathbf{X}_n) \mathbf{x} \mu(d\mathbf{x}). \end{aligned}$$

Since

$$\int \ln \mathbf{b}^* \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int \ln \hat{\mathbf{b}}(\mathbf{X}_1 \mathbf{X}_2 \dots, \mathbf{X}_n) \mathbf{x} \hat{\mu}_n(d\mathbf{x}) \leq 0$$

by the definition of $\hat{\mathbf{b}}(\cdot)$, and

$$\lim_{n \rightarrow \infty} \int \ln \mathbf{b}^* \mathbf{x} \mu(d\mathbf{x}) - \int \ln \mathbf{b}^* \mathbf{x} \hat{\mu}_n(d\mathbf{x}) = 0 \quad \text{a. s.}$$

by ergodicity, we have,

$$\begin{aligned} 0 &\leq \limsup_{n \rightarrow \infty} \int \ln \mathbf{b}^* \mathbf{x} \mu(d\mathbf{x}) - \int \ln \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) \mathbf{x} \mu(d\mathbf{x}) \leq \\ &\limsup_{n \rightarrow \infty} \int \ln \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int \ln \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) \mathbf{x} \mu(d\mathbf{x}) \leq \\ &\limsup_{n \rightarrow \infty} \sup_{\mathbf{b} \in B} \int \ln \mathbf{b} \mathbf{x} \hat{\mu}_n(d\mathbf{x}) - \int \ln \mathbf{b} \mathbf{x} \mu(d\mathbf{x}) \leq 0 \quad \text{a. s.} \end{aligned}$$

where the last step follows from Lemma 7.

Thus we have,

$$\liminf_{n \rightarrow \infty} \int \ln \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) \mathbf{x} \mu(d\mathbf{x}) \geq \int \ln \mathbf{b}^* \mathbf{x} \mu(d\mathbf{x}) \quad \text{a. s.},$$

and

$$\int \ln \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) \mathbf{x} \mu(d\mathbf{x}) \leq \int \ln \mathbf{b}^* \mathbf{x} \mu(d\mathbf{x})$$

hence

$$\lim_{n \rightarrow \infty} \int \ln \hat{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) \mathbf{x} \mu(d\mathbf{x}) = \int \ln \mathbf{b}^* \mathbf{x} \mu(d\mathbf{x}) \quad \text{a. s.}$$

Now the statement follows from Lemma 1.

Lemma 9. Let the process $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be stationary and ergodic. Suppose that

$$-\infty < \sup_{\mathbf{b}} E \ln \mathbf{b} \mathbf{x} < \infty.$$

Consider a portfolio selector $\tilde{\mathbf{b}}(\cdot)$ such that $P(\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i = 0) = 0$ for all i and the accumulation points of $\tilde{\mathbf{b}}(\cdot)$ are log-optimum with probability one. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i = E \ln \mathbf{b}^* \mathbf{X} \quad \text{a. s.}$$

Proof. Let $\mathbf{b}^{*'}$ be a log-optimum portfolio such that $b_j^{*'} = 0 \Rightarrow b_j^* = 0$ for $j = 1, 2, \dots, m$ and for all $\mathbf{b}^* \in B^*$, where B^* denotes the set of log-optimal portfolios. Such a portfolio exists, since suppose $b_{1,j}^* = 0$ and $b_{2,j}^* \neq 0$ for some j . Then for any $\lambda \in (0, 1)$, $\lambda \mathbf{b}_1^* + (1 - \lambda) \mathbf{b}_2^* \in B^*$ and contains less number of zeros than \mathbf{b}_1^* does. (Note $E(\lambda \mathbf{b}_1^* \mathbf{X} + (1 - \lambda) \mathbf{b}_2^* \mathbf{X}) = E \ln \mathbf{b}_1^* \mathbf{X} = E \ln \mathbf{b}_2^* \mathbf{X}$ by Lemma 2.) If this new portfolio does not satisfy the condition we can repeat this procedure. After at most m steps we get a proper portfolio.

Since $\mathbf{b}^* \mathbf{X} = \mathbf{b}^{*'} \mathbf{X}$ a. s. (see Lemma 2),

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i}{\mathbf{b}^* \mathbf{X}_i} &= \frac{1}{n} \sum_{i=1}^n \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i}{\mathbf{b}^{*'} \mathbf{X}_i} = \\ &= \frac{1}{n} \sum_{i=1}^n \ln \left(\frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i}{\mathbf{b}^{*'} \mathbf{X}_i} + \right. \\ &\quad \left. \frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i - \tilde{\mathbf{b}}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i}{\mathbf{b}^{*'} \mathbf{X}_i} \right) \end{aligned}$$

where $\tilde{\mathbf{b}}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})$ denotes the closest log-optimal portfolio to $\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})$ in Euclidean distance. (Such a portfolio exists since the set of log-optimal portfolios is closed by Lemma 4.) Thus

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})\mathbf{X}_i}{\mathbf{b}^*\mathbf{X}_i} = \\ & \frac{1}{n} \sum_{i=1}^n \ln \left(1 + \frac{(\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) - \tilde{\mathbf{b}}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}))\mathbf{X}_i}{\mathbf{b}^*\mathbf{X}_i} \right) \geq \\ & \frac{1}{n} \sum_{i=1}^{k(\omega)} \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1(\omega), \mathbf{X}_2(\omega), \dots, \mathbf{X}_{i-1}(\omega))\mathbf{X}_i(\omega)}{\mathbf{b}^*\mathbf{X}_i(\omega)} + \frac{1}{n} \sum_{i=k(\omega)+1}^n \ln \left(1 - \frac{\epsilon\mathbf{a}\mathbf{X}_i(\omega)}{\mathbf{b}^*\mathbf{X}_i(\omega)} \right), \end{aligned}$$

where $a_j = 1$ if $b_j^* \neq 0$, $a_j = 0$ if $b_j^* = 0$ and $k(\omega)$ is an integer such that $\|\tilde{\mathbf{b}}(\mathbf{X}_1(\omega), \mathbf{X}_2(\omega), \dots, \mathbf{X}_i(\omega)) - \tilde{\mathbf{b}}^*(\mathbf{X}_1(\omega), \mathbf{X}_2(\omega), \dots, \mathbf{X}_i(\omega))\| < \epsilon$ for $i > k(\omega)$, where $0 < \epsilon < 0.5 \min_{j \in I} b_j^*$, $I = \{j : b_j^* \neq 0\}$.

Thus

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})\mathbf{X}_i}{\mathbf{b}^*\mathbf{X}_i} \geq \\ & \frac{1}{n} \sum_{i=1}^{k(\omega)} \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1(\omega), \mathbf{X}_2(\omega), \dots, \mathbf{X}_{i-1}(\omega))\mathbf{X}_i(\omega)}{\mathbf{b}^*\mathbf{X}_i(\omega)} \\ & - \frac{1}{n} \sum_{i=1}^{k(\omega)} \ln \left(1 - \frac{\epsilon\mathbf{a}\mathbf{X}_i(\omega)}{\mathbf{b}^*\mathbf{X}_i(\omega)} \right) + \frac{1}{n} \sum_{i=1}^n \ln \left(1 - \frac{\epsilon\mathbf{a}\mathbf{X}_i(\omega)}{\mathbf{b}^*\mathbf{X}_i(\omega)} \right). \end{aligned}$$

Thus

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})\mathbf{X}_i}{\mathbf{b}^*\mathbf{X}_i} \geq E \ln \left(1 - \frac{\epsilon\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X}} \right) \quad \text{a. s.}$$

Expanding the function $\ln \left(1 - \frac{y\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X}} \right)$ into Taylor series around 0 in the interval $[0, \epsilon]$, we have, $\left| \ln \left(1 - \frac{y\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X}} \right) \right| = \left| \ln(1) + \frac{-y\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X} - t\mathbf{a}\mathbf{X}} \right|$ for some $t \in [0, \epsilon]$.

Thus

$$\left| \ln \left(1 - \frac{\epsilon\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X}} \right) \right| \leq \frac{\epsilon\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X} - t\mathbf{a}\mathbf{X}} \leq \frac{\epsilon\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X} - \epsilon\mathbf{a}\mathbf{X}} \leq \frac{\epsilon\mathbf{a}\mathbf{X}}{0.5\mathbf{b}^*\mathbf{X}} = \frac{2\epsilon\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X}}.$$

Since $E \frac{X_j}{\mathbf{b}^*\mathbf{X}} \leq 1$ for $j = 1, 2, \dots, m$ by log-optimality (see Bell and Cover [7]), hence

$$E \frac{2\epsilon\mathbf{a}\mathbf{X}}{\mathbf{b}^*\mathbf{X}} \leq 2\epsilon m < \infty.$$

Since ϵ was arbitrary,

$$\lim_{\epsilon \rightarrow 0} E \ln \left(1 - \frac{\epsilon \mathbf{aX}}{\mathbf{b}^* \mathbf{X}} \right) = E \ln 1 = 0$$

by the Lebesgue dominated convergence theorem. The upper bound follows similarly,

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \frac{\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i}{\mathbf{b}^* \mathbf{X}_i} = \\ & \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \left(1 + \frac{(\tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) - \tilde{\mathbf{b}}^*(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1})) \mathbf{X}_i}{\mathbf{b}^* \mathbf{X}_i} \right) \leq \\ & \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \left(1 + \frac{\epsilon \mathbf{eX}_i}{\mathbf{b}^* \mathbf{X}_i} \right) = E \ln \left(1 + \frac{\epsilon \mathbf{eX}}{\mathbf{b}^* \mathbf{X}} \right), \end{aligned}$$

where $\mathbf{e} = (1, 1, \dots, 1)$. Since ϵ was arbitrary,

$$\lim_{\epsilon \rightarrow 0} E \ln \left(1 + \frac{\epsilon \mathbf{eX}}{\mathbf{b}^* \mathbf{X}} \right) = 0 \quad \text{a. s.,}$$

by the Lebesgue dominated convergence theorem. Hence

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \ln \tilde{\mathbf{b}}(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{i-1}) \mathbf{X}_i = E \ln \mathbf{b}^* \mathbf{X} \quad \text{a. s.}$$

Proof of Theorem 1. The accumulation points of $\tilde{\mathbf{b}}(\cdot)$ are log-optimal by Lemma 8. Then the theorem follows from Lemma 9.

References

1. Kelly, J., A New Interpretation of Information Rate. Bell System Technical Journal, **35**, 1956.
2. Breiman, L., Investment Policies for Expanding Businesses Optimal in a Long-Run Sense. Naval Research Logistics Quarterly, Office of the Naval Research Navexos P-1278, Vol. 7, No. 4, December 1960.
3. Breiman, L., Optimal Gambling Systems for Favorable Games. Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1961.
4. Algoet, A. H., Cover, T. M., Asymptotic Optimality and Asymptotic Equipartition Properties of Log-Optimum Investment. Ann. Probab. No. 16, 1988.
5. Finkelstein, M., Whitley, R., Optimal Strategies for Repeated Games. Adv. Appl. Prob., **13**, 1981.

6. *Barron, A. R., Cover, T. M.*, A Bound on the Financial Value of Information. *IEEE Trans. Inform. Theory*, Vol. **34**, No. 5, 1988.
7. *Bell, R., Cover T. M.*, Game-Theoretic Optimal Portfolios. *Management Science*, Vol. **34**, No. 6, 1988.
8. *Cover, T. M.*, An Algorithm for Maximizing Expected Log Investment Return. *IEEE Trans. Inform. Theory*, Vol. **IT-30**, No. 2, 1984.
9. *Cover, T. M.*, Universal Portfolios. *Mathematical Finance*, **16**, January 1991.

Эмпирическая лог-оптимальная селекция портфеля

Г. МОРВАИ

(Будапешт)

Показано, что эмпирический лог-оптимальный портфель ведет себя асимптотически как оптимально при некоторых условиях.

G. Morvai
Technical University of Budapest
H-1521 Budapest
Stoczek u. 2.
Hungary

PATTERN CLASSIFICATION FROM DISTORTED SAMPLE

GÁBOR LUGOSI

(*Budapest*)

(Received October 1, 1991)

In nonparametric pattern classification the optimal (Bayesian) decision on the category of the observed vector is designed from a long training sequence, that is, independent pairs of observations and corresponding labels. In many practical situations, however, due to feature extraction, quantization, or noise, the observed vector and the training sequence may be distorted. In this paper we show how asymptotically optimal decisions can be derived from distorted training or made from slightly distorted observation.

1. Introduction

The usual pattern classification problem is the following: Let the random variable pair (X, Y) be such that the *observation* X takes its values from \mathbf{R}^d , the set of d -dimensional real vectors, while the value of the *label* Y is from the set $\{0, 1\}$. The task is to estimate the value of the label Y knowing only the observation X , that is, to find a measurable *decision function* $g : \mathbf{R}^d \rightarrow \{0, 1\}$ so as the *error probability* of the decision $P_g(X, Y) = \Pr\{g(X) \neq Y\}$ be minimal. It is well known that the optimal solution is given by the *Bayes-decision*:

$$g^*(x) = \arg \max_{i=0,1} p_i(x),$$

where $p_i(x) = \Pr\{Y = i \mid X = x\}$, $i = 0, 1$ are the *a posteriori* probabilities. The error probability of this decision is the *Bayes-risk*: $P^B(X, Y) = \Pr\{g^*(X) \neq Y\}$. It is well known that

$$P_g(X, Y) = 1 - E(p_g(X)),$$

and

$$P^B(X, Y) = 1 - E(\max_i p_i(X)).$$

If the *a posteriori* probabilities are not known, then we have to approximate the optimal decision. Assume that we are given a training sequence $\xi_n =$

$((X_1, Y_1), \dots, (X_n, Y_n))$, where the pairs (X_i, Y_i) are independent and have the same distribution as (X, Y) , and ξ_n is independent from (X, Y) . In this case we estimate Y in the form $g_n(X, \xi_n)$, a measurable function of the observation and the training sequence. The error probability is $\Pr\{g_n(X, \xi_n) \neq Y\}$. Due to the results of nonparametric pattern recognition and regression estimation (e.g. Stone [7], Devroye, Györfi [1]) it is well known that there exist decision rules such that $\lim_{n \rightarrow \infty} \Pr\{g_n(X, \xi_n) \neq Y\} - P^B(X, Y) = 0$ regardless of the underlying distribution of (X, Y) . In many practical cases, however, either the observation X or the training vectors are available only in distorted form: $T(X, \mu)$ or $T(X_i, \mu_i)$ ($i = 1, \dots, n$), respectively. Here μ, μ_1, \dots, μ_n are independent random variables taking their values from a measurable space (S, \mathcal{S}) and, by assumption, independent from (X, Y, ξ_n) , while T is an \mathbb{R}^d -valued mapping defined on $(\mathbb{R}^d \times S)$. Our question is whether it is possible to obtain a decision rule with error probability close to the Bayes-risk if the distortion is small, that is, if $E\rho(X, T(X, \mu))$ is small, where ρ denotes the Euclidean metric. The next theorem is an important good news.

THEOREM 1 (Faragó, Györfi [2]). Given $\epsilon > 0$ there exists a $\delta > 0$ such that for every function T and random variable μ satisfying $E\rho(X, T(X, \mu)) < \delta$,

$$P^B(T(X, \mu), Y) - P^B(X, Y) < \epsilon \quad \text{holds.}$$

The theorem states that the risk of the best decision from distorted observation is close to the optimum if the distortion is sufficiently small. This optimum can be approximated arbitrarily well if the training pairs are of the form $(T(X_i, \mu_i), Y_i)$ ($i = 1, \dots, n$). We are interested if the asymptotic error probability can be close to the Bayes-risk when the training is errorless but the observation is distorted, (Section 3) and when the training is distorted but the observation is not (Section 4) for sufficiently small distortion. As we will see, the answer is affirmative in both cases if the decision rule is based on the proposed randomization of the training.

2. Preliminary results

Before stating our results we need some key lemmas. Assume that the random variable ξ takes its values from the measurable space (S, \mathcal{S}) and it is independent from the pair (X, Y) (typically ξ plays the role of the training sequence) and let the measurable real valued functions $q_i(x, s)$, $i = 0, 1$, be defined on $\mathbb{R}^d \times S$. Define the decision g as follows:

$$g(x, s) = \arg \max_i q_i(x, s).$$

Using these notations we have the following:

Lemma 1 (Devroye, Györfi [1]).

$$\Pr\{g(X, \xi) \neq Y\} - P^B(X, Y) \leq E \left(\sum_{i=0}^1 |p_i(X) - q_i(X, \xi)| \right).$$

The statement of the lemma indicates that the error probability of decision g is very close to the Bayes-risk if the q_i are good L_1 approximations of the a posteriori probabilities.

The next lemma states that every decision based on maximization of measurable functions can be arbitrarily approximated by approximating the functions in L_1 sense.

Lemma 2 (Lugosi [5]). Let $q_0(x)$ and $q_1(x)$ be real valued measurable functions defined on \mathbb{R}^d . Let the decision function g be the following:

$$g(x) = \arg \max_i q_i(x).$$

If this maximum is unique almost everywhere (mod P_X), then for any sequence of measurable functions $\tilde{q}_i^{(n)}(x, s)$ ($i = 0, 1; n = 1, 2, \dots$), for which

$$\begin{aligned} \lim_{n \rightarrow \infty} E \left(\sum_{i=0}^1 |q_i(x) - \tilde{q}_i^{(n)}(x, s)| \right) &= 0, \\ \lim_{n \rightarrow \infty} |P_g(X, Y) - P_{\tilde{g}^{(n)}}(X, Y)| &= 0 \end{aligned}$$

holds, where

$$\tilde{g}^{(n)}(x, s) = \arg \max_i \tilde{q}_i^{(n)}(x, s)$$

(P_Z denotes the measure induced by a random variable Z).

We need one more technical lemma:

Lemma 3 (Faragó, Györfi [2]). Let $q : \mathbb{R}^d \rightarrow \mathbb{R}^+$ be a bounded continuous function. Then for all $\epsilon > 0$ there is a $\delta > 0$ such that $E\rho(X, T(X, \mu)) < \delta$ implies $E|q(X) - q(T(X, \mu))| < \epsilon$.

3. Errorless training, distorted observation

In this section we deal with the following problem: a decision designed from the training $\xi_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ is made from the distorted observation $T(X, \mu)$. First we consider the decision rule obtained by the maximization of an L_1 -consistent estimator $p_{in}(x)$ of the a posteriori probabilities $p_i(x) = \Pr\{Y = i \mid X =$

$x\}$, $i = 0, 1$, (by L_1 -consistency we mean that $\lim_{n \rightarrow \infty} E \left(\sum_{i=0}^1 |p_i(X) - p_{in}(X)| \right) = 0$), that is, our decision is of the form

$$g_n(x) = \arg \max_{i=0,1} p_{in}(x).$$

(In the notation we suppressed the dependence of g_n and p_{in} on ξ_n .) Our goal is to analyze the asymptotic behavior of the decisions, therefore, by Lemma 2 it is enough to investigate the performance of the decision

$$g(x) = \arg \max_{i=0,1} p_i(x).$$

The following lemma is a good news for smooth a posteriori probability functions:

Lemma 4. If the functions $p_i(x)$ are continuous ($i = 0, 1$), then for all $\epsilon > 0$ there is a $\delta > 0$ such that if $E\rho(X, T(X, \mu)) < \delta$, then

$$P_g(T(X, \mu), Y) - P^B(X, Y) < \epsilon.$$

Proof. It follows from Lemma 1 that

$$P_g(T(X, \mu), Y) - P^B(X, Y) \leq E \left(\sum_{i=0}^1 |p_i(X) - p_i(T(X, \mu))| \right),$$

the continuity and Lemma 3 give the desired result.

The following counterexample shows the necessity of the smoothness condition in Lemma 4.

Example. Let X be a real valued random variable and

$$Y = \begin{cases} 1 & \text{if } X \text{ is irrational} \\ 0 & \text{otherwise.} \end{cases}$$

Assume furthermore that $\Pr\{Y = 0\} = \Pr\{Y = 1\} = 1/2$. Consider the following sequence of transformations:

$$T_n(x) = \begin{cases} x - \frac{\pi}{n} & \text{if } X \text{ is rational} \\ \frac{[nx]}{n} & \text{otherwise.} \end{cases}$$

It is clear that on one hand $P^B(X, Y) = 0$, and on the other hand

$$P_g(T_n(X), Y) = \Pr\{g(T_n(X)) \neq Y\} = 1$$

for every n , while $\lim_{n \rightarrow \infty} E\rho(X, T_n(X)) = 0$.

In the remaining part of the section we show that there exists a randomized classification rule with asymptotic error probability close to the Bayes-risk for all distributions if the distortion is sufficiently small. The idea of the method is adding small "noise" to the training observations, estimating the a posteriori probabilities from the "noisy" sample and making the decision by maximizing them. This means that instead of the training ξ_n we use the data

$$\omega_n = ((X_1 + \nu_1, Y_1), \dots, (X_n + \nu_n, Y_n))$$

to estimate the functions

$$\hat{p}_i(x) = \Pr\{Y = i \mid X + \nu = x\} \quad (i = 0, 1),$$

where the \mathbb{R}^d -valued random variables ν, ν_1, \dots, ν_n are i.i.d., independent from (ξ_n, X, Y) with expected value zero, uniformly continuous density and

$$\sqrt{E(\|\nu\|^2)} \leq E\rho(X, T(X, \mu)).$$

If the applied estimation of the functions $\hat{p}_i(x)$ is L_∞ -consistent, that is, for the estimator $\hat{p}_i^{(n)}(x, \omega_n)$

$$\sup_x \sum_{i=0}^1 \left| \hat{p}_i(x) - \hat{p}_i^{(n)}(x, \omega_n) \right| \rightarrow 0$$

holds almost surely, then for all $\epsilon > 0$ there is a δ and n such that if $E\rho(X, T(X, \mu)) < \delta$ then

$$\sum_{i=0}^1 E|\hat{p}_i(T(X, \mu)) - \hat{p}_i^{(n)}(T(X, \mu), \omega_n)| < \epsilon.$$

In this case, by Lemma 2, the asymptotic error probability of the decision obtained by the maximization of the estimated functions is equal to that of decision

$$\hat{g}(x) = \arg \max_{i=0,1} \hat{p}_i(x),$$

thus, it is enough to investigate its error probability $P_{\hat{g}}(T(X, \mu), Y)$. Note that the uniform continuity of the density of ν implies that the distribution of $X + \nu$ and the conditional distributions of $X + \nu$ given $Y = i, i = 0, 1$, are absolutely continuous with uniformly continuous density (e.g. Wheeden, Zygmund [8]), therefore density estimation methods L_∞ -consistent for uniformly continuous densities are appropriate for our purpose. Such methods are kernel density estimation, histogram estimation and k-NN estimation (see Mack, Rosenblatt [6], Härdle, Janssen,

Serfling [4], Györfi, Härdle, Sarda, Vieu [3]). The main result of this section is the following theorem:

THEOREM 2. For every $\epsilon > 0$ there is a $\delta > 0$ such that if $E\rho(X, T(X, \mu)) < \delta$, then

$$P_{\hat{g}}(T(X, \mu), Y) - P^B(X, Y) < \epsilon.$$

Proof. Since $E\rho(X + \nu, X) \leq E\rho(X, T(X, \mu))$, using Theorem 1, for every ϵ there is a $\delta_1 > 0$ such that if $E\rho(X, T(X, \mu)) < \delta_1$, then

$$P^B(X + \nu, Y) - P^B(X, Y) < \epsilon/2.$$

The continuity of the density of ν implies the continuity of $\hat{p}_i(x)$, $i = 0, 1$, therefore, using Lemma 4, we conclude that there is a $\delta_2 > 0$ such that $E\rho(X, T(X, \mu)) < \delta$ implies

$$P_{\hat{g}}(T(X, \mu), Y) - P^B(X + \nu, Y) < \epsilon/2,$$

since $E\rho(X + \nu, T(X, Y)) \leq 2E\rho(X, T(X, \mu))$ by the triangle inequality. The choice $\delta = \min(\delta_1, \delta_2)$ completes the proof.

4. Distorted training, errorless observation

In the sequel we deal with the situation when the observation X is known but instead of knowing the training sequence ξ_n we have its distorted form $\zeta_n = (T(X_1, \mu_1), Y_1), \dots, (T(X_n, \mu_n), Y_n)$. First we show under certain conditions that we can get an asymptotically good decision by estimating the probabilities $\Pr\{Y = i \mid T(X, \mu) = x\}$ from ζ_n . However, this method does not work in general, but, as we will see, the randomization of the training helps just as in Section 3. Introduce some notations: Let $T \subset \mathbb{R}^d$ be the set of the possible values of $T(X, \mu)$, furthermore

$$q_i(x) = \Pr\{Y = i \mid T(X, \mu) = x\} \quad (i = 0, 1; x \in T),$$

$$g(x) = \begin{cases} \arg \max_{i=0,1} q_i(x) & \text{if } x \in T \\ -1 & \text{otherwise.} \end{cases}$$

Note that $X \notin T$ means error in the decision ($g(X) \neq Y$). Introducing the following function

$$\hat{g}(x, s) = \arg \max_{i=0,1} q_i(T(x, s))$$

it is clear that $\Pr\{\hat{g}(X, \mu) \neq Y\} = P^B(T(X, \mu), Y)$.

Lemma 5. If $\Pr\{X \in T\} = 1$, then for all $\epsilon > 0$ there exists a $\delta > 0$ such that from $E\rho(T(X, \mu), X) < \delta$

$$P_g(X, Y) - P^B(X, Y) < \epsilon \quad \text{follows.}$$

Proof.

$$\begin{aligned} & P_g(X, Y) - P^B(X, Y) \\ & \leq |P_g(X, Y) - P^B(T(X, \mu), Y)| + P^B(T(X, \mu), Y) - P^B(X, Y) \end{aligned} \quad (1)$$

Fix an arbitrary $\epsilon > 0$. Theorem 1 implies that there exists a $\delta_1 > 0$ such that if $E\rho(T(X, \mu), X) < \delta_1$, then $P^B(T(X, \mu), Y) - P^B(X, Y) < \epsilon/2$. Thus, we have to prove that the first term on the right hand side of (1) is small too, that is, there is a $\delta_2 > 0$ such that $E\rho(T(X, \mu), X) < \delta_2$ implies

$$|P_g(X, Y) - P^B(T(X, \mu), Y)| < \epsilon/2. \quad (2)$$

Let $\tilde{q}_0(x)$ and $\tilde{q}_1(x)$ be nonnegative continuous functions defined on \mathbf{R}^d . Introduce the following notations:

$$\begin{aligned} h(x) &= \arg \max_{i=0,1} \tilde{q}_i(x), \\ \hat{h}(x, s) &= \arg \max_{i=0,1} \tilde{q}_i(T(x, s)). \end{aligned}$$

Then applying the triangle inequality we have

$$\begin{aligned} & |P_g(X, Y) - P^B(T(X, \mu), Y)| \\ & \leq |P_g(X, Y) - P_h(X, Y)| + |P_h(X, Y) - \Pr\{\hat{h}(X, \mu) \neq Y\}| \\ & \quad + |\Pr\{\hat{h}(X, \mu) \neq Y\} - P^B(T(X, \mu), Y)| \end{aligned} \quad (3)$$

In the remaining part of the proof we show that the continuous functions $\tilde{q}_i(x)$ can be chosen such that for some $\delta_2 > 0$ all three terms on the right hand side of (3) are smaller than $\epsilon/6$ if $E\rho(T(X, \mu), X) < \delta_2$. We look at the right hand side of (3) term by term.

(i) The first term: By Lemma 2 there exists a $\delta_3 > 0$ such that if

$$E\left(\sum_{i=0}^1 |q_i(X) - \tilde{q}_i(X)|\right) < \delta_3, \quad (4)$$

then $|P_g(X, Y) - P_h(X, Y)| < \epsilon/6$.

(ii) The third term: By Lemma 1

$$|\Pr\{\hat{h}(X, \mu) \neq Y\} - P^B(T(X, \mu), Y)| \leq E\left(\sum_{i=0}^1 |q_i(T(X, \mu)) - \tilde{q}_i(T(X, \mu))|\right).$$

Thus, we have to show that the $\tilde{q}_i(x)$ can be chosen such that

$$E\left(\sum_{i=0}^1 |q_i(T(X, \mu)) - \tilde{q}_i(T(X, \mu))|\right) < \epsilon/6 \quad (5)$$

and (4) hold. This is possible since the set of continuous functions is dense in the space of integrable functions with respect to $P_X + P_{T(X, \mu)}$. Assume, therefore, that the functions $\tilde{q}_i(\mathbf{x})$ satisfy (4) and (5).

(iii) The second term: By Lemma 2 there is a $\delta_4 > 0$ such that

$$|P_h(X, Y) - \Pr\{\hat{h}(X, \mu) \neq Y\}| < \epsilon/6 \quad \text{if} \quad E\left(\sum_{i=0}^1 |\tilde{q}_i(X) - \tilde{q}_i(T(X, \mu))|\right) < \delta_4. \quad (6)$$

Now, because of the continuity of $\tilde{q}_i(\mathbf{x})$ we can use Lemma 3 which states that for this δ_4 there exists a $\delta_5 > 0$ such that (6) holds if $E\rho(T(X, \mu), X) < \delta_5$. Now, we can see that by choosing $\delta_2 = \delta_5$ (2) holds, which completes the proof. Finally, we note that the condition of the applicability of Lemma 2 is that decision h is unique almost surely (mod P_X). However, this can always be achieved by an arbitrarily small change in $\tilde{q}_i(\mathbf{x})$.

The condition in Lemma 5 ($\Pr\{X \in \mathcal{T}\} = 1$) does not hold in many important cases, e.g. if $T(X, \mu)$ is a quantization of X (does not depend on μ). To overcome this difficulty we can apply the same randomization as in Section 3, that is, instead of ζ_n we can use the “noisy” training

$$\vartheta_n = ((T(X_1, \mu_1) + \nu_1, Y_1), \dots, (T(X_n, \mu_n) + \nu_n, Y_n)),$$

where ν, ν_1, \dots, ν_n are i.i.d. random variables with zero mean, everywhere positive density and $\sqrt{E\|\nu\|^2} \leq E\rho(X, T(X, \mu))$. If we use ϑ_n to estimate the functions $\hat{q}_i(\mathbf{x}) = \Pr\{Y = i | T(X, \mu) + \nu = \mathbf{x}\}$ in an L_1 -consistent way, then by Lemma 2 it is enough to deal with the error probability $P_{\hat{g}}(X, Y)$ of the decision rule

$$\hat{g}(\mathbf{x}) = \arg \max_{i=0,1} \hat{q}_i(\mathbf{x}).$$

It is clear that $\hat{q}_i(\mathbf{x}), i = 0, 1$, and $\hat{g}(\mathbf{x})$ are defined everywhere. The following theorem states that, without any additional condition the asymptotic error probability of the randomized decision is close to the Bayes-risk.

THEOREM 3. Given $\epsilon > 0$ there exists a $\delta > 0$ such that $E\rho(X, T(X, \mu)) < \delta$ implies

$$P_{\hat{g}}(X, Y) - P^B(X, Y) < \epsilon.$$

Proof. Introducing the notation $\tilde{T}(X, (\mu, \nu)) = T(X, \mu) + \nu$, it is clear that

$$E\rho(X, \tilde{T}(X, (\mu, \nu))) \leq 2E\rho(X, T(X, \mu)),$$

from which using Lemma 5 the statement follows.

Acknowledgments

The author wishes to express his thanks to Laci Györfi and Tamás Linder for their help and support.

References

1. Devroye, L., Györfi, L., Nonparametric Density Estimation: The L_1 -View. Wiley, New York 1985.
2. Faragó, T., Györfi, L., On the continuity of the error distortion function for multiple hypotheses decisions. IEEE Trans. on Information Theory, **IT-21** (1975), pp. 458-460
3. Györfi, L., Härdle, W., Sarda, P., Vieu, P., Nonparametric Curve Estimation From Time Series. Springer-Verlag 1989
4. Härdle, W., Janssen, P., Serfling R., Strong uniform consistency for estimators of conditional functionals. Sonderforschungsbereich. In: Information und die Koordination wirtschaftlicher Aktivitäten, 1986.
5. Lugosi, G., Learning with an unreliable teacher. Pattern Recognition, to appear
6. Mack, Y. P., Rosenblatt, M., Multivariate k -nearest neighbor density estimates. J. Multivariate Analysis, **9** (1979), pp. 1-15.
7. Stone, C. J., Consistent nonparametric regression. Annals of Statistics, Vol. 8 (1977), pp. 1348-1360.
8. Wheeden, R. L., Zygmund, A., Measure and Integral. Marcel Dekker, New York, 1977.

Классификация образов из искаженных выборок

Г. ЛУГОШИ

(Будапешт)

В задачах непараметрической классификации образов решение о категории наблюдаемого вектора производится из длинной обучаемой последовательности. Однако во многих практических случаях наблюдаемый вектор и последовательность обучения могут быть искажены. В статье рассматриваются асимптотические вероятности ошибок решений в таких случаях.

G. Lugosi
Technical University of Budapest
H-1521 Budapest
Stoczek u. 2.
Hungary

ON ASYMPTOTICALLY OPTIMAL COMPANDING QUANTIZATION

TAMÁS LINDER

(*Budapest*)

(Received October 1, 1991)

The validity of Bennett's formula for companding quantizers is shown under precise conditions for r th power distortion measures. Using these conditions it is shown rigorously that certain companders are asymptotically optimal, i.e., their distortion and the distortion of optimal quantizers decrease to zero at the same rate, as the number of quantization levels increases to infinity. Some defects in previous derivations concerning companders are pointed out.

1. Introduction

The design of optimal N -level scalar quantizers for mean-squared distortion measure was first considered by Lloyd [8] and Max [9]. In general, the resulting iterative algorithms give suboptimal quantizers. Necessary conditions for the so-called Lloyd–Max algorithm to converge to the global optimum was given by Thrushkin [11] and Kieffer [7], the later proving exponential rate of convergence.

A parallel approach for scalar quantization is Bennett's companding quantizer. Bennett [1] modeled a nonuniform N -level scalar quantizer by a memoryless nonlinearity $G(\cdot)$ followed by an N -level uniform quantizer $Q_{N,U}$, which is followed by the inverse of the nonlinearity G^{-1} . Formally, the N -level companding quantizer (also called compander) $Q_{N,G}$ is defined by

$$Q_{N,G}(x) = G^{-1}(Q_{N,U}[G(x)]) \quad (1)$$

where $G : \mathbf{R} \rightarrow [0, 1]$ is onto and increasing, and $Q_{N,U}$ is the N -level uniform quantizer on $[0, 1]$, i.e.,

$$Q_{N,U}(x) = \frac{n-1}{N} + \frac{1}{2N} \quad \text{if } x \in \left(\frac{n-1}{N}, \frac{n}{N} \right]$$

for $n = 1, \dots, N-1$. Clearly, all N -level scalar quantizers can be implemented this way. From now on we assume that the quantized random variable X has a density

f . Bennett demonstrated that for large N the mean-squared error $D(Q_{N,G}) = E|X - Q_{N,G}(X)|^2$ satisfies

$$D(Q_{N,G}) \approx \frac{1}{N^2} \frac{1}{12} \int_{-L}^L \frac{f(x)}{[g(x)]^2} dx, \quad (2)$$

where g is the derivative of G , and $[-L, L]$ is the (bounded) support of f , the density of the random variable being quantized. Note that g is a probability density function, i.e., $g \geq 0$ and $\int g = 1$.

Bennett's formula, when formally generalized for r th power distortions $D(Q_{N,G}) = E|X - Q_{N,G}(X)|^r$, $r > 0$, gives

$$\lim_{N \rightarrow \infty} N^r D(Q_{N,G}) = \frac{1}{(r+1)2^r} \int_{\mathbb{R}} \frac{f(x)}{[g(x)]^r} dx. \quad (3)$$

(See Gish and Pierce [4], or Gray and Gray [5] for details.) Although this formula has been widely used in the engineering literature, the only results claiming to give sufficient conditions for it to hold appeared in Cambanis and Gorr [3], and Bucklew and Wise [2]. It is easily seen via Hölder's inequality that the right-hand side of (3) is minimal iff $g(x) = f(x)^{1/(r+1)} / \int_{\mathbb{R}} f(z)^{1/(r+1)} dz$, suggesting that if this generalized

Bennett's formula holds, then the companding quantizers with this characteristics are nearly optimal for large N . This near optimality of the quantizers $Q_{N,G}$ was first rigorously dealt with by Cambanis and Gorr [3]. They called a sequence of quantizers Q_N^* *asymptotically optimal* if the distortion of Q_N^* tends to zero at the same rate as the distortion of the N -level optimal quantizer, as $N \rightarrow \infty$. Formally

$$\lim_{N \rightarrow \infty} \frac{D(Q_N^*)}{\inf_{Q_N} D(Q_N)} = 1,$$

where the infimum is taken over all N -level quantizers Q_N . Now, from Zador [14] (c.f. [13]), we have

$$\lim_{N \rightarrow \infty} N^r \inf_{Q_N} D(Q_N) = \frac{1}{(r+1)2^r} \left(\int_{\mathbb{R}} [f(x)]^{1/(r+1)} dx \right)^{r+1}, \quad (4)$$

if X has density f . While the precise conditions for this are hard to deduce from the paper, Bucklew and Wise [2, Theorem 2] give the following simple and general condition for the validity of (4): $E|X|^{r+\epsilon} < \infty$, for some $\epsilon > 0$. If we substitute $g(x) = f(x)^{1/(r+1)} / \int_{\mathbb{R}} f(z)^{1/(r+1)} dz$ into (3), then we obtain

$$\lim_{N \rightarrow \infty} N^r D(Q_{N,G}) = \frac{1}{(r+1)2^r} \left(\int_{\mathbb{R}} [f(x)]^{1/(r+1)} dx \right)^{r+1}.$$

This shows that the optimal choice of compander characteristics results in asymptotically optimal quantizers if (3) with this choice holds.

In what follows we derive precise conditions for (3) to hold, thus giving sufficient conditions for a sequence of companding quantizers being asymptotically optimal.

2. Main result

Bennett's integral (2) is a formula that can be found in most of the engineering literature dealing with quantization. The companding approach to scalar quantization and the derivation of the optimal compander characteristics is studied from a more practical point of view in Jayant and Noll [6]. Interestingly enough, only two results [2] and [3] present sufficient conditions for (3). In [3, Theorem 1] these sufficient conditions were the following:

- (a) $f(x)$ and $G'(x) = g(x)$ are continuous,
- (b) $E|X|^r < \infty$,
- (c) $f(x)/[g(x)]^r$ is Riemann integrable on \mathbb{R} ,

$$\lim_{y \rightarrow \infty} \left(\int_y^\infty |x - y|^r f(x) dx \right) / \left(\int_y^\infty g(x) dx \right)^r = 0$$

- (d) and

$$\lim_{y \rightarrow -\infty} \left(\int_{-\infty}^y |x - y|^r f(x) dx \right) / \left(\int_{-\infty}^y g(x) dx \right)^r = 0.$$

The conditions that f be continuous and (c) are rather restrictive. In addition, there is a gap in the proof concerning the convergence of Riemann sums with increasing support to an improper Riemann integral.

In Bucklew and Wise [2, Theorem 1] the conditions for (3) were the following:

- (A) $G'(x) = g(x)$ is continuous and positive,
- (B) there exists an $M > 0$ such that $g(x)$ is increasing if $x < -M$ and $g(x)$ is decreasing if $x > M$,
- (C) $\int_{\mathbb{R}} f(x)/[g(x)]^{r+\epsilon} dx < \infty$ for some $\epsilon > 0$

These conditions do not involve the smoothness and Riemann integrability of f . Although they are not directly comparable, conditions (A)–(C) are more appealing

than conditions (a)–(d). However, the proof in [2, Theorem 1] is not complete either. Namely, they use the following proposition at the end of the proof:

Let $p(x) \geq 0$ be integrable on $[0, 1]$, and let $q(x) > 0$ be continuous and monotone decreasing in $(0, 1)$. Define the function $q_N(x)$ by

$$q_N(x) = \begin{cases} q(x) & \text{if } x \in \left(0, \frac{1}{N}\right) \\ \sup_{x \in \left[\frac{n-1}{N}, \frac{n}{N}\right)} q(x) & \text{if } x \in \left[\frac{n-1}{N}, \frac{n}{N}\right), \quad n = 2, \dots, N, \end{cases}$$

and assume that

$$\int_0^1 q(x)p(x) dx < \infty. \quad (5)$$

Then it is asserted that

$$\lim_{N \rightarrow \infty} \int_0^1 q_N(x)p(x) dx = \int_0^1 q(x)p(x) dx. \quad (6)$$

Were the integrands in (5) and (6) $q_N(x)$ and $q(x)$ alone, this claim would surely hold, for the shift of $q_N(x)$ to the left by $1/N$ would allow us to use the dominated convergence theorem. But $p(x)$ is not translation invariant, thus this trick does not apply. Indeed, consider $p(x) = e^{-1/x}$ and $q(x) = e^{1/x}$. Then $\int_0^1 q(x)p(x) dx = 1$ thus (5) is satisfied. But, since $p(x) = e^{-1/x}$ is convex in $(0, 1/2]$, it is lower bounded by

$$p(1/N) + p'(1/N)(x - 1/N) = e^{-N} + (x - 1/N)N^2 e^{-N}$$

in the interval $[1/N, 2/N]$. Thus we have

$$\begin{aligned} \int_{1/N}^{2/N} q_N(x)p(x) dx &\leq \int_{1/N}^{2/N} [1 + (x - 1/N)n^2] dx \\ &= \frac{1}{2} + \frac{1}{N} \rightarrow \frac{1}{2} \end{aligned} \quad (7)$$

as $N \rightarrow \infty$. For (6) to hold it is necessary that the left-hand side of (7) tend to zero, thus (6) does not hold with the above conditions in general.

In what follows we give rather general sufficient conditions for (3). Our conditions will be almost the same as (A)–(C), except that we relax (C) and impose a new tail condition to avoid the convergence problem above.

Denote the inverse of G by S , and let $S' = s$. A simple change of variables shows that

$$\int_{\mathbb{R}} \frac{f(x)}{[g(x)]^r} dx = \int_0^1 [s(x)]^r p(x) dx$$

where $p(x) = f(S(x))/g(S(x))$, a probability density function with support $[0, 1]$. We need the following conditions:

(C')
$$\int_{\mathbb{R}} \frac{f(x)}{[g(x)]^r} dx < \infty,$$

For some $\epsilon > 0$

$$\int_0^\epsilon [s(x/2)]^r p(x) dx < \infty$$

(D) and

$$\int_{1-\epsilon}^1 [s((x+1)/2)]^r p(x) dx < \infty.$$

THEOREM 1. Suppose that the conditions (A), (B), (C') and (D) hold. Then

$$\lim_{N \rightarrow \infty} N^r D(Q_{N,G}) = \frac{1}{(r+1)2^r} \int_{\mathbb{R}} \frac{f(x)}{[g(x)]^r} dx. \tag{8}$$

Proof. Note that the quantization intervals of $Q_{N,G}$ are $I_{1,N} = (-\infty, S(1/N))$, $I_{n,N} = \left[S\left(\frac{n-1}{N}\right), S\left(\frac{n}{N}\right) \right]$, $n = 2, \dots, N-1$, and $I_{N,N} = \left[S\left(\frac{N-1}{N}\right), \infty \right)$.

The corresponding levels are $y_{n,N} = S\left(\frac{2n-1}{N}\right)$, $n = 1, \dots, N$. First we consider the quantizer $\bar{Q}_{N,G}$ with the same quantization intervals as $Q_{N,G}$, but with levels at the midpoints of the intervals, except for the two unbounded intervals, where the levels are unchanged. Thus

$$\bar{Q}_{N,G}(x) = \begin{cases} \bar{y}_{n,N} = \frac{1}{2} \left[S\left(\frac{n}{N}\right) + S\left(\frac{n-1}{N}\right) \right], & \text{if } x \in I_{n,N} \ n = 2, \dots, N-1, \\ \bar{y}_{1,N} = S(1/2N) & \text{if } x \in I_{1,N}, \\ \bar{y}_{N,N} = S\left(\frac{2N-1}{2N}\right) & \text{if } x \in I_{N,N}. \end{cases}$$

Let us define the function $s_N : (0, 1) \rightarrow (0, \infty)$ by

$$s_N(x) = \begin{cases} \sup_{x \in [\frac{n-1}{N}, \frac{n}{N})} s(x) & \text{if } x \in \left[\frac{n-1}{N}, \frac{n}{N} \right), n = 2, \dots, N-1, \\ s(1/2N) & \text{if } x \in \left[\frac{1}{2N}, \frac{1}{N} \right). \\ s\left(\frac{2N-1}{2N}\right) & \text{if } x \in \left[\frac{N-1}{N}, \frac{2N-1}{2N} \right). \\ s(x) & \text{if } x \in \left(0, \frac{1}{N}\right) \cup \left[\frac{2N-1}{2N}, 1\right) \end{cases}$$

First we show that

$$\lim_{N \rightarrow \infty} \int_{\mathbb{R}} [s_N(G(x))]^r f(x) dx = \int_{\mathbb{R}} [s(G(x))]^r f(x) dx. \tag{9}$$

Note that $s_N(x) \rightarrow s(x)$ as $N \rightarrow \infty$ for all $x \in (0, 1)$ by the continuity of s . By change of variables

$$\int_{\mathbb{R}} [s_N(G(x))]^r f(x) dx = \int_0^1 [s_N(y)]^r p(y) dy.$$

Let $0 < \epsilon < 1/2$ be such that (D) is satisfied and $s(x)$ is decreasing in $(0, \epsilon)$ and $s(x)$ is increasing in $(1-\epsilon, 1)$. The inequalities $\frac{1}{2} \frac{i+1}{N} \leq \frac{i}{N}$ and $\frac{1}{2} \left(\left[1 - \frac{i+1}{N} \right] + 1 \right) \geq 1 - \frac{i}{N}$ valid for $i \geq 1$ show that $s(x/2) \geq s_N(x)$ and $s((x+1)/2) \geq s(x)$ if $x \in (0, \epsilon)$ or $x \in (1-\epsilon, 1)$, respectively. Then by (D) and the continuity of s , the dominated convergence theorem implies that

$$\lim_{N \rightarrow \infty} \int_0^\epsilon [s_N(x)]^r p(x) dx = \int_0^\epsilon [s(x)]^r p(x) dx, \tag{10}$$

and

$$\lim_{N \rightarrow \infty} \int_{1-\epsilon}^1 [s_N(x)]^r p(x) dx = \int_{1-\epsilon}^1 [s(x)]^r p(x) dx. \tag{11}$$

Since s is bounded on $[\epsilon, 1-\epsilon]$, $\lim_{N \rightarrow \infty} \int_{\epsilon}^{1-\epsilon} s_N^r p = \int_{\epsilon}^{1-\epsilon} s^r p$ clearly holds. (10), (11) and this observation proves (9).

The remaining part of the proof is done in two steps. The first step is to prove that

$$\lim_{N \rightarrow \infty} N^r D(\bar{Q}_{N,G}) = \frac{1}{(r+1)2^r} \int_{\mathbb{R}} \frac{f(x)}{[g(x)]^r} dx. \tag{12}$$

Let $x \in I_{n,N}$ for $n = 1, \dots, N$. Then by the mean-value theorem of differentiation

$$\begin{aligned} N^r |x - \bar{Q}_{N,G}(x)|^r &\leq N^r [\lambda(I_{n,N})]^r \\ &\leq [s_N(G(x))]^r, \end{aligned} \tag{13}$$

where λ stands for the Lebesgue measure. Furthermore, some simple calculations using the monotonicity of $s(x)$ near 0 and 1 show that (13) holds for all $x \in (0, 1)$, if N is large enough. Let $A_N \stackrel{\text{def}}{=} \int_{\mathbb{R} \setminus (I_{1,N} \cup I_{N,N})} f(x) dx$. Then $A_N \uparrow 1$ as $N \rightarrow \infty$.

Define the piecewise constant density f_N by

$$f_N(x) = \begin{cases} \frac{1}{A_N \lambda(I_{n,N})} \int_{I_{n,N}} f(y) dy & \text{if } x \in I_{n,N}, n = 2, \dots, N-1, \\ 0 & \text{if } x \in I_{1,N} \cup I_{N,N}. \end{cases}$$

Now by (13) we have

$$\begin{aligned} \left| N^r \int_{\mathbb{R}} |x - \bar{Q}_{N,G}(x)|^r f(x) dx - N^r \int_{\mathbb{R}} |x - \bar{Q}_{N,G}(x)|^r f_N(x) dx \right| &\leq \\ &\int_{\mathbb{R}} [s_N(G(x))]^r |f(x) - f_N(x)| dx. \end{aligned} \tag{14}$$

But from the definition of f_N and from the fact that s_N is piecewise constant on the support of f_N , we have

$$\begin{aligned} \int_{\mathbb{R}} [s_N(G(x))]^r |f(x) - f_N(x)| dx &\leq \left(1 + \frac{1}{A_N}\right) \int_{\mathbb{R}} [s_N(G(x))]^r f(x) dx \\ &\leq 3 \int_{\mathbb{R}} [s_N(G(x))]^r f(x) dx, \end{aligned} \tag{15}$$

if N is large enough. From Lebesgue's differentiation theorem [12] $f_N \rightarrow f$ as $N \rightarrow \infty$ a.e. λ . Considering (9), (15), (14) and this fact, a generalization of the dominated convergence theorem (see Royden [10]) implies that

$$\lim_{N \rightarrow \infty} N^r \int_{\mathbb{R}} |x - \bar{Q}_{N,G}(x)|^r f(x) dx = \lim_{N \rightarrow \infty} N^r \int_{\mathbb{R}} |x - \bar{Q}_{N,G}(x)|^r f_N(x) dx, \tag{16}$$

provided that the limit on the right-hand side exists. To prove that the right-hand side is $\frac{1}{(r+1)2^r} \int f/g^r$, define \hat{s}_N by

$$\hat{s}_N(x) = \begin{cases} N\lambda(I_{n,N}) & \text{if } x \in \left[\frac{n-1}{N}, \frac{n}{N} \right) \quad n = 2, \dots, N-1, \\ 0 & \text{if } x \in \left(0, \frac{1}{N} \right) \cup \left[\frac{N-1}{N}, 1 \right). \end{cases}$$

Then $\hat{s}_N(x) \leq s_N(x)$ for all $x \in (0, 1)$ by (13). Since the $\bar{y}_{n,N}$, $n = 2, \dots, N-1$ are the midpoints of the corresponding $I_{n,N}$, and since $f_N(x)$ is constant on $I_{n,N}$, we obtain

$$N^r \int_{\mathbb{R}} |x - \bar{Q}_{N,G}(x)|^r f_N(x) dx = \frac{1}{(r+1)2^r} \frac{1}{A_N} \int_{\mathbb{R}} [\hat{s}_N(G(x))]^r f(x) dx. \quad (17)$$

Now since $\hat{s}_N(x) \rightarrow s(x)$ as $N \rightarrow \infty$ for all $x \in (0, 1)$, the generalized dominated convergence theorem implies that (17) tends to

$$\frac{1}{(r+1)2^r} \int_{\mathbb{R}} [s(G(x))]^r f(x) dx = \frac{1}{(r+1)2^r} \int_{\mathbb{R}} \frac{f(x)}{[g(x)]^r} dx,$$

which, when combined with (16) proves (12).

The next step is to prove that

$$\lim_{N \rightarrow \infty} N^r D(Q_{N,G}) = \lim_{N \rightarrow \infty} N^r D(\bar{Q}_{N,G}). \quad (18)$$

Clearly, the quantity of interest is

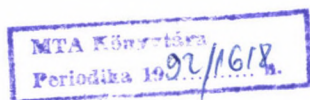
$$N^r \left| |x - \bar{Q}_{N,G}(x)|^r - |x - Q_{N,G}(x)|^r \right| = N^r \left| |x - \bar{y}_{n,N}|^r - |x - y_{n,N}|^r \right|, \quad (19)$$

if $x \in I_{n,N}$. By a first order Taylor expansion

$$\left| |x - \bar{y}_{n,N}|^r - |x - y_{n,N}|^r \right| \leq |\bar{y}_{n,N} - y_{n,N}| r \zeta_{n,N}^{r-1} \quad (20)$$

where $0 \leq \zeta_{n,N} \leq \lambda(I_{n,N}) \leq \frac{1}{N} s_N(G(x))$. The difference $|\bar{y}_{n,N} - y_{n,N}|$ can be estimated as follows. If $n = 1$ or $n = N$, the difference is 0. Otherwise we have $y_{n,N} = S\left(\frac{n-1}{N}\right) + \frac{1}{2N} s(\xi_{n,N})$ and $\bar{y}_{n,N} = S\left(\frac{n-1}{N}\right) + \frac{1}{2N} s(\eta_{n,N})$ for some $\eta_{n,N}, \xi_{n,N} \in \left[\frac{n-1}{N}, \frac{n}{N}\right]$ by the mean-value theorem of differentiation. Thus we have

$$\begin{aligned} N^r \left| |x - \bar{y}_{n,N}|^r - |x - y_{n,N}|^r \right| &\leq \frac{r}{2} |s(\eta_{n,N}) - s(\xi_{n,N})| [s_N(G(x))]^{r-1} \\ &\leq r [s_N(G(x))]^r. \end{aligned} \quad (21)$$



The continuity of $s(x)$ and the first inequality in (21) show that the right-hand side tends to zero as $N \rightarrow \infty$. On the other hand, (9) and the second inequality in (21) allow us to use again the generalized dominated convergence theorem to conclude that

$$\lim_{N \rightarrow \infty} \int_{\mathbb{R}} N^r \left| |x - \hat{Q}_{N,G}(x)|^r - |x - Q_{N,G}(x)|^r \right| f(x) dx = 0,$$

which implies (18). This and (12) yield the theorem. \square

To obtain provenly asymptotically optimal companders one should only check whether f and the optimal choice of g satisfy the conditions of the theorem. Admittedly, there is a discrepancy between the conditions for f and g , since for the former the conditions are much less restrictive. Unfortunately, condition (D) can not be interpreted in an easy way, in general.

It should be mentioned, that Bennett's formula is conjectured to be true with the only condition (C'), but no proof has been given to date.

For some regular densities such as Gaussian, Rayleigh, and Laplacian, the asymptotically optimal quantizers were computed for different values of r in [3]. These numerical results show that the performance of these quantizers compare favorably with the performance of optimal quantizers even for moderately low number of quantization levels, thus showing the applicability of the asymptotic theory.

3. Conclusion

The validity of Bennett's formula was established under precise sufficient conditions, giving a justification for the claim that the optimal compander characteristics yield nearly optimal quantizers. These quantizers are easy to compute if the density of the random variable being quantized is known, and their implementation is straightforward.

References

1. Bennett, W. R., Spectrum of quantized signals. *Bell. Syst. Tech. J.*, Vol. 27 (1948), pp. 446-472.
2. Bucklew, J. A. and Wise, G. L., Multidimensional asymptotic quantization theory with r th power distortion measures. *IEEE Trans. Inform. Theory*, Vol. IT-28 (March 1982), pp. 239-247.
3. Cambanis, S. and Gerr, N. L., A simple class of asymptotically optimal quantizers. *IEEE Trans. Inform. Theory*, Vol. IT-29 (Sept. 1983), pp. 664-676.
4. Gish, H. and Pierce, J. N., Asymptotically efficient quantizing. *IEEE Trans. Inform. Theory*, Vol. IT-14 (Sept. 1968), pp. 676-683.

5. Gray, R. M. and Gray, A. H., Jr., Asymptotically optimal quantizers. IEEE Trans. Inform. Theory, Vol. IT-23 (Jan. 1977), pp. 143-144.
6. Jayant, N. S. and Noll, P., Digital Coding of Waveforms. Prentice-Hall, Englewood Cliffs, 1984.
7. Kieffer, J. C., Exponential rate of convergence for Lloyd's method I. IEEE Trans. Inform. Theory, Vol. IT-28 (March 1982), pp. 205-210.
8. Lloyd, S. P., Least squares quantization in PCM. IEEE Trans. Inform. Theory, Vol. IT-28 (March 1982), pp. 129-137 (originally a 1957 Bell Labs memorandum).
9. Max, J., Quantizing for minimum distortion. IEEE Trans. Inform. Theory, Vol. IT-6 (March 1960), pp. 7-12.
10. Royden, H. L., Real Analysis. Collier Macmillan, New York, 1968.
11. Trushkin, A. V., Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting function. IEEE Trans. Inform. Theory, Vol. IT-28 (March 1982), pp. 187-198.
12. Wheeden, R. L. and Zygmund, A. Z., Measure and Integral. Marcel Dekker, New York, 1977.
13. Zador, P., Topics in the asymptotic quantization of continuous random variables. Unpublished memorandum, Bell Laboratories, Murray Hill, NJ, Feb. 1966.
14. Zador, P., Asymptotic quantization error of continuous signals and the quantization dimension. IEEE Trans. Inform. Theory, Vol. IT-28 (March 1982), pp. 139-149.

Об асимптотически оптимальном нелинейном квантовании

Т. ЛИНДЕР

(Будапешт)

Справедливость формулы Беннетта для нелинейных квантователей показана для точных условий для мер искажений мощности порядка r . С использованием этих условий строго доказывается, что некоторые компандеры являются асимптотически оптимальными.

T. Linder
 Technical University of Budapest
 H-1521 Budapest
 Stoczek u. 2.
 Hungary

NOTE TO CONTRIBUTORS

Two copies of the *manuscript* (each complete with figures, tables and references) are to be sent 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 V. STREJC
UTIA ČSAV
182 08 Prague 8
Pod vodárenskou věží 4, Czechoslovakia

or to L. GYÖRFI
Technical University of Budapest
H-1111 Budapest, Stoczek u. 2, Hungary

Authors are requested to retain a third copy of the submitted typescript to be able to check the proofs.

The papers, preferably in English or Russian, should be typed double spaced on one side of good-quality paper with wide margins (4–5 cm). The first page of the paper should carry the title, the author(s)' names and the name of the town where they are active. The name and address of the author to whom the proofs should be sent should be given at the end of the paper. An *abstract* should head the paper. English papers should also have a Russian abstract.

The papers should not exceed 15 pages (25 × 50 characters per page) including tables and references. The proper location of the tables and figures must be indicated on the margin.

Mathematical notations should follow up-to-date usage. Equations longer than half a line should not be incorporated in the text. In-text equations must be typed on a single line except that one level of subscripting and/or superscripting is permissible. Use / instead of horizontal bars. Displayed equations should be written so as to require the fewest possible lines. Therefore use "exp" for the exponential function whenever the exponent requires more than a single line. Matrices should, if possible, not be written in full. Use subscript notations instead such as $A = ||a_{ij}||$. Write diagonal matrices as $\text{diag}(d_1, d_2, \dots, d_n)$.

The authors will be sent galley proofs to be returned by next mail. Rejected manuscripts will be returned. Authors will receive 100 reprints free of charge. Additional reprints may be ordered.

К СВЕДЕНИЮ АВТОРОВ

Рукописи статей в трех экземплярах на русском языке и в трех на английском следует направлять по адресу: 117312 Москва В-312, просп. 60 летия Октября, 9, МНИИПУ. Редакция журнала «Проблемы управления и теории информации» (зав. редакцией Н. И. Родионова).

Объем статьи не должен превышать 15 печатных страниц (25 строк по 50 букв). Статья должна предшествовать аннотация объемом 50–100 слов и приложено резюме–реферат объемом не менее 10–15% объема статьи на русском языке в трех экземплярах, на котором напечатан служебный адрес автора (фамилия, название учреждения, адрес).

При написании статьи авторам надо строго придерживаться следующей формы: введение (постановка задачи), основное содержание, примеры практического использования, обсуждение результатов, выводы и литература.

Статьи должны быть отпечатаны с промежутком в два интервала, последовательность таблиц и рисунков должна быть отмечена на полях. Математические обозначения рекомендуется давать в соответствии с современными требованиями и традициями. Разметку букв следует производить только во втором экземпляре и русского, и английского варианта статьи.

Авторам высылается верстка, которую необходимо незамедлительно проверить и вернуть в редакцию.

После публикации авторам высылаются бесплатно 100 оттисков их статей.

Рукописи непринятых статей возвращаются авторам.

CONTENTS · СОДЕРЖАНИЕ

<i>Vesely, V., Barč, V., Hindi, K. S.</i> : A decentralized control scheme for continuous-time systems through partial aggregation (<i>Веселы В., Барч В., Хинди К. С.</i> Децентрализованный алгоритм управления многосвязными системами)	373
<i>Faragó, A., Linder, T., Lugosi, G.</i> : Nearest neighbor search and classification in $O(1)$ time (<i>Фараго А., Линдер Т., Лугоши Г.</i> Поиск ближайшего соседа и классификация за время $O(1)$)	383
<i>En-hui Yang</i> : Universal almost sure data compression for abstract alphabets and arbitrary fidelity criterions (<i>Эн-хуи Янг</i> , Универсальное почти всюду сжатие данных для абстрактных алфавитов у произвольных критериев верности)	397
<i>Gabasov, R., Kirillova, F. M., Gaishun, P. V., Prischepova, S. V.</i> : Synthesis of optimal controls on nonexact measurements of output signals (<i>Габасов Р., Кириллова Ф. М., Гайшун П. В., Прищепова С. В.</i> Синтез оптимальных управлений по неточным измерениям выходных сигналов)	409
<i>Otáhal, A.</i> : Parameter estimation for nearest neighbor Gaussian random fields in the plane (<i>Отахал А.</i> Оценка параметров марковских гауссовских случайных полей в плоскости)	429
<i>Carbonez, A., Györfi, L., van der Meulen, E. C.</i> : Nonparametric entropy estimation based on randomly censored data (<i>Карбонез А., Дьёрфи Л., ван дер Майлен Э. К.</i> Непараметрическая оценка энтропии на основе случайно цензурированных данных)	441
<i>Morvai, G.</i> : Empirical log-optimal portfolio selection (<i>Морваи Г.</i> Эмпирическая лог-оптимальная селекция портфеля)	453
<i>Lugosi, G.</i> : Pattern classification from distorted sample (<i>Лугоши Г.</i> Классификация образов из искаженных выборок)	465
<i>Linder, T.</i> : On asymptotically optimal companding quantization (<i>Линдер Т.</i> Об асимптотически оптимальном нелинейном квантовании)	475