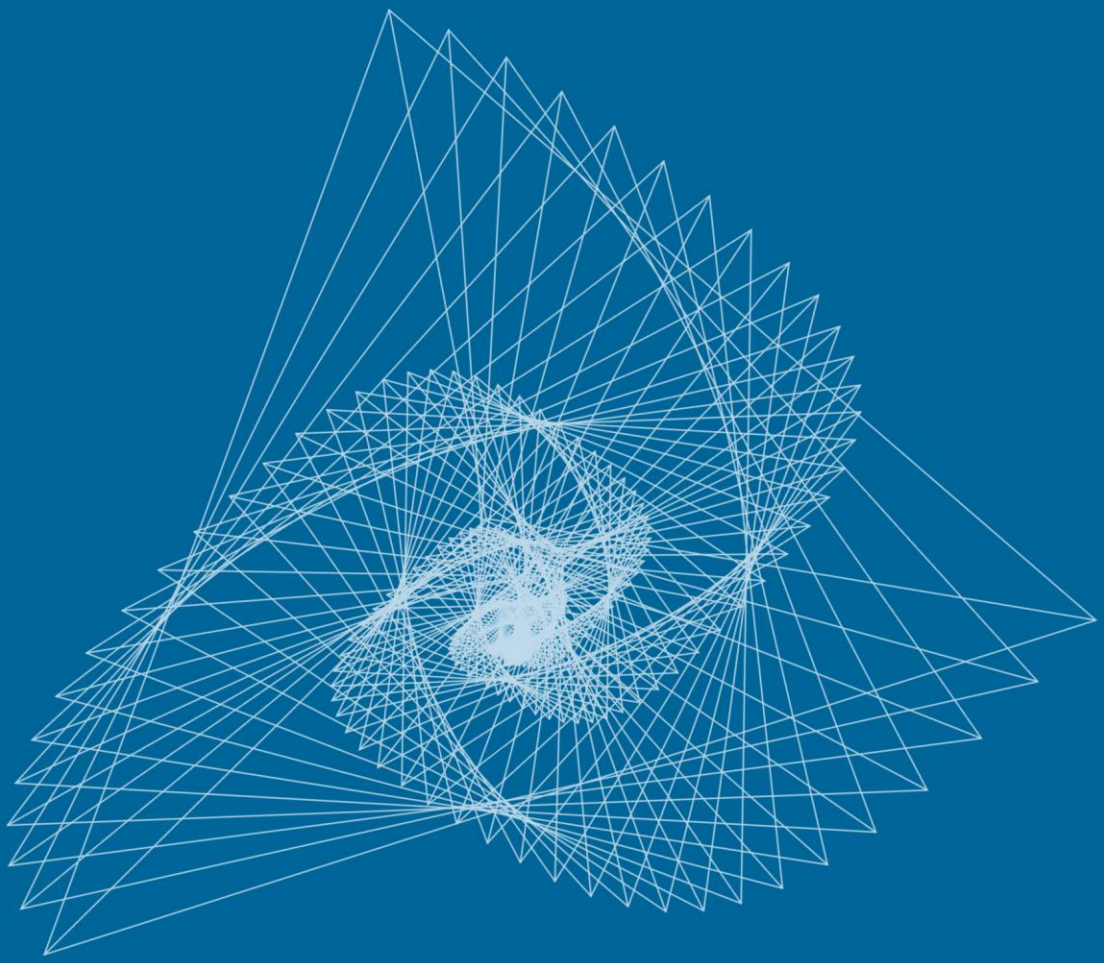
VOL. 2, NO. 2

URAL MATHEMATICAL JOURNAL

N.N. Krasovskii Institute of Mathematics and Mechanics of
the Ural Branch of Russian Academy of Sciences and
Ural Federal University named after the first President of Russia B.N.Yeltsin

ISSN: 2414-3952



UMJURAN.RU
UMJ

2016

ISSN 2414-3952
(Online)

Ural Mathematical Journal

Vol. 2, No. 2, 2016

https://umjuran.ru

Electronic Periodical Reviewed Scientific Journal
Founded in 2015

The Journal is registered by the Federal Service for Supervision in the Sphere of Communication, Information Technologies and Mass Communications
Certificate of Registration of the Mass Media El № FS77-61719 of 07.05.2015

Founders

N.N.Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N.Yeltsin

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990
Phone: +7 (343) 375-34-73 Fax: +7 (343) 374-25-81
Email: secretary@umjuran.ru
Web-site: https://umjuran.ru

EDITORIAL TEAM

TABLE OF CONTENTS

PART I. INTERNATIONAL CONFERENCE "SYSTEMS ANALYSIS: MODELING AND CONTROL"

PART II. GENERAL TOPICS

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

IN MEMORY OF ARKADY VIKTOROVICH KRYAZHIMSKIY (1949–2014)

Sergei M. Aseev[a,b], Alexander G. Chentsov[c], Alexey A. Davydov[b,d,e], Nikolai L. Grigorenko[e], Vyacheslav I. Maksimov[c,f], Elena A. Rovenskaya[b,e], Alexander M. Tarasiev[c]

[a] Steklov Mathematical Institute, Moscow, Russia

[b] International Institute for Applied Systems Analysis, Laxenburg, Austria

[c] Krasovskii Institute of Mathematics and Mechanics, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia

[d] Moscow University of Science and Technology MISiS, Moscow, Russia

[e] Lomonosov Moscow State University, Moscow, Russia

[f] Email: maksimov@imm.uran.ru

Abstract: The article is devoted to the description of Academician Arkady Kryazhimskiy's life path. The facts of the scientific biography of Acad. Kryazhimskiy are presented with the emphasis on his outstanding contribution into the theory of dynamic inversion, the theory of differential games, and control theory. His personal talents in different spheres are also marked out.

Key words: Arkady Viktorovich Kryazhimskiy.

Arkady Viktorovich Kryazhimskiy was born on January 2, 1949, in Qingdao, China. In 1971 he graduated from the Department of Mathematics and Mechanics of Gor'kii Ural State University in Sverdlovsk (now, Ekaterinburg) and entered a postgraduate program under the supervision of Yurii Sergeevich Osipov. In 1971 Osipov completed the development of the foundations of positional game theory for control systems with delayed argument and suggested Kryazhimskiy to study an pursuit–evasion differential game for a target set given in an infinite-dimensional phase space of a delay system. That was the time when convex analysis in Hilbert spaces, a division of functional analysis, was actively developed. Kryazhimskiy used technique from this research area to design solution methods for the described problem. He carried out a comprehensive study of an pursuit–evasion game with a functional target. Based on the results of these studies, Kryazhimskiy defended his candidate's dissertation "Some Game Problems of Pursuit–Evasion" in 1974.

In July 1972, the Laboratory (later, Department) of Differential Equations, headed by Osipov, was created at the Institute of Mathematics and Mechanics of the Ural Scientific Center of the Russian Academy of Sciences (in Sverdlovsk). Kryazhimskiy worked at this department from its creation till the beginning of the 1990s. In the 1970s, after defending his candidate's dissertation, he abandoned his work on differential games for delayed systems and turned to studying differential games for "ordinary" systems with incomplete information as well as infinite-dimensional control systems. Numerous workshops were held on these topics at the Laboratory of Differential Equations. One of difficult important problems was to extend the basic principles of the theory of positional differential games to "ordinary" systems whose right-hand sides did not satisfy the

Lipschitz condition in the phase variable. Working on this problem, Kryazhimskiy designed a "universal" implementation of the extremal shift principle, which was independent of the specifics of a control system's phase space. The implementation was based on an interesting idea: if a control system is looked upon as a "control–trajectory" transformation, then the extremal shift rule can be specified in the space of "inputs," i.e., controls, rather than in the space of "outputs," i.e., trajectories. This idea showed a way that later led Kryazhimskiy to the following solution of a differential game for non-Lipschitz "ordinary" systems: passage to the infinite-dimensional functional control space, search for an adequate criterion for the deviation between the "true" and "target" controls, and implementation of extremal shift in terms of this criterion. These results served as a base for Kryazhimskiy's doctoral dissertation "Differential Games for Non-Lipschitz Systems" (1981).

By the beginning of the 1980s, the development of fundamental issues of the theory of positional differential games related to finding general solvability criteria of game problems and describing the general structure of their solutions has been mostly completed. Scientists at the Institute of Mathematics and Mechanics who worked in the area of differential games had to decide on the directions on further studies. One of the possible directions was the development of the theory of positional differential games "into the depth" by designing new solution methods for differential game problems; another, the search for new problems and the development of new theoretical approaches. Kryazhimskiy and his colleagues from the department chose the new research area. They intended to search for topical problems at the interface of subject areas. At that time, along with the studies on control theory and differential games, other research directions were developed successfully at the institute; one of them was concerned with the theory of ill-posed problems. Despite the remoteness of this theory from differential games, specialists within game theory were familiar with the idea of regularization, which played an important role in the theory of ill-posed problems. In particular, the widely known method of positional control with a guide proposed by Nikolai Nikolaevich Krasovskii in the first half of the 1970s was based on the effect of regularization, i.e., elimination of instability in the presence of small information noise. Kryazhimskiy and Osipov set a goal to find an application direction for methods of the theory of differential games in the area of ill-posed problems. Finding a specific direction of applications was an extremely difficult task, since it required problems of a principally new class. In the theory of ill-posed problems, the so-called inverse problems of control systems are closest to objects studied in control theory. A typical inverse problem consists in finding a control implementing a specified trajectory of a system or a given signal from a trajectory. Similar problems in the presence of trajectory perturbations are close to the process of observing a real trajectory of a system generated by an unknown control, which in this case loses the traditional meaning of "control," i.e., rational influence aimed at the optimization of motion, since the control is replaced by an unobservable and uncontrolled "input" fed to the system from the environment. According to the ideology of the theory of ill-posed problems, the unobservable input is to be recovered, and the recovery error must be arbitrarily small for sufficiently small observation error. Since a direct observation of perturbing inputs, as a rule, was not possible, a new problem of dynamic inversion arose, which consisted in the real-time recovery of current values of unobservable inputs from an available signal about the trajectory. Later, the problem of dynamic inversion became an "inversion block" in the general "inversion–control" scheme, in which, in the process of operation of a control system exposed to the action of unobservable inputs, current values of inputs are recovered approximately in real time from current, generally speaking, inaccurate observations of states of the system (the "inversion block"); these values, together with the results of direct observations of the system's states are fed to an automatic regulator, which produces current values of the control parameter (the "control block").

In the problem of dynamic inversion, an important requirement on the solution algorithm is its dynamic property, i.e., the real–time mode of operation. With reference to the theory of ill-posed problems, this requirement restricts the class of admissible regularizing algorithms, and the problem

of dynamic inversion can be referred to as a dynamic regularization problem. Kryazhimskiy and Osipov proposed a new methodological approach to dynamic regularization, which became known as the principle of regularized extremal shift. It is based on the procedure of control with a guide from the theory of positional differential games and consists in the following. The process of dynamic recovery of an unobservable input is interpreted as the process of control of an auxiliary dynamic system (model). The model, which is often a copy of the original system, is essentially different from the latter in that it is controllable: the uncontrolled input is replaced by a control parameter. Current values of the model control are formed in real time by means of the feedback principle, as a reaction to the "real" (inaccurate) information on the current state of the original system and to the accurate information on the current state of the model. The feedback in the control loop of the model is chosen so that the implementation of the model control as a function of time track accurately enough the implementation of the input of the original system.

The described scheme was developed initially for "ordinary" finite-dimensional systems affine in the input variable. For such systems, the choice of a model feedback guaranteeing the proximity of trajectories of the model and of the system was not difficult: it was sufficient to use the standard rule of extremal shift of the model's current state toward the current signal about the system's state. The crucial step, which shaped the further development of the approach, was the understanding that an appropriate regularization of the extremal shift rule provides the required much stronger property—the proximity of the model control to the input of the system in the mean-square metric. The proposed regularization involves the combination of the basic criterion—the extremal shift—with an auxiliary criterion—the minimum criterion for the norm of the current value of the control parameter. The basic version of the method includes the auxiliary criterion by adding to the main, linear, shift criterion a quadratic smoothing function multiplied by a small regularization parameter. The regularized extremal shift, which consists in the minimization of the resulting criterion in the control variable, corresponds exactly to the application of Tikhonov's regularization method to the extremal shift method. Thus, at the interface of the theory of ill-posed problems and the theory of positional control, a new range of problems was found—dynamic regularization problems—and an approach to their solution was proposed—the method of regularized extremal shift.

The studies on the dynamic inversion of "ordinary" finite-dimensional systems carried out by Kryazhimskiy mainly in the 1980s were summarized in the famous monograph, which presents to the reader a deep theory covering a wide range of issues, from the formulation of dynamic inversion problems, investigation of their solvability, and comparison of the possibilities of dynamic and a posteriori methods to the construction of optimal algorithms and detailed implementation of the "inversion–control" scheme, which played an important motivating role at the initial stage of research. The theory is based on the methods of regularized extremal shift, which combine, as mentioned above, approaches from the theory of positional differential games and the theory of ill-posed problems. Certain divisions of the theory involve methods of the theory of differential equations, control theory (in particular, the techniques of generalized controls), estimation theory, functional analysis, convex analysis, and function approximation theory. Explicit descriptions of algorithms of inversion and inversion–control, which are ready for immediate application and are accompanied by accuracy estimates, are combined with the study of delicate theoretical issues, such as regularizability, order optimality, asymptotic optimality, etc.

In the process of creating the theory of dynamic inversion, its authors developed a new approach to the investigation of some divisions of the theory of solution of operator equations, function approximation theory, etc. One of the strongest developments concerned the application of the dynamic inversion ideology to problems of the classical infinite-dimensional optimization. Studies in this direction began in the second half of the 1980s, when a new iterative algorithm for solving a linear–convex problem of optimal control under phase constraints was proposed. The algorithm was based on the principle of regularized extremal shift applied to an artificially designed dynamic

model with discrete time. An original optimal control problem is interpreted as an optimization problem in the space of artificially created processes under constraints of the type of equality (provided by the equation of the system) and inclusion (provided by the original phase constraints and the original constraints on the control). A special Lyapunov functional related to the Tikhonov regularization method was introduced and stabilized by means of extremal shift; as a result, the model's states converge to the required solution. In further studies, similar algorithmic schemes, based on regularization ideas, were refined and extended to various classes of extremal problems. This series of papers was mainly concerned with the advance into the area of methods of global nonconvex optimization. One of the papers was devoted to studying a wide class of nonconvex optimization problems with constraints; for these problems, the regularized extremal shift principle produces a converging iterative solution algorithm. Problems of this class are characterized by a geometric condition on the separability of the graph of the "perturbed optimal value" function.

In the middle of the 1980s, Kryazhimskiy started investigations related to defense projects. Until the collapse of the Soviet Union in 1991, the researchers of the Department of Differential Equations of the Institute of Mathematics and Mechanics who worked in the sector headed by Kryazhimskiy took part in joint studies with their colleagues from NPO Energiya (Korolev, Moscow region) and NPO Avtomatika (Sverdlovsk). These studies were devoted to processes of interaction of dynamic systems under incomplete and varying information.

In the beginning of the 1990s, Kryazhimskiy moved to Laxenburg, Austria, where he started to work at the International Institute of Applied Systems Analysis. Until the end of 2012, he headed the Dynamic Systems Program (later integrated into the Advanced Systems Analysis Program). The systemic, comprehensive, approach to the solution of difficult interdisciplinary problems, which is a sort of trademark of the institute, was natural to Kryazhimskiy to the full extent. His role in the investigation of certain large-scale problems, such as various applied game problems, economic growth modeling, finding optimal ways of sustainable development on the global scale, modeling of innovation market dynamics, optimal gas transportation, etc., cannot be overestimated.

Since 1996, Kryazhimskiy had worked at the Steklov Institute of Mathematics of the RAS: first, as a leading researcher and, from 1997, as a chief researcher. Simultaneously, he had been a lecturer at the Department of Optimal Control of the Faculty of Computational Mathematics and Cybernetics at Moscow State University. His lectures were very popular with the students because of their rich content, informative value, and clarity of presentation. The lecturer's personal charm was of no less importance.

The wide scientific scope and industriousness at the highest intellectual level were Kryazhimskiy's important traits. He was successful at solving problems from the most diverse divisions of mathematics and borderline disciplines. His main motives in choosing new problems were the synthesis of disciplines and rich practical content.

A result of Kryazhimskiy's fruitful work and acknowledgement of his remarkable contribution to the development of Russian science was his election to the Academy of Sciences. He has been a corresponding member of the academy since May 1997 and a full member since May 2006.

A talented person is talented in everything. This popular saying is Kryazhimskiy's best characterization. He had been attracted to music and literature since his childhood, being a brilliant guitar player and an author of poetry and songs. Arkady Viktorovich was a responsive and warm-hearted person. He was open to people and did not dominate because of his position and well-deserved authority. His relatives and colleagues noticed his tact, enthusiasm, polymathy, and amazing willingness to both share his ideas and appreciate and discuss ideas of other people.

All who knew him were shocked to receive the news of Arkady Viktorovich's untimely death on 3rd of November 2014.

The International Conference in memory of Arkady Viktorovich Kryazhimskiy was organized by the Krasovskii Institute of Mathematics and Mechanics and the Ural Federal University. The

conference named "System analysis: modeling and control" was held in Ekaterinburg in October 2016 to get together the colleagues of Kryazhimskiy from different countries for discussing actual scientific problems. The event was a success: more than fifty participants, more than thirty plenary reports. Several selected papers presented at the conference are published in this issue of the journal.

KRYAZHIMSKIY'S MAIN SCIENTIFIC PAPERS

1. *A.V. Kryazhimskii and Yu.S. Osipov*, Differential–difference game of encounter with a functional target set, J. Appl. Math. Mech. 37 (1), pp. 1–10 (1973).

2. *A.V. Kryazhimskii*, A differential-difference game of evasion from a functional target, Izv. Akad. Nauk SSSR, Ser. Tekhn. Kibernet., No. 4, pp. 71–79 (1973).

3. *A.V. Kryazhimskii*, Some Game Problems of Pursuit–Evasion, Candidates Dissertation in Physics and Mathematics (Sverdlovsk, 1974).

4. *A.V. Kryazhimskii*, Differential games of pursuit in conditions of imperfect information about the system, Ukr. Math. J. 27 (4), pp. 425–429 (1975).

5. *A.V. Kryazhimskii*, On the admissibility of an optimal strategy, in Differential Games and Control Problems: Collection of Papers (UNTs AN SSSR, Sverdlovsk, 1975), Issue 15, pp. 125–130 [in Russian].

6. *A.V. Kryazhimskii*, An alternative in a linear pursuit-evasion game with incomplete information, Dokl. Akad. Nauk SSSR 230 (4), pp. 773–776 (1976).

7. *A.V. Kryazhimskii and S.D. Filippov*, On a game problem on the convergence of two points on a plane under incomplete information, in Control Problems with Incomplete Information (IMM UNTs AN SSSR, Sverdlovsk, 1976), Issue 19, pp. 62–77 [in Russian].

8. *A.V. Kryazhimskii*, On the problem of the deviation of a linear system with aftereffect from a functional target, in Game Problems of Control: Collection of Papers (IMM UNTs AN SSSR, Sverdlovsk, 1977), Issue 24, pp. 46–52 [in Russian].

9. *A.V. Kryazhimskii*, On the theory of positional differential games of pusruit-evasion. Dokl. Akad. Nauk SSSR 239 (4), pp. 779–782 (1978).

10. *A.V. Kryazhimskii*, On stochastic approximation in differential games, Sov. Math. Dokl. 19, pp. 955–959 (1978).

11. *A.V. Kryazhimskii and V.I. Maksimov*, Approximation in linear differencedifferential games, J. Appl. Math. Mech. 42 (2), pp. 212–219 (1978).

12. *Yu.S. Osipov, A.V. Kryazhimskii, and S. P. Okhezin*, Control problems in systems with distributed parameters, in Dynamics of Control Systems: Proceedings of the Third All-Union Chetaev Conference, Irkutsk, Russia, 1977 (Nauka, Novosibirsk, 1979), pp. 199–208 [in Russian].

13. *A.V. Kryazhimskii*, Differential Games for Non-Lipschitz Systems, Doctoral Dissertation in Physics and Mathematics (Sverdlovsk, 1980).

14. *A.V. Kryazhimskii*, On some stable bridges for linear controlled systems, in Optimal Control of Systems with Uncertain Information: Collection of Papers (UNTs AN SSSR, Sverdlovsk, 1980), pp. 35–41 [in Russian].

15. *A.V. Kryazhimskii*, On stable position control in differential games, J. Appl. Math. Mech. 42 (6), pp. 1055–1060 (1980).

16. *A.V. Kryazhimskii*, Deviation of a linear system with aftereffect from a functional target, J. Dynamic Systems Measurement Control 103 (2), pp. 43–48 (1981).

17. *A.V. Kryazhimskii*, Game evasion problem for a partially continuous system, in Control and Estimation in Dynamical Systems: Collection of Papers (UNTs AN SSSR, Sverdlovsk, 1982), pp. 25–41 [in Russian].

18. *Yu.S. Osipov and A.V. Kryazhimskii*, On the dynamic solution of operator equations, Sov. Math. Dokl. 27, pp. 382–386 (1983).

19. *A.V. Kryazhimskii and Yu.S. Osipov*, Modelling of a control in a dynamic system, Engrg. Cybernetics 21 (2), pp. 38–47 (1984).

20. *A.V. Kriazhimskii, V.I. Maksimov, and Iu.S. Osipov*, On positional simulation in dynamic systems, J. Appl. Math. Mech. 47 (6), pp. 709–714 (1985).

21. *O. Abdyrakhmanov and A.V. Kryazhimskii*, On the question of the well–posedness of an optimal control problem, Differents. Uravneniya 20 (10), pp. 1659–1665 (1984).

22. *O. Abdyrakhmanov and A.V. Kryazhimskii*, On regularization of an optimal control problem for a system with nonuniqueness, Izv. Akad. Nauk TSSR, Ser. Fiz.–Tekh. Khim. Geol. Nauk, No. 4, pp. 3–6 (1984).

23. *O. Abdyrakhmanov and A.V. Kryazhimskii*, On regularization of an optimal control problem for a system with nonuniqueness. II, Izv. Akad. Nauk TSSR, Ser. Fiz.–Tekh. Khim. Geol. Nauk, No. 6, pp. 7–11 (1984).

24. *A.V. Kryazhimskii and Yu.S. Osipov*, On positional calculation of Ω–normal controls in dynamical system, Probl. Control Inform. Theory 13 (6), pp. 425–436 (1984).

25. *A.V. Kim and A.V. Kryazhimskii*, Dynamics of the running point of a trajectory with variable initial condition, in Problems of Control and Modeling in Dynamical Systems: Collection of Papers (UNTs AN SSSR, Sverdlovsk, 1984), pp. 19–27 [in Russian].

26. *A.V. Kryazhimskii and Yu.S. Osipov*, Modeling of parameters of a dynamic system, in Problems of Control and Modeling in Dynamical Systems: Collection of Papers (UNTs AN SSSR, Sverdlovsk, 1984), pp. 47–68 [in Russian].

27. *A.V. Kryazhimskii and Yu.S. Osipov*, Best approximation of the differentiation operator in the class of nonanticipatory operators, Math. Notes 37 (2), pp. 109–114 (1985).

28. *Yu.S. Osipov and A.V. Kryazhimskii*, The Lyapunov function method in a motion simulation problem, in Motion Stability: Collection of Papers (Nauka, Novosibirsk, 1985), pp. 53–56 [in Russian].

29. *M.S. Gabrielyan and A.V. Kryazhimskii*, The convergence–evasion differential game with m goal sets, Sov. Math. Dokl. 33, pp. 691–694 (1986).

30. *A.V. Kryazhimskii and K.E. Lovtskii*, Weak continuity of motions with respect to control for controllable differential inclusions and systems with discontinuous right–hand side, Differents. Uravneniya 22 (11), pp. 1895–1905 (1986).

31. *Yu.S. Osipov and A.V. Kryazhimskii*, Positional modeling of a stochastic control in dynamical systems, in Stochastic Optimization: Proceedings of the International Conference, Kiev, Ukraine, 1984 (Springer, Berlin, 1986), Ser. Lecture Notes in Control and Information Sciences 81, pp. 696–704.

32. *A.V. Kryazhimskii*, On the positional regularizing algorithms for control dynamical systems, in Differential Equations and Applications (Angel Kanchev Univ., Ruse, 1987), pp. 767–770.

33. *A.V. Kryazhimskii*, Optimization of the ensured result for the dynamical systems, in Proceedings of the International Congress of Mathematicians, Berkeley, CA, USA, 1986 (Amer. Math. Soc., Providence, RI, 1987), Vol. 2, pp. 1171–1179.

34. *A.V. Kryazhimskii and Yu.S. Osipov*, On the regularization of a convex extremal problem with inexactly given constraints. Application to an optimal control problem with phase constraints, in Some Methods of Positional and Program Control: Collection of Research Papers (UNTs AN SSSR, Sverdlovsk, 1987), pp. 34–54 [in Russian].

35. *A.V. Kryazhimskii and Yu.S. Osipov*, Inverse problems of dynamics and controllable models, in Mechanics and Scientific – Technological Progress, Vol. 1: General and Applied Mechanics (Nauka, Moscow, 1987), pp. 196–211 [in Russian].

36. *A.V. Kryazhimskii and Yu.S. Osipov*, Stable solutions of inverse problems of the dynamics of controlled systems, Proc. Steklov Inst. Math. 185, pp. 143–164 (1988).

37. *A.V. Kryazhimskii and Yu.S. Osipov*, On the methods of positional modeling of control in dynamical systems, in Qualitative Questions in the Theory of Differential Equations and Control Systems: Collection of Research Papers (UNTs AN SSSR, Sverdlovsk, 1988), pp. 34–44 [in Russian].

38. *A.V. Kryazhimskii and Yu.S. Osipov*, On a stable positional recovery of control from measurements of a part of coordinates, in Some Problems of Control and Stability: Collection of Papers (UNTs AN SSSR, Sverdlovsk, 1989), pp. 33–47 [in Russian].

39. *A.V. Kryazhimskii*, On the continuity of Lebesgue sets in an optimal control problem, in Optimization and Stability Problems in Control Systems: Collection of Research Papers (UNTs AN SSSR, Sverdlovsk, 1990), pp. 54–73 [in Russian].

40. *A.Yu. Vdovin and A.V. Kryazhimskii*, On the recovery the perturbation set from measurements of the trajectory, in Studies in System Analysis and Applications: Collection of Research Papers (Izd. Ural. Gos. Univ., Sverdlovsk, 1990), pp. 15–35 [in Russian].

41. *A.V. Kryazhimskii*, The problem of optimization of the ensured result: Unimprovability of full-memory strategies, in Constantin Carath'eodory: An International Tribute (World Sci., Teaneck, NJ, 1991), Vol. 1, pp. 636–675.

42. *A.Yu. Vdovin and A.V. Kryazhimskii*, On a lower estimate for a positional regularization method for a disturbance recovery problem, in Modeling and Optimization Problems: Collection of Research Papers (UNTs AN SSSR Sverdlovsk, 1991), pp. 3–13 [in Russian].

43. *A.V. Kryazhimskii*, Dynamical regularizibility of inverse problems for control systems, in System Modelling and Optimization: Proceedings of the 15th IFIP Conference, Zurich, Switzerland, 1991 (Springer, Berlin, 1992), Ser. Lecture Notes in Control and Information Sciences 180, pp. 384–393.

44. *A.V. Kryazhimskii*, Order optimal real-time observers for completely observable control systems, Appl. Math. Comput. Sci. 2 (1), pp. 149–154 (1992).

45. *A.V. Kryazhimskiy and V.B. Savinov*, The traveling-salesman problem with moving objects, J. Comput. Syst. Sci. Int. 33 (3), pp. 144–148 (1993).

46. *A.V. Kryazhimskii*, Conditions for stable nonanticipatory motion approximation, Proc. Steklov Inst. Math. 211, pp. 221–233 (1995).

47. *A.V. Kryazhimskii and Yu.S. Osipov*, On differential-evolutionary games, Proc. Steklov Inst. Math. 211, pp. 234–261 (1995).

48. *A.V. Kryazhimskii and V.B. Savinov*, On a model of conflict interaction with aftereffect in controls, in RoutingDistributional Problems: Collection of Papers (UGTU, Ekaterinburg, 1995), pp. 44–53 [in Russian].

49. *A.V. Kryazhimskii, V.I. Maksimov, and E. A. Samarskaia*, On Estimation of Forcing Functions in Parabolic Systems: IIASA Working Paper WP–95–75 (IIASA, Laxenburg, 1995).

50. *G. Hutschenreiter, Yu.M. Kaniovski, and A.V. Kryazhimskii*, Endogenous Growth, Absorptive Capacities and International R&D Spillovers: IIASA Working Paper WP–95–92 (IIASA, Laxenburg, 1995).

51. *A. Kryazhimskii*, An Endogenous Growth Model for Technological LeadingFollowing: An Asymptotical Analysis: IIASA Working Paper WP–95–93 (IIASA, Laxenburg, 1995).

52. *Yu.S. Osipov and A.V. Kryazhimskii*, Inverse Problems for Ordinary Differential Equations: Dynamical Solutions (Gordon and Breach, London, 1995).

53. *Yu.S. Osipov, A.V. Kryazhimskii, and V.I. Maksimov*, Dynamical inverse problems for systems with distributed parameters, J. Inverse Ill–Posed Probl. 4 (4), pp. 267–282 (1996).

54. *V.I. Heymann and A.V. Kryazhimskii*, On finite–dimensional parametrizations of attainability sets, Appl. Math. Comput. 78 (2–3), pp. 137–151 (1996).

55. *A.V. Kryazhimskii, V.I. Maksimov, A.A. Solov'ev, and A.G. Chentsov*, On a probabilistic approach to the quantitative description of the dynamics of natural processes, Problem. Control Inform., Nos. 1–2, pp. 192–210 (1996).

56. *A.V. Kryazhimskii, V.I. Maksimov, and Yu.S. Osipov*, Reconstruction of Boundary Sources through Sensor Observations: IIASA Working Paper WP–96–97 (IIASA, Laxenburg, 1996).

57. *A.V. Kryazhimskii, V.I. Maksimov, and E.A. Samarskaya*, On reconstruction of inputs in parabolic systems, Mat. Model. 9 (3), pp. 51–72 (1997).

58. *Yu.M. Ermoliev, A.V. Kryazhimskii, and A. Ruszczynski*, Constraint aggregation principle in convex optimization, Math. Programm., Ser. B, 76 (3), pp. 353–372 (1997).

59. *A.V. Kryazhimskii, V.I. Maksimov, and Yu.S. Osipov*, Reconstruction of extremal perturbations in parabolic equations, Comp. Math. Math. Phys. 37 (3), pp. 288–298.

60. *A.V. Kryazhimskii,* Convex optimization via feedbacks, SIAM J. Contr. and Optimiz. 37 (1), 278302 (1998).

61. *A.V. Kryazhimskii and A.M. Tarasyev*, Equilibrium and Guaranteeing Solutions in Evolutionary Nonzero Sum Games, IIASA Interim Report IR–98–003 (IIASA, Laxenburg, 1998).

62. *A. Kryazhimskii, A. Nentjes, S. Shibayev, and A. Tarasyev*, Searching Market Equilibria under Uncertain Utilities, IIASA Interim Report IR–98–007 (IIASA, Laxenburg, 1998).

63. *A.V. Kryazhimskii and V.I. Maksimov*, An iterative procedure for solving a control problem with phase constraints, Comp. Math. Math. Phys. 38 (9), pp. 1423–1428 (1998).

64. *B.V. Digas, Yu.M. Ermoliev, and A.V. Kryazhimskii*, Guaranteed Optimization in Insurance of Catastrophic Risks, IIASA Interim Report IR–98–082 (IIASA, Laxenburg, 1998).

65. *A.F. Kleimenov and A.V. Kryazhimskii*, Normal Behavior, Altruism and Aggression in Cooperative Game Dynamics, IIASA Interim Report IR–98–076 (IIASA, Laxenburg, 1998).

66. *A.V. Kryazhimskii and Yu.S. Osipov*, Approximate linear reduction in guidance and evasion differential game, Proc. Steklov Inst. Math. 220, pp. 170–191 (1998).

67. *A.V. Kryazhimskii and Yu.S. Osipov*, On two-dimensional reduction in a differential game of quality, Proc. Steklov Inst. Math. 224, pp. 198–211 (1999).

68. *A.V. Kryazhimskii, A. Nentjes, S.V. Shibayev, and A.M. Tarasyev*, A game model of negotiations and market equilibria, J. Math. Sci. 100 (6), pp. 2601–2612 (2000).

69. *V.F. Borisov, G. Hutschenreiter, and A.V. Kryazhimskii*, Asymptotic growth rates in knowledge-exchanging economies, Ann. Oper. Res. 89, pp. 61–73 (1999).

70. *F. Kappel, A. Kryazhimskii, and V. Maksimov*, Constraint aggregation principle in the problem of optimal control of distributed parameter systems, in Nonsmooth and Discontinuous Problems of Control and Optimization: Proceedings of the International IFAC Workshop, Chelyabinsk, Russia, 1998 (IIASA, Laxenburg, 1999), pp. 137–141.

71. *A.V. Kryazhimskii and G. Sonnevend*, Dynamics for bimatrix games via analytic centers, in Dynamics and Control (Gordon and Breach, London, 1999), Ser. Stability and Control: Theory, Methods, and Applications, Vol. 9, pp. 129–138.

72. *Yu.M. Kaniovski, A.V. Kryazhimskii, and H. P. Young*, Adaptive dynamics in games played by heterogeneous populations, Games Econom. Behav. 31 (1), pp. 50–96 (2000).

73. *F. Kappel, A.V. Kryazhimskii, and V.I. Maksimov*, Dynamic reconstruction of states and guaranteeing control of a reactiondiffusion system, Dokl. Math. 61 (1), pp. 143–145 (2000).

74. *V. Borisov, G. Feichtinger, and A. Kryazhimskii*, Optimal enforcement on a pure sellers market of illicit drugs, J. Optim. Theory Appl. 106 (1), pp. 1–22 (2000).

75. *Yu.S. Osipov, A.V. Kryazhimskii, and V.I. Maksimov*, Dynamic inverse problems for parabolic systems, Differential Equations 36 (5), pp. 643–661 (2000).

76. *A.V. Kryazhimskii and Yu.S. Osipov*, On an algorithmic criterion of the solvability of game problems for linear controlled systems, Proc. Steklov Inst. Math., Suppl. 1, pp. S154–S162 (2000).

77. *A.V. Kryazhimskii and B.V. Digas*, Insurance optimization for catastrophic risks: a guarantee approach, in Information Technologies in Economics: Theory, Models, and Methods (Izd. Ural. Gos. Ekon. Univ., Yekaterinburg, 2000), pp. 98–105 [in Russian].

78. *S.M. Aseev, A.V. Kryazhimskii, and A.M. Tarasyev*, The Pontryagin maximum principle and transversality conditions for an optimal control problem with infinite time interval, Proc. Steklov Inst. Math. 233, pp. 64–80 (2001).

79. *A.V. Kryazhimskii and S.V. Paschenko*, On the problem of optimal compatibility, J. Inverse Ill-Posed Probl. 9 (3), pp. 283300 (2001).

80. *A.V. Kryazhimskii and A. Ruszczynski*, Constraint aggregation in infinite-dimensional spaces and applications, Math. Oper. Res. 26 (4), pp. 769–795 (2001).

81. *A.V. Kryazhimskii*, Optimization problems with convex epigraphs. Application to optimal control, Intern. J. Appl. Math. Comput. Sci. 11 (4), pp. 773–801 (2001).

82. *A. Kryazhimskii, A. Nentjes, S. Shibayev, and A. Tarasyev*, Modeling market equilibrium for transboundary environmental problem, Nonlinear Anal. 47 (2), pp. 991–1002 (2001).

83. *A.V. Kryazhimskii*, Optimization problems with convex epigraphs. Application to optimal control, Intern. J. Appl. Math. Comput. Sci. 11 (4), pp. 773–801 (2001).

84. *A.V. Kryazhimskii and S.V. Pashchenko*, On the solution of the linear time-optimal control problem with mixed constraints, J. Math. Sci. 114 (3), pp. 1345–1362 (2003).

85. *A. Kryazhimskii, C. Watanabe, and Y. Tou*, Dynamic model of market of patents and equilibria in technology stocks, Comput. Math. Appl. 44 (7), pp. 979–995 (2002).

86. *A.V. Kryazhimskii and Yu.S. Osipov*, Extremum problems with separable graphs, Cybern. Syst. Anal. 38 (2), pp. 175–194 (2002).

87. *G. Klaassen, A. Kryazhimskii, O. Nikonov and Ya. Minullin*, On a game of gas pipeline projects competition, in Game Theory and Applications: Proceedings of the ICM 2002 Satellite Conference, Qingdao, China, 2002 (Qingdao Publ., Qingdao, 2002), pp. 327–334.

88. *S.M. Aseev, A.V. Kryazhimskii and G. Hutschenreiter*, A dynamic model of optimal investment in research and development, in Modern Mathematics and Applications (Inst. Kibernet. AN Gruzii, Tbilisi, 2005), Vol. 9, pp. 3–43 [in Russian].

89. *A.V. Kryazhimskiy and C. Watanabe*, Optimization of Technological Growth (Gendaitosho, Kanagawa, 2004).

90. *A.V. Kryazhimskii and R.A. Usachev*, On a convex two–level optimization problem, in Nonlinear Dynamics and Control: Collection of Papers (Fizmatlit, Moscow, 2004), Issue 4, pp. 257–286 [in Russian].

91. *A.V. Kryazhimskiy and V.I. Maksimov*, A solution algorithm for problems of optimal control in Hilbert spaces, J. Math. Sci. 121 (2), pp. 2226–2247 (2004).

92. *A. Kryazhimskiy and V. Maksimov*, On exact stabilization of an uncertain dynamical system, J. Inverse Ill–Posed Probl. 12 (2), pp. 145–182 (2004).

93. *A.V. Kryazhimskii and Yu.S. Osipov*, The method of extremal shift and optimization problems, Proc. Steklov Inst. Math., Suppl. 2, pp. S91–S114 (2004).

94. *G. Klaassen, A.V. Kryazhimskii and A. M. Tarasyev*, Multiequilibrium game of timing and competition of gas pipeline projects, J. Optim. Theory Appl. 120 (1), pp. 147–179 (2004).

95. *S.M. Aseev and V. Kryazhimskii*, The Pontryagin maximum principle for an optimal control problem with a functional specified by an improper integral, Dokl. Math. 69 (1), pp. 89–91 (2004).

96. *A. Kryazhimskiy and V. Maksimov*, Parallelization in an algorithm of multi–dimensional nonconvex optimization: an application to insurance network design, in Parallel Processing and Applied Mathematics (Springer, Berlin, 2004), Ser. Lecture Notes in Computer Science, Vol. 3019, pp. 754–761.

97. *S.M. Aseev and V. Kryazhimskiy*, The Pontryagin maximum principle and transversality conditions for a class of optimal control problems with infinite time horizons, SIAM J. Control Optim. 43 (3), pp. 1094–1119 (2004).

98. *A. Kryazhimskiy, Ya. Minullin and L. Schrattenholzer*, Global long–term energy–economy–environment scenarios with an emphasis on Russia, Perspectives in Energy 9, pp. 119–137 (2005).

99. *S.A. Brykalov, O.N. Golovina and A.V. Kryazhimskii*, Nash equilibrium in multi–player games with the choice of time instants and integral cost functionals, J. Math. Sci. 140 (6), pp. 796–807 (2007).

100. *A. Kryazhimskii and V. Maksimov*, On identification of nonobservable contamination inputs, Environmental Modelling and Software 20 (8), pp. 1057–1061 (2005).

101. *S.M. Aseev, G. Hutschenreiter and A. V. Kryazhimskii*, A dynamical model of optimal investment in R&D, J. Math. Sci. 126 (6), pp. 1495–1535 (2005).

102. *A. Kryazhimskiy*, A system-robust stabilization technique with application to an uncertain model of global carbon cycle, in Modeling and Control of Autonomous Decision Support Based Systems: Proceedings of the 13th International Workshop on Dynamics and Control, Wiesensteig, Germany, 2005, Ed. by E. Hofer and E. Reithmeier (Shaker, Aachen, 2005), pp. 149–156.

103. *S. Aseev, G. Hutschenreiter, A. Kryazhimskiy and A. Lysenko*, A dynamic model of optimal investment in research and development with international knowledge spillovers, Math. Comput. Model. Dyn. Syst. 11 (2), pp. 125–133 (2005).

104. *A. V. Kryazhimskii and V. I. Maksimov*, A solution algorithm for problems of optimal control in Hilberts space, J. Math. Sci. 121 (2), pp. 2226–2247 (2004).

105. *Yu.S. Osipov, A.V. Kryazhimskii and E.A. Rovenskaya*, An optimal compatibility parameter problem: Constructive regularization method, Izv. Ural. Gos. Univ., No. 10, pp. 128–166 (2006).

106. *A.V. Kryazhimskii, O. Nikonov and Ya. Minullin*, Game of timing in gas pipeline projects competition: Simulation software and generalized equilibrium solutions, in Advances in Dynamic Games: Applications to Economics, Management Science, Engineering, and Environmental Management (Birkhäuser, Boston, 2006), Ser. Annals of the International Society of Dynamic Games, Vol. 8, pp. 237–252.

107. *Yu.S. Osipov and A.V. Kryazhimskii*, Problems of dynamic inversion, Herald Russ. Acad. Sci. 76 (4), pp. 352–360 (2006).

108. *A.V. Kryazhimskiy and V.I. Maksimov*, Dynamical state reconstruction and guaranteeing control for a system of parabolic equations, Proc. Steklov Inst. Math. 253 (Suppl. 1), pp. S168–S184 (2006).

109. *A.V. Kryazhimskiy*, Yu.S. Osipov's work in mathematical control theory, Russ. Math. Surv. 61 (4), pp. 593–610 (2006).

110. *S.M. Aseev and A.V. Kryazhimskii*, The Pontryagin maximum principle and optimal economic growth problems, Proc. Steklov Inst. Math. 257, pp. 1–255 (2007).

111. *S.M. Aseev and A.V. Kryazhimskii*, On a class of optimal control problems arising in mathematical economics, Proc. Steklov Inst. Math. 262, pp. 10–25 (2008).

112. *A.V. Kryazhimskii, V.I. Maksimov, E.A. Rovenskaya and M.V. Rodkin*, On a regime of repetition of rare strong events (catastrophes): New approaches and results of their application, in Change in the Environment and Climate: Natural Catastrophes and Induced Technogenic Catastrophes (Inst. Geogr. RAN, Moscow, 2008), Vol. 3, pp. 158–189 [in Russian].

113. *B.D. Fath, A.V. Kryazhimskiy, H. Liljenstroem and E. Rovenskaya*, Introduction: Towards the design of an integrated socio-environmental assessment model for the Baltic Sea region, in Evolutionary and Deterministic Methods for Design, Optimization and Control: Applications to Industrial and Societal Problems (Int. Center Numer. Meth. Eng., Barcelona, 2008), pp. 425–429.

114. *A.V. Kryazhimskii and R.A. Usachev*, Convex two–level optimization problem, Comput. Math. Model. 19 (1), pp. 73–101 (2008).

115. *A.V. Kryazhimskiy and V.I. Maksimov*, On rough inversion of a dynamical system with a disturbance, J. Inverse Ill-Posed Probl. 16 (6), pp. 587–600 (2008).

116. *S.M. Aseev and A.V. Kryazhimskiy*, Shadow prices in infinite-horizon optimal control problems with dominating discounts, Appl. Math. Comput. 204 (2), pp. 519–531 (2008).

117. *A. Kryazhimskiy, M. Obersteiner and A. Smirnov*, Infinite-horizon dynamic programming and application to management of economies effected by random natural hazards, Appl. Math. Comput. 204 (2), pp. 609–620 (2008).

118. *Yu.S. Osipov, A.V. Kryazhimskii and V.I. Maksimov*, N.N. Krasovskiis extremal shift method and problems of boundary control, Autom. Remote Control 70 (4), pp. 577–588 (2009).

119. *A. Kryazhimskiy*, On a Boundedly Rational Pareto–Optimal Trade in Emission Reduction, IIASA Interim Report IR–09–017 (IIASA, Laxenburg, 2009).

120. *A.V. Kryazhimskii and Yu. S. Osipov*, Idealized program packages and problems of positional control with incomplete information, Proc. Steklov Inst. Math. 268 (Suppl. 1), pp. S155–S174 (2010).

121. *A.V. Kryazhimskii and Yu.S. Osipov*, On dynamical regularization under random noise, Proc. Steklov Inst. Math. 271, pp. 125–137 (2010).

122. *A.V. Kryazhimskii*, Models of economic growth. Authoritarian planning and up-to-the-minute solutions, in Problems of Dynamical Control (MAKS, Moscow, 2010), Issue 5, pp. 157–165 [in Russian].

123. *A. Kryazhimskiy*, On a decentralized boundedly rational emission reduction strategy, in Dynamic Systems, Economic Growth and the Environment, Ed. by J. Crespo-Cuaresma, T. Palokangas, and A. Tarasyev (Springer, Berlin, 2010), Ser. Dynamic Modeling and Econometrics in Economics and Finance, Vol. 12, pp. 215–235.

124. *A.V. Kryazhimskii, S.P. Konovalov, and M.S. Nikolskii*, A simplified model of tax collection from enterprises in the presence of legal and shadow capital, Comp. Math. Modeling 24 (3), pp. 378–403 (2010).

125. *A. Kryazhimskiy*, Two–step win–stay, lose–shift and learning to cooperate in the repeated prisoners dilemma, Int. Game Theory Rev. 12 (4), pp. 437–451 (2010).

126. *A. Kryazhimskiy and V. Maksimov*, Resource-saving infinite-horizon tracking under uncertain input, Appl. Math. Comput. 217 (3), pp. 1135–1140 (2010).

127. *Yu.S. Osipov, A. V. Kryazhimskii, and V. I. Maksimov*, Dynamic Recovery Methods for Inputs of Control Systems (UrO RAN, Ekaterinburg, 2011) [in Russian].

128. *A.V. Kryazhimskii and V.I. Maksimov*, Extremal control methods and dynamical inversion problems, Vestn. Nizhegorod. Univ. 4, pp. 184–185 (2011).

129. *A.V. Kryazhimskiy and V.I. Maksimov*, Resource-saving tracking problem with infinite time horizon, Differential Equations 47 (7), pp. 1004–1013 (2011).

130. *Yu.S. Osipov, A.V. Kryazhimskii, and V.I. Maksimov*, Some algorithms for the dynamic reconstruction of inputs, Proc. Steklov Inst. Math. 275 (Suppl. 1), pp. S86–S120 (2011).

131. *A.V. Kryazhimskii*, Numerical encoding of sampled controls and an approximation metric criterion for the solvability of a guidance game problem, Proc. Steklov Inst. Math. 276 (Suppl. 1), pp. S106–S125 (2012).

132. *A.V. Kryazhimskiy and Yu.S. Osipov*, On the solvability of problems of guaranteeing control for partially observable linear dynamical systems, Proc. Steklov Inst. Math. 277, pp. 144–159 (2012).

133. *S.M. Aseev, K.O. Besov and A.V. Kryazhimskii*, Infinite-horizon optimal control problems in economics, Russ. Math. Surv. 67 (2), pp. 195–253 (2012).

134. *A.V. Kryazhimskii and A.V. Raigorodskaya*, On uniform behavior strategies in infinite repeating games, in Problems of Dynamical Control (MAKS, Moscow, 2012), Issue 6, pp. 133–159 [in Russian].

135. *A. Frank, M.G. Collins, M. Clegg, U. Dieckmann, V. Kremenyuk, A. Kryazhimskiy, J. Linnerooth-Bayer, S. Levin, A. Lo, B. Ramalingam, J. Ramo, S. Roy, D. Saari, Z. Shtauber, K. Sigmund, J. Tepperman, S. Thurner, W. Yiwei and D. von Winterfeldt*, Security in the Age of Systemic Risk: Strategies, Tactics and Options for Dealing with Femtorisks and Beyond, IIASA Interim Report IR–12–010 (IIASA, Laxenburg, 2012).

136. *A. Kryazhimskiy and A. Puchkova*, Towards Detection of Early Warning Signals on Financial Crises, IIASA Interim Report IR–12–001 (IIASA, Laxenburg, 2012).

137. *A.V. Kryazhimskiy*, Relaxation of optimal control problems and linearquadratic systems, Dynamics of Continuous, Discrete and Impulsive Systems, Ser. B: Appl. and Algorithms 19 (1–2), pp. 17–42 (2012).

138. *A.V. Kryazhimskii and V. I. Maksimov*, On combination of the processes of reconstruction and guaranteeing control, Automat. Remote Control 74 (8), pp. 1235–1248 (2013).

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

AN APPLICATION OF MOTION CORRECTION METHODS TO THE ALIGNMENT PROBLEM IN NAVIGATION[1]

Boris I. Ananyev

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of Russian Academy of Sciences, Ekaterinburg, Russia,
abi@imm.uran.ru

Abstract: In this paper, we apply some motion correction methods to the alignment problem in navigation. This problem consists in matching two coordinate systems having the common origins. As a rule, one of the systems named as basic coordinate system is located at a ship or airplane. The dependent coordinate system belongs to another object (e.g. missile) that starts from the ship. The problem is considered with incomplete information on state coordinates which can be measured with disturbances without statistical description.

Key words: Alignment problem, Motion correction, Incomplete information, Set-membership description of uncertainty.

Introduction

Alignment is the process whereby the orientation of the axes of an inertial navigation system is determined with respect to the reference axis system. The basic concept of aligning an inertial navigation system is quite simple and straightforward. However, there are many complications that make alignment both time consuming and complex. Consider a simulated transport ship-airplane system. Suppose that the base coordinate system (BCS) of the ship is correct. Let $\overrightarrow{\Omega}_1$ be the absolute angular velocity of the BCS in the motionless coordinate system η_1, η_2, η_3. The projection Ω_1^2 on vertical 2_1 equals zero. This system is shown on Fig. 1. The axis 1_1 is directed along the

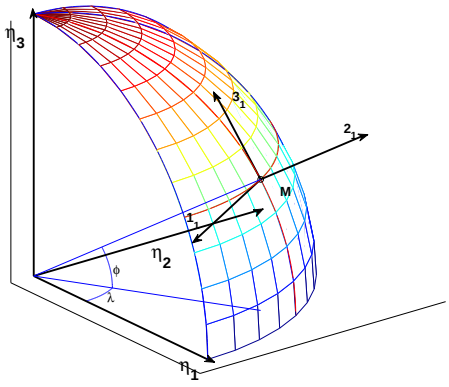


Figure 1. The section of Earth sphere and the base coordinate system.

parallel to the west. The axis 2_1 is the local vertical. The axis 3_1 is directed along the meridian to the north. The position of the dependent coordinate system (DCS) related to the airplane or

[1]The research was supported by Russian Science Foundation (RSF), project No. 16-11-10146.

the missile with respect to the BCS is estimated by the Krylov angles. In Fig. 2, one can see the sequence of clockwise rotations: θ^1 around axis 1, θ^3 around new axis 3, and θ^2 around new axis 2 coinciding now with 2_1.

Thus, the transition of coordinates of a vector $\vec{f}$ in the DCS to new coordinates in the BCS is occurred by the formula $\vec{f_1} = \mathbf{M}(\theta)\vec{f}$, where the matrix of direction cosines is of the form

$$\mathbf{M}(\theta) = \begin{pmatrix} \cos\theta^2 & 0 & -\sin\theta^2 \\ 0 & 1 & 0 \\ \sin\theta^2 & 0 & \cos\theta^2 \end{pmatrix} \cdot \begin{pmatrix} \cos\theta^3 & -\sin\theta^3 & 0 \\ \sin\theta^3 & \cos\theta^3 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta^1 & -\sin\theta^1 \\ 0 & \sin\theta^1 & \cos\theta^1 \end{pmatrix} = (m_{ij}).$$

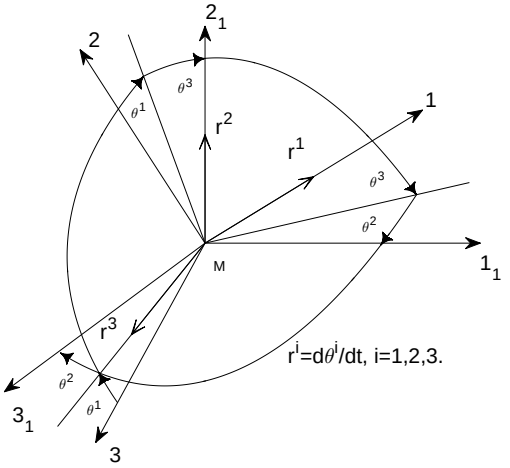


Figure 2. The sequence of clockwise rotations.

Projecting the equality $\vec{\omega} = \dot{\vec{\theta}}^1 + \dot{\vec{\theta}}^3 + \dot{\vec{\theta}}^2$ for the angular velocities on the axes of the DCS, we obtain the kinematic Krylov equations

$$\dot\theta^1 = \omega^1 - \dot\theta^2 \sin\theta^3, \quad \dot\theta^2 = (\omega^2 \cos\theta^1 - \omega^3 \sin\theta^1)/\cos\theta^3, \quad \dot\theta^3 = \omega^2 \sin\theta^1 + \omega^3 \cos\theta^1, \tag{0.1}$$

where ω^i are the projections of the relative angular velocity. These projections are related with the absolute velocities by the formulas

$$\omega^i = \Omega^i - m_{1i}\Omega_1^1 - m_{3i}\Omega_1^3 + \varepsilon^i, \quad i \in 1:3, \tag{0.2}$$

where ε^i are the projections of an uncertain drift.

For measurements, the differences of accelerometer readings in the DCS and BCS are used. These accelerometers are on the axes and gage the nongravity acceleration $\vec{a} = \overrightarrow{w}_M - \vec{g}$. Let a^i be accelerometers readings in the DCS and a_1^i be gage readings in the BCS. Therefore, the measurement equations are of the form

$$\begin{gathered} y^1 = (m_{11}-1)a_1^1 + m_{21}\,a_1^2 + m_{31}\,a_1^3 + w^1, \quad y^2 = m_{12}\,a_1^1 + (m_{22}-1)a_1^2 + m_{32}\,a_1^3 + w^2, \\ y^3 = m_{13}\,a_1^1 + m_{23}\,a_1^2 + (m_{33}-1)a_1^3 + w^3, \end{gathered} \tag{0.3}$$

where w^i are uncertain leavings of zero. About drifts ε^i in (0.2), the assumption is accepted that they are constant but unknown. Uncertain functions in relations (0.3) satisfy the integral inequalities

$$\int_0^T (w^i)^2 dt \le \gamma_i^2 T, \quad i \in 1:3. \tag{0.4}$$

Let $\vec{i}_1, \vec{i}_2, \vec{i}_3$ be the unit direction vectors of the BCS. The velocity of point M equals

$$4\vec{v}_M = \overrightarrow{\Omega}_1 \times \overrightarrow{R} = \begin{vmatrix} \vec{i}_1 & \vec{i}_2 & \vec{i}_3 \\ \Omega_1^1 & 0 & \Omega_1^3 \\ 0 & R & 0 \end{vmatrix},$$

where R is the radius of Earth. From here we find the projections of velocity on the BCS axes: $v_1^1 = -R\,\Omega_1^3$, $v_1^2 = 0$, $v_1^3 = R\,\Omega_1^1$. Computing the derivative of $\vec{v}_M$, we get the acceleration $\overrightarrow{w}_M = \widetilde{\overrightarrow{w}}_M + \overrightarrow{\Omega}_1 \times \vec{v}_M$ in the form of the sum of relative and translation accelerations. So, the accelerometers readings in BCS are of the form:

$$a_1^1 = -R\,\dot{\Omega}_1^3, \quad a_1^2 = g - v^2/R, \quad a_1^3 = R\,\dot{\Omega}_1^1, \tag{0.5}$$

where v is the velocity magnitude. As $R = 6370\ km$ and the velocity of the ship on water is no more than 20 m/c, we assume $a_1^2 = g$.

Further we consider some approaches from motion correction for solving the alignment problem. This problem in inertial navigation was first in detail considered in [1]. Russian books devoted to this topic are [2–5]. The alignment problem was mostly solved in [1–5] by statistical methods with the help of Kalman filter or its modifications. On the other hand, in [2, 6] it was noted that the statistics of disturbances often happens incomplete or completely absent. Therefore, it is natural to use here the minimax methods from books [7, 8]. Thus, all the disturbances in our paper are deterministic.

Consider only the case of small angular deviations (no more than several degrees). Equations (0.1) are replaced by the follwing ones:

$$\begin{gathered} \dot{\theta}^1 = u^1 + \varepsilon^1 - \theta^2\Omega_1^3 - \theta^3 u^2, \quad \dot{\theta}^2 = u^2 + \varepsilon^2 + \theta^3\Omega_1^1 - \theta^1\Omega_1^3 - \theta^1 u^3, \\ \dot{\theta}^3 = u^3 + \varepsilon^3 + \theta^2\Omega_1^1 + \theta^1 u^2. \end{gathered} \tag{0.6}$$

Here, $u^i = \Omega^i - \Omega_1^i$, $i \in 1:3$. In the linear approximation, the differences of accelerometer readings in (0.3) are equal to

$$y^1 = g\theta^3 + a_1^3\theta^2 + w^1, \quad y^2 = -a_1^1\theta^3 + a_1^3\theta^1 + w^2, \quad y^3 = -a_1^1\theta^2 - g\theta^1 + w^3. \tag{0.7}$$

Equations (0.6) contain the multiplications of controls and state variables, but, in the case of small angles and angular velocities these terms may be neglected. In the specific case of movement on the equator under condition $\theta^1 = \theta^2 \equiv 0$, we assume $\theta = \theta^3$ as shown on Fig. 3. The angular velocity $\Omega_1 = \Omega_1^3 \neq 0$ under given movement and the rest projections of absolute angular velocity are equal to zero. We have

$$\dot{\theta} = u + \varepsilon, \quad \dot{\varepsilon} = 0, \quad y = g\theta + w, \tag{0.8}$$

where the first equation from (0.7) is taken as the output.

1. Set-membership background

So, we consider a determinate n-dimensional linear system of the form

$$\dot{x}(t) = A(t)x + B(t)u + C(t)v, \quad t \in [0, T], \tag{1.1}$$

assuming that the initial state x_0 of system (1.1) is completely unknown, the matrices $A(t)$, $B(t)$, $C(t)$, and $G(t)$ below are continuous. In comparison with equations (0.6), the term with disturbance v in

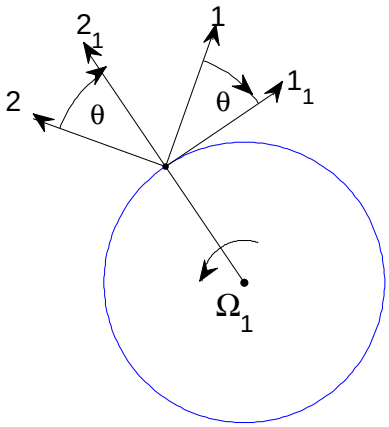


Figure 3. System deviation in the simple model.

the system is added here. It corresponds to the case when drifts are not constants. The unknown function $v(\cdot)$ and the disturbance $w(\cdot)$ in the m-dimensional equation of measurement

$$y(t) = G(t)x(t) + w(t) \tag{1.2}$$

are bounded by the constraint

$$\int_0^T \left(|v(t)|^2_{Q(t)} + |w(t)|^2_{R(t)} \right) dt \le 1, \tag{1.3}$$

where the symbol $|x|^2_P$ equals $x'Px$, prime $'$ means the transposition, $Q(t)$, $R(t)$ are symmetrical, positive-defined, and continuous matrices having suitable dimension. Constraint (1.3) involves that the elements of vector functions $v(\cdot)$ and $w(\cdot)$ belong to the space $L_2[0,T]$. We need the following

Assumption 1. *The system* (1.1), (1.2) *under* $u \equiv 0$, $v \equiv 0$, $w \equiv 0$ *is completely observable* [7] *on any subinterval* $[s,\tau] \subset [0,T]$.

Assumption 1 means that the vector $x(s)$ can be uniquely restored from the signal observed on $[s,\tau]$ if the disturbances are absent. Moreover, Assumption 1 holds if and only if

$$\int_s^\tau X'(t,s)G'(t)G(t)X(t,s)dt > 0,$$

where $X(t,s)$ is the fundamental matrix of system (1.1).

We use piecewise-constant functions $u(t)$, for which

$$u(t) \in P \subset \mathbb{R}^p, \tag{1.4}$$

where P is a compact convex set. Constraint (1.4) is more realistic than integral constraints in [9]. The aim of the control is to minimize the terminal function $|Dx(T)|$, where $|\cdot|$ is the Euclidean norm and $D \in \mathbb{R}^{d\times n}$ is a matrix. The choice of uncertain parameters $\{x_0, v(\cdot), w(\cdot)\}$ may impede the minimization.

1.1. Informational and compatible sets

At first, let us consider a set-membership estimation scheme for system (1.1), (1.2) under constraint (1.3).

Definition 1. *A set* $\mathbf{X}(t,y,u) \subset \mathbb{R}^n$ *is said to be the informational if it consists of all vectors* $x = x(t)$, *which may realize in system* (1.1), (1.2) *with given signal* $y(\tau)$, $0 \le \tau \le t$, *the control* $u(\tau)$, *and some disturbances satisfying constraint* (1.3).

To describe the informational set, we introduce the Bellman function

$$V(t,x) = \inf_{v(\cdot)} \left\{ \int_0^t \left(|v(s)|^2_{Q(s)} + |y(s) - G(s)x(s)|^2_{R(s)} \right) ds \right\}, \quad x(t) = x.$$

The Bellman equation for $V(t,x)$ is of the form:

$$V_t = \min_v \left\{ -(A(t)x + B(t)u(t) + C(t)v)'V_x + |v|^2_{Q(t)} + |y(t) - G(t)x|^2_{R(t)} \right\}, \quad V(0,x) = 0. \quad (1.5)$$

If the solution of equation (1.5) in any sense is found, the *informational set* $\mathbf{X}(t,y,u)$ is written as the inequality $\mathbf{X}(t,y,u) = \{x : V(t,x) \le 1\}$. Let us seek a solution of equation (1.5) in the form

$$V(t,x) = |x|^2_{P(t)} - 2x'd(t) + g(t), \quad (1.6)$$

where $P(t)$ is a positive definite and continuously differentiable matrix, $d(t)$ and $g(t)$ are a continuously differentiable vector function and a function respectively. Substituting (1.6) into (1.5), we get

$$|x|^2_{\dot P(t)} - 2x'\dot d(t) + \dot g(t) = |y(t) - G(t)x|^2_{R(t)} - |P(t)x - d(t)|^2_{C(t)Q^{-1}(t)C'(t)} -$$
$$-2(A(t)x + B(t)u(t))'(P(t)x - d(t)).$$

Therefore, the parameters of (1.6) must satisfy the equations

$$\begin{aligned} \dot P(t) &= G'(t)R(t)G(t) - P(t)C(t)Q^{-1}(t)C'(t)P(t) - A'(t)P(t) - P(t)A(t), \quad P(0) = 0, \\ \dot d(t) &= G'(t)R(t)y(t) - (P(t)C(t)Q^{-1}(t)C'(t) + A'(t))d(t) + P(t)B(t)u(t), \quad d(0) = 0, \\ \dot g(t) &= |y(t)|^2_{R(t)} - |d(t)|^2_{C(t)Q^{-1}(t)C'(t)} + 2d'(t)B(t)u(t), \quad g(0) = 0. \end{aligned} \quad (1.7)$$

It is known [10] that the matrix $P(t)$ is non-singular for any $t > 0$ under Assumption 1. Then the ellipsoid (informational set)

$$\mathbf{X}(t,y,u) = \left\{ x \in \mathbb{R}^n : V(t,x) = |x|^2_{P(t)} - 2x'd(t) + g(t) = |x - \hat x(t)|^2_{P(t)} + h(t) \le 1 \right\} \quad (1.8)$$

is bounded for any $t > 0$ with the center $\hat x(t) = P^{-1}(t)d(t)$ and the function $h(t) = g(t) - |d(t)|^2_{P^{-1}(t)}$. Differentiating the $\hat x(t)$ and $h(t)$, we obtain the equations

$$\begin{aligned} \dot{\hat x}(t) &= A(t)\hat x(t) + B(t)u(t) + P^{-1}(t)G'(t)R(t)(y(t) - G(t)\hat x(t)), \\ \dot h(t) &= |y(t) - G(t)\hat x(t)|^2_{R(t)}. \end{aligned} \quad (1.9)$$

Let us introduce the function

$$f(t) = y(t) - G(t)\hat x(t), \quad t \in (0,T], \quad (1.10)$$

that is similar to the innovation process in theory of Kalman filtering [10].

Lemma 1. *Function* (1.10) *does not depend on the control* $u(t)$*, belongs to the space* $L_2^m[0,T]$*, and we have*

$$h(t) = \int_0^t |y(s) - G(s)\hat{x}(s)|^2_{R(s)} ds \le 1, \quad t \in [0,T].$$

On the other hand, let the instant $\tau \in (0,T)$ *and* $f(\cdot)$ *be any function from* $L_2^m[\tau,T]$ *with*

$$\int_\tau^T |f(s)|^2_{R(s)} ds \le 1 - h(\tau).$$

Then we obtain

$$\mathbf{X}(t,y,u) = \left\{ x \in \mathbb{R}^n : |x - \check{x}(t)|^2_{P(t)} + h(t) \le 1 \right\}, \quad t \in [\tau,T], \tag{1.11}$$

where

$$\dot{\check{x}}(t) = A(t)\check{x}(t) + B(t)u(t) + P^{-1}(t)G'(t)R(t)f(t),$$
$$\check{x}(\tau) = \hat{x}(\tau); \quad h(t) = h(\tau) + \int_\tau^t |f(s)|^2_{R(s)} ds.$$

Here, we set $y(t) = f(t) + G(t)\check{x}(t)$, $t \in [\tau,T]$.

P r o o f. As $0 \le h(t) \le g(t)$ and $g(0) = 0$, we conclude that $h(0) = 0$. From (1.8) and (1.9) we obtain the formula for $h(t)$. The signal $y(\cdot)$ may realize in system (1.1), (1.2) on $[\tau,t]$, $t \in (\tau,T]$, under closed-loop disturbance $v(s) = Q^{-1}(s)C'(s)(P(s)x(s) - d(s))$ that gives minimum to the functional according to (1.5), (1.6), and (1.7) with any final state $x(t) = x \in \mathbf{X}(t,y,u)$. As the formulas for $\check{x}(t)$ coincide with (1.9), formula (1.11) holds. □

From now on, the narrowings of a measurable vector-function $x(s)$, $s \in [0,T]$, on intervals $[0,t]$ and $[t,T]$ are denoted by $x^t(\cdot)$ and $x_t(\cdot)$ respectively. The narrowing on $[t,s]$ is denoted by $x_t^s(\cdot)$. Let the dimension of the disturbance v be equal q.

Definition 2. *A set* $\mathbf{V}(t,y,u) \subset \mathbb{R}^n \times L_2^q[t,T] \times L_2^m[t,T]$ *is said to be the compatible if it consists of all triples* $\{(x(t), v_t(\cdot), w_t(\cdot))\}$*, for which there exist functions* $(v(\cdot), w(\cdot))$ *satisfying* (1.3) *such that output* (1.2) *on* $[0,t]$ *with final state* $x = x(t)$ *almost everywhere coincides with the given signal* $y^t(\cdot)$.

Note that the sets $\mathbf{X}(t,y,u)$ and $\mathbf{V}(t,y,u)$ depend only on $y^t(\cdot)$ and $u^t(\cdot)$. Suppose that we have the compatible set $\mathbf{V}(t,y,u)$, and on the interval $[t,s]$ a signal $y_t^s(\cdot)$ and a control $u_t^s(\cdot)$ are realized. Similarly to Definitions 1 and 2, we can define the sets $\mathbf{X}(s, y_t^s, u_t^s \mid \mathbf{V}(t,y,u))$ and $\mathbf{V}(s, y_t^s, u_t^s \mid \mathbf{V}(t,y,u))$. The following assertion seems to be obvious.

Lemma 2. *The relation between compatible and information sets is given by the equality* $\mathbf{X}(t,y,u) = \mathrm{proj}_{\mathbb{R}^n}\mathbf{V}(t,y,u)$*. The compatible set is described by the formula*

$$\mathbf{V}(t,y,u) = \left\{ (x, v_t, w_t) : \int_t^T \left(|v_t(s)|^2_{Q(s)} + |w_t(s)|_{R(s)} \right) ds + V(t,x) \le 1 \right\}, \tag{1.12}$$

where $V(t,x)$ *is defined in* (1.6) *or* (1.8)*. Under Assumption* 1*, set* (1.12) *is weakly compact in the space* $\mathbb{R}^n \times L_2^q[t,T] \times L_2^m[t,T]$ *when* $t \in (0,T)$*. Moreover, compatible sets posses the semigroup property:* $\mathbf{V}(s, y_t^s, u_t^s \mid \mathbf{V}(t,y,u)) = \mathbf{V}(s,y,u)$*, where* $0 < t < s \le T$*. As a consequence, we have* $\mathbf{X}(s, y_t^s, u_t^s \mid \mathbf{V}(t,y,u)) = \mathbf{X}(s,y,u)$.

The final reachable set of system (1.1) from the compatible set $V(t,y,u)$ is denoted further by $\mathbf{X}_T(u_t|\mathbf{V}(t,y,u))$. This set consists of all vectors $x(T)$ under searching in (1.12) for the set $V(t,y,u)$ with $w_t = 0$.

2. Problems formulation

Let $\lambda : 0 < t_1 < \cdots < t_{N+1} = T$ be a partition of the interval $[0, T]$. The times t_i are called the *instants of control correction*. It is easily seen that the compatible set $\mathbf{V}(t, y, u)$ depends only on the pair $(\hat{x}(t), h(t))$ which is called the *position at the instant* t. The transition between two adjacent positions $(\hat{x}(t_i), h(t_i))$ and $(\hat{x}(t_{i+1}), h(t_{i+1}))$ depends on the control $u_i(\cdot)$ and the innovation function $f_i(\cdot)$ on the interval $[t_i, t_{i+1})$ according to Lemma 1. Consider two problems.

Problem 1. *Find a piecewise-constant control* $u^*(t)$ $(u^*(t) = u_i^*$ *on* $[t_i, t_{i+1})$, $i \in 1 : N)$ *that gives the value*

$$J^* = \min_{u_1 \in P} \max_{f_1(\cdot)} \ldots \min_{u_N \in P} \max_{f_N(\cdot)} \max_{x \in \mathbf{X}_T(u_N | \mathbf{V}(t_N, y, u))} |Dx|, \tag{2.1}$$

where

$$\sum_{i=1}^{N} \int_{t_i}^{t_{i+1}} |f_i(s)|^2_{R(s)} ds \le 1 - h(t_1).$$

Remark 1. As equations (1.1), (1.2) are linear, we have $\mathbf{X}(t, y, u) = z(t) + \mathbf{X}(t, \tilde{y}, 0)$, where $\tilde{y}(t) = y(t) - G(t)z(t)$ and

$$\dot{z}(t) = A(t)z(t) + B(t)u(t), \quad z(0) = 0. \tag{2.2}$$

Similarly, we have $\mathbf{V}(t, y, u) = (z(t), 0, 0) + \mathbf{V}(t, \tilde{y}, 0)$. From now on, we write the sets with $\tilde{y}(\cdot)$ and $u(\cdot) = 0$ as $\mathbf{X}(t, \tilde{y})$ and $\mathbf{V}(t, \tilde{y})$, respectively. Therefore, $\mathbf{X}_T(u_t \mid \mathbf{V}(t, y, u)) = z(T) + \mathbf{X}_T(0 \mid \mathbf{V}(t, \tilde{y}))$ and value (2.1) may be rewritten as

$$J^* = \min_{u_1 \in P} \max_{f_1(\cdot)} \ldots \min_{u_N \in P} \max_{f_N(\cdot)} \max_{x \in \mathbf{X}_T(0 | \mathbf{V}(t_N, \tilde{y}))} |D(z(T) + x)|. \tag{2.3}$$

Remark 2. We obtain as a fact that controls u_i in (2.1) and (2.3) depend on the positions $(\hat{x}(t_i), h(t_i))$. Problem 1 may be generalized if we seek non-constant functions $u_i(\cdot)$ on the interval $[t_i, t_{i+1})$.

Problem 2. *At the any instant* t_i, $i \in 1 : N$, *we find open loop minimax control* $u_i^{T*}(\cdot)$ *that give a solution of the problem:*

$$\max_{f_i(\cdot)} \max_{x \in \mathbf{X}_T(u_i | \mathbf{V}(t_i, y, u))} |Dx| \to \min_{u_i(t) \in P} = j_i(y), \tag{2.4}$$

where

$$\int_{t_i}^{T} |f_i(s)|^2_{R(s)} ds \le 1 - h(t_i),$$

and do one-step forecasting

$$J_i(y, u_i) = \max_{f_i(\cdot)} j_{i+1}(y), \tag{2.5}$$

where

$$\int_{t_i}^{t_{i+1}} |f_i(s)|^2_{R(s)} ds \le 1 - h(t_i).$$

If $J_i(y, u_i^{T*}) < j_i(y)$ *we keep the control* u_i^{T*} *on the interval* $[t_i, t_{i+1}]$. *Otherwise, we pass to the control* u_i^{i+1*} *that minimizes value* (2.5). *Of course, the controls may be not unique. If so, we choose any minimizers.*

3. Minimax solutions

For brevity we denote $\hat{x}(t_i) = \hat{x}_i$ and $h(t_i) = h_i$. Introduce the function of next losses

$$W_i(\hat{x}_i, h_i) = \min_{u_i \in P} \max_{f_i(\cdot)} \dots \min_{u_N \in P} \max_{f_N(\cdot)} \max_{x \in \mathbf{X}_T(u_{t_N} | \mathbf{V}(t_N, y, u))} |Dx|,$$

where

$$\sum_{j=i}^{N} \int_{t_j}^{t_{j+1}} |f_j(s)|^2_{R(s)} ds \le 1 - h_i$$

for Problem 1. It is easily seen that the functions $W_i(\hat{x}_i, h_i)$ satisfy the following recurrent relations

$$W_i(\hat{x}_i, h_i) = \min_{u_i \in P} \max_{f_i(\cdot)} W_{i+1}(\hat{x}_{i+1}, h_{i+1}), \tag{3.1}$$

where

$$\int_{t_i}^{t_{i+1}} |f_i(s)|^2_{R(s)} ds \le 1 - h_i.$$

Relations (3.1) have the boundary condition

$$W_{N+1}(\hat{x}(T), h(T)) = \max_{|x - \hat{x}(T)|^2_{P(T)} \le 1 - h(T)} |Dx| = \max_{|l| \le 1} \left\{ l' D\hat{x}(T) + (1 - h(T))^{1/2} |D'l|_{P^{-1}(T)} \right\}.$$

Consider the last stage of relations (3.1) when $i = N$. Using boundary condition, we obtain

$$W_N(\hat{x}_N, h_N) = \max_{|l| \le 1} \left\{ r(l; t_N) \hat{x}_N + \min_{u \in P} \int_{t_N}^{T} r(l; s) B(s) ds u + \Big((1 - h_N) \Big(\lambda(t_N)(1 - |l|^2) + |D'l|^2_{P(T, t_N)} \Big) \Big)^{1/2} \right\},$$

where

$$\begin{gathered} r(l; s) = l' DX(T, s), \quad \partial P(t, s)/\partial t = A(t) P(t, s) + P(t, s) A'(t) + C(t) Q^{-1}(t) C'(t), \\ P(s, s) = P^{-1}(s), \quad \lambda(s) = \max_{|l| \le 1} |D'l|^2_{P(T, s)}. \end{gathered} \tag{3.2}$$

Here, the term with integral must be replaced on

$$\int_{t_N}^{T} \min_{u \in P} r(l; s) B(s) u ds$$

if the control is not piecewise-constant. Let us explain the formula for $W_N(\hat{x}_N, h_N)$. It is obtained with the help of elementary equality

$$\max_{k \in [0, 1 - h_N]} \left\{ k^{1/2} A + (1 - h_N - k)^{1/2} B \right\} = (1 - h_N)^{1/2} (A^2 + B^2)^{1/2},$$

where $A \ge 0$, $B \ge 0$, and the maximum is achieved at $r^* = (1 - h_N) A^2 (A^2 + B^2)^{-1/2}$. Besides, the optimization over $f(\cdot)$ is fulfilled under the constraint $\int_{t_N}^{T} |f(s)|^2_{R(s)} ds = k$. If $\lambda(s)$ is the maximal eigenvalue of the matrix $DP(T, s)D'$, we use the fact that $\mathrm{conc}|l|_Q$ on unite ball is equal to $\left(\lambda_{\max}(1 - |l|^2) + |l|^2_Q \right)^{1/2}$, see [7]. Hereinafter, the symbol $\mathrm{conc}\varphi(l)$ means a minimal concave function majorizing $\varphi(l)$ on unite ball. At last, we apply the minimax theorem.

Continuing calculations on the subsequent stages, we come to the conclusion.

Theorem 1 (Conditions of the optimality in Problem 1). *On the stage i, we have*

$$W_i(\hat{x}_i,h_i)=\max_{|l|\le 1}\Big\{r(l;t_i)\hat{x}_i+\min_{u\in P}\int_{t_i}^{t_{i+1}}r(l;s)B(s)dsu+\varphi_i(l)\Big\},\quad \textit{where}$$
$$\varphi_i(l)=\operatorname{conc}\Big\{\min_{u\in P}\int_{t_{i+1}}^{t_{i+2}}r(l;s)B(s)u+\max_{f_i(\cdot)}\Big\{\int_{t_i}^{t_{i+1}}r(l;s)P^{-1}(s)G'(s)R(s)f_i(s)ds \tag{3.3}$$
$$+\varphi_{i+1}(l)\Big\}\Big\},\quad i\in 1:N-1.$$

Here

$$\int_{t_i}^{t_{i+1}}|f_i(s)|^2_{R(s)}ds\le 1-h_i.$$

The optimal controls necessarily satisfy the relation

$$\int_{t_i}^{t_{i+1}}r(l^*;s)B(s)dsu_i^*=\min_{u\in P}\int_{t_i}^{t_{i+1}}r(l^*;s)B(s)dsu\quad or\quad \int_{t_i}^{t_{i+1}}r(l^*;s)B(s)u_i^*(s)ds \tag{3.4}$$
$$=\int_{t_i}^{t_{i+1}}\min_{u\in P}r(l^*;s)B(s)uds\quad \textit{if the control is not piecewise-constant,}$$

where l^* is a maximizer in problem (3.3).

P r o o f. For the first two stages, we have

$$\varphi_N(l)=\Big((1-h_N)\Big(\lambda(t_N)(1-|l|^2)+|D'l|^2_{P(T,t_N)}\Big)\Big)^{1/2},$$
$$\varphi_{N-1}(l)=\operatorname{conc}\Big\{\min_{u\in P}\int_{t_N}^{T}r(l;s)B(s)dsu+\max_{f_{N-1}(\cdot)}\Big\{\int_{t_{N-1}}^{t_N}r(l;s)P^{-1}(s)G'(s)R(s)f_{N-1}(s)ds$$
$$+\varphi_N(l)\Big\}\Big\}=\operatorname{conc}\Big\{\min_{u\in P}\int_{t_N}^{T}r(l;s)B(s)dsu+\Big((1-h_{N-1})\Big(\lambda(t_N)(1-|l|^2)+|D'l|^2_{P(T,t_{N-1})}\Big)\Big)^{1/2}\Big\}.$$

For derivation of the last relation, we use the same reasoning as for $W_N(\hat{x}_N,h_N)$. The subsequent considerations are obtained by induction with the help of the minimax theorem. □

To solve Problem 2, we need to calculate values (2.4), (2.5). Doing as above we get

$$j_i(y)=\max_{|l|\le 1}\Big\{r(l;t_i)\hat{x}_i+\int_{t_i}^{T}\min_{u\in P}r(l;s)B(s)uds+\Big((1-h_i)\Big(\lambda(t_i)(1-|l|^2)$$
$$+|D'l|^2_{P(T,t_i)}\Big)\Big)^{1/2}\Big\},$$
$$J_i(y,u_i)=\max_{f_i(\cdot)}j_{i+1}(y)=\max_{|l|\le 1}\Big\{r(l;t_i)\hat{x}_i+\int_{t_i}^{t_{i+1}}r(l;s)B(s)u_i(s)ds \tag{3.5}$$
$$+\int_{t_{i+1}}^{T}\min_{u\in P}r(l;s)B(s)uds+\Big((1-h_i)\Big(\lambda(t_{i+1})(1-|l|^2)+|D'l|^2_{P(T,t_i)}\Big)\Big)^{1/2}\Big\}.$$

Theorem 2 (Properties of controls in Problem 2). *The control procedure in Problem 2 begins from $i=1$ and leads to a sequence of positions, where $j_1(y)\ge j_2(y)\ge\cdots\ge j_N(y)$.*

P r o o f. Let us compare the values $j_i(y)$ and $j_{i+1}(y)$. If $J_i(y, u_i^{T*}) < j_i(y)$, we get $j_i(y) > j_{i+1}(y)$. Otherwise, we use the control u_i^{i+1*} that minimizes the value $J_i(y, u_i)$. Therefore,

$$\min_{u_i(\cdot)} J_i(y, u_i) = \max_{|l|\le 1}\left\{ r(l; t_i)\hat{x}_i + \int_{t_i}^{t_{i+1}} \min_{u\in P} r(l; s)B(s)uds \right.$$

$$\left. +\mathrm{conc}\left\{ \int_{t_{i+1}}^{T} \min_{u\in P} r(l; s)B(s)uds + \Big((1-h_i)\Big(\lambda(t_{i+1})(1-|l|^2) + |D'l|^2_{P(T,t_i)}\Big)\Big)^{1/2} \right\} \right\} \le j_i(y),$$

as $\lambda(t_{i+1}) \le \lambda(t_i)$. The last inequality implies the relation

$$\partial P(T, s)/\partial s = -X(T, s)P^{-1}(s)G'(s)R(s)G(s)P^{-1}(s)X'(T, s),$$

whence the norm of the matrix $P(T, s)$ decreases on s. □

Remark 3. The procedure of calculation of optimal controls in Problem 1 is more difficult than in Problem 2. But we can simplify it if by a slight increase of the function of future losses. Namely, we have $W_i(\hat{x}_i, h_i) \le j_i(y)$. This inequality follows by induction from relations (3.3)–(3.5). One can find the controls in this simplified procedure by formulas (3.4).

To illustrate the different approaches to optimal control, consider a simple

Example. Given the one-dimensional system $\dot{x} = u + v, \quad 0 \le t \le 3$, with the measurement $y(t) = x(t) + w$ and the constraints

$$x_0^2 + \int_0^3 (v^2(t) + w^2(t))dt \le 1,$$

$|u| \le 1/2$, we suppose $y(t) \equiv 1$ on $[0, 3]$. Let $t_1 = 1$, $t_2 = 2$ be two correction instants. Here, we add the limitation on initial state for simplicity.

We have $P \equiv 1$, $\hat{x}(t) = 1 - e^{-t}$, $h(t) = (1 - e^{-2t})/2$ on $[0, 3]$ under $u \equiv 0$, as follows from (1.7), $P(T, s) = 4 - s$. The unknown real movement $x(t) \equiv 1$ under $u \equiv 0$. Formula (3.3) gives $W_1(\hat{x}_1, h_1) = 1.0655$ and optimal control on $[1, 2]$ equals $u_1 = -0.5$. Here, the choice of control is not unique. At the next stage $W_2(\hat{x}_2, h_2) = 1.0091$ and the optimal control on $[2, 3]$ equals $u_2 = -\hat{x}_2 = e^{-2} - 1/2 = -0.3647$. In Problem 2, we have $j_1(y) = 1.3050$ and we obtain the same sequence of optimal controls. At last, consider the partition of $[1, 3]$ with step 0.25, $N = 8$, and we use the procedure of Remark 3. This procedure leads us to the sequence of control $u_i = -0.5$ at each step. The final value of the functional equals 0.7577.

4. Numerical simulation of alignment process

We restrict ourself by the consideration of the simple case of system (0.8) and the procedure of Remark 3. The qualitative sense does not change in the common case.

The following data are used: $|\theta^i| \le 3$ *grad*, $|\varepsilon^i| \le 0.1$ *grad/sec*, $|u_i| \le 0.1$ *rad/sec*, $T = 100$ *sec*. In integral constraint (0.4) the constants are $\gamma_i = 0.1$ *m/sec*^{2}. The signal is given by $w(t) = \sin(t)/\sqrt{55}$. The alignment process is shown on the figures.

5. Conclusion

In this paper, we consider the application of motion correction methods to the alignment problem in inertial navigation. We use the deterministic approach with set-membership description of uncertainty. The Theorems 1, 2 and the procedure in Remark 3 are new. The investigation of the influence of ship movement on the accuracy of alignment was not performed. It will be done in subsequent papers.

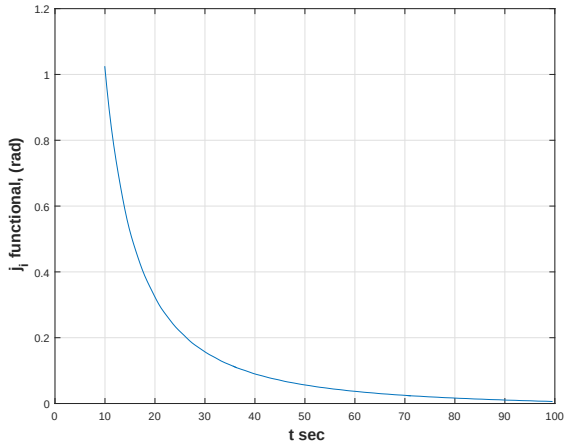


Figure 4. Alteration of the functional in the simple model.

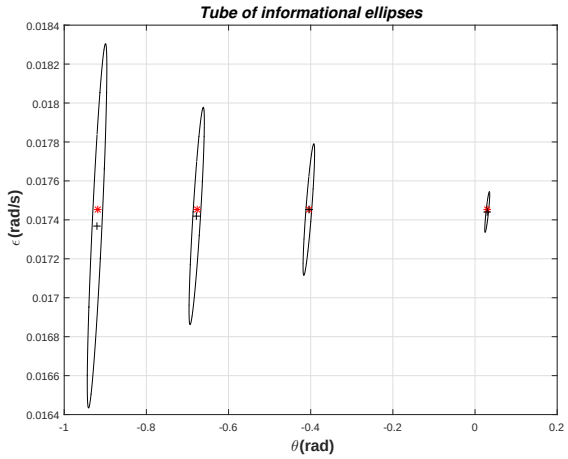


Figure 5. Informational ellipses at the instants $t = 44$, 58, 72, and 100.

REFERENCES

1. **Lipton A.H.** Alignment of Inertial Systems on a Moving Base. Washington D.C.: NASA, 1967. 178 p.
2. **Boguslavski I.A.** Applied Problems of Filtering and Control. M.: Nauka, 1983. 314 p. [in Russian]
3. **Bromberg P.V.** The Theory of Inertial Navigation Systems. M.: Nauka, 1979. 245 p. [in Russian]
4. **Klimov D.M.** Inertial Navigation on the Sea. M.: Nauka, 1984. 211 p. [in Russian]
5. **Parusnikov N.A., Morozov V.M, Borzov V.I.** Correction Problem in Inertial Navigation. M.: MGU, 1982. 256 p. [in Russian]
6. **Bachshiyan B.Ts., Nazirov R.R., Eliyasberg P.E.** Determination and Correction of Motion. M.: Nauka, 1980. 402 p. [in Russian]
7. **Kurzhanski A.B.** Control and Observation under Conditions of Uncertainty. M.: Nauka, 1977. 392 p. [in Russian]
8. **Krasovskiĭ N.N. and Subbotin A.I.** Game-Theoretical Control Problems. Springer–Verlag, New York, 1988. 517 p.
9. **Ananyev B.I. and Gredasova N.V.** The Alignment Problem of Inertial Systems and Motion Correction Procedure // Bulletin of Buryatian State University, 2011. No. 9. P. 203–208. old.bsu.ru/content/pages2/1074/2011/AnanevBI.pdf [in Russian]
10. **Liptser R.Sh. and Shiryayev A.N.** Statistics of Random Processes, V.1 General Theory, V.2 Applications, Springer–Verlag, New York, 2000.

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

SIMPLIFIED MODEL OF THE HEAT EXCHANGE PROCESS IN ROTARY REGENERATIVE AIR PRE–HEATER[1]

Abdulla A. Azamov

Institute of Mathematics, National University of Uzbekistan,
Tashkent, Uzbekistan,
abdulla.azamov@gmail.com

Mansur A. Bekimov

Institute of Mathematics, National University of Uzbekistan,
Tashkent, Uzbekistan,
mansu@mail.ru

Abstract: A simplified mathematical model of a rotary regenerative air pre-heater (RRAP) is suggested and studied based on the averaged dynamics of the heat exchange process between nozzles and a heat carrier (i.e. air or gas-smoke mixture). Averaging in both spatial coordinates and time gives a linear discrete system that allows deriving explicit formulas for determining the characteristics of the air heater and establishing some properties such as periodicity, stability, ergodicity and others.

Key words: Heat exchange, Cyclic process, Averaging, Linear discrete system, Stability, Ergodicity, Inverse problem.

Introduction

A rotary regenerative air pre-heater (RRAP) is a special unit, usually applied in thermal power plants (TPP) in order to increase its efficiency by heating the air which is blowing into a boiler of the plant by means of an exhaust of hot mixture of smoke and gas (from now on simply gas) generated in fuel combustion. The RRAP essentially reduces the thermal pollution of atmosphere [6], [8–12].

Currently, several types of RRAPs are used in TPP. In the present paper, we consider the case of a unit with the main part consisting of a rotating cylindrical drum with metal nozzles of high thermal conductivity. The region of space occupied by the wheel of the RRAP is divided into two parts B_A and B_G by a fixed conditional plane passing through the axis of the cylinder. During the work of the RRAP, the atmospheric air passes through the part B_A in one direction parallel to the axis of the drum, being heated by the nozzles; as a result, the temperature of the nozzles is reduced. Through the part B_G, the gas flows in the opposite direction, being cooled by the heat output to the nozzles. The final transfer of heat from the hot gas to the cool air occurs due to the rotation of the drum around its axis.

Monitoring and control of the temperatures of air and gas leaving the RRAP and especially the temperature of the nozzles is the important problem for the effective exploitation of the RRAP [3], [15], [17–19]. The direct measurement of the temperatures of air and gas of both inlet and outlet is easily carried out; in contrast, the observation and control of the nozzle temperature require to use sophisticated measuring tools and represent an important problem. Therefore, the treatment for mathematical modeling of the heat exchange process in the RRAP is reasonable.

[1]This work was supported by Committee for Coordination Science and Technology Development Under Cabinet of Ministers of Uzbekistan (project no F4–FA–F014).

In recent decades, a variety of mathematical models of the heat exchange process in RRAPs were suggested [1–3], [5] and see also [8]. It should be noted that the thermodynamic process in RRAP nozzles can be in principle described by the classical heat conduction equation, but, without additional simplifying assumptions, it is difficult to write out the boundary conditions since the nozzle system has a rather complex geometry. On the other hand, modeling of the thermodynamic process of air heating and gas cooling passing through the drum of the RRAP is even more challenging problem pertaining to thermos-aerodynamics [8]. Furthermore, the necessity of considering the turbulence of flow that arises when air and gas pass through the RRAP drum also adds complexity. Because of these and other features, all mathematical models of the RRAP are built under essential simplifying assumptions.

In the present paper, we propose a mathematical model of the thermodynamic process in the RRAP based on averaging the quantities associated with the heat exchange process between the nozzles, air and gas in both spatial coordinates and time interval. As a result, we obtain rather simple linear discrete equation. This allows us to write out explicit calculating formulas for the current values of parameters, and find steady and periodical states, establish the ergodicity and other properties. Then, we provide appropriate calculations for the case when averaging over the time variable is carried out on a period of time equal to a half-cycle of rotation of the RRAP drum. In subsequent parts of the work, models with averaging performed over a small time interval are considered, a comparative analysis of numerical results and experimental data is fulfilled.

1. Derivation of equations

Let the RRAP drum be of the form of a cylinder $x^2+y^2\le R^2$, $0\le z\le H$. Assume that the parts B_A and B_G are described by conditions $y\ge 0$ and $y\le 0$, respectively. Let $\Theta(t,x,y,z)$ denote the temperature at the point (x,y,z) of the drum occupied by the nozzles at time t, $t\ge 0$, and let $T(t,x,y,z)$ denote the temperature of heat transfer (air or gas) at the point (x,y,z) of the drum outside the nozzles. The pair of quantities $\Theta(t,x,y,z)$ and $T(t,x,y,z)$ completely characterizes the heat change process in the RRAP drum. However, as mentioned above, due to the complexity of the configuration of the nozzles, the initial-boundary problem for corresponding system of equations of thermo-aerodynamics is too hard to be investigated by analytical methods. One of the ways to overcome such complexity is the method of averaging [4], [7]. To this end, we consider the corresponding average values

$$\Theta_A(t)=\frac{1}{V(B_A^\bullet)}\int_{B_A^\bullet}\Theta_A(t,x,y,z)dv,\quad \Theta_G(t)=\frac{1}{V(B_G^\bullet)}\int_{B_G^\bullet}\Theta_G(t,x,y,z)dv,$$

$$T_A(t)=\frac{1}{V(B_A^\circ)}\int_{B_A^\circ}T_A(t,x,y,z)dv,\quad T_G(t)=\frac{1}{V(B_G^\circ)}\int_{B_G^\circ}T_G(t,x,y,z)dv,$$

where $B_A^\bullet=B_A\cap\bar B$, $B_G^\bullet=B_A\cap\bar B$, $\bar B$ is the part of the drum occupied by the nozzles, $B_A^\circ=B_A\backslash B_A^\bullet$, $B_G^\circ=B_G\backslash B_G^\bullet$, dv is the volume element, V denotes the volume of the corresponding part of the drum.

Next, we perform averaging over the time intervals $I(n)=[nh,\ \ (n+1)h)$ as well, where h is the half-turn time of the RRAP drum, $n=0,1,2,\dots$. Denote the average temperatures obtained in this way by

$$x_n=\frac{1}{h}\int_{I(n)}\Theta_A(t)dt,\quad y_n=\frac{1}{h}\int_{I(n)}\Theta_G(t)dt,$$

$$u_n=\frac{1}{h}\int_{I(n)}T_A(t)dt,\quad v_n=\frac{1}{h}\int_{I(n)}T_G(t)dt.$$

Equations connecting these quantities are derived under the following simplifying assumptions:

1. for $nh \le t < (n+1)h$, the RRAP drum is motionless, the portion of air (respectively, gas) filling the part B_A° (respectively, B_G°) is also motionless, the heat exchange between the heat carrier and nozzles occurs in accordance with the linear Newton law [8];

2. at the time $t = (n+1)h$, the heated portion of air with temperature u_n leaves the area B_A°, the gas portion cooled down to temperature v_n leaves the region B_G°, the drum turns to 180° with jump (i.e., the part $B_A^{\bullet}$ goes to the part $B_G^{\bullet}$ and vice versa); then the part B_A° is filled in by a new portion of air from outward (or from calorifer in case of its connection to the RRAP, [12]) of temperature p_n, and the part B_G° is filled in by a new portion of gas of temperature q_n.

In accordance with the Newton law, we have the following relations:

$$x_{n+1} = (1-\beta h)y_n + \beta h q_n, \quad y_{n+1} = (1-\alpha h)x_n + \alpha h p_n, \tag{1.1}$$

$$u_n = p_n + \gamma h(x_n - p_n), \quad v_n = q_n + \delta h(y_n - q_n), \tag{1.2}$$

where α, β, γ, δ are parameters depending on characteristics of the heat exchange process in the RRAP (the geometry, the heat capacity of the carcass of drum, the system of nozzles and their structure, the humidity of air and gas and their thermodynamic characteristics, the coefficients of heat conductivity and diffusion, parameters that characterize the heat exchange process on the contact surface of the nozzles with air and gas and etc.).

Relations (1.1) and (1.2) have been obtained as a result of extremely simplifying assumptions on the heat exchange process in the RRAP. Nevertheless, due to such simplification, system (1.1) allows a fairly complete analysis; therefore, it can serve as a basic model to describe the work of the RRAP.

2. Solution of the system

Introducing the vectors $z_n = (x_n, y_n)^T$ and $r_n = h(\beta q_n, \alpha p_n)^T$, where T is the transpose sign to transform the row-vector to the column-vector, and the matrix

$$A = \begin{pmatrix} 0 & 1-\beta h \\ 1-\alpha h & 0 \end{pmatrix},$$

one can rewrite system (1.1) as follows:

$$z_{n+1} = Az_n + r_n, \quad n = 0, 1, 2, \ldots . \tag{2.1}$$

All further reasoning is conducted under the assumption $0 < \alpha h,\ \beta h < 1$, called the physical realizability of the model. This condition implies the fact that the eigenvalues of the matrix A, which are $\pm\mu$ where $\mu = \sqrt{(1-\alpha h)(1-\beta h)}$, belong to the interval $(-1,\ 1)$ and therefore all the solutions of system (1.1) are asymptotically stable [13, 14], [16].

The solution of equation (2.1) can be written by the Cauchy formula [13]

$$z_n = A^n z_0 + \sum_{k=0}^{n-1} A^{n-1-k} r_k. \tag{2.2}$$

Since $A^n = \mu^n E$ for even numbers n, and $A^n = \mu^{n-1}A$ for odd numbers n, then equation (2.2) can be transformed to the form

$$z_n = \mu^n z_0 + A \sum_{j=0}^{n/2-1} \mu^{n-2-2j} r_{2j} + \sum_{j=0}^{n/2-1} \mu^{n-2-2j} r_{2j+1},$$

for even n, and to the form

$$z_n = \mu^{n-1}Az_0 + A\sum_{j=1}^{(n-1)/2}\mu^{n-1-2j}r_{2j-1} + \sum_{j=0}^{(n-1)/2}\mu^{n-1-2j}r_{2j}$$

for odd n.

Based on these relations, one can easily obtain explicit formulas for x_n and y_n. In practice, however, it is more convenient to calculate them, in case of need, directly from equations (1.1) that are accommodated to computer calculations. In the rest of the paper, some properties of solution of systems (1.1), (1.2) are established.

3. Steady state and periodic regimes

Let us first consider the case of a steady state, which occurs under the assumption of constant incoming flows. Such a state can be established in the periods of time measured in hours, if the parameters of the air and the energy load at the TPP remain virtually unchanged. Since h is measured in minutes, the work of the RRAP actually consists of long periods of steady states and relatively short transition intervals from one steady state to another. Therefore, it is important to determine the parameters of the RRAP under steady state conditions.

Thus, let

$$p_n \equiv \bar{p}, \quad q_n \equiv \bar{q}, \quad n = 0, 1, 2, \dots, \quad \bar{r} = (\beta\bar{q}, \quad \alpha\bar{p})^T.$$

Then formula (2.2) can be simplified even more: for even n

$$z_n = \mu^n z_0 + (1-\mu^n)(E-A)^{-1}\bar{r}, \tag{3.1}$$

or in the component form

$$x_n = \mu^n x_0 + \frac{1-\mu^n}{\alpha+\beta-\alpha\beta h}[p + q(1-\beta h)],$$
$$y_n = \mu^n y_0 + \frac{1-\mu^n}{\alpha+\beta-\alpha\beta h}[q + p(1-\alpha h)],$$

and for odd n

$$z_n = \mu^{n-1}Az_0 + \frac{1}{\alpha+\beta-\alpha\beta h}(E-\mu^n A)(E-A)^{-1}r,$$

or

$$x_n = \mu^{n-1}(1-\beta h)y_0 + \frac{1}{\alpha+\beta-\alpha\beta h}[p(1-\mu^{n+2}) + q(1-\beta h)(1-\mu^{n-1})],$$
$$y_n = \mu^{n-1}(1-\alpha h)x_0 + \frac{1}{\alpha+\beta-\alpha\beta h}[q(1-\mu^{n+2}) + p(1-\alpha h)(1-\mu^{n-1})].$$

For the steady state, we obtain $\bar{z} = (E-A)^{-1}\bar{r}$ from the equation $\bar{z} = A\bar{z} + \bar{r}$ (the invertibility of the matrix $E-A$ follows from the condition of physical realizability) or in the component form

$$\begin{aligned}\bar{x} &= \frac{1}{\alpha+\beta-\alpha\beta h}\,(\alpha p + \beta q(1-\beta h)),\\ \bar{y} &= \frac{1}{\alpha+\beta-\alpha\beta h}\,(\beta q + \alpha p(1-\alpha h)).\end{aligned} \tag{3.2}$$

Due to the fact noticed above, steady state (3.2) is asymptotically stable. On time intervals measured in weeks, a periodic state for the RRAP can be formed.

Theorem 1. *Let the sequence r_n be periodic with the period m, $m \geq 2$, i.e. $r_{n+m} = r_n$, where $n = n_0,\ n_0 + 1,\ n_0 + 2,\ \ldots;\ n_0$ is some fixed value of n. Then system* (2.1) *has a unique periodic solution of period m, which is asymptotically stable.*

P r o o f. The uniqueness and asymptotic stability follows from the condition of physical realizability. To establish the existence, we set

$$z_n^* = (E - A^m)^{-1} \sum_{k=0}^{m-1} A^{m-1-k} r_k \tag{3.3}$$

(it follows from the condition of physical realizability that the matrix $E - A^m$ is invertible).

Let z_n^* be a trajectory with the initial point z_0^*. Rewrite relation (3.3) as follows:

$$z_n^* = A^m z_0^* + \sum_{k=0}^{m-1} A^{m-1-k} r_k = z_m^*.$$

Then,

$$z_{n+m}^* = A^{n+m} z_0^* + \sum_{k=0}^{m-1} A^{n+m-1-k} r_k + \sum_{k=m}^{n+m-1} A^{n+m-1-k} r_k. \tag{3.4}$$

The sum of the first two summands in the right hand side of (3.4) is $A^n z_m^*$, the third one is equal to

$$\sum_{k=0}^{n-1} A^{n-1-k} r_{k+m} = \sum_{k=0}^{n-1} A^{n-1-k} r_k$$

since r_k is periodic. Thus,

$$\sum_{k=0}^{n-1} A^{n-1-k} r_k = z_n - A^n z_0.$$

Now, in accordance with (3.4), we obtain

$$z_{n+m}^* = A^n z_m^* + z_n - A^n z_0^* = z_n,$$

and the proof is complete. □

4. Boundedness and ergodicity of the solutions

In time intervals of longer duration, in a changeable external environment and energy load at the TPP, the sequence r_n is not periodic, a fortiori, it is not stationary. In this regard, we give two properties of the solution for more general classes of systems (2.1), which models more or less irregular regime of the heat exchange process.

Theorem 2. *If r_n is a bounded sequence, then each solution of equation* (2.1) *is also bounded.*

This statement is also a special case of more general theorem [11].

Theorem 3. *Let $\lim\limits_{n\to\infty} r_n = l$. Then each solution z_n approaches the limit $(E-A)^{-1}l$ as $n \to \infty$ independently of z_0.*

P r o o f. First assume that $l = 0$. Let $M = \max\limits_n |r_n|$. Given any $\varepsilon > 0$, choose $n_0(\varepsilon)$ such that $|r_n| < \dfrac{1}{2}(1 - ||A||)\varepsilon$ at $n \geq n_0(\varepsilon)$. Then choose $N(n_0(\varepsilon))$ such that

$$\mu^n < \frac{1 - ||A||}{2M\,(1 - \|A^{n_0}\|)}\varepsilon$$

for all $n \geq N(n_0(\varepsilon))$. Clearly, $N(n_0(\varepsilon))$ can be chosen to satisfy $N(n_0(\varepsilon)) \geq n_0(\varepsilon)$. Then, for $n \geq N(n_0(\varepsilon))$, we have

$$|z_n| < \|A^n\|\,|z_0| + \left|\sum_{k=0}^{n_0-1} A^{n-1-k} r_k\right| + \left|\sum_{k=n_0}^{n-1} A^{n-1-k} r_k\right|.$$

Since $0 < \alpha h,\ \beta h < 1$, the first summand tends to zero as $n \to \infty$. Denote two other terms by S_n^I and S_n^{II}. Then, in view of $\|A\| = \mu \in (0,\, 1)$, we get

$$S_n^I < \left\|A^{n-n_0}\right\| \sum_{k=0}^{n_0-1} \left\|A^{n_0-1-k}\right\| |r_k| < M \left\|A^{n-n_0}\right\| \sum_{k=0}^{n_0-1} \left\|A^{n_0-1-k}\right\| < M\mu^n \frac{1 - \|A^{n_0}\|}{1 - \|A\|} < \frac{\varepsilon}{2},$$

$$S_n^{II} < \sum_{k=n_0}^{n-1} \left\|A^{n-1-k}\right\| |r_k| < \frac{1 - \|A\|}{2}\varepsilon \sum_{k=0}^{\infty} \left\|A^k\right\| = \frac{1 - \|A\|}{2}\varepsilon \frac{1}{1 - \|A\|} < \frac{\varepsilon}{2}.$$

Hence, $|z_n| \leq S_n^I + S_n^{II} < \varepsilon$ at $n \geq N(n_0(\varepsilon))$, i.e. $z_n \to 0$.

Let now $r_n \to l$ as $n \to \infty$, where $l \neq 0$. We make the change of variables

$$z_n = \bar{z}_n + (E - A)^{-1} l, \quad \bar{r}_n = r_n - l.$$

Then

$$\bar{z}_{n+1} + (E - A)^{-1} l = A\left[\bar{z}_n + (E - A)^{-1} l\right] + \bar{r}_n + l = A\bar{z}_n + \bar{r}_n,$$

where $\bar{r}_n \to 0$. As proved, $\bar{z}_n \to 0$. Therefore, $z_n \to (E - A)^{-1} l$. □

In general, the change of the parameters characterizing the state of air, as well as the variation of the load at the TPP is random with a hard-determinable distribution function.

Considering the work of the RRAP as a stochastic process, we leave it for the next part of the paper, and we now present another property, taking into account irregular characters of values of phase variables of system (1.1), (1.2).

A sequence a_n is called almost-periodic, if it can be represented in the form $b_n + c_n$, where b_n is periodic and $c_n \to 0$ as $n \to \infty$.

Corollary 1. *If r_n is almost-periodic, then each solution of* (2.2) *is also almost-periodic.*

Definition. A sequence x_n is called ergodic, if the sequence of Cesaro means

$$\sigma_n = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

converges as $n \to \infty$.

Theorem 4. *If the sequence r_n is ergodic, namely*

$$\rho_n = \frac{r_0 + r_1 + \cdots + r_{n-1}}{n} \to l,$$

then each solution z_n is also ergodic with

$$S_n = \frac{z_1 + z_2 + \cdots + z_n}{n} \to (E - A)^{-1} l.$$

P r o o f. Express S_N as the sum of two terms: $S_N = \Sigma_I + \Sigma_{II}$, where

$$\Sigma_I = \frac{1}{N}\sum_{n=1}^{N} A^n z_0, \quad \Sigma_{II} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=0}^{n-1} A^{n-1-k} r_k.$$

Then

$$\Sigma_I = \frac{1}{N}\sum_{n=1}^{N} A^n z_0 = \frac{1}{N} A(E-A)^{-1}(E-A^N) z_0 \to 0$$

as $n \to \infty$.

For Σ_{II}, we have

$$\Sigma_{II} = \frac{1}{N}\sum_{k=0}^{N-1}\sum_{n=k+1}^{N} A^{n-1-k} r_k = \frac{1}{N}\sum_{k=0}^{N-1}\left(\sum_{n=0}^{N-1-k} A^n\right) r_k$$

$$= \frac{(E-A)^{-1}}{N}\left[\sum_{k=0}^{N-1} r_k - \sum_{k=0}^{N-1} A^{N-1-k} r_k\right]$$

$$= (E-A)^{-1}\rho_n - \frac{(E-A)^{-1}}{N}\sum_{k=0}^{N-1} A^{N-1-k} r_k.$$

By hypothesis of the theorem, $(E-A)^{-1}\rho_n \to (E-A)^{-1} l$ as $N \to \infty$. Assuming $\rho_k = \xi_k + l$, where $\xi_k \to 0$, we arrive at the equation

$$r_k = (k+1)\rho_{k+1} - k\rho_k = (k+1)\xi_{k+1} - k\xi_k + l.$$

Thus,

$$\frac{(E-A)^{-1}}{N}\sum_{k=0}^{N-1} A^{N-1-k} r_k = \frac{(E-A)^{-1}}{N}\left[\sum_{k=0}^{N-1} A^{N-1-k}\left[(k+1)\xi_{k+1} - k\xi_k\right] - \sum_{k=0}^{N-1} A^{N-1-k} l\right].$$

Obviously,

$$\sum_{k=0}^{N-1} A^{N-1-k} l \to 0$$

as $N \to \infty$. Let

$$S = \frac{1}{N}\sum_{k=0}^{N-1} A^{N-k-1} k\xi_k.$$

Show that $S \to 0$ as $N \to \infty$. Indeed,

$$|S| \le \sum_{k=0}^{N-1} \mu^{N-1-k}\frac{k}{N}|\xi_k| \le \sum_{k=0}^{N-1} \mu^{N-1-k}|\xi_k|.$$

By Theorem 3, the right-hand side of this inequality tends to 0 as $N \to \infty$. Thus, finally we obtain

$$S_N = \frac{1}{N}\sum_{n=1}^{N} z_n \to (E-A)^{-1} l,$$

which is our claim. □

5. Finding the coefficients of the system

As mentioned above, besides the temperature and velocity of incoming air and gas, the drum rotation speed, the heat exchange process in the RRAP depends on many parameters expressing the thermodynamic characteristics of air and gas, the material and geometry of the nozzles of the RRAP, thermal properties of the drum casing etc.

There are a number of works devoted to the mathematical modeling of the RRAP where formulas are given to calculate the values of the parameters above on the basis of molecular physics laws [1–3], [8], [9]. In addition, finding the numerical parameters characterizing the RRAP is only possible with a certain accuracy. From this point of view, it is much easier to define them with a satisfactory accuracy by solving the inverse problem for system (1.1), (1.2) on the basis of empirical data. With regard to model (1.1), (1.2), such a problem consists in calculating the values of α, β, γ, and δ based on the results of measurements of temperatures of outgoing air u_n and gas v_n. Here, there is a wide field of application of the least squares method and tools of mathematical statistics. Here, we confine ourselves to the simplest case, when α, β, γ, and δ are found by measuring u_1, u_2, v_1, and v_2 assuming that incoming streams $p_n = p,\ q_n = q$ are stationary. We can assume that $u_0 = p,\ v_0 = q$, and the values of u_1, u_2, v_1, and v_2 are found by the direct measurements. As a result, we arrive at the problem of finding the unknowns α, β, γ, δ based on the given values h, p, q, u_1, u_2, v_1, v_2, without measuring the values x_n, y_n (the nozzle temperature).

We have

$$x_2 = (1 - \beta h)y_1 + \beta hq, \quad y_2 = (1 - \alpha h)x_1 + \alpha hp, \tag{5.1}$$

$$u_1 = p + \gamma h(x_1 - p), \quad v_1 = q + \delta h(y_1 - q), \tag{5.2}$$

$$u_2 = p + \gamma h(x_2 - p), \quad v_2 = q + \delta h(y_2 - q). \tag{5.3}$$

Set $\bar{u}_k = u_k - p, \quad \bar{v}_k = v_k - q, k = 1, 2,$ (these values have the clear physical meaning). Substituting the values of x_2 and y_2 from (5.1) into (5.3), we obtain the following system

$$\gamma h(x_1 - p) = \bar{u}_1, \quad \delta h(y_1 - q) = \bar{v}_1, \tag{5.4}$$

$$\gamma h[(1 - \beta h)y_1 + \beta hq - p] = \bar{u}_2, \quad \delta h[(1 - \alpha h)x_1 + \alpha hp - q] = \bar{v}_2, \tag{5.5}$$

with 6 unknowns α, β, γ, δ, x_1, y_1. It is nonlinear and, in general, cannot be solved explicitly.

Therefore, we use the fact that, for the values of α, β, γ, δ, there is a priori estimate $0.1 \div 0.6$ and $h < 1$. This allows us to neglect the terms containing $\beta\gamma h^2$ and $\alpha\delta h^2$.

Then equations (5.5) take the form $\gamma h(y_1 - p) = \bar{u}_2$, $\delta h(x_1 - q) = \bar{v}_2$. As a result, for the intermediate unknowns x_1, y_1, we obtain the linear system

$$\frac{x_1 - p}{y_1 - p} = \frac{\bar{u}_1}{\bar{u}_2}, \quad \frac{y_1 - q}{x_1 - q} = \frac{\bar{v}_2}{\bar{v}_1}, \tag{5.6}$$

with the determinant equal to $\bar{u}_1\bar{v}_1 - \bar{u}_2\bar{v}_2$. We call the quantity $\chi = |\bar{u}_1\bar{v}_1 - \bar{u}_2\bar{v}_2|$ the divergence coefficient of the RRAP. The deviation of χ from zero is a characteristic of the RRAP that expresses how the rates of air heating and gas cooling differ. Further, we assume $\chi \neq 0$. It follows from (5.6) that

$$x_1 = \frac{(\bar{u}_1 - \bar{u}_2)\bar{v}_2 p + (\bar{v}_1 - \bar{v}_2)\bar{u}_1 q}{\bar{u}_1\bar{v}_1 - \bar{u}_2\bar{v}_2}, \quad y_1 = \frac{(\bar{u}_1 - \bar{u}_2)\bar{v}_1 p + (\bar{v}_1 - \bar{v}_2)\bar{u}_2 q}{\bar{u}_1\bar{v}_1 - \bar{u}_2\bar{v}_2}.$$

Substituting the values $x_1,\ y_1$ into (5.4), (5.5), we obtain the final formulas

$$\alpha h = \frac{x_1 - q}{x_1 - p}\bar{v}_1 - \frac{y_1 - q}{x_1 - p}\bar{v}_2, \quad \beta h = \frac{y_1 - p}{y_1 - q}\bar{u}_1 - \frac{x_1 - p}{y_1 - q}\bar{u}_2$$

$$\gamma h = \frac{\bar{u}_1}{x_1 - p}, \quad \delta h = \frac{\bar{v}_1}{y_1 - q}.$$

Table 1 shows posteriori values of parameters α, β, γ, δ calculated by these formulas under the assumption that $p = 32^\circ$, $q = 282^\circ$, $h = 0.25$ [4, 5].

$\bar{u}_1$	$\bar{v}_1$	$\bar{u}_2$	$\bar{v}_1$	α	β	γ	δ
3.61	−2.58	6.75	−4.82	0.048	0.068	0.041	0.029

Table 1. Posteriori values of parameters α, β, γ, δ.

6. Conclusion

In the present paper, we have proposed the mathematical model of thermodynamic process of the RRAP, which is described by linear discrete equations. To obtain this, we have used averaging the quantities associated with the heat exchange process between the nozzles, air and gas in both the spatial coordinates and time interval. We have found steady and periodical states, established the ergodicity and other properties. Next, we have studied the cases when the time averaging is performed over the period of time equal to the half-cycle of rotation of the RRAP drum as well as when the time averaging is performed over a small time interval. Finally, we have provided the comparative analysis of the numerical results obtained and the experimental data.

Acknowledgements

Authors express their gratitude to G.I. Ibragimov for useful discussion and help.

REFERENCES

1. **Alagic S., Kovacevic A. and Buljubasic I.** A numerical analysis of heat transfer and fluid flow in rotary regenerative air pre-heaters // Strojniški Vestnik, 2005. Vol. 51, no. 7–8, P. 411–417.
2. **Armin Heidari-Kaydan, Ebrahim Hajidavalloo.** Three-dimensional simulation of rotary air pre-heater in steam power plant // Applied Thermal Engineering, 2014. Vol. 73. P. 397–405.
3. **Boštjan Drobnič, Janez Oman, Matija Tuma.** A numerical model for the analysis of heat transfer and leakages in a rotary air preheater //Int. J. of Heat and Mass Transfer, 2006. Vol. 49. P. 5001–5009.
4. **Burd V.Sh.** Method of averaging for differential equations on an infinite interval: Theory and Applications. Chapman and Hall/CRC, 2007. 360 p.
5. **Chi-Liang Lee.** Regenerative air preheaters with four channels in a power plant system // J. of Chinese Institute of Engineers. 2009. Vol. 32, no. 5, P. 703–710.
6. **Grebennikov E.A.** Method of averaging in applied problems. Moscow: Nauka, 1986. 256 p. [in Russian]
7. **Kirsanov Yu.A.** Cyclic thermal processes and the theory of thermal conductivity in regenerative air heaters. Moscow: Fizmatlit, 2007. 240 p. [in Russian]
8. **Kovalevskii V.P.** Simulation of heat and aerodynamic processes in regenerators of continuous and periodic operation. I. Nonlinear mathematical model and numerical algorithm // J. of Engineering Physics and Thermophysics, 2004. Vol. 77, no. 6, P. 1096–1109.
9. **Kudinov A.A.** The study of heat transfer processes in rotary regenerative air power boilers // J. of Energetics, 2012. No 6. P. 32–34. [in Russian]
10. **Kudinov A.A., Ziganshina S.K.** Energy savings in power and heat technologies. Moscow: Mashinostroenie, 2011. 374 p. [in Russian]
11. **Liu Fuguo, Zhou Xingang.** Heat transfer model of tri-section rotary air preheater and experimental verification // J. of Mechanical Engineering, 2010. Vol. 46, no 22, P. 144–150.

12. **Ismatkhodjayev S.K., Nuraliyev E.N., Abduraimov F.E., Mokrushev V.A., Azamov A.A., Bekimov M.A.** A method for automatic adjusting the air temperature in air heater. Patent No IAP 05072. [in Russian]
13. **Romanko V.K.** The course of difference equations. oscow: Fizmatlit, 2012. 200 p. [in Russian]
14. **Saber Elaydi.** An introduction to difference equations. Third edition, Springer Science+Business Media, New York, 2005.
15. **Shruti G., Ravinarayan Bhat, Gangadhar Sheri.** Performance evaluation and optimization of air preheater in thermal power plant // International J. of Mechanical Engineering and Technology (IJMET), 2014. Vol. 5, iss. 9. P. 22–30.
16. **Srochko V.A.** Numerical methods: Lectures. Irkutsk: Irkutsk University Press, 2004. 205 p. [in Russian]
17. **Wang H.Y.** Analysis on thermal stress deformation of rotary air-preheater in a thermal power plant // Korean J. Chem. Eng., 2009. Vol. 26, P. 833–839.
18. **Wang H.Y., Zaho L.L., Xu Z.G., Chun W.G., Kim H.T.** The study on heat transfer model of tri-sectional rotary air preheater based on the semi-analytical method// Appl. Therm. Eng. 2008. Vol. 28, no. 14–15. P. 1882–1888.
19. **Zhuang Wu, Roderick V.N. Melnik, Finn Borup.** Model-based analysis and simulation of regenerative heat wheel // Energy and Buildings, 2006. No. 28. P. 502–514.

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

GROUP CLASSIFICATION FOR A GENERAL NONLINEAR MODEL OF OPTION PRICING[1]

Vladimir E. Fedorov

Laboratory of Quantum Topology, Mathematical Analysis Department, Chelyabinsk State University, Chelyabinsk, Russia, kar@csu.ru

Mikhail M. Dyshaev

Mathematical Analysis Department, Chelyabinsk State University, Chelyabinsk, Russia, MikhailDyshaev@gmail.com

Abstract: We consider a family of equations with two free functional parameters containing the classical Black–Scholes model, Schönbucher–Wilmott model, Sircar–Papanicolaou equation for option pricing as partial cases. A five-dimensional group of equivalence transformations is calculated for that family. That group is applied to a search for specifications' parameters specifications corresponding to additional symmetries of the equation. Seven pairs of specifications are found.

Key words: Nonlinear partial differential equation, Group analysis, Group of equivalency transformations, Group classification, Nonlinear Black–Scholes equation, Pricing options, Dynamic hedging, Feedback effects of hedging.

Introduction

In the paper a nonlinear model

$$u_t + \frac{w(t,x)u_{xx}}{2\left(1 - xv(u_x)u_{xx}\right)^2} + r(xu_x - u) = 0. \tag{0.1}$$

from the theory of financial markets is considered. In the case of $v \equiv 0$ it is generalized Black—Scholes equation [1], if, besides, $w(t,x) = \sigma^2 x^2$ (0.1) is the classical Black—Scholes model [2]. For arbitrary v and $w(t,x) = \sigma^2 x^2$ (0.1) is the Sircar—Papanicolaou nonlinear feedback pricing equation [1]. If v is arbitrary, $w(t,x) = \sigma^2 x^2$ and $r = 0$, it is the equilibrium pricing model or Schönbucher—Wilmott nonlinear feedback pricing model [3–6]. The last two models take into account a feedback effect of the presence of two types of traders. The programm traders are the portfolio insurers and the reference traders are the Black—Scholes uploaders.

The aim of the paper is to obtain a group classification [7] of equation (0.1) with free parameters v and w. The group of equivalence transformations [7,8] of equation (0.1) will be found. By means of this group symmetries for the equation with all specifications will be calculated. Further these results will be applied to the theory of financial markets, particularly, they will allow to calculate various exact solutions of equation (0.1).

The groups of classical Black—Scholes model and their accordance to the groups of the heat equation were found in [9]. Research of symmetries of Schönbucher—Wilmott model and of some other nonlinear pricing models was made in [10–13].

[1]The work is partially supported by Laboratory of Quantum Topology of Chelyabinsk State University (Russian Federation government grant 14.Z50.31.0020).

1. Group of the equivalence transformations

Let us find the continuous group of equivalence transformations of equation (0.1) for the applying to the search of specifications of the functions $v = v(u_x)$, $w = w(t, x)$ in the equation, that corresponds to additional symmetries for the symmetries of the kernel of principal Lie group for the equation. We rewrite equation (0.1) in the form

$$u_t + \frac{wu_{xx}}{2\left(1 - xvu_{xx}\right)^2} + r(xu_x - u) = 0, \tag{1.1}$$

where v, w are the additional variables, depending on t, x, u, u_t and u_x. Generators of a continuous group of equivalence transformations will be searched in the form $Y = \tau\partial_t + \xi\partial_x + \eta\partial_u + \mu\partial_v + \nu\partial_w$, where the functions τ, ξ, η depend on t, x, u, and μ ν depend on t, x, u, u_t, u_x, v, w. For brevity hereafter $\frac{\partial}{\partial t} \equiv \partial_t$ and similar notations are used. We add to (1.1) the equations

$$v_t = 0, \quad v_x = 0, \quad v_u = 0, \quad v_{u_t} = 0, \tag{1.2}$$

$$w_u = 0, \quad w_{u_t} = 0, \quad w_{u_x} = 0, \tag{1.3}$$

meaning that in the statement of the problem the function v depends only on u_x and the function w depends on t, x.

We consider the system of equations (1.1)–(1.3) as a manifold $\mathfrak{N}$ in an expanded space of corresponding variables. Let us act on the left-hand side of system (1.1)–(1.3) by the extended operator

$$\widetilde{Y} = Y + \varphi^t\partial_{u_t} + \varphi^{xx}\partial_{u_{xx}} + \mu^t\partial_{v_t} + \mu^x\partial_{v_x} + \mu^u\partial_{v_u} + \mu^{u_t}\partial_{v_{u_t}} + \nu^u\partial_{w_u} + \nu^{u_t}\partial_{w_{u_t}} + \nu^{u_x}\partial_{w_{u_x}},$$

we restrict a result of the action on $\mathfrak{N}$ and we obtain the equations

$$\varphi^t + \frac{vwu_{xx}^2\xi}{\left(1 - xvu_{xx}\right)^3} + \frac{w(1 + xvu_{xx})\varphi^{xx}}{2\left(1 - xvu_{xx}\right)^3} + \frac{u_{xx}\nu}{2\left(1 - xvu_{xx}\right)^2} + \frac{wxu_{xx}^2\mu}{\left(1 - xvu_{xx}\right)^3} +$$

$$+ r(u_x\xi + x\varphi^x - \eta)\Big|_{\mathfrak{N}} = 0, \tag{1.4}$$

$$\mu^t|_{\mathfrak{N}} = 0, \quad \mu^x|_{\mathfrak{N}} = 0, \quad \mu^u|_{\mathfrak{N}} = 0, \quad \mu^{u_t}|_{\mathfrak{N}} = 0, \tag{1.5}$$

$$\nu^u|_{\mathfrak{N}} = 0, \quad \nu^{u_t}|_{\mathfrak{N}} = 0, \quad \mu^{u_x}|_{\mathfrak{N}} = 0. \tag{1.6}$$

From (1.2) and (1.3) it follows that

$$\widetilde{D}_t = \partial_t + w_t\partial_w + w_{tt}\partial_{w_t} + w_{tx}\partial_{w_x} + \ldots, \quad \widetilde{D}_x = \partial_x + w_x\partial_w + w_{tx}\partial_{w_t} + w_{xx}\partial_{w_x} + \ldots,$$

$$\widetilde{D}_u = \partial_u, \quad \widetilde{D}_{u_t} = \partial_{u_t}, \quad \widetilde{D}_{u_x} = \partial_{u_x} + v'(u_x)\partial_v + v''(u_x)\partial_{v'(u_x)} + \ldots,$$

$$\mu^t = \mu_t + w_t\mu_w - v'(u_x)\varphi_t^x = \mu_t + w_t\mu_w - v'(u_x)(\eta_{tx} + u_x\eta_{tu} - u_t\tau_{tx} - u_tu_x\tau_{tu} - u_x\xi_{tx} - u_x^2\xi_{tu}),$$

$$\mu^x = \mu_x + w_x\mu_w - v'(u_x)\varphi_x^x = \mu_x + w_x\mu_w - v'(u_x)(\eta_{xx} + u_x\eta_{xu} - u_t\tau_{xx} - u_tu_x\tau_{xu} - u_x\xi_{xx} - u_x^2\xi_{xu}),$$

$$\mu^u = \mu_u - v'(u_x)\varphi_u^x = \mu_u - v'(u_x)(\eta_{xu} + u_x\eta_{uu} - u_t\tau_{xu} - u_tu_x\tau_{uu} - u_x\xi_{xu} - u_x^2\xi_{uu}),$$

$$\mu^{u_t} = \mu_{u_t} - v'(u_x)\varphi_{u_t}^x = \mu_{u_t} + v'(u_x)\tau_x + u_xv'(u_x)\tau_u,$$

$$\nu^u = \nu_u - w_t\tau_u - w_x\xi_u, \quad \nu^{u_t} = \nu_{u_t}, \quad \nu^{u_x} = \nu_{u_x} + v'(u_x)\nu_v.$$

Therefore, equations (1.5) and (1.6) have the form

$$\mu_t + w_t\mu_w - v'(u_x)(\eta_{tx} + u_x\eta_{tu} - u_t\tau_{tx} - u_tu_x\tau_{tu} - u_x\xi_{tx} - u_x^2\xi_{tu})|_{\mathfrak{N}} = 0, \tag{1.7}$$

$$\mu_x + w_x\mu_w - v'(u_x)(\eta_{xx} + u_x\eta_{xu} - u_t\tau_{xx} - u_tu_x\tau_{xu} - u_x\xi_{xx} - u_x^2\xi_{xu})|_{\mathfrak{N}} = 0, \tag{1.8}$$

$$\mu_u - v'(u_x)(\eta_{xu} + u_x\eta_{uu} - u_t\tau_{xu} - u_tu_x\tau_{uu} - u_x\xi_{xu} - u_x^2\xi_{uu})|_{\mathfrak{N}} = 0, \tag{1.9}$$

$$\mu_{u_t} + v'(u_x)\tau_x + u_xv'(u_x)\tau_u = 0, \tag{1.10}$$

$$\nu_u - w_t\tau_u - w_x\xi_u = 0, \quad \nu_{u_t} = 0, \quad \nu_{u_x} + v'(u_x)\nu_v = 0. \tag{1.11}$$

By equality (1.1) equations (1.7)–(1.9) can be rewritten in the form

$$\mu_t + w_t\mu_w - v'(u_x)\left(\eta_{tx} + u_x\eta_{tu} - u_x\xi_{tx} - u_x^2\xi_{tu} + \right.$$
$$\left. + \frac{wu_{xx}(\tau_{tx} + u_x\tau_{tu})}{2\left(1 - xvu_{xx}\right)^2} + (rxu_x - ru)(\tau_{tx} + u_x\tau_{tu})\right) = 0, \tag{1.12}$$

$$\mu_x + w_x\mu_w - v'(u_x)\left(\eta_{xx} + u_x\eta_{xu} - u_x\xi_{xx} - u_x^2\xi_{xu} + \right.$$
$$\left. + \frac{wu_{xx}(\tau_{xx} + u_x\tau_{xu})}{2\left(1 - xvu_{xx}\right)^2} + (rxu_x - ru)(\tau_{xx} + u_x\tau_{xu})\right) = 0, \tag{1.13}$$

$$\mu_u - v'(u_x)\left(\eta_{xu} + u_x\eta_{uu} - u_x\xi_{xu} - u_x^2\xi_{uu} + \right.$$
$$\left. + \frac{wu_{xx}(\tau_{xu} + u_x\tau_{uu})}{2\left(1 - xvu_{xx}\right)^2} + (rxu_x - ru)(\tau_{xu} + u_x\tau_{uu})\right) = 0. \tag{1.14}$$

By means of the equality

$$\varphi^{xx} = \eta_{xx} + 2u_x\eta_{xu} + {u_x}^2\eta_{uu} + u_{xx}\eta_u - u_t\tau_{xx} - 2u_tu_x\tau_{xu} - 2u_{tx}\tau_x - u_t{u_x}^2\tau_{uu} -$$
$$-2u_xu_{tx}\tau_u - u_tu_{xx}\tau_u - u_x\xi_{xx} - 2{u_x}^2\xi_{xu} - 2u_{xx}\xi_x - {u_x}^3\xi_{uu} - 3u_xu_{xx}\xi_u$$

equation (1.4) is rewritten as

$$\eta_t + u_t\eta_u - u_t\tau_t - u_t^2\tau_u - u_x\xi_t - u_tu_x\xi_u + \frac{1}{2\left(1 - xvu_{xx}\right)^3}\left(2vwu_{xx}^2\xi + 2xwu_{xx}^2\mu + \right.$$
$$+u_{xx}\nu - xu_{xx}^2v\nu + w(1 + xvu_{xx})(\eta_{xx} + 2u_x\eta_{xu} + {u_x}^2\eta_{uu} + u_{xx}\eta_u -$$
$$-u_t\tau_{xx} - 2u_tu_x\tau_{xu} - 2u_{tx}\tau_x - u_t{u_x}^2\tau_{uu} - 2u_xu_{tx}\tau_u -$$
$$-u_tu_{xx}\tau_u - u_x\xi_{xx} - 2{u_x}^2\xi_{xu} - 2u_{xx}\xi_x - {u_x}^3\xi_{uu} - 3u_xu_{xx}\xi_u)) +$$
$$+ru_x\xi + rx(\eta_x + u_x\eta_u - u_t\tau_x - u_tu_x\tau_u - u_x\xi_x - u_x^2\xi_u) - r\eta|_{\mathfrak{N}} =$$
$$= \eta_t + \frac{wu_{xx}(\tau_t - \eta_u)}{2\left(1 - xvu_{xx}\right)^2} + (rxu_x - ru)(\tau_t - \eta_u) - \frac{w^2u_{xx}^2\tau_u}{4\left(1 - xvu_{xx}\right)^4} -$$
$$-(rxu_x - ru)^2\tau_u + \frac{wu_{xx}(rxu_x - ru)\tau_u}{\left(1 - xvu_{xx}\right)^2} - u_x\xi_t +$$
$$+\frac{wu_xu_{xx}\xi_u}{2\left(1 - xvu_{xx}\right)^2} + (rxu_x - ru)u_x\xi_u + \frac{1}{2\left(1 - xvu_{xx}\right)^3}\left(2vwu_{xx}^2\xi + 2xwu_{xx}^2\mu + \right. \tag{1.15}$$
$$+u_{xx}\nu - xu_{xx}^2v\nu + w(1 + xvu_{xx})\left(\eta_{xx} + 2u_x\eta_{xu} + {u_x}^2\eta_{uu} + u_{xx}\eta_u + \right.$$
$$+\frac{wu_{xx}\tau_{xx}}{2\left(1 - xvu_{xx}\right)^2} + (rxu_x - ru)\tau_{xx} + \frac{wu_xu_{xx}\tau_{xu}}{\left(1 - xvu_{xx}\right)^2} + 2(rxu_x - ru)u_x\tau_{xu} -$$
$$-2u_{tx}\tau_x + \frac{wu_x^2u_{xx}\tau_{uu}}{2\left(1 - xvu_{xx}\right)^2} + (rxu_x - ru)u_x^2\tau_{uu} - 2u_xu_{tx}\tau_u + \frac{wu_{xx}^2\tau_u}{2\left(1 - xvu_{xx}\right)^2} +$$
$$+(rxu_x - ru)u_{xx}\tau_u - u_x\xi_{xx} - 2{u_x}^2\xi_{xu} - 2u_{xx}\xi_x - {u_x}^3\xi_{uu} - 3u_xu_{xx}\xi_u)) +$$
$$+ru_x\xi + rx(\eta_x + u_x\eta_u - u_x\xi_x - u_x^2\xi_u) - r\eta +$$
$$+\frac{rxwu_{xx}(\tau_x + u_x\tau_u)}{2\left(1 - xvu_{xx}\right)^2} + rx(rxu_x - ru)(\tau_x + u_x\tau_u) = 0.$$

We differentiate the last equations with respect to u_{tx} and obtain $w(1+xvu_{xx})(\tau_x+u_x\tau_u)=0$, consequently, $\tau=\tau(t)$, if $w\neq 0$. Therefore, equations (1.11)–(1.15) have the form

$$\mu_{u_t}=0,\quad \nu_u-w_x\xi_u=0,\quad \nu_{u_t}=0,\quad \nu_{u_x}+v'(u_x)\nu_v=0, \tag{1.16}$$

$$\mu_t+w_t\mu_w-v'(u_x)\left(\eta_{tx}+u_x\eta_{tu}-u_x\xi_{tx}-u_x^2\xi_{tu}\right)=0, \tag{1.17}$$

$$\mu_x+w_x\mu_w-v'(u_x)\left(\eta_{xx}+u_x\eta_{xu}-u_x\xi_{xx}-u_x^2\xi_{xu}\right)=0, \tag{1.18}$$

$$\mu_u-v'(u_x)\left(\eta_{xu}+u_x\eta_{uu}-u_x\xi_{xu}-u_x^2\xi_{uu}\right)=0, \tag{1.19}$$

$$\begin{gathered}\eta_t+\frac{wu_{xx}(\tau'(t)-\eta_u)}{2\left(1-xvu_{xx}\right)^2}+(rxu_x-ru)(\tau'(t)-\eta_u)-u_x\xi_t+\\ +\frac{wu_xu_{xx}\xi_u}{2\left(1-xvu_{xx}\right)^2}+(rxu_x-ru)u_x\xi_u+\frac{1}{2\left(1-xvu_{xx}\right)^3}\left(2vwu_{xx}^2\xi+2xwu_{xx}^2\mu+\right.\\ +u_{xx}\nu-xu_{xx}^2v\nu+w(1+xvu_{xx})\left(\eta_{xx}+2u_x\eta_{xu}+{u_x}^2\eta_{uu}+u_{xx}\eta_u-u_x\xi_{xx}-\right.\\ \left.\left.-2{u_x}^2\xi_{xu}-2u_{xx}\xi_x-{u_x}^3\xi_{uu}-3u_xu_{xx}\xi_u\right)\right)+ru_x\xi+\\ +rx(\eta_x+u_x\eta_u-u_x\xi_x-u_x^2\xi_u)-r\eta=0.\end{gathered}$$

We multiply by $2(1-xvu_{xx})^3$ the last equation, then

$$\begin{gathered}2(1-xvu_{xx})^3(\eta_t+(rxu_x-ru)(\tau'(t)-\eta_u+u_x\xi_u))+(1-xvu_{xx})wu_{xx}(\tau'(t)-\eta_u)-\\ -2(1-xvu_{xx})^3u_x\xi_t+(1-xvu_{xx})wu_xu_{xx}\xi_u+2vwu_{xx}^2\xi+2xwu_{xx}^2\mu+\\ +u_{xx}\nu-xu_{xx}^2v\nu+w(1+xvu_{xx})\left(\eta_{xx}+2u_x\eta_{xu}+{u_x}^2\eta_{uu}+u_{xx}\eta_u-u_x\xi_{xx}-\right.\\ \left.-2{u_x}^2\xi_{xu}-2u_{xx}\xi_x-{u_x}^3\xi_{uu}-3u_xu_{xx}\xi_u\right)+\\ +2(1-xvu_{xx})^3(ru_x\xi+rx(\eta_x+u_x\eta_u-u_x\xi_x-u_x^2\xi_u)-r\eta)=0.\end{gathered} \tag{1.20}$$

Equation (1.20) for the case $v\neq 0$ has at u_{xx}^3 multiplier

$$\eta_t+rxu_x\tau'(t)-ru(\tau'(t)-\eta_u+u_x\xi_u)-u_x\xi_t+ru_x\xi+rx(\eta_x-u_x\xi_x)-r\eta,$$

after its splitting with respect to u_x, we obtain two equations

$$\eta_t+rx\eta_x+ru\eta_u-r\eta-ru\tau'(t)=0, \tag{1.21}$$

$$rx\tau'(t)-\xi_t-rx\xi_x-ru\xi_u+r\xi=0. \tag{1.22}$$

After the splitting with respect to u_x of the multiplier at u_{xx} in zero degree it follows that

$$\xi=A(t,x)u+B(t,x),$$

$$\eta=A_x(t,x)u^2+C(t,x)u+D(t,x)$$

and by (1.21), (1.22)

$$\begin{gathered}2\eta_t-2ru\tau'+2ru\eta_u+2rx\eta_x-2r\eta+w\eta_{xx}=w\eta_{xx}=0,\\ 2rx\tau'-2ru\xi_u-2\xi_t+2r\xi-2rx\xi_x-2w\xi_{xx}+4w\eta_{xu}=-w\xi_{xx}+2w\eta_{xu}=0.\end{gathered} \tag{1.23}$$

The last equality implies that

$$A_{xx}=0,\quad A(t,x)=A_1(t)x+A_0(t),\quad C(t,x)=\frac{1}{2}B_x(t,x)+E(t),$$

$$\xi=A_1(t)xu+A_0(t)u+B(t,x),\quad \eta=A_1(t)u^2+\frac{1}{2}B_x(t,x)u+E(t)u+D(t,x).$$

Then from (1.23) it follows that

$$B_{xxx}=0,\quad D_{xx}=0,\quad \xi=A_1(t)xu+A_0(t)u+B_2(t)x^2+B_1(t)x+B_0(t),$$
$$\eta=A_1(t)u^2+B_2(t)xu+\frac{1}{2}B_1(t)u+E(t)u+D_1(t)x+D_0(t).$$

Now the equality (1.22) implies that $A_1(t)=Fe^{-rt}$, $A_0(t)$ is a constant,

$$B_2(t)=Ge^{-rt},\quad B_1(t)=r\tau(t)+H,\quad B_0(t)=Je^{rt},$$
$$\xi=Fe^{-rt}xu+A_0u+Ge^{-rt}x^2+r\tau(t)x+Hx+Je^{rt},$$
$$\eta=Fe^{-rt}u^2+Ge^{-rt}xu+\frac{1}{2}(r\tau(t)+H)u+E(t)u+D_1(t)x+D_0(t).$$

By (1.21) D_1 is a constant,

$$D_0(t)=Ke^{rt},\quad E(t)=\frac{1}{2}r\tau(t)+P,$$
$$\eta=Fe^{-rt}u^2+Ge^{-rt}xu+r\tau(t)u+Pu+D_1x+Ke^{rt}.$$

From (1.16) it follows that $\nu=w_x(Fe^{-rt}x+A_0)u+S(t,x,u_x,v,w)$.

The coefficient at u_{xx} in equation (1.20) is equated to zero and we obtain the equation

$$-6xv\eta_t-6xv(rxu_x\tau'(t)-ru\tau'(t)+ru\eta_u-ruu_x\xi_u)+w\tau'(t)+$$
$$+6xvu_x\xi_t+wu_x\xi_u+\nu-2w\xi_x-3wu_x\xi_u+2xvwu_x\eta_{xu}+xvwu_x^2\eta_{uu}-$$
$$-xvwu_x\xi_{xx}-2xvwu_x^2\xi_{xu}-6xv(ru_x\xi+rx\eta_x-rxu_x\xi_x-r\eta)=0.$$

Let us substitute in it the expressions for ξ, η, ν that were found before, and splitting with respect to the variable u leads to the equations

$$-2Fe^{-rt}w+w_x(Fe^{-rt}x+A_0)=0, \tag{1.24}$$
$$S=4Ge^{-rt}xw-w\tau'+2rw\tau+2Hw+2Fe^{-rt}xu_xw+2A_0u_xw.$$

The last of them implies the equalities $\nu_v=S_v=0$, consequently, by (1.16) we obtain

$$\nu_{u_x}=S_{u_x}=2Fe^{-rt}xw+2A_0xw=0,\quad A_0=F=0.$$

Thus,

$$\xi=Ge^{-rt}x^2+r\tau(t)x+Hx+Je^{rt},$$
$$\eta=Ge^{-rt}xu+r\tau(t)u+Pu+D_1x+Ke^{rt},$$
$$\nu=4Ge^{-rt}xw-w\tau'+2rw\tau+2Hw.$$

Analogous calculations are made with the coefficient at u_{xx}^2 in equation (1.20), we obtain the equation

$$6x^2v^2(\eta_t+rxu_x\tau'-ru\tau'+ru\eta_u)-xvw(\tau'-\eta_u)-6x^2v^2u_x\xi_t+2vw\xi+2xw\mu-xv\nu+$$
$$+xvw(\eta_u-2\xi_x)+6x^2v^2(ru_x\xi+rx\eta_x-rxu_x\xi_x-r\eta)=0,$$

that implies the equality $\mu=v\left(H-P-\frac{J}{x}e^{rt}+2Ge^{-rt}x\right)$. Therefore $\mu_u=\mu_w=0$, and for the case $v'\neq 0$ obtain $G=0$ from equation (1.19). Then equation (1.18) implies that $\mu_x=0$, hence $J=0$. From equation (1.17) it follows that $\mu_t=0$, it corresponds to the resulting formula $\mu=(H-P)v$. Thus, $\tau(t)$ is an arbitrary function,

$$\xi=Hx+r\tau(t)x,\quad \eta=Ke^{rt}+D_1x+Pu+r\tau(t)u,$$
$$\mu=(H-P)v,\quad \nu=2Hw+(2r\tau(t)-\tau'(t))w.$$

Let us formulate the result in the form of theorem.

Theorem 1. *The Lie algebra of infinitesimal generators of the equivalency transformations groups for equation* (0.1)*, is generated by operators*

$$Y_1 = x\partial_u, \quad Y_2 = e^{rt}\partial_u, \quad Y_3 = x\partial_x + u\partial_u + 2w\partial_w, \quad Y_4 = x\partial_x + v\partial_v + 2w\partial_w,$$

$$Y_5 = \tau(t)\partial_t + r\tau(t)x\partial_x + r\tau(t)u\partial_u + (2r\tau(t) - \tau'(t))w\partial_w,$$

when v', w are identically unequal to zero.

Remark 1. It is easy to check that the infinitely-dimensional part of the Lie algebra from Theorem 1 consists of operators of the form Y_5 only.

The extensions of the operators Y_k, $k = 1, 2, 3, 4, 5$, are

$$\begin{gathered} \widetilde{Y}_1 = x\partial_u + \partial_{u_x}, \quad \widetilde{Y}_2 = e^{rt}\partial_u, \quad \widetilde{Y}_3 = x\partial_x + u\partial_u + 2w\partial_w, \\ \widetilde{Y}_4 = x\partial_x + v\partial_v + 2w\partial_w - u_x\partial_{u_x}, \quad \widetilde{Y}_5 = \tau\partial_t + r\tau x\partial_x + r\tau u\partial_u + (2r\tau - \tau')w\partial_w. \end{gathered} \tag{1.25}$$

Therefore, the kernel of the principal Lie algebras for equation (0.1) is one-dimensional with the basis Y_2, because the corresponding group only doesn't transform the additional variables v, w and their arguments t, x, u_x.

Corollary 1. *The kernel of the principal Lie algebras for equation* (0.1) *is spanned by the operator $X_1 = e^{rt}\partial_u$ when v', w are identically unequal to zero.*

2. Group classification

Consider Lie algebra of projections of operators (1.25) on the subspace of the variables t, x, u_x, v, w, i. e. the algebra generated by

$$\begin{gathered} Z_1 = \partial_{u_x}, \quad Z_2 = v\partial_v - u_x\partial_{u_x}, \\ Z_3 = x\partial_x + 2w\partial_w, \quad Z_4 = \tau\partial_t + r\tau x\partial_x + (2r\tau - \tau')w\partial_w. \end{gathered} \tag{2.1}$$

It is the direct sum of subalgebras $\langle Z_1, Z_2\rangle$ and $\langle Z_3, Z_4\rangle$ that corresponds to two different functions v and w and their different arguments. Therefore, the subalgebras can be considered separately.

Nonzero structure constants of $\langle Z_1, Z_2\rangle$ are $c^1_{12} = -1$, $c^1_{21} = 1$. Therefore, the inner automorphisms are $E_1 : \bar{e}^1 = e^1 - e^2 a_1$, $E_2 : \bar{e}^1 = e^1 e^{a_2}$. Here e^i, $i = 1, 2$ are the coefficients at Z_i respectively in the basis decomposition of Z. If $e^2 \neq 0$, then $e^1 = 0$ by the acting of E_1. Therefore the optimal system of one-dimensional subalgebras consists of subalgebras with bases Z_1 and Z_2.

In the subalgebra $\langle Z_3, Z_4\rangle$ there are no nontrivial inner automorphisms, consequently, the optimal system of one-dimensional subalgebras has a form $\Theta_1 = \{\langle Z_2\rangle, \langle bZ_2 + Z_4\rangle, b \in \mathbb{R}\}$.

For operators Z from optimal systems we calculate the expressions

$$Z(V(u_x) - v)|_{v=V} = 0, \quad Z(W(t, x) - w)|_{w=W} = 0.$$

Note, that if Z contains Z_1 with a nonzero coefficient and doesn't contain Z_2, then $v' = 0$. Such case doesn't correspond to the conditions of Theorem 1. If an operator Z has nonzero coefficients at Z_1 and at Z_3, then by E_1 the coefficient at Z_1 can be equated to zero for equivalent operator to Z. Therefore, the operator Z_1 can be excluded from further considerations.

We have

$$Z_2(V(u_x) - v)|_{w=W} = -V - u_x V' = 0, \quad V = \beta/u_x$$

for arbitrary $\beta \in \mathbb{R}$. Further,

$$Z_3(W(t,x)-w)|_{w=W} = xW_x - 2W = 0, \quad W = D(t)x^2$$

for arbitrary function $D(t)$. Finally,

$$(bZ_3 + Z_4(W(t,x)-w)|_{w=W} = \tau(t)W_t + (r\tau(t)+b)xW_x - (2r\tau(t) - \tau'(t) + 2b)W = 0,$$

$$W = \frac{e^{2rt+2b\int \frac{dt}{\tau(t)}}}{\tau(t)}\varphi(xe^{-rt-b\int \frac{dt}{\tau(t)}})$$

for arbitrary functions $\varphi \neq 0$, $\tau \neq 0$.

Optimal system of two-dimensional subalgebras consists of $\langle Z_2, Z_3\rangle$, $\langle Z_2, bZ_3 + Z_4\rangle$, $\langle Z_3, Z_4\rangle$. In the first two cases we have the simultaneous specifications for v and w that are already known. In the last one specification we have the form $W = \gamma x^2/\tau(t)$.

For the Lie algebra $\langle Z_2, Z_3, Z_4\rangle$ the specifications are $V = \beta/u_x$, $W = \gamma x^2/\tau(t)$.

For every basis operator from the optimal systems calculate the projection of the corresponding generator of the group of equivalency transformations on the space of the variables t, x, u. Then Z_2 corresponds to $\mathrm{pr}_{(t,x,u)}(Y_4 - Y_3) = -u\partial_u$, for the operator Z_3 it will be $\mathrm{pr}_{(t,x,u)}Y_3 = x\partial_x + u\partial_u$, and $\mathrm{pr}_{(t,x,u)}(bY_3 + Y_5) = \tau(t)\partial_t + (r\tau(t)+b)x\partial_x + (r\tau(t)u + b)\partial_u$ corresponds to $bZ_3 + Z_4$. It implies the next theorem.

Theorem 2. *Let v', w be identically unequal to zero i. Then next assertions are true.*

1. *The principal Lie algebra of the equation*

$$u_t + \frac{w(t,x)u_{xx}}{2\left(1 - \beta x\dfrac{u_{xx}}{u_x}\right)^2} + r(xu_x - u) = 0, \quad \beta \neq 0,$$

is generated by the operators $X_1 = e^{rt}\partial_u$, $X_2 = u\partial_u$.

2. *The principal Lie algebra of the equation*

$$u_t + \frac{T'(t)e^{2rt+2bT(t)}\varphi(xe^{-rt-bT(t)})u_{xx}}{2\left(1 - xv(u_x)u_{xx}\right)^2} + r(xu_x - u) = 0, \quad T'(t) \neq 0, \quad \varphi(z) \neq 0,$$

is generated by the operators $X_1 = e^{rt}\partial_u$, $X_2 = \dfrac{1}{T'(t)}\partial_t + \left(\dfrac{r}{T'(t)} + b\right)x\partial_x + \left(\dfrac{r}{T'(t)} + b\right)u\partial_u$.

3. *The principal Lie algebra of the equation*

$$u_t + \frac{T'(t)e^{2rt+2bT(t)}\varphi(xe^{-rt-bT(t)})u_{xx}}{2\left(1 - \beta x\dfrac{u_{xx}}{u_x}\right)^2} + r(xu_x - u) = 0, \quad T'(t) \neq 0, \quad \varphi(z) \neq 0, \quad \beta \neq 0,$$

is generated by the operators

$$X_1 = e^{rt}\partial_u, \quad X_2 = u\partial_u, \quad X_3 = \frac{1}{T'(t)}\partial_t + \left(\frac{r}{T'(t)} + b\right)x\partial_x + \left(\frac{r}{T'(t)} + b\right)u\partial_u.$$

4. *The principal Lie algebra of the equation*

$$u_t + \frac{D(t)x^2u_{xx}}{2\left(1 - xv(u_x)u_{xx}\right)^2} + r(xu_x - u) = 0, \quad D(t) \neq 0,$$

is generated by the operators

$$X_1 = e^{rt}\partial_u, \quad X_2 = x\partial_x + u\partial_u, \quad X_3 = \frac{1}{D(t)}\partial_t + \frac{r}{D(t)}x\partial_x + \frac{r}{D(t)}u\partial_u.$$

5. *The principal Lie algebra of the equation*

$$u_t + \frac{D(t)x^2u_{xx}}{2\left(1 - \beta x\dfrac{u_{xx}}{u_x}\right)^2} + r(xu_x - u) = 0, \quad D(t) \neq 0,$$

is generated by the operators

$$X_1 = e^{rt}\partial_u, \quad X_2 = x\partial_x, \quad X_3 = u\partial_u, \quad X_4 = \frac{1}{D(t)}\partial_t + \frac{r}{D(t)}x\partial_x + \frac{r}{D(t)}u\partial_u.$$

Remark 2. Theorem 1 and Theorem 2 are valid for the case $r = 0$.

3. Conclusion

Further Theorem 2 will be applied to the search of exact solutions of the option pricing nonlinear models. Specification $W(t, x) = \sigma^2x^2$ as partial case of $D(t)x^2$ corresponds to the Scönbucher—Wilmott model, if $r = 0$, and to Circar—Papanicolaou model for $r \neq 0$.

REFERENCES

1. **Sircar K.R., Papanicolaou G.** General Black–Scholes models accounting for increased market volatility from hedging strategies // Appl. Math. Finance, 1998. Vol. 5. P. 45–82.
2. **Black F., Scholes M.** The pricing of options and corporate liabilities // J. of Political Economy, 1973. Vol. 81. P. 637–659.
3. **Frey R., Stremme A.** Market volatility and feedback effects from dynamic hedging // Mathematical Finance, 1997. Vol. 7, no. 4. P. 351–374.
4. **Frey R.** Perfect option replication for a large trader // Finance and Stochastics. 1998. Vol. 2. P. 115–148.
5. **Jarrow R.A.** Derivative securities markets, market manipulation and option pricing theory // J. of Financial and Quantitative Analysis, 1994. Vol. 29. P. 241–261.
6. **Schönbucher P., Wilmott P.** The feedback-effect of hedging in illiquid markets // SIAM J. on Applied Mathematics, 2000. Vol. 61. P. 232–272.
7. **Ovsyannikov L.V.** Group analysis of differential equations. New York: Academic press, 1982.
8. **Chirkunov Yu.A., Khabirov S.V.** Elements of symmetry analysis for differential equations of continuum mechanics. Novosibirsk: Novosibirsk State Technical University, 2012. 659 p. [In Russian]
9. **Gazizov R.K., Ibragimov N.H.** Lie symmetry analysis of differential equations in finance // Nonlinear Dynamics, 1998. Vol. 17. P. 387–407.
10. **Bordag L.A., Chmakova A.Y.** Explicit solutions for a nonlinear model of financial derivatives // International Journal of Theoretical and Applied Finance, 2007. Vol. 10, no. 1. P. 1–21.
11. **Bordag L.A.** On option-valuation in illiquid markets: invariant solutions to a nonlinear model. Mathematical Control Theory and Finance, eds. A. Sarychev, A. Shiryaev, M. Guerra and M. R. Grossinho. Springer, 2008. P. 71–94.
12. **Mikaelyan A.** Analytical study of the Schönbucher–Wilmott model of the feedback effect in illiquid markets. Master's thesis in financial mathematics, Halmstad: Halmstad University, 2009. viii+67 p.
13. **Bordag L.A., Mikaelyan A.** Models of self-financing hedging strategies in illiquid markets: symmetry reductions and exact solutions // J. Letters in Mathematical Physics. 2011. Vol. 96, no. 1–3. P. 191–207.

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

PARALLEL ALGORITHM FOR CALCULATING GENERAL EQUILIBRIUM IN MULTIREGION ECONOMIC GROWTH MODELS

Nikolai B. Melnikov

Lomonosov Moscow State University;
Central Economics and Mathematics Institute, RAS, Moscow, Russia
melnikov@cs.msu.ru

Arseniy P. Gruzdev

Lomonosov Moscow State University, Moscow, Russia
gruzdev@cs.msu.ru

Michael G. Dalton

National Oceanic and Atmospheric Administration, Seattle WA, USA
michael.dalton@noaa.gov

Brian C. O'Neill

National Center for Atmospheric Research, Boulder CO, USA
Email: boneill@ucar.edu

Abstract: We develop and analyze a parallel algorithm for computing a solution in a multiregion dynamic general equilibrium model. The algorithm is based on an iterative method of the Gauss – Seidel type and exploits a special block structure of the model. Calculation of prices and input-output ratios in production for different time steps is carried out in parallel. We implement the parallel algorithm using the OpenMP interface for systems with shared memory. The efficiency of the algorithm is studied with the numbers of cores varying in the full range from one to the number of time steps of the model.

Key words: Computable general equilibrium, Economic growth, Iterative methods, High-performance computing, OpenMP.

AMS Classification: 91B50, 91B62, 91B66, 91B74, 68W10

1. Introduction

Dynamic computable general equilibrium (CGE) models are widely used for estimating the effects of demographic and technological changes on energy use and carbon dioxide (CO_2) emissions. The equilibrium is described in the framework of the Arrow – Debreu theory, which leads to a systems of nonlinear equations. Usually large-scale nonlinear systems are solved by one of the "general-purpose" Krylov subspace solvers, which can deal effectively with sparse matrices (see, e.g., [1]).

In our paper [2], we presented a parallel algorithm based on an iterative method of the Gauss – Seidel type [3]. We exploited the special block structure of the nonlinear system of equations in dynamic CGE models. We implemented the algorithm using parallel programming environments for the one-region version of the Population-Environmental-Technology (PET) model [4, 5]. The

numerical results showed that the speed of our algorithm is comparable to the one of Krylov methods solvers such as NITSOL [6].

In this paper we extend the algorithm to models with international trade and apply it to the multiregion PET model [7]. We implement the parallel algorithm using the OpenMP interface for systems with shared memory. To demonstrate the effectiveness of the parallel algorithm we use the PET model calibrated to reproduce major outcomes for the socioeconomic scenarios from the Shared Socioeconomic Pathways (SSP) database (see, e.g. [8]). The calibration of the PET model to the SSPs is described in the supplementary material to [8].

The paper is organized as follows. In Sect. 2 we present a description of the multiregion PET model. In particular, we explain in detail how the intermediate goods demand is calculated in the presence of the international trade. In Sect. 3 we present the numerical method for calculating the equilibrium and explain the parallel algorithm. In Sect. 4 we discuss the calculation results.

2. Structure of the CGE model

In this section we describe the multiregion PET model (for description of the one-region PET model, see, e.g. [2, 4, 5]).

The PET model is a forward-looking CGE model with tree types of agents: consumers, producers, and government. Consumers maximize their lifetime utility function taking prices as given (Subsec. 2.1). Producers maximize profits supported by the prices as described in Subsec. 2.2. Government redistributes capital through taxes and transfers (for details see, e.g. [5]). International trade is described by the Armington model as described in Subsec. 2.3. Prices are determined by the markets clearing conditions for production factors, intermediate and final goods (Subsec. 2.4). The first-order optimality conditions for the agents and supply-equals-demand conditions for markets form a system of nonlinear equations. A solution to this system of equation is called the *general equilibrium*.

2.1. Consumers side

In each of the N_R regions the utility function of the representative household is given by the discounted lifetime consumption

$$U(c) = \frac{1}{\psi} \sum_{t=0}^{\infty} \beta^t n_t \left(\sum_{j=1}^{N_C} (\mu_{jt} c_{jt})^{\rho} \right)^{\frac{\psi}{\rho}},$$

where $t = 0, 1, 2, \ldots$ is time, index $j = \overline{1, N_C}$ labels consumer goods, c_{jt} is consumption, n_t is the size of population, $\psi \in (-\infty, 1) \setminus \{0\}$ is the intertemporal substitution parameter, $\beta \in (0, 1)$ is the discount rate, $\sigma = 1/(1-\rho)$ is the electivity ($\rho \in (-\infty, 1) \setminus \{0\}$ is the substitution parameter) and μ_{jt} is the preference coefficient (for details of calculating μ_{jt} see, e.g., [5]).

The capital dynamics is

$$(1 + \nu_t)\, k_{t+1} = (1 - \delta)\, k_t + x_t, \tag{2.1}$$

where k_t is capital ($k_0 > 0$), x_t is investment, $\delta \in (0, 1)$ is the capital depreciation coefficient, $1 + \nu_t = n_{t+1}/n_t$ is the population growth coefficient (ν_t is the growth rate).

The budget constraint is

$$\sum_{j=1}^{N_C} p_{jt} c_{jt} + q_t x_t = (1 - \theta_t)\, \omega_t l_t + (1 - \phi_t)\, r_t k_t + g_t, \tag{2.2}$$

where p_{jt} is the price of the jth consumer good, q_t is the prices of investments, ω_t is the wage rate, r_t is the rental rate of capital, g_t is the government transfers, l_t is the labor supply, θ_t and ϕ_t are the tax rates on capital and labor incomes, respectively. Here the quantities c_{jt}, k_t, x_t, l_t and g_t are given in *per capita* terms.

Taking prices as given, the representative household maximizes the utility,

$$U(c) \to \max, \tag{2.3}$$

subject to constraints (2.1) and (2.2). The first-order optimality condition for problem (2.1), (2.2) and (2.3) gives the Euler equation

$$\frac{q_t}{\overline{p}_t}\overline{c}_t{}^{\psi-1} = \beta\frac{q_{t+1}(1-\delta)+(1-\phi_{t+1})r_{t+1}}{\overline{p}_{t+1}}\overline{c}_{t+1}{}^{\psi-1},$$

where the consumption composite and price index are

$$\overline{c}_t = \left[\sum_{j=1}^{N_C}(\mu_{jt}c_{jt})^{\rho}\right]^{\frac{1}{\rho}}, \quad \bar{p}_t = \left[\sum_{j=1}^{N_C}\left(\frac{p_{jt}}{\mu_{jt}}\right)^{\frac{\rho}{\rho-1}}\right]^{\frac{\rho-1}{\rho}}$$

such that

$$\sum_{j=1}^{N_C} p_{jt}c_{jt} = \bar{p}_t\overline{c}_t.$$

The transversality conditions

$$\lim_{t\to\infty}\lambda_t k_t = 0,$$

where λ_t is the Lagrange multiplier, guarantees that the optimal trajectory (c_t, k_t, x_t) exists and is unique (see, e.g., [9]).

2.2. Producers side

Firms are aggregated into sectors that produce final goods (N_C consumer goods and one "investment good") and intermediate goods (N_E energy goods and the rest, which we call materials). The total number of production sector is $N_X = N_C + 1 + N_E + 1$.

Production level of the good X is defined by the constant elasticity of substitution (CES) function

$$X = \gamma_X\left(\alpha_K(G_K K)^{\rho_X} + \alpha_L(G_L L)^{\rho_X} + \alpha_{\bar{E}}(G_{\bar{E}}\bar{E})^{\rho_X} + \alpha_{\widehat{M}}(G_{\widehat{M}}\widehat{M})^{\rho_X}\right)^{\frac{1}{\rho_X}}, \tag{2.4}$$

where K is capital, L is labor, $\bar{E}$ is energy composite and $\widehat{M}$ is materials (unlike small letters that indicate the *per capita* values, capital letters denote the totals). Here G_I, $I = K, L, \bar{E}, \widehat{M}$, are the productivity factors and the coefficient γ_X normalizes the production shares α_I to unity. Both productivity factors and production shares can be sector- and time-dependent. (Current version of the PET model [8] also has land as a production factor but, for simplicity, we do not consider it here.)

At each time moment, the producer of the good X maximizes profit, or equivalently, minimizes costs

$$P_K K + P_L L + P_{\bar{E}}\bar{E} + (1+\tau_{\widehat{M}})P_{\widehat{M}}\widehat{M} \to \min_{K,L,\bar{E},\widehat{M}}, \tag{2.5}$$

given the level of production (2.4). Here P_I is the corresponding price and $\tau_{\widehat{M}}$ is the tax on the use of materials (for brevity, we omit the time index).

The minimal cost for problem (2.4) and (2.5) is given by $P_X X$, where

$$P_X = \frac{1}{\gamma_X}\left(\alpha_K^{\frac{1}{1-\rho_X}}\left(\frac{P_K}{G_K}\right)^{\frac{\rho_X}{\rho_X-1}} + \alpha_L^{\frac{1}{1-\rho_X}}\left(\frac{P_L}{G_L}\right)^{\frac{\rho_X}{\rho_X-1}} + \alpha_{\bar{E}}^{\frac{1}{1-\rho_X}}\left(\frac{P_{\bar{E}}}{G_{\bar{E}}}\right)^{\frac{\rho_X}{\rho_X-1}} + \alpha_{\widehat{M}}^{\frac{1}{1-\rho_X}}\left(\frac{(1+\tau_{\widehat{M}})P_{\widehat{M}}}{G_{\widehat{M}}}\right)^{\frac{\rho_X}{\rho_X-1}}\right).$$

The cost minimizing input-output ratios $A_X^I = I/X$ for $I = K, L, \bar{E}$ are given by

$$A_X^I = \left(\frac{1}{\alpha_I(\gamma_X G_I)^{\rho_X}}\frac{P_I}{P_X}\right)^{\frac{1}{\rho_X-1}},$$

and for $I = \widehat{M}$ the ratio is given by

$$A_X^{\widehat{M}} = \left(\frac{1}{\alpha_{\widehat{M}}(\gamma_X G_{\widehat{M}})^{\rho_X}}\frac{(1+\tau_{\widehat{M}})P_{\widehat{M}}}{P_X}\right)^{\frac{1}{\rho_X-1}}.$$

Since the PET model is primarily intended for energy economics analysis, it is detailed in the energy sector,

$$\bar{E} = \gamma_{\bar{E}}\left(\sum_{i=1}^{N_E}\alpha_{E_i}(G_{E_i}E_i)^{\rho_{\bar{E}}}\right)^{\frac{1}{\rho_{\bar{E}}}}, \tag{2.6}$$

where E_i, $i = \overline{1, N_E}$ are different energy types. Solving the cost–minimization problem

$$\sum_{i=1}^{N_E}(1+\tau_{E_i})P_{E_i}E_i \to \min_{E_i}$$

given the level of production (2.6), we derive the price of the energy composite,

$$P_{\bar{E}} = \frac{1}{\gamma_{\bar{E}}}\left(\sum_{i=1}^{N_E}\alpha_{E_i}^{\frac{1}{1-\rho_{\bar{E}}}}\left(\frac{(1+\tau_{E_i})P_{E_i}}{G_{E_i}}\right)^{\frac{\rho_{\bar{E}}}{\rho_{\bar{E}}-1}}\right)^{\frac{\rho_{\bar{E}}-1}{\rho_{\bar{E}}}}$$

and the input-output ratios $A_{\bar{E}}^{E_i} = E_i/\bar{E}$,

$$A_{\bar{E}}^{E_i} = \left(\frac{1}{\alpha_{E_i}(\gamma_{\bar{E}}G_{E_i})^{\rho_{\bar{E}}}}\frac{(1+\tau_{E_i})P_{E_i}}{P_{\bar{E}}}\right)^{\frac{1}{\rho_{\bar{E}}-1}},$$

where τ_{E_i}, $i = 1, \ldots, N_E$, are the taxes on the use of energy.

2.3. Intermediate goods demand

Production has a nested structure. Therefore, calculation of the intermediate goods demand requires a recursive procedure. We derive the necessary formulae first for the one–region model and then for the multiregion case.

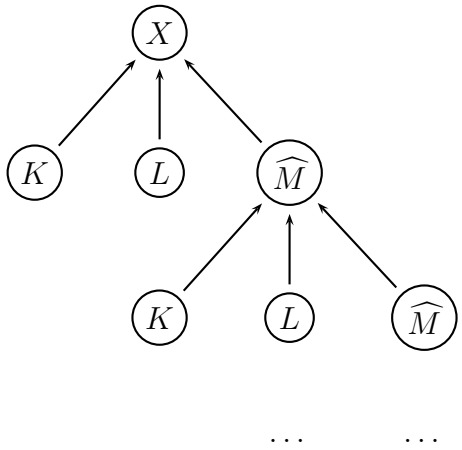


Figure 1. Nested production structure with one intermediate good.

2.3.1. One-region case

To explain the main idea, we first consider the market in which intermediate goods are aggregated into one good, which we call materials $\widehat{M}$. In this case, according to the nested production structure shown in Fig. 1, demand for materials is given by

$$\widehat{M} = A_X^{\widehat{M}} X + A_{\widehat{M}}^{\widehat{M}}(A_X^{\widehat{M}} X) + \left(A_{\widehat{M}}^{\widehat{M}}\right)^2 (A_X^{\widehat{M}} X) + \dots,$$

where the first term corresponds to the portion of materials used in production of the final good $X(K, L, \widehat{M})$, the second term corresponds to the portions of materials used in production of materials $\widehat{M}(K, L, \widehat{M})$ one level down, etc. Calculating the sum of the geometric series, we obtain

$$\widehat{M} = \left(1 - A_{\widehat{M}}^{\widehat{M}}\right)^{-1} A_X^{\widehat{M}} X \tag{2.7}$$

or, equivalently,

$$\widehat{M} = A_X^{\widehat{M}} X + A_{\widehat{M}}^{\widehat{M}} \widehat{M}. \tag{2.8}$$

The latter means that demand for materials is equal to the amount of materials needed to produce the final good and amount needed to produce the materials themselves. Denoting $Z = \widehat{M}$, $A = A_{\widehat{M}}^{\widehat{M}}$ and $Y = A_X^{\widehat{M}} X$, we write (2.7) as

$$Z = (1 - A)^{-1} Y. \tag{2.9}$$

Next, we consider the production (2.4) with two intermediate goods, energy and materials. In this case, the aggregate demand for materials is given by

$$\begin{aligned} \widehat{M} \;=\; & A_X^{\widehat{M}} X + \left[A_{\widehat{M}}^{\widehat{M}} A_X^{\widehat{M}} X + A_{\widehat{E}}^{\widehat{M}} A_X^{\widehat{E}} X\right] \\ & + \left[A_{\widehat{M}}^{\widehat{M}} \left(A_{\widehat{M}}^{\widehat{M}} A_X^{\widehat{M}} X + A_{\widehat{E}}^{\widehat{M}} A_X^{\widehat{E}} X\right) + A_{\widehat{E}}^{\widehat{M}} \left(A_{\widehat{M}}^{\widehat{E}} A_X^{\widehat{M}} X + A_{\widehat{E}}^{\widehat{E}} A_X^{\widehat{E}} X\right)\right] + \dots \end{aligned}$$

This formula describes the sum over layer of the nested production structure (Fig. 2). Each expression in square brackets corresponds to a particular layer. Rearrangement of the terms in square brackets gives

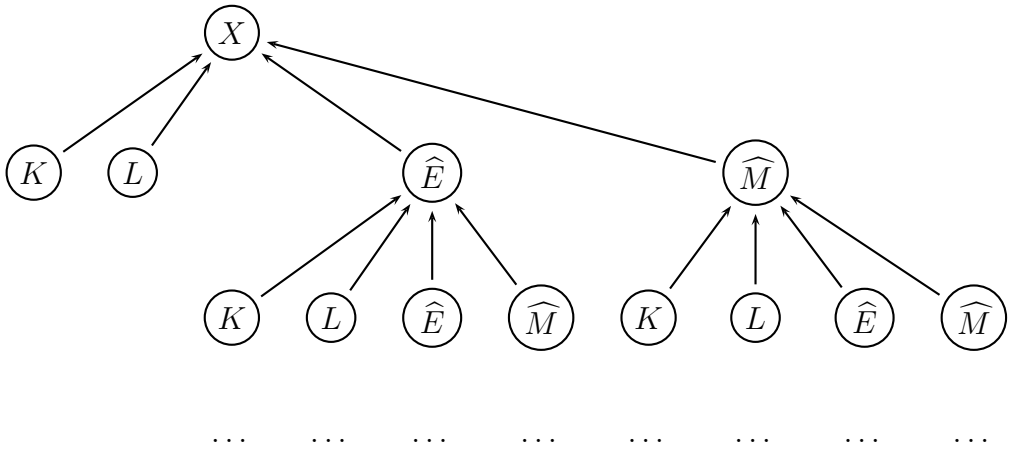


Figure 2. Nested production structure with energy and materials.

$$\begin{aligned}\widehat{M} = A_X^{\widehat{M}}X + \left[A_{\widehat{M}}^{\widehat{M}}A_X^{\widehat{M}}X + A_{\widehat{E}}^{\widehat{M}}A_X^{\widehat{E}}X\right] \\ + \left[\left(A_{\widehat{M}}^{\widehat{M}}A_{\widehat{M}}^{\widehat{M}} + A_{\widehat{E}}^{\widehat{M}}A_{\widehat{M}}^{\widehat{E}}\right)A_X^{\widehat{M}}X + \left(A_{\widehat{M}}^{\widehat{M}}A_{\widehat{E}}^{\widehat{M}} + A_{\widehat{E}}^{\widehat{M}}A_{\widehat{E}}^{\widehat{E}}\right)A_X^{\widehat{E}}X\right] + \ldots\end{aligned} \tag{2.10}$$

Similarly, for energy we obtain

$$\begin{aligned}\widehat{E} = A_X^{\widehat{E}}X + \left[A_{\widehat{M}}^{\widehat{E}}A_X^{\widehat{M}}X + A_{\widehat{E}}^{\widehat{E}}A_X^{\widehat{E}}X\right] \\ + \left[\left(A_{\widehat{M}}^{\widehat{E}}A_{\widehat{M}}^{\widehat{M}} + A_{\widehat{E}}^{\widehat{E}}A_{\widehat{M}}^{\widehat{E}}\right)A_X^{\widehat{M}}X + \left(A_{\widehat{M}}^{\widehat{E}}A_{\widehat{E}}^{\widehat{M}} + A_{\widehat{E}}^{\widehat{E}}A_{\widehat{E}}^{\widehat{E}}\right)A_X^{\widehat{E}}X\right] + \ldots\end{aligned} \tag{2.11}$$

Defining $y = (A_X^{\widehat{M}}X,\ A_X^{\widehat{E}}X)^T$ and

$$A = \begin{pmatrix} A_{\widehat{M}}^{\widehat{M}} & A_{\widehat{E}}^{\widehat{M}} \\ A_{\widehat{M}}^{\widehat{E}} & A_{\widehat{E}}^{\widehat{E}} \end{pmatrix},$$

we write expressions (2.10) and (2.11) as a matrix series:

$$\begin{pmatrix} \widehat{M} \\ \widehat{E} \end{pmatrix} = (I + A + A^2 + A^3 + \ldots)\begin{pmatrix} A_X^{\widehat{M}}X \\ A_X^{\widehat{E}}X \end{pmatrix},$$

where I is the unity 2×2-matrix. Summing the series, we have

$$\begin{pmatrix} \widehat{M} \\ \widehat{E} \end{pmatrix} = (I - A)^{-1}\begin{pmatrix} A_X^{\widehat{M}}X \\ A_X^{\widehat{E}}X \end{pmatrix}. \tag{2.12}$$

Equation (2.12) can be written in the form

$$\mathbf{Z} = (I - A)^{-1}\,\mathbf{Y}, \tag{2.13}$$

where

$$\mathbf{Z} = \begin{pmatrix} \widehat{M} \\ \widehat{E} \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} A_X^{\widehat{M}}X \\ A_X^{\widehat{E}}X \end{pmatrix}.$$

Note that equation (2.13) is the same as equation (2.9) we obtained with one intermediate good. It is the dimensionality of this equation and form of the vectors $\mathbf{Z}$ and $\mathbf{Y}$ and matrix A that change when we change the number of intermediate goods.

2.3.2. Multiregion case

In this subsection we obtain the intermediate goods demand in the multiregion economy with trade.

International trade is described by the Armington model (see, e.g. [10]). It is based on the assumption that the same goods produced in different regions are not perfect substitutes but can be aggregated according a certain rule (usually a CES function). The Armington model enables the representation of markets in which domestically produced goods keep a share of domestic markets even though their price is higher than the price in other regions, and in which different exporters co-exist even if they have different prices.

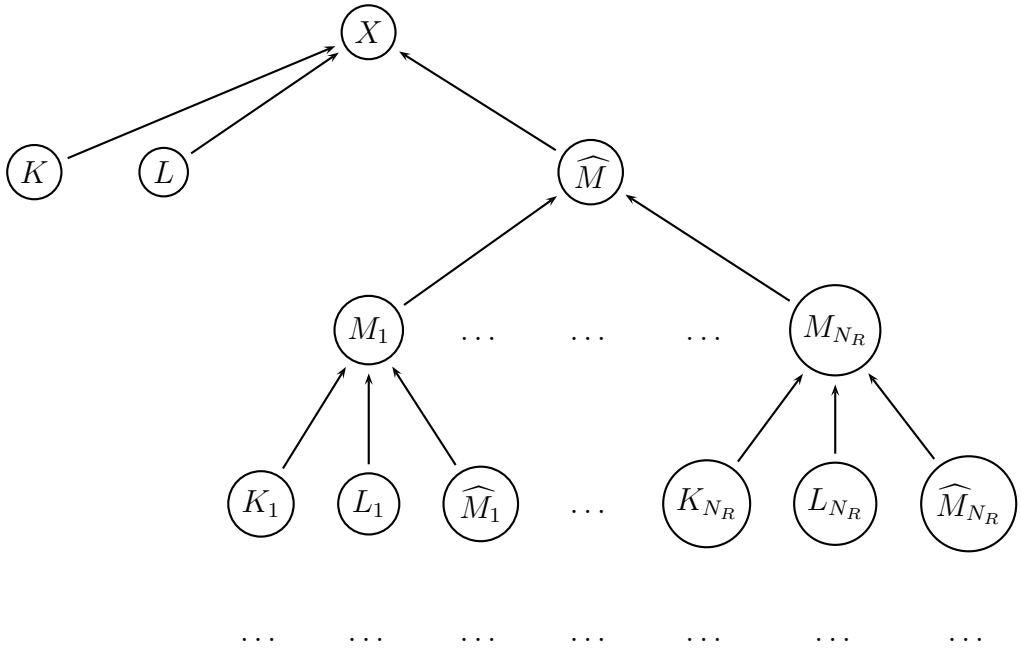


Figure 3. Nested production structure with one intermediate good for the multiregion case.

Same as in the previous subsection, first we consider the market with only one intermediate good (Fig. 3). Then $\widehat{M}(M_1, \ldots, M_{N_R})$ aggregates materials $M_1, \ldots, M_{N_R}$ from N_R regions (Fig. 4).

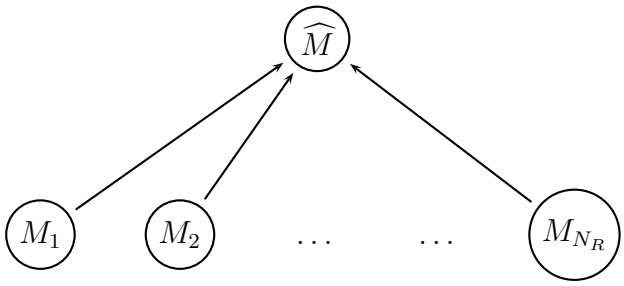


Figure 4. Armington trade structure for materials.

Similarly to the problem (2.5) and (2.4), we consider

$$P_1 M_1 + \ldots + P_{N_R} M_{N_R} \to \min,$$

subject to

$$\gamma_{\widehat{M}} \left(\sum_{i=1}^{N_R} \alpha_i M_i^{\rho_{\widehat{M}}} \right)^{\frac{1}{\rho_{\widehat{M}}}} = \widehat{M},$$

where $P_1, \ldots, P_{N_R}$ are the export prices. Then the minimum of the cost function is equal to $P_{\widehat{M}}\widehat{M}$, where

$$P_{\widehat{M}} = \frac{1}{\gamma_{\widehat{M}}} \sum_{i=1}^{N_R} \left(\alpha_i^{\frac{1}{1-\rho_{\widehat{M}}}} P_i^{\frac{\rho_{\widehat{M}}}{\rho_{\widehat{M}}-1}} \right).$$

The cost minimizing input-output ratios are given by

$$b_{\widehat{M}}^{M_i} = \frac{M_i}{\widehat{M}} = \left(\frac{1}{\alpha_I \gamma^{\rho_{\widehat{M}}}} \frac{P_i}{P_{\widehat{M}}} \right)^{\frac{1}{\rho_{\widehat{M}}-1}}.$$

Similarly to relation (2.8), we have

$$\widehat{M}_i = \sum_{j=1}^{N_R} b_{ij}^M \left(A_{X_j}^{\widehat{M}_j} X_j + A_{\widehat{M}_j}^{\widehat{M}_j} \widehat{M}_j \right).$$

Denoting $B = \left(b_{ij}^M\right)$, $A^X = \operatorname{diag}\left(A_{X_i}^{\widehat{M}_i}\right)$ and $A = \operatorname{diag}\left(A_{\widehat{M}_i}^{\widehat{M}_i}\right)$, we write

$$\mathbf{Z} = (I - BA)^{-1} B\mathbf{Y} \tag{2.14}$$

where we set $\mathbf{Z} = (\widehat{M}_1, \ldots, \widehat{M}_{N_R})^T$, $\mathbf{X} = (X_1, \ldots, X_{N_R})^T$, $\mathbf{Y} = A^X\mathbf{X}$ and I is the unity $N_R \times N_R$-matrix.

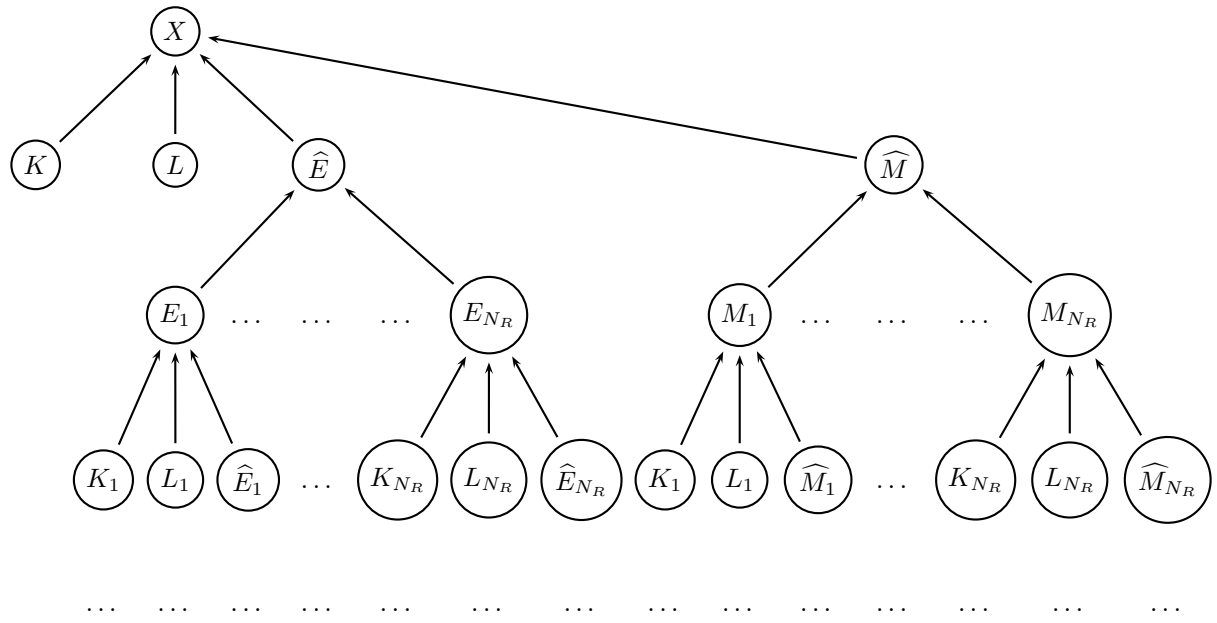


Figure 5. Nested production structure with energy and materials in the multiregion case.

In the case of production (2.4) with two intermediate goods, energy and materials (Fig. 5), the vector $\mathbf{Z}$ has the form

$$\mathbf{Z} = (\widehat{M}_1, \widehat{E}_1, \ldots, \widehat{M}_{N_R}, \widehat{E}_{N_R})^T$$

and the components of equation (2.14) will have the block structure

$$B = \left(B_j^i\right), \quad B_j^i = \begin{pmatrix} b_{ij}^M & 0 \\ 0 & b_{ij}^E \end{pmatrix},$$

where $b^E_{ij} = b^{E_i}_{\widehat{E}} \,/\, b^{E_j}_{\widehat{E}}$. Matrices A^X and A will consist of the input-output ratios for materials and energy,

$$A^X = \operatorname{diag}\left(A^X_i\right), \quad A^X_i = \begin{pmatrix} A^{\widehat{M}_i}_{X_i} \\ A^{\widehat{E}_i}_{X_i} \end{pmatrix};$$

$$A = \operatorname{diag}\left(A_i\right), \quad A_i = \begin{pmatrix} A^{\widehat{M}_i}_{\widehat{M}_i} & 0 \\ 0 & A^{\widehat{E}_i}_{\widehat{E}_i} \end{pmatrix}.$$

In the PET model the energy composite is the aggregate (2.6) of N_E energy types. In this case, $\mathbf{Z}$ will be a vector of dimensions $(N_E + 1)N_R$ (N_E energy types plus materials per region). Matrix elements b^E_{ij} and $A^{\widehat{E}_i}_{\widehat{E}_i}$ will be diagonal matrices and A^X_i will have $N_E + 1$ elements.

2.4. Market equilibrium

Aggregate supply for capital K^{AS} and labor L^{AS} are determined by the sums over all regions of $n_t k_t$ and $n_t l_t$, respectively. Aggregate demand for capital and labor are

$$K^{AD} = \sum_{j=1}^{N_X} A^K_{X_j} X_j + A^K_{GP} GP,$$

$$L^{AD} = \sum_{j=1}^{N_X} A^L_{X_j} X_j + A^L_{GP} GP,$$

where GP is government purchases and A^K_{GP} and A^L_{GP} are the government sector input-output ratios of capital and labor, respectively.

An equilibrium is defined by the markets clearing conditions. That means aggregate demand is equal to aggregate supply (in each region and each time t) for the factors of production and final goods,

$$K^{AD} = K^{AS},$$
$$L^{AD} = L^{AS},$$
$$X^{AD} = X^{AS}.$$

Here X^{AD} is equal to the sums of $n_t c_t$ (or $n_t x_t$) over all regions, and X^{AS} is the production output. For the government sector, we require that revenues are equal to expenditures,

$$G^{REV} = G^{EXP}.$$

The set of the optimality conditions for consumers and producers and markets clearing conditions form a system of nonlinear equations that need to be solved. This system of equations depends on consumer quantities, i.e. capital, investment, consumption and government transfers, on the one hand and production costs (prices) and input-output ratios on the other.

3. Parallel algorithm

Since all other quantities can be obtained explicitly if we know capital K and prices P, the system of equations describing the general equilibrium can be written as

$$f(K, P) = 0.$$

The block structure of the system and parallel algorithm for solving such systems were described in detail in our paper [2]. Here we briefly recall the main ideas before describing the implementation of the parallel algorithm.

```
    input  : K^0, P^0
    output : K, P
 1  marker:  if diff > tol and it < numIt then
 2  |  omp parallel default(private)
 3  |  omp shared(dyn arrays, stor arrays)
 4  |  omp copyin(parameters)
 5  |  omp for
 6  |  for t ← 0 to T do
 7  |  |  Calculate prices P for time moment t (inner loop);
 8  |  |  Update dyn arrays;
 9  |  end
10  |  omp end parallel
11  |  Update stor arrays;
12  |  it ← (it + 1);
13  |  diff ← update (K, P);
14  end
15  goto marker
```

Figure 6. The OpenMP implementation.

The Fair – Taylor method [3] works as follows. Let K^s be the sth iterate of capital. To obtain the next iterate of prices P^{s+1} it is necessary to solve the system

$$f(K^s, P) = 0 \tag{3.1}$$

with respect to P. To obtain the next iterate of capital K^{s+2} it is necessary to solve the system

$$f(K, P^{s+1}) = 0 \tag{3.2}$$

with respect to K, and so on.

The part of the algorithm that calculates the next iterate of capital (3.2) is implemented as the *outer loop*. The part that calculates the next iterate of prices (3.1) is implemented as the *inner loop*. Blocks of the system (3.1) that correspond to different time-periods can be calculated in parallel. To improve the convergence, solution of each block is broken down into two nested loops: the NewtonA-loop for factor prices (P_K and P_L in N_R regions) and the NewtonB-loop for all other prices (goods prices in each region and export prices). The NewtonA-loop has a smaller dimensions, therefore we can use the classical Newton method with backtracking as a solver. For the NewtonB-loop we use a more advanced Krylov subspace method NITSOL (see, e.g. [6, 11]), because it has much larger dimensions and it is called more often to calculate the Jacobian for the NewtonA-loop.

The algorithm is described in Fig. 6. The input data of the algorithm is the initial approximations of capital K^0 and prices P^0 and the output is the equilibrium capital K and prices P. The general parameters are the tolerance *tol* and number of iterations *numIt*. Parameter T is the time horizon of the model.

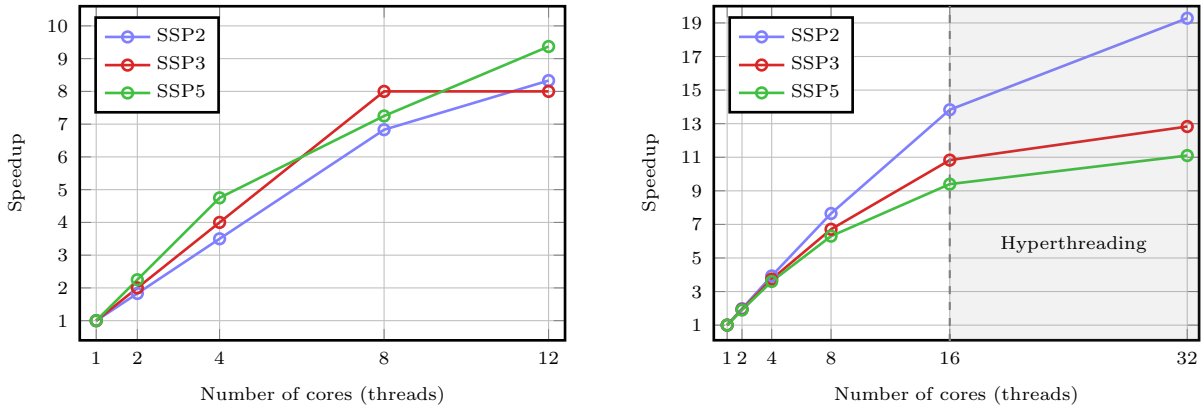
There are two types of arrays for storing and processing the economic data: *dyn* and *stor*. The first group of arrays corresponds to data at the current time and is used by the inner loop (Fig. 6, lines 6–9), the second is used for storing data over the iterations of the algorithm (outer loop). The

variable *it* is the iteration index and *diff* is the target error for the outer loop. The lines 8 and 11 in Fig. 6 correspond to the implementation of economic equations and line 13 computes the error using current iterates of capital and prices.

In the OpenMP version, the time steps of the inner loop are performed in the parallel region. All *dyn* and *stor* arrays are shared. The arrays with parameters are distributed using *copyin* clause (Fig. 6, line 4).

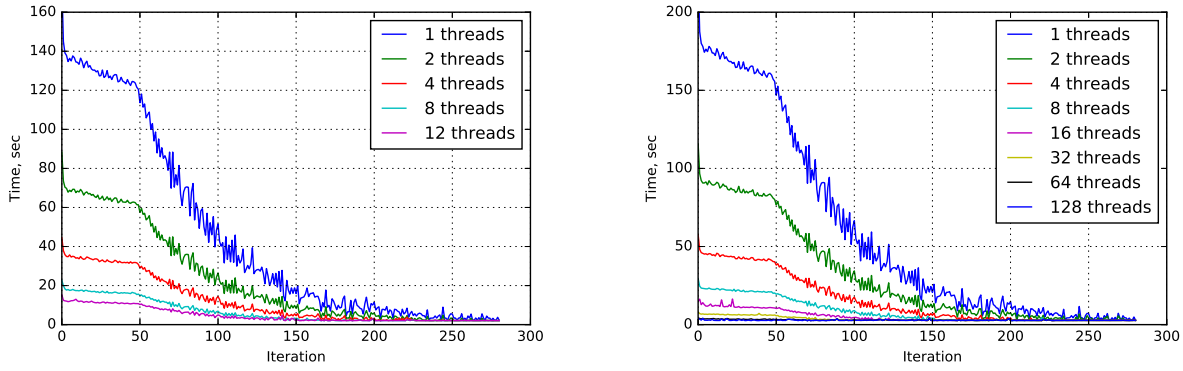
4. Results and discussion

For calculations we use the PET model with $N_R = 9$ regions and time horizon $T = 105$ years. The total number of production sectors is $N_X = 10$ in each region. As inputs the PET model uses national production and household survey data at the baseyear and long-term population and technical change projections over the whole time period. We use three sets of input data that correspond to socioeconomic scenarios from the Shared Socioeconomic Pathways (SSP) database (for the implementation of SSPs in the PET model, see [8]).



(a) Lomonosov (Intel Xeon X5670 2.93 GHz, 12 Gb) (b) Yellowstone (Intel Xeon E5-2670 2.6 GHz, 64 Gb)

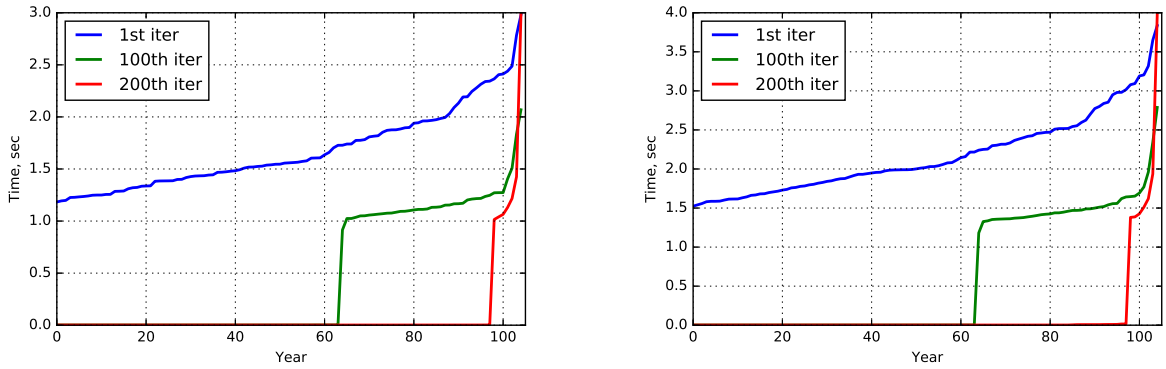
Figure 7. Speedup of the model runs for different SSPs.



(a) Regular node (b) SMP node

Figure 8. Timing of the outer loop iterations for the SSP3 obtained at the Lomonosov supercomputer.

We use two supercomputer systems for the model runs. The first one is the Lomonosov supercomputer [12]. We use two types of nodes at the Lomonosov: regular node with 12 cores (Intel



(a) Regular node (4 threads)

(b) SMP node (16 threads)

Figure 9. Timing of year-blocks in the inner loop for the SSP3 obtained at the Lomonosov supercomputer.

Xeon X5670 2.93 GHz, 1 Gb/core) and a node with 128 cores (16 Gb/core) with shared memory, the Symmetric MultiProcessing (SMP) node. The second system is the Yellowstone supercomputer [13]. At the Yellowstone we use regular node with 16 cores (Intel Xeon E5-2670 2.6 GHz, 4 Gb/core), up to 32 cores with hyperthreading. The model is implemented using Fortran. For the algorithm implementation we use BLAS [14], LAPACK [15] and Fortran implementation of NITSOL [6]. For compiling the libraries and our code we use the Intel Fortran Compiler 15 with optimization flag `-O3` and standard make-file techniques for building the project.

To study *strong scalability* of the parallel algorithm we need to increase the computing power while keeping the total problem size constant. This is achieved by running the model with the same initial approximations K_0 and P_0 and same set of numerical parameters (for each SSP) with increasing number of threads. The results show that the speedup of the parallel algorithm grows almost linearly at both supercomputers as the number of threads grows from 1 to about 12–16 (Fig. 7). Overall, we obtain the speedup of about 10 times for a regular node. With further increase of the number of nodes the speedup slows down and saturates (Table 1). Once the number of nodes becomes greater than the time horizon of the model each thread solves one year-block of the inner loop an no more speed up is possible with this algorithm. From Table 1 we see that the maximum speedup is about 22 times but using 64 nodes we already get very close to it.

Fig. 8 shows that, for the number of threads from about one to ten, there is a visible monotone decreases in timing of the outer loop as the algorithm converges (especially after the 50th iteration). This effect can be explained if we look at the timings of different year-blocks of the inner loop (Fig. 9). As the number of iterations increase the algorithm stops computing the Jacobian in the NewtonA-loop using the one from the previous iteration. The number of these "fast" year-blocks of the inner loop is increasing from iteration to iteration. For the 100th iteration of the outer loop the calculation times of more than 60 first year-blocks are close to zero. For the 200th the "fast" year-blocks span almost the whole time horizon $T = 105$.

Table 1. Speedup of the model runs for the SSP3 at the SMP node of the Lomonosov supercomputer.

Threads	1	2	4	8	16	32	64	128
Speedup	1	1.8	3.3	7.3	11	17	21.5	22

From Fig. 8 we also see that the timings of the outer loops uniformly decrease with the number of threads increasing. But as the number of threads increases above ten, the timings of the outer loops level out. The reason is that the timing of the inner loop in the parallel algorithm cannot get

smaller than the timing of the slowest year-block. From Fig. 9 we see that the number of "fast" year-blocks is increasing as the algorithm converges but there are always some "slow" year-blocks close to the end period T.

Aknowledgments

We are grateful to R. Loft and M. Weitzel for useful discussions of the results.

REFERENCES

1. **Kelley C.** Iterative Methods for Linear and Nonlinear Equations. SIAM: Philadelphia, 1995.
2. **Melnikov N., Gruzdev A., Dalton M. and O'Neill B.** Parallel algorithm for solving large-scale dynamic general equilibrium models // Russian Supercomputing Days, Moscow, 2015. P. 84–95.
3. **Fair R., Taylor J.** Solution and maximum likelihood estimation of dynamic nonlinear rational expectations models // Econometrica, 1983. Vol. 51. P. 1169–1185.
4. **Dalton M., O'Neill B., Prskawetz A., Jiang L. and Pitkin J.** Population aging and future carbon emissions in the United States // Energy economics, 2008. Vol. 30, P. 642–675.
5. **Melnikov N., O'Neill B. and Dalton M.** Accounting for the household heterogeneity in dynamic general equilibrium models // Energy economics, 2012. Vol. 34, P. 1475–1483.
6. **Pernice M., Walker H.** NITSOL: a Newton iterative solver for nonlinear systems // SIAM J. Sci. Comput., 1998. Vol. 19, P. 302–318.
7. **O'Neill B., Dalton D., Fuchs R., Jiang L., Pachauri S. and Zigova K.** Global demographic trends and future carbon emissions // Proc. Natl. Acad. Sci. U.S.A., 2010. Vol. 107, P. 17521–17526.
8. **Ren X., Weitzel M., O'Neill B.C., Lawrence P., Meiyappan P., Levis S., Balistreri E.J. and Dalton M.** Avoided economic impacts of climate change on agriculture: integrating a land surface model (CLM) with a global economic model (iPETS)// Climatic Change, 2016. P. 1–15. DOI: 10.1007/s10584-016-1791-1
9. **Stokey N., Lucas R. and Prescott E.** Recursive Methods in Economic Dynamics. Harvard University Press: Cambridge MA, 1989. 608 p.
10. **Armington P.** A theory of demand for products distinguished by place of production // IMF Staff Papers, 1969. Vol. 16, P. 170–201.
11. **Eisenstat S. and Walker H.** Globally convergent inexact Newton methods // SIAM J. Optimization, 1994 Vol. 4, P. 393–422.
12. **Sadovnichy V., Tikhonravov A., Voevodin Vl. and Opanasenko V.** "Lomonosov": Supercomputing at Moscow State University. In Contemporary High Performance Computing: From Petascale toward Exascale. Chapman & Hall/CRC Computational Science, 2013. P. 283–307.
13. Computational and Information Systems Laboratory, 2012. Yellowstone: IBM iDataPlex System (Climate Simulation Laboratory). Boulder, CO: National Center for Atmospheric Research. http://n2t.net/ark:/85065/d7wd3xhc.
14. Basic Linear Algebra Subprograms. Available from: `http://www.netlib.org/blas/` Accessed 10 October 2016.
15. Linear Algebra Package. Available from: `http://www.netlib.org/lapack/` Accessed 10 October 2016.

DEGENERATE DISTRIBUTED CONTROL SYSTEMS WITH FRACTIONAL TIME DERIVATIVE[1]

Marina V. Plekhanova

Computational Mechanics Department, South Ural State University;
Laboratory of Quantum Topology, Mathematical Analysis Department,
Chelyabinsk State University, Chelyabinsk, Russia,
mariner79@mail.ru

Abstract: The existence of a unique strong solution for the Cauchy problem to semilinear nondegenerate fractional differential equation and for the generalized Showalter–Sidorov problem to semilinear fractional differential equation with degenerate operator at the Caputo derivative in Banach spaces is proved. These results are used for search of solution existence conditions for a class of optimal control problems to a system described by the degenerate semilinear fractional evolution equation. Abstract results are applied to the research of an optimal control problem solvability for the equations system of Kelvin–Voigt fractional viscoelastic fluids.

Key words: Fractional differential calculus, Caputo deivative, Mittag–Leffler function, Partial differential equation, Degenerate evolution equation, (L,p)-bounded operator, Optimal control, Fractional viscoelastic fluid.

Introduction

Let $\mathcal{X}$, $\mathcal{Y}$ be Banach spaces, $L, M : \mathcal{X} \to \mathcal{Y}$ be linear operators, $\ker L \neq \{0\}$, $\alpha > 0$, $m \in \mathbb{N}$, $m-1 < \alpha \leq m$, $r \in \{0, 1, \ldots, m-1\}$, $N : (t_0, T) \times \mathcal{X}^{r+1} \to \mathcal{Y}$. Denote by D_t^α the Caputo fractional derivative [1]. The main purpose of the paper is to study the initial value problems unique solvability to the fractional order differential equation

$$LD_t^\alpha x(t) = Mx(t) + N(t, x(t), x^{(1)}(t), \ldots, x^{(r)}(t)), \quad t \in (t_0, T), \tag{0.1}$$

in the sense of the strong solutions and the solvability of optimal control problems for systems with the state that described by (0.1). Such equations are called degenerate because of degeneracy of the operator L at the highest derivative. The equation with left-hand side in the form $D^\alpha Lx$ is considered also. It has different properties beginning with the definition of a solution.

The theory of fractional differentiation in the last decades is actively used in the engineering and science problems. At first in the paper the existence of a unique solution is proved for the Cauchy problem to the nondegenerate fractional differential equation ($\mathcal{X} = \mathcal{Y}$, $L = I$ in (0.1)). These results are used for research of the unique solvability for the generalized Showalter–Sidorov initial value problem to the degenerate fractional differential equations. Applying the obtained statements solution existence conditions are found for a class of optimal control problems to a distributed systems described by equation (0.1) with initial conditions. Abstract results are illustrated on an optimal control problem for the equations system of Kelvin–Voigt fractional viscoelastic fluids [2].

The main condition on the operators L, M in this paper is (L,p)-boundedness of M. It was introduced in [3] for the investigation of the first order degenerate equations. The conditions of the unique solution existence for the semilinear first order degenerate differential equations under this condition were studied in [4]. The solvability in the classical sense of the linear degenerate fractional equations with (L,p)-bounded operator M was studied in the works [5, 6] and in [7] in the case of strongly (L,p)-sectorial operator. Initial boundary value problem for the linearized

[1]The work is supported by Laboratory of Quantum Topology of Chelyabinsk State University (Russian Federation government grant 14.Z50.31.0020).

system of Kelvin–Voigt fractional fluids was investigated in [8]. The equations of form (0.1) with (L,p)-bounded operator M and with $\alpha = m \in \mathbb{N}$ were investigated in [9]. The solvability in the sense of the classical solution for another class of degenerate fractional equations (0.1) in Banach spaces with restriction on the image of N was studied in [10]. Related problems in Banach and locally convex spaces for degenerate and nondegenerate fractional order evolution equations were explored by M. Kostić [11] but for other classes of operators and using mild solution and similar notions. Note papers by A.V. Glushak [12, 13] devoted to some differential equations in Banach spaces with the Riemann–Liouville, Euler–Poisson–Darboux and other derivatives. In contrast to the mentioned works the results of the present paper concern the existence of a unique strong solution for semilinear degenerate evolution fractional order equations that previously were not investigated.

In the present paper, when studying optimal control problems for equations of form (0.1), we use the general scheme suggested in the monograph [14, p. 16]. It was earlier applied to optimal control problems for a degenerate distributed systems of the first order in papers [15–17]. Optimal control problems for fractional equations are poorly understood. Most of them devoted to nondegenerate equations [18, 19], stochastic equations [20] and others. Here a research of control problems for semilinear degenerate evolution equations that has previously not been studied is presented.

1. Nondegenerate linear equation of fractional order

Let $\mathcal{Z}$ be Banach space. Introduce the Lebesgue spaces $L_q(0,T;\mathcal{Z})$ and for $q \in (1,\infty)$, $k \in \mathbb{N}$ Sobolev spaces

$$W_q^k(0,T;\mathcal{Z}) = \{f \in L_q(0,T;\mathcal{Z}) : f^{(k)} \in L_q(0,T;\mathcal{Z})\}.$$

Denote $g_\delta(t) = \Gamma(\delta)^{-1} t^{\delta-1}$,

$$J_t^\delta h(t) = (g_\delta * h)(t) = \int_0^t g_\delta(t-s)h(s)ds, \quad \text{for} \quad \delta > 0, \quad t > 0.$$

Let $\alpha > 0$, m be the smallest positive number not exceeding α, D_t^m is a usual derivative of the order $m \in \mathbb{N}$, J_t^0 is the identical operator,

$$D_t^\alpha f(t) = D_t^m J_t^{m-\alpha}\Big(f(t) - \sum_{k=0}^{m-1} f^{(k)}(0) g_{k+1}(t)\Big)$$

is the Caputo derivative [1, p. 11].

Consider the Cauchy problem

$$z^{(k)}(0) = z_k, \quad k = 0, 1, \dots, m-1, \tag{1.1}$$

for the inhomogeneous differential equation

$$D_t^\alpha z(t) = Az(t) + f(t), \quad t \in (0,T), \tag{1.2}$$

where $A \in \mathcal{L}(\mathcal{Z})$ (linear and bounded operator from $\mathcal{Z}$ to $\mathcal{Z}$), the function $f : (0,T) \to \mathcal{Z}$ is given for $T > 0$.

A strong solution of the problem (1.1)–(1.2) is a function $z \in C^{m-1}([0,T];\mathcal{Z})$, such that

$$g_{m-\alpha} * \left(z - \sum_{k=0}^{m-1} z^{(k)}(0) g_{k+1}\right) \in W_q^m(0,T;\mathcal{Z}),$$

conditions (1.1) are valid and equality (1.2) holds almost everywhere on $(0,T)$.

For $\alpha,\beta>0$ denote the Mittag–Leffler function

$$E_{\alpha,\beta}(z)=\sum_{n=0}^{\infty}\frac{z^n}{\Gamma(\alpha n+\beta)}.$$

Theorem 1. *Let $A\in\mathcal{L}(\mathcal{Z})$, $f\in L_q(0,T;\mathcal{Z})$, $q\in(\max\{1,1/\alpha\},\infty)$. Then for any $z_k\in\mathcal{Z}$, $k=0,1,\dots,m-1$, there exists a unique strong solution of the problem (1.1)–(1.2), it has the form*

$$z(t)=\sum_{k=0}^{m-1}t^kE_{\alpha,k+1}(At^\alpha)z_k+\int_0^t(t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)f(s)ds. \tag{1.3}$$

P r o o f. For $k=1,2,\dots,m-1$, $l=1,2,\dots,k$ we have

$$\frac{d^l}{dt^l}t^kE_{\alpha,k+1}(At^\alpha)=\sum_{n=0}^{\infty}\frac{A^nt^{\alpha n+k-l}}{\Gamma(\alpha n+k+1-l)}=t^{k-l}E_{\alpha,k+1-l}(At^\alpha), \tag{1.4}$$

and for $l=k+1,k+2,\dots,m-1$

$$\frac{d^l}{dt^l}t^kE_{\alpha,k+1}(At^\alpha)=\sum_{n=1}^{\infty}\frac{A^nt^{\alpha n+k-l}}{\Gamma(\alpha n+k+1-l)}=t^{\alpha+k-l}AE_{\alpha,\alpha+k+1-l}(At^\alpha).$$

So for $l=1,2,\dots,m-1$

$$\frac{d^l}{dt^l}\sum_{k=0}^{m-1}t^kE_{\alpha,k+1}(At^\alpha)z_k\Big|_{t=0}=E_{\alpha,1}(At^\alpha)z_l\Big|_{t=0}=z_l.$$

Then, using formula (1.4), we get with $l=0,1,\dots,m-1$

$$\frac{d^l}{dt^l}\Big|_{t=0}\int_0^t(t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)f(s)ds=0,$$

therefore

$$D_t^\alpha\int_0^t(t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)f(s)ds=$$

$$=D_t^m\int_0^t\frac{s^{m-\alpha-1}}{\Gamma(m-\alpha)}ds\int_0^{t-s}(t-s-\sigma)^{\alpha-1}E_{\alpha,\alpha}(A(t-s-\sigma)^\alpha)f(\sigma)d\sigma=$$

$$=D_t\int_0^t f(\sigma)d\sigma\sum_{n=0}^{\infty}\int_0^{t-\sigma}\frac{A^n(t-s-\sigma)^{\alpha(n+1)-m}s^{m-\alpha-1}}{\Gamma(m-\alpha)\Gamma(\alpha(n+1)-m+1)}ds=$$

$$=D_t\int_0^t f(\sigma)d\sigma\sum_{n=0}^{\infty}(t-\sigma)^{\alpha n}A^n\int_0^1\frac{(1-\tau)^{\alpha(n+1)-m}\tau^{m-\alpha-1}}{\Gamma(m-\alpha)\Gamma(\alpha(n+1)-m+1)}d\tau=$$

$$=D_t\int_0^t f(\sigma)E_{\alpha,1}(A(t-\sigma)^\alpha)d\sigma=A\int_0^t(t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)f(s)ds+f(t)$$

almost everywhere on $(0,T)$.

From Hölder's inequality it follows that

$$\int_0^T \left\| A \int_0^t (t-s)^{\alpha-1} E_{\alpha,\alpha}(A(t-s)^\alpha) f(s) ds \right\|_{\mathcal{Z}}^q dt \le$$

$$\le \left(\frac{q-1}{\alpha q - 1} \right)^{q-1} T^{\alpha q} \left(\|A\|_{\mathcal{L}(\mathcal{Z})} E_{\alpha,\alpha}(T^\alpha \|A\|_{\mathcal{L}(\mathcal{Z})}) \right)^q \|f\|^q_{L_q(0,T;\mathcal{Z})}$$

because $q > 1/\alpha$. Thus, function (1.3) is a strong solution of problem (1.1), (1.2).

If there exist strong solutions y_1 and y_2 of the problem (1.1)–(1.2), then their difference $z = y_1 - y_2$ is the solution of the Cauchy problem (1.1) with the initial data $z_k = 0$, $k = 0, 1, \dots, m-1$, for a homogeneous equation $D_t^\alpha z(t) = Az(t)$. Act on both sides of this equation by the operator J_t^α and obtain

$$z(t) = \int_0^t \frac{(t-s)^{\alpha-1}}{\Gamma(\alpha)} Az(s) ds, \tag{1.5}$$

because [1, p. 12]

$$J_t^\alpha D_t^\alpha z = z + \sum_{k=0}^{m-1} z^{(k)}(0) g_{k+1}.$$

By definition of a strong solution we have $z \in C([0,T];\mathcal{Z})$ even for $\alpha \in (0,1)$. Then

$$\max_{t\in[0,t_A]} \left\| \int_0^t \frac{(t-s)^{\alpha-1}}{\Gamma(\alpha)} Az(s) ds \right\|_{\mathcal{Z}} \le \frac{t_A^\alpha \|A\|_{\mathcal{L}(\mathcal{Z})}}{\Gamma(\alpha+1)} \|z\|_{C([0,t_A];\mathcal{Z})}.$$

Therefore, the integral operator defined by the right-hand side of equality (1.5) is a contraction operator in the space $C([0,t_A];\mathcal{Z})$ if

$$t_A < \left(\Gamma(\alpha+1)/\|A\|_{\mathcal{L}(\mathcal{Z})} \right)^{1/\alpha}.$$

Consequently, the unique fixed point of the integral operator is the solution $z \equiv 0$ on $[0,t_A]$. On the segment $[t_A, t_{2A}]$ repeat the reasoning. After finite number of steps the uniqueness of the zero solution will be obtained for the homogeneous Cauchy problem on the interval $(0,T)$. □

2. The Cauchy problem for the semilinear equation

Let $A \in \mathcal{L}(\mathcal{Z})$, $m \in \mathbb{N}$, $m-1 < \alpha \le m$. Operator $B : (t_0,T) \times \mathcal{Z}^m \to \mathcal{Z}$ be Caratheodory mapping, i.e. for all $z_0, z_1, \dots, z_{m-1} \in \mathcal{Z}$ it sets measurable mapping on (t_0,T) and for almost all $t \in (t_0,T)$ it is continuous with respect to $z_0, z_1, \dots, z_{m-1} \in \mathcal{Z}$. Consider the Cauchy problem

$$z^{(k)}(t_0) = z_k, \quad k = 0,\ 1,\ \dots,\ m-1, \tag{2.1}$$

for the semilinear equation

$$D_t^\alpha z(t) = Az(t) + B(t, z(t), z^{(1)}(t), \dots, z^{(m-1)}(t)), \quad t \in (t_0,T). \tag{2.2}$$

A strong solution of the problem (2.1)–(2.2) on the interval (t_0,T) is a function $z \in C^{m-1}([t_0,T];\mathcal{Z})$, such that

$$g_{m-\alpha} * \left(z - \sum_{k=0}^{m-1} z^{(k)}(t_0) g_{k+1} \right) \in W_q^m(t_0,T;\mathcal{Z}),$$

conditions (2.1) hold and almost everywhere on (t_0, T) equality (2.2) is true, (here $g_{k+1}=(t-t_0)^k/k!$, $k = 0, 1, \dots, m-1$).

Lemma 1. *Let $A \in \mathcal{L}(\mathcal{Z})$, $z_0, z_1, \dots, z_{m-1} \in \mathcal{Z}$, $B : (t_0, T) \times \mathcal{Z}^m \to \mathcal{Z}$ be Caratheodory mapping, for all $y_0, y_1, \dots y_{m-1} \in \mathcal{Z}$ and almost all $t \in (t_0, T)$ the estimate*

$$\|B(t, y_0, y_1, \dots, y_{m-1})\|_{\mathcal{Z}} \le a(t) + c \sum_{k=0}^{m-1} \|y_k\|_{\mathcal{Z}}, \tag{2.3}$$

be satisfied, where $a \in L_q(t_0, T; \mathbb{R})$, $c > 0$. Then the function z is a strong solution of the problem (2.1)–(2.2) if and only if $z \in C^{m-1}([t_0, T]; \mathcal{Z})$ and on $[t_0, T]$ we have

$$\begin{aligned} z(t) = \sum_{k=0}^{m-1} (t-t_0)^k E_{\alpha,k+1}(A(t-t_0)^\alpha) z_k + \\ + \int_{t_0}^{t} (t-s)^{\alpha-1} E_{\alpha,\alpha}(A(t-s)^\alpha) B(s, z(s), z^{(1)}(s), \dots, z^{(m-1)}(s)) ds. \end{aligned} \tag{2.4}$$

P r o o f. Let z be a solution of the problem (2.1)–(2.2), then $z \in C^{m-1}([t_0, T]; \mathcal{Z})$. In view of condition (2.3) the operator B is bounded and continuous as mapping from $W_q^{m-1}(t_0, T; \mathcal{Z})$ (and also from $C^{m-1}([t_0, T]; \mathcal{Z})$) to $L_q(t_0, T; \mathcal{Z})$. Arguing as in the proof of Theorem 1, we find that the solution satisfies equation (2.4).

Let $z \in C^{m-1}([t_0, T]; \mathcal{Z})$ on $[t_0, T]$ satisfies equation (2.4), then the function $B(\cdot, z(\cdot), \dots, z^{(m-1)}(\cdot)) \in L_q(t_0, T; \mathcal{Z})$ and by analogy with Theorem 1 we can verify that z is a strong solution of the problem (2.1)–(2.2). □

The bar over a symbol will mean an ordered set of m elements with indexes from 0 to $m-1$, for example, $\overline{z} = (z_0, z_1, \dots, z_{m-1})$. A mapping $B : (t_0, T) \times \mathcal{Z}^m \to \mathcal{Z}$ is called uniformly Lipschitz continuous in $\overline{y}$, if there exists $l > 0$, such that the inequality

$$\|B(t, \overline{y}) - B(t, \overline{z})\|_{\mathcal{Z}} \le l \sum_{k=0}^{m-1} \|y_k - z_k\|_{\mathcal{Z}}$$

is true for almost all $t \in (t_0, T)$ and for all $\overline{y}, \overline{z}$ of $\mathcal{Z}^m$.

Theorem 2. *Let $A \in \mathcal{L}(\mathcal{Z})$, $B : (t_0, T) \times \mathcal{Z}^m \to \mathcal{Z}$ be Caratheodory mapping, uniformly Lipschitz continuous in $\overline{y}$, $q \in (\max\{1, 1/\alpha\}, \infty)$, for some $\overline{v} \in \mathcal{Z}^m$ $B(\cdot, \overline{v}) \in L_q(t_0, T; \mathcal{Z})$. Then for any $z_0, z_1, \dots, z_{m-1} \in \mathcal{Z}$ the problem (2.1)–(2.2) has a unique strong solution on (t_0, T).*

P r o o f. The uniformly Lipschitz continuity implies that for any $\overline{y} \in \mathcal{Z}^m$ for almost all $t \in (t_0, T)$ we have

$$\|B(t, \overline{y})\|_{\mathcal{Z}} \le \|B(t, \overline{v})\|_{\mathcal{Z}} + l \sum_{k=0}^{m-1} \|v_k\|_{\mathcal{Z}} + l \sum_{k=0}^{m-1} \|y_k\|_{\mathcal{Z}},$$

therefore condition (2.3) is performed with

$$a(t) = \|B(t, \overline{v})\|_{\mathcal{Z}} + l \sum_{k=0}^{m-1} \|v_k\|_{\mathcal{Z}}, \quad c = l.$$

According to the statement of Lemma 1 it is sufficient to show that the equation (2.4) has a unique solution $z \in C^{m-1}([t_0,T];\mathcal{Z})$. In the space $C^{m-1}([t_0,T];\mathcal{Z})$ define an operator F as

$$F(y)(t) = \sum_{k=0}^{m-1}(t-t_0)^k E_{\alpha,k+1}(A(t-t_0)^\alpha)z_k +$$
$$+\int_{t_0}^{t}(t-s)^{\alpha-1}E_{\alpha,\alpha}(A(t-s)^\alpha)B(s,y(s),y^{(1)}(s),\dots,y^{(m-1)}(s))\,ds.$$

By the proof of Theorem 1 $F : C^{m-1}([t_0,T];\mathcal{Z}) \to C^{m-1}([t_0,T];\mathcal{Z})$.

We denote by F^r the r-th power of the operator F, $r \in \mathbb{N}$, and in further reasoning if $T-t_0 < 1$ we will replace $T-t_0$ by 1. For $t \in [t_0,T]$, $n = 0,1,\dots,m-1$, $r \in \mathbb{N}$, $y,z \in C^{m-1}([t_0,T];\mathcal{Z})$ by induction the inequality

$$\|[F^r(y)]^{(n)}(t) - [F^r(z)]^{(n)}(t)\|_{\mathcal{Z}} \le \frac{K^r(t-t_0)^{\alpha-m+r}\|y-z\|_{C^{m-1}([t_0,T];\mathcal{Z})}}{m(r-1)!} \tag{2.5}$$

can be proved, where

$$K = ml(\alpha-m+1)^{-1}(T-t_0)^\alpha \max_{n=0,\dots,m-1} E_{\alpha,\alpha-n}((T-t_0)^\alpha\|A\|_{\mathcal{L}(\mathcal{Z})}).$$

For $r=1$, $n=0,1,\dots,m-1$ Hölder's inequality implies that

$$\|[F(y)]^{(n)}(t) - [F(z)]^{(n)}(t)\|_{\mathcal{Z}} \le E_{\alpha,\alpha-n}((t-t_0)^\alpha\|A\|_{\mathcal{L}(\mathcal{Z})})\times$$
$$\times\int_{t_0}^{t}(t-s)^{\alpha-1-n}\|B(s,y(s),\dots,y^{(m-1)}(s)) - B(s,z(s),\dots,z^{(m-1)}(s))\|_{\mathcal{Z}}ds \le$$
$$\le \frac{l(t-t_0)^{\alpha-m+1}(T-t_0)^{m-1-n}}{\alpha-m+1}E_{\alpha,\alpha-n}((T-t_0)^\alpha\|A\|_{\mathcal{L}(\mathcal{Z})})\|y-z\|_{C^{m-1}([t_0,T];\mathcal{Z})}.$$

If for $r-1$ inequality (2.5) is valid, then

$$\|[F^r(y)]^{(n)}(t) - [F^r(z)]^{(n)}(t)\|_{\mathcal{Z}} \le \frac{K}{m}\int_{t_0}^{t}\sum_{k=0}^{m-1}\|[F^{r-1}(y)]^{(k)}(s) - [F^{r-1}(z)]^{(k)}(s)\|_{\mathcal{Z}}ds \le$$
$$\le K\int_{t_0}^{t}\frac{K^{r-1}(s-t_0)^{\alpha-m+r-1}\|y-z\|_{C^{m-1}([t_0,T];\mathcal{Z})}}{m(r-2)!}ds \le$$
$$\le \frac{K^r(t-t_0)^{\alpha-m+r}\|y-z\|_{C^{m-1}([t_0,T];\mathcal{Z})}}{m(\alpha-m+r)(r-2)!} < \frac{K^r(t-t_0)^{\alpha-m+r}\|y-z\|_{C^{m-1}([t_0,T];\mathcal{Z})}}{m(r-1)!}.$$

From (2.5) it follows that for $r \in \mathbb{N}$ we have

$$\|[F^r(y)] - [F^r(z)]\|_{C^{m-1}([t_0,T];\mathcal{Z})} \le \frac{K^r(T-t_0)^{\alpha-m+r}\|y-z\|_{C^{m-1}([t_0,T];\mathcal{Z})}}{(r-1)!}.$$

Therefore, if r is sufficiently large, then F^r is a strict contraction in $C^{m-1}([t_0,T];\mathcal{Z})$, so it has in this space a unique fixed point. It is the unique solution of the equation (2.4) in the space $C^{m-1}([t_0,T];\mathcal{Z})$, and, therefore, a unique strong solution of the problem (2.1)–(2.2) on the interval (t_0,T). □

We will need solutions of the problem (2.1)–(2.2) with additional smoothness. For fractional $\alpha > 1$ conditions of their existence were found in the case of incomplete equation only (without $(m-1)$-th derivative under the sign of operator B).

Theorem 3. *Let $\alpha > 1$, $q > (\alpha + 1 - m)^{-1}$, $A \in \mathcal{L}(\mathcal{Z})$, $n \in \mathbb{N}$, $B \in C^n([t_0, T] \times \mathcal{Z}^{m-1}; \mathcal{Z})$ be uniformly Lipschitz continuous in $z_0, z_1, \dots, z_{m-2} \in \mathcal{Z}$, $f \in W_q^n(t_0, T; \mathcal{Z})$ and let for z satisfying conditions* (2.1) *and the equation*

$$D_t^\alpha z(t) = Az(t) + B(t, z(t), z^{(1)}(t), \dots, z^{(m-2)}(t)) + f(t) \tag{2.6}$$

the equalities

$$D_t^k \Big|_{t=t_0} [B(t, z(t), z^{(1)}(t), \dots, z^{(m-2)}(t))] = -f^{(k)}(t_0), \quad k = 0, 1, \dots, n-1, \tag{2.7}$$

hold. Then for every $z_0, z_1, \dots, z_{m-1} \in Z$ there exists a unique strong solution z of the problem (2.1), (2.6). *Besides, $z \in C^{m-1+n}([t_0, t_1]; \mathcal{Z})$.*

P r o o f. For $\alpha > 1$ we have $m \geq 2$. Using equalities (2.7) and sequentially computing the derivatives of the right-hand side of (2.4), we obtain for $k \in \mathbb{N}_0$

$$D_t^{m+k} \int_{t_0}^{t} (t-s)^{\alpha-1} E_{\alpha,\alpha}(A(t-s)^\alpha) B(s, z(s), z^{(1)}(s), \dots, z^{(m-2)}(s)) ds =$$

$$= \int_{t_0}^{t} (t-s)^{\alpha-m} E_{\alpha,\alpha-m+1}(A(t-s)^\alpha) D_s^{k+1} [B(s, z(s), \dots, z^{(m-2)}(s)) + f(s)] ds.$$

□

Remark 1. The form of the integral in the solution formula (2.4) implies the existence of singularity of a solution at $t = t_0$ in case of fractional α, if conditions (2.7) isn't used.

3. Degenerate semilinear equation

Let an operator $L \in \mathcal{L}(\mathcal{X}; \mathcal{Y})$ (linear and continuous from a Banach space $\mathcal{X}$ to a Banach space $\mathcal{Y}$), $M \in \mathcal{C}l(\mathcal{X}; \mathcal{Y})$ (linear, closed and densely defined in $\mathcal{X}$ with image in $\mathcal{Y}$), D_M is a domain of an operator M, endowded by the graph norm $\| \cdot \|_{D_M} = \| \cdot \|_{\mathcal{X}} + \|M \cdot \|_{\mathcal{Y}}$. Define L-resolvent set $\rho^L(M) = \{\mu \in \mathbb{C} : (\mu L - M)^{-1} \in \mathcal{L}(\mathcal{Y}; \mathcal{X})\}$ of an operator M and introduce the denotations $R_\mu^L(M) = (\mu L - M)^{-1} L$, $L_\mu^L = L(\mu L - M)^{-1}$.

An operator M will be called (L, σ)-*bounded*, if

$$\exists a > 0 \quad \forall \mu \in \mathbb{C} \quad (|\mu| > a) \Rightarrow (\mu \in \rho^L(M)).$$

Lemma 2 [3]. *Let an operator M be (L, σ)-bounded, $\gamma = \{\mu \in \mathbb{C} : |\mu| = r > a\}$. Then the operators*

$$P = \frac{1}{2\pi i} \int_\gamma R_\mu^L(M)\, d\mu \in \mathcal{L}(\mathcal{X}), \quad Q = \frac{1}{2\pi i} \int_\gamma L_\mu^L(M)\, d\mu \in \mathcal{L}(\mathcal{Y})$$

are projections.

Put $\mathcal{X}^0 = \ker P$, $\mathcal{X}^1 = \mathrm{im}P$, $\mathcal{Y}^0 = \ker Q$, $\mathcal{Y}^1 = \mathrm{im}Q$. Denote by L_k (M_k) the restriction of the operator L (M) on $\mathcal{X}^k$ ($D_{M_k} = D_M \cap \mathcal{X}^k$), $k = 0, 1$.

Theorem 4 [3]. *Let an operator M be (L,σ)-bounded. Then*

(i) $M_1 \in \mathcal{L}\big(\mathcal{X}^1;\mathcal{Y}^1\big)$, $M_0 \in \mathcal{C}l\big(\mathcal{X}^0;\mathcal{Y}^0\big)$, $L_k \in \mathcal{L}\big(\mathcal{X}^k;\mathcal{Y}^k\big)$, $k = 0, 1$;

(ii) *there exist operators* $M_0^{-1} \in \mathcal{L}\big(\mathcal{Y}^0;\mathcal{X}^0\big)$, $L_1^{-1} \in \mathcal{L}\big(\mathcal{Y}^1;\mathcal{X}^1\big)$.

Denote $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$, $G = M_0^{-1}L_0$. For $p \in \mathbb{N}_0$ the operator M is called (L,p)-bounded, if it is (L,σ)-bounded, $G^p \neq \mathbb{O}$, $G^{p+1} = \mathbb{O}$.

For $m-1 < \alpha \leq m$, $r \in \{0, 1, \dots, m-1\}$ consider the semilinear evolution equation

$$D_t^\alpha Lx(t) = Mx(t) + N(t, x(t), x^{(1)}(t), \dots, x^{(r)}(t)) + f(t), \quad t \in (t_0, T), \tag{3.1}$$

with operators $L \in \mathcal{L}(\mathcal{X};\mathcal{Y})$, $\ker L \neq \{0\}$, $M \in \mathcal{C}l(\mathcal{X};\mathcal{Y})$, with a nonlinear operator $N : (t_0, T) \times \mathcal{X}^{r+1} \to \mathcal{Y}$ and a function $f : (t_0, T) \to \mathcal{Y}$.

A strong solution of equation (3.1) on the interval (t_0, T) is a function $x \in W_q^r(t_0, T; \mathcal{X}) \cap L_q(t_0, T; D_M)$, $q \in (1, \infty)$, such that $Lx \in C^{m-1}([t_0, T]; \mathcal{Y})$,

$$g_{m-\alpha} * \left(Lx - \sum_{k=0}^{m-1} (Lx)^{(k)}(t_0) g_{k+1} \right) \in W_q^m(t_0, T; \mathcal{Y}),$$

and almost everywhere on (t_0, T) equality (3.1) is true.

Let operator M be (L,σ)-bounded. Consider the generalized Showalter—Sidorov problem [21, 22]

$$(Px)^{(k)}(t_0) = x_k, \ k = 0, 1, \dots, m-1, \tag{3.2}$$

for equation (3.1) on the interval (t_0, T).

Remark 2. We have the equalities $Px = L_1^{-1}L_1Px = L_1^{-1}QLx$. Therefore, the smoothness of Px is not smaller than for the function Lx.

Denote by $[\beta]$ the integer part of $\beta \in \mathbb{R}$.

Theorem 5. *Let $\alpha > 0$, $q \in (\max\{1, 1/\alpha\}, \infty)$, $r = [(m-1)/2]$, an operator M be $(L,0)$-bounded, an operator $N : [t_0, T] \times \mathcal{X}^{r+1} \to \mathcal{Y}$ be Caratheodory mapping, the equality*

$$N(t, z_0, z_1, \dots, z_r) = N_1(t, Pz_0, Pz_1, \dots, Pz_r) \tag{3.3}$$

with some $N_1 : [t_0, T] \times (\mathcal{X}^1)^{r+1} \to \mathcal{Y}$ be valid for all $z_0, z_1, \dots, z_r \in \mathcal{X}$, almost all $t \in (t_0, T)$. Let QN_1 be uniformly Lipschitz continuous in $\overline{v} = (v_0, v_1, \dots, v_r) \in (\mathcal{X}^1)^{r+1}$, for some $\overline{v} \in (\mathcal{X}^1)^{r+1}$, $QN_1(\cdot, v_0, \dots, v_r) \in L_q(t_0, T; \mathcal{Y})$, $(I-Q)N_1 \in C^r([t_0, T] \times (\mathcal{X}^1)^{r+1}; \mathcal{Y})$, $(I-Q)f \in W_q^r(t_0, T; \mathcal{Y})$, $Qf \in L_q(t_0, T; \mathcal{Y})$. Then for any $x_0, x_1, \dots, x_{m-1} \in \mathcal{X}^1$ the problem (3.1)–(3.2) *has a unique strong solution on the interval (t_0, T).*

P r o o f. Multiply (3.1) from the left by the operators $L_1^{-1}Q$ or $M_0^{-1}(I-Q)$ and obtain the problem

$$\begin{gathered} D_t^\alpha v(t) = S_1 v(t) + L_1^{-1}QN_1(t, v(t), v^{(1)}(t), \dots, v^{(r)}(t)) + L_1^{-1}Qf(t), \\ v^{(k)}(t_0) = Px_k, \ k = 0, 1, \dots, m-1, \end{gathered} \tag{3.4}$$

$$0 = w(t) + M_0^{-1}(I-Q)N_1(t, v(t), v^{(1)}(t), \dots, v^{(r)}(t)) + M_0^{-1}(I-Q)f(t) \tag{3.5}$$

for the pair of functions $v(t) \equiv Px(t)$, $w(t) \equiv (I-P)x(t)$. Here the notations $S_1 = L_1^{-1}M_1$, $G = M_0^{-1}L_0$ are used.

By Theorem 2 the problem (3.4) has a unique strong solution, since the operator S_1 is bounded by Theorem 4. Knowing v, obtain

$$w(t) = -M_0^{-1}(I-Q)N_1(t, v(t), v^{(1)}(t), \dots, v^{(r)}(t)) - M_0^{-1}(I-Q)f(t)$$

from equation (3.5). Here $w \in W_q^r(t_0, T; \mathcal{X}) \cap L_q(0, T; D_M)$, $Lw \equiv 0$. Thus, there exists a unique strong solution $x = v + w$ of the problem (3.1)–(3.2). □

A function $x \in C^{m-1}([t_0, T]; \mathcal{X}) \cap L_q(t_0, T; D_M)$, $q \in (1, \infty)$, is a strong solution of equation

$$LD_t^\alpha x(t) = Mx(t) + N(t, x(t), x^{(1)}(t), \dots, x^{(r)}(t)) + f(t) \tag{3.6}$$

on the interval (t_0, T) if

$$g_{m-\alpha} * \Big(x - \sum_{k=0}^{m-1} x^{(k)}(t_0) g_{k+1}\Big) \in W_q^m(t_0, T; \mathcal{X}),$$

and almost everywhere on (t_0, T) the equality (3.6) is valid.

Theorem 6. *Let $\alpha > 1$, $q > (\alpha + 1 - m)^{-1}$, $r = 0$, operator M be $(L, 0)$-bounded, suppose that $N : [t_0, T] \times \mathcal{X} \to \mathcal{Y}$ for all $z \in \mathcal{X}$, $t \in [t_0, T]$ satisfies the equality $N(t, z) = N_1(t, Pz)$ for some mapping $N_1 \in C^1([t_0, T] \times \mathcal{X}^1; \mathcal{Y})$, $(I-Q)N_1 \in C^m([t_0, T] \times \mathcal{X}^1; \mathcal{Y})$, QN_1 is uniformly Lipschitz continuous in $v \in \mathcal{X}^1$, $f \in W_q^1(t_0, T; \mathcal{Y})$, $q > (\alpha + 1 - m)^{-1}$, $(I-Q)f \in C^m([t_0, T]; \mathcal{Y})$, $x_0, x_1, \dots, x_{m-1} \in \mathcal{X}^1$, the equality $QN_1(t_0, Px_0) + Qf(t_0) = 0$ is valid. Then there exists a unique strong solution of the problem* (3.2), (3.6).

P r o o f. Arguing as in the proof of Theorem 5, obtain the unique solution $x = v + w$, where v is an unique solution of the Cauchy problem for the equation $D_t^\alpha v(t) = S_1 v(t) + L_1^{-1}QN_1(t, v(t)) + L_1^{-1}Qf(t)$ and the function $w(t) = -M_0^{-1}(I-Q)N_1(t, v(t)) - M_0^{-1}(I-Q)f(t)$. By Theorem 3 we have $v \in C^m([t_0, T]; \mathcal{X})$, therefore $w \in C^m([t_0, T]; \mathcal{X})$ and there exists $D_t^\alpha x \in L_q(t_0, T; \mathcal{X})$. □

The proof of the next statement for the equation of an order $\alpha > 2$, with $r \in \{1, 2, \dots, m-2\}$ is similar to the previous one.

Theorem 7. *Let $\alpha > 2$, $q > (\alpha+1-m)^{-1}$, $r \in \{1, 2, \dots, m-2\}$, operator M be $(L, 0)$-bounded, suppose that $N : [t_0, T] \times \mathcal{X}^{r+1} \to \mathcal{Y}$ for all $z_0, z_1, \dots z_r \in \mathcal{X}$, $t \in [t_0, T]$ satisfies condition* (3.3) *with some $N_1 \in C^{r+1}([t_0, T] \times (\mathcal{X}^1)^{r+1}; \mathcal{Y})$; a mapping QN_1 is uniformly Lipschitz continuous in $\overline{v} \in \mathcal{X}^{r+1}$, $(I-Q)N_1 \in C^m([t_0, T] \times (\mathcal{X}^1)^{r+1}; \mathcal{Y})$, $f \in W_q^{r+1}(t_0, T; \mathcal{Y})$, $(I-Q)f \in C^m([t_0, T]; \mathcal{Y})$, $x_0, \dots, x_{m-1} \in \mathcal{X}^1$; when $k = 0, 1, \dots, m-1$ for the solution of problem*

$$\begin{gathered} D_t^\alpha v(t) = S_1 v(t) + L_1^{-1}QN_1(t, v(t), v^{(1)}(t), \dots, v^{(r)}(t)) + L_1^{-1}Qf(t), \\ v^{(l)}(t_0) = Px_l, \ l = 0, 1, \dots, m-1, \end{gathered} \tag{3.7}$$

conditions

$$D_t^k\Big|_{t=t_0} Q(N_1(t, v(t), v^{(1)}(t), \dots, v^{(r)}(t)) + f(t)) = 0, \quad k = 0, 1, \dots, r, \tag{3.8}$$

hold. Then problem (3.2), (3.6) *has a unique strong solution on the interval (t_0, T).*

P r o o f. The proof is similar to the previous one. Here we have $v \in C^{m+r}([t_0, T]; \mathcal{X})$ by Theorem 3. □

4. Optimal control problem

Now let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{U}$ be Banach spaces, $L \in \mathcal{L}(\mathcal{X};\mathcal{Y})$, $\ker L \neq \{0\}$, $B \in \mathcal{L}(\mathcal{U};\mathcal{Y})$, $M \in \mathcal{C}l(\mathcal{X};\mathcal{Y})$ is (L,p)-bounded operator, $N : [t_0,T] \times \mathcal{X} \to \mathcal{Y}$. Consider the control problem

$$LD_t^\alpha x(t) = Mx(t) + N(t,x(t)) + Bu(t), \tag{4.1}$$

$$(Px)^{(k)}(t_0) = x_k,\ k = 0,1,\dots,m-1, \tag{4.2}$$

$$u \in \mathcal{U}_\partial, \tag{4.3}$$

$$J(x,u) \to \inf, \tag{4.4}$$

where $\mathcal{U}_\partial$ is a set of admissible controls, the cost functional J will be described below.

Taking into account the form of equation (4.1), we will seek its strong solutions in the linear space

$$\mathcal{Z}_{\alpha,q} = \left\{ x \in L_q(t_0,T;D_M) \cap C^{m-1}([t_0,T];\mathcal{X}) :\ g_{m-\alpha} * \left(x - \sum_{k=0}^{m-1} x^{(k)}(t_0) g_{k+1} \right) \in W_q^m(t_0,T;\mathcal{X}) \right\}.$$

Lemma 3. *For $q \in (\max\{1, 1/\alpha\}, \infty)$ $\mathcal{Z}_{\alpha,q}$ is a Banach space with the norm*

$$\|x\|_{\mathcal{Z}} = \|x\|_{L_q(t_0,T;D_M)} + \|x\|_{C^{m-1}([t_0,T];\mathcal{X})} + \|D_t^\alpha x\|_{L_q(t_0,T;\mathcal{X})}.$$

P r o o f. Prove the closedness of the operator $D_t^\alpha : L_q(t_0,T;D_M) \cap C^{m-1}([t_0,T];\mathcal{Z}) \to L_q(t_0,T;\mathcal{Z})$ with the domain $\mathcal{Z}_{\alpha,q}$. By definition of the Caputo fractional derivative $D_t^\alpha = {}^{RL}D_t^\alpha S_m$, where ${}^{RL}D_t^\alpha$ is the Riemann—Liouville fractional derivative [1], we have

$$S_m z \equiv z - \sum_{k=0}^{m-1} z^{(k)}(t_0) g_{k+1}.$$

It is evident that the operator S_m acts continuously from $\mathcal{Z}_{\alpha,q}$ with the norm of $C^{m-1}([t_0,T];\mathcal{Z})$ into the space

$$\mathcal{R}_{\alpha,q,0} \equiv \{ z \in L_q(t_0,T;\mathcal{Z}) : g_{m-\alpha} * z \in W_{q,0}^m(t_0,T;\mathcal{Z}) \},$$

endowed with the norm of $L_q(t_0,T;\mathcal{Z})$. And the operator ${}^{RL}D_t^\alpha : \mathcal{R}_{\alpha,q,0} \to L_q(t_0,T;\mathcal{Z})$ is closed by Lemma 1.8 (a) [1, p. 15]. □

Introduce the continuous operator $\gamma_0 : C([t_0,T];\mathcal{X}) \to \mathcal{X}$, $\gamma_0 x = x(t_0)$.

The set of pairs (x,u) will be called as admissible pairs set $\mathcal{W}$ of the problem (4.1)–(4.4) if $u \in \mathcal{U}_\partial$, $x \in \mathcal{Z}_{\alpha,q}$ is a strong solution of (4.1), (4.2), $J(x,u) < \infty$. Problem (4.1)–(4.4) is the problem of finding pairs $(\hat{x},\hat{u}) \in \mathcal{W}$, which minimize the cost functional, i. e. $J(\hat{x},\hat{u}) = \inf\limits_{(x,u)\in\mathcal{W}} J(x,u)$.

Theorem 8. *Let $\alpha > 1$, $q > (\alpha + 1 - m)^{-1}$, an operator M be $(L,0)$-bounded, $N : (t_0,T) \times \mathcal{X} \to \mathcal{Y}$, for all $z \in \mathcal{X}$, $t \in (t_0,T)$ $N(t,z) = N_1(t,Pz)$ for some $N_1 \in C^1([t_0,T] \times \mathcal{X}^1;\mathcal{Y})$, QN_1 be uniformly Lipschitz continuous in $x \in \mathcal{X}^1$, $(I-Q)N_1 \in C^m([t_0,T] \times \mathcal{X}^1;\mathcal{Y})$. Suppose that $\mathcal{U}_\partial$ is a non-empty closed convex subset of $L_q(t_0,T;\mathcal{U})$, there exists $u_0 \in \mathcal{U}_\partial \cap W_q^1(t_0,T;\mathcal{U})$ such that $(I-Q)Bu_0 \in C^m([t_0,T];\mathcal{U})$, $QBu_0(t_0) = -QN_1(t_0,Px_0)$; $\mathcal{Z}_{\alpha,q}$ is continuously embedded in Banach space $\mathfrak{Y}$, $\mathfrak{Y}$ is continuously embedded in $L_q(t_0,T;\mathcal{X})$, cost functional J is convex, lower semicontinuous, and bounded from below on $\mathfrak{Y} \times L_q(t_0,T;\mathcal{U})$, and J is coercive on $\mathcal{Z}_{\alpha,q} \times L_q(t_0,T;\mathcal{U})$, $x_k \in \mathcal{X}^1$, $k = 0,1,\dots,m-1$. Then there exists a solution $(\hat{x},\hat{u}) \in \mathcal{Z}_{\alpha,q} \times \mathcal{U}_\partial$ of the problem (4.1)–(4.4).*

P r o o f. The operator N and the function $f = Bu_0$ satisfy the conditions of Theorem 6. Hence, Theorem 6 implies the existence of a strong solution of problem (4.1), (4.2) with $u = u_0 \in \mathcal{U}_\partial$. So, the set of admissible pairs $\mathcal{W}$ is nonempty.

Further we will use Theorem 1.2.4 [14]. Put $\mathfrak{Y}_1 = \mathcal{Z}_{\alpha,q}$, $\mathfrak{U} = L_q(t_0, T; \mathcal{U})$, $\mathfrak{V} = L_q(t_0, T; \mathcal{Y}) \times \mathcal{X}^m$, $\mathfrak{F}(x(\cdot)) = -(N(\cdot, x(\cdot)), x_0, x_1, \ldots, x_{m-1})$, $\mathfrak{L}(x, u) = (LD_t^\alpha x - Mx - Bu, \gamma_0 Px, \ldots, \gamma_0 Px^{(m-1)})$. The continuity of the linear operator $\mathfrak{L} : \mathfrak{Y}_1 \times \mathfrak{U} \to \mathfrak{V}$ follows from the inequalities

$$\|(LD_t^\alpha x - Mx - Bu, \gamma_0 Px, \gamma_0 Px^{(1)}, \ldots, \gamma_0 Px^{(m-1)})\|_{L_q(t_0,T;\mathcal{Y})\times\mathcal{X}^m} \le$$

$$\le C_1 \left(\|x\|_{\mathcal{Z}_{\alpha,q}} + \|u\|_{L_q(t_0,T;\mathcal{U})} + \|x\|_{C^{m-1}([t_0,T];\mathcal{X})}\right) \le C_2 \|(x, u)\|_{\mathcal{Z}_{\alpha,q}\times\mathcal{U}}.$$

From the relation $\|x_n - x\|_{\mathcal{Z}_{\alpha,q}} \to 0$ for $n \to \infty$ it follows that

$$\|N(\cdot, x_n(\cdot)) - N(\cdot, x(\cdot))\|_{L_q(t_0,T;\mathcal{Y})} \le C_1 \|x_n - x\|_{C([t_0,T];\mathcal{X})} \to 0,$$

therefore the operator $\mathfrak{F}$ is continuous.

After choosing $\mathfrak{Y}_{-1} = L_q(t_0, T; \mathcal{X})$, check the remaining conditions of Theorem 1.2.4 [14]. From Rellich—Kondrashov theorem it follows that $\mathcal{Z}_{\alpha,q}$ enclosed to $W_q^{m-1}(t_0, T; \mathcal{X})$ and compactly enclosed to $L_q(t_0, T; \mathcal{X})$. For $v \in (L_q(t_0, T; \mathcal{Y}))^*$ the uniform Lipschitz continuity of the operator N implies the inequality

$$|v(N(t, x_n(t)) - N(t, x(\cdot))| \le C_1 \|v\|_{(L_q(t_0,T;\mathcal{Y}))^*} \|x_n - x\|_{L_q(t_0,T;\mathcal{X})}.$$

It gives the continuous extension of the functional $f(\cdot) = v(\mathfrak{F}(\cdot))$ from $\mathcal{Z}_{\alpha,q}$ to $L_q(t_0, T; \mathcal{X})$. □

In applications the condition of the uniform Lipschitz continuity of N is too strong. But the nonemptyness of $\mathcal{W}$ is often evident. Consider the optimal control problem in such case.

A mapping $N \in C([t_0, T] \times \mathcal{X}; \mathcal{Y})$ will be called locally Lipschitz continuous in $x \in \mathcal{X}$, uniformly with respect to $t \in [t_0, T]$, if for every $x \in \mathcal{X}$ there exists $\delta > 0$ and $l > 0$ such that for every $y \in \mathcal{Y}$ the inequality $\|y - x\|_{\mathcal{X}} < \delta$ implies that $\|N(t, y) - N(t, x)\|_{\mathcal{Y}} \le l\|y - x\|_{\mathcal{X}}$ for all $t \in [t_0, T]$.

Theorem 9. *Let $\alpha, q > 1$, an operator M be (L, p)-bounded, the mapping $N \in C([t_0, T] \times \mathcal{X}; \mathcal{Y})$ be locally Lipschitz continuous in $z \in \mathcal{X}$, uniformly with respect to $t \in [t_0, T]$. Suppose that $x_k \in \mathcal{X}^1$, $k = 0, 1, \ldots, m-1$, $\mathcal{U}_\partial$ is a non-empty closed convex subset of $L_q(t_0, T; \mathcal{U})$, for some $u_0 \in \mathcal{U}_\partial$ there exists a solution of the problem (4.1)–(4.2); $\mathcal{Z}_{\alpha,q}$ is continuously embedded in Banach space $\mathfrak{Y}$, $\mathfrak{Y}$ is continuously embedded in $L_q(t_0, T; \mathcal{X})$, cost functional J is convex, lower semicontinuous, and bounded from below on $\mathfrak{Y} \times L_q(t_0, T; \mathcal{U})$, and J is coercive on $\mathcal{Z}_{\alpha,q} \times L_q(t_0, T; \mathcal{U})$. Then there exists a solution $(\hat{x}, \hat{u}) \in \mathcal{Z}_{\alpha,q} \times \mathcal{U}_\partial$ of the problem (4.1)–(4.4).*

P r o o f. The set $\mathcal{W}$ is non-empty by the conditions of the theorem. The conditions on the mapping N are sufficient for repeating the previous proof. □

5. Optimal control for fractional Kelvin—Voigt fluid

Consider a control problem

$$(1 - \chi\Delta)D_t^\alpha v(s, t) = \nu\Delta v(s, t) - (v \cdot \nabla)v(s, t) - r(s, t) + u(s, t),\ (s, t) \in \Omega \times [0, T], \tag{5.1}$$

$$\nabla \cdot v(s, t) = 0, \quad (s, t) \in \Omega \times [0, T], \tag{5.2}$$

$$v(s, t) = 0, \quad (s, t) \in \partial\Omega \times [0, T], \tag{5.3}$$

$$\frac{\partial^k v}{\partial t^k}(s,0) = \psi_k(s), \quad k = 0,1,\dots,m-1, \quad s \in \Omega, \tag{5.4}$$

$$\|u\|_{L_q(0,T;\mathbb{L}_2)} \le R, \tag{5.5}$$

$$J(v,r,u) = \|v - v_d\|_{C^{m-1}([0,T];\mathbb{H}^2_\sigma)} + \|r - r_d\|_{C^{m-1}([0,T];\mathbb{H}_\pi)} +$$
$$+\|D^\alpha_t v - D^\alpha_t v_d\|^q_{L_q(0,T;\mathbb{H}^2_\sigma)} + \|D^\alpha_t r - D^\alpha_t r_d\|^q_{L_q(0,T;\mathbb{H}_\pi)} + \|u - u_d\|^q_{L_q(0,T;\mathbb{L}_2)} \to \inf. \tag{5.6}$$

Here, $\Omega \subset \mathbb{R}^3$ is a domain with a smooth boundary $\partial\Omega$, $\chi, \nu \in \mathbb{R}$, $T > 0$. The vector-functions $\psi_k = (\psi_{k1}, \psi_{k2}, \psi_{k3}) : \Omega \to \mathbb{R}^3$, $k = 0,1,\dots,m-1$, are set. Vector-functions $v = (v_1, v_2, v_3)$ of the velocity and $r = (r_1, r_2, r_3) = (p_{s_1}, p_{s_2}, p_{s_3})$ of the pressure p gradient are unknown. An external source $u = (u_1, u_2, u_3) : \Omega \times [0,T] \to \mathbb{R}^3$ is a control function. The system models the dynamics of a fractional viscoelastic incompressible Kelvin — Voigt fluid [2].

To reduce the optimal control problem (5.1)–(5.6) to problem (4.1)–(4.4), denote the Lebesgue space $\mathbb{L}_2 = (L_2(\Omega))^3$, and the Sobolev spaces $\mathbb{H}^1 = (W^1_2(\Omega))^3$, $\mathbb{H}^2 = (W^2_2(\Omega))^3$ of vector-functions $w = (w_1, w_2, w_3)$, defined in Ω. A closure of the lineal $\mathcal{L} = \{w \in (C^\infty_0(\Omega))^3 : \nabla \cdot w = 0\}$ by the norm in $\mathbb{L}_2$ is denoted by $\mathbb{H}_\sigma$; $\mathbb{H}^1_\sigma$ is its closure by the norm in $\mathbb{H}^1$. Also, we use $\mathbb{H}^2_\sigma = \mathbb{H}^1_\sigma \cap \mathbb{H}^2$. An orthogonal complement to $\mathbb{H}_\sigma$ in $\mathbb{L}_2$ is denoted by $\mathbb{H}_\pi$. The corresponding orthoprojectors are $\Sigma : \mathbb{L}_2 \to \mathbb{H}_\sigma$, $\Pi = I - \Sigma : \mathbb{L}_2 \to \mathbb{H}_\pi$.

Consider an operator $A = \Sigma\Delta$ in $\mathcal{L}$. The operator A, extended to a closed operator in $\mathbb{H}_\sigma$, with a domain $\mathbb{H}^2_\sigma$, is known (see [23]) to have a real, negative discrete spectrum of finite multiplicity, condensing at $-\infty$ only. Its eigenvalues are denoted by $\{\lambda_k\}$, numbered in non-increasing, counting their multiplicities. The orthonormal system of corresponding eigenfunctions $\{\varphi_k\}$ is known to form a basis in $\mathbb{H}_\sigma$.

Choose spaces and operators as

$$\mathcal{X} = \mathbb{H}^2_\sigma \times \mathbb{H}_\pi, \quad \mathcal{Y} = \mathbb{L}_2 = \mathbb{H}_\sigma \times \mathbb{H}_\pi, \quad \mathcal{U} = \mathbb{L}_2, \tag{5.7}$$

$$L = \begin{pmatrix} I - \chi A & \mathbb{O} \\ -\chi\Pi\Delta & \mathbb{O} \end{pmatrix}, \quad M = \begin{pmatrix} \nu A & \mathbb{O} \\ \nu\Pi\Delta & -I \end{pmatrix} \in \mathcal{L}(\mathcal{X};\mathcal{Y}). \tag{5.8}$$

Lemma 4. *Let spaces $\mathcal{X}$ and $\mathcal{Y}$ be defined in* (5.7)*, and operators L and M be defined in* (5.8)*, $\nu, \chi \ne 0$, $\chi^{-1} \notin \sigma(A)$. Then M is $(L,0)$-bounded operator, and*

$$P = \begin{pmatrix} I & \mathbb{O} \\ \nu\Pi\Delta(I - \chi A)^{-1} & \mathbb{O} \end{pmatrix}, \quad Q = \begin{pmatrix} I & \mathbb{O} \\ -\chi\Pi\Delta(I - \chi A)^{-1} & \mathbb{O} \end{pmatrix}. \tag{5.9}$$

Denote

$$\Psi(s,t) = \psi_0(s) + \psi_1(s)t + \dots + \psi_{m-1}(s)\frac{t^{m-1}}{(m-1)!}.$$

Theorem 10. *Let $\nu, \chi \ne 0$, $\chi^{-1} \notin \sigma(A)$, $\alpha, q > 1$, $\psi_k \in \mathbb{H}^2_\sigma$, $k = 0,1,\dots,m-1$, the inequality*

$$\|(1 - \chi\Delta)D^\alpha_t\Psi - \nu\Delta\Psi + (\Psi \cdot \nabla)\Psi\|_{L_q(0,T;\mathbb{L}_2)} \le R$$

is true. Then there exists a solution of the problem (5.1)–(5.6)*.*

P r o o f. From the form of the projector P it follows that (5.4) are Showalter — Sidorov conditions. Besides, there exists a control

$$u_0 = (1 - \chi\Delta)D^\alpha_t\Psi - \nu\Delta\Psi + (\Psi \cdot \nabla)\Psi \in \mathcal{U}_\partial = \{u \in L_q(0,T;\mathbb{L}_2) : \|u\|_{L_q(0,T;\mathbb{L}_2)} \le R\},$$

such that $(\Psi, 0)$ $(r = 0)$ is a strong solution of the problem (5.1)–(5.4) with $u = u_0$, i. e. $(\Psi, 0, u_0) \in \mathcal{W}$.

Define $N(v) = -(v \cdot \nabla)v$, hence, by Sobolev's embedding theorem

$$\|N(v)\|^2_{\mathbb{L}_2} \le C_1\|v\|^4_{\mathbb{W}^1_4} \le C_2\|v\|^4_{\mathbb{H}^2},$$

where $\mathbb{W}^1_4 = (W^1_4(\Omega))^3$. Besides, N doesn't depend on r and is locally Lipschitzian mapping.

Choose $\mathfrak{Y} = \{(v, r) \in C^{m-1}([0, T]; \mathcal{X}) : (D^\alpha_t v, D^\alpha_t r) \in L_q(0, T; \mathcal{X})\}$ with the norm

$$\|x\|_{\mathfrak{Y}} = \|x\|_{C^{m-1}([t_0,T];\mathcal{X})} + \|D^\alpha_t x\|_{L_q(t_0,T;\mathcal{X})}, \quad x = (v, r).$$

The completeness of $\mathfrak{Y}$ can be shown as in the proof of Lemma 4. The functional J is coercive on $\mathcal{Z}_{\alpha,q}$ because of the estimate

$$\|Mx\|_{L_q(0,T;\mathbb{L}_2)} \le C_1\|D^\alpha_t v\|_{L_q(0,T;\mathbb{H}^2_\sigma)} + \|u\|_{L_q(0,T;\mathbb{L}_2)} + \max_{\|v\|_{\mathbb{H}^2_\sigma} \le 1} \|N(v)\|_{\mathbb{L}_2}.$$

The required statement follows from Theorem 9. □

REFERENCES

1. **Bajlekova E.G.** Fractional Evolution Equations in Banach Spaces // PhD thesis, Eindhoven University of Technology, University Press Facilities, 2001.
2. **Mainardi F., Spada G.** Creep, relaxation and viscosity properties for basic fractional models in rheology // The European Physics Journal, Special Topics, 2011. Vol. 193. P. 133–160.
3. **Sviridyuk G.A., Fedorov V.E.** Linear Sobolev Type Equations and Degenerate Semigroups of Operators. VSP, Utrecht, Boston, 2003.
4. **Fedorov V.E., Davydov P.N.** On nonlocal solutions of semilinear equations of the Sobolev type // Differential Equations. 2013. Vol. 49, no. 3. P. 338–347. DOI: 10.1134/S0012266113030087
5. **Fedorov V.E., Gordievskikh D.M.** Resolving operators of degenerate evolution equations with fractional derivative with respect to time // Russian Math., 2015. Vol. 59. P. 60–70. DOI:10.3103/S1066369X15010065
6. **Fedorov V.E., Gordievskikh D.M., Plekhanova M.V.** Equations in Banach spaces with a degenerate operator under a fractional derivative // Differential Equations, 2015. Vol. 51. P. 1360–1368.
7. **Fedorov V.E., Debbouche A.** A class of degenerate fractional evolution systems in Banach spaces // Differential Equations, 2013. Vol. 49, no. 12. P. 1569–1576. DOI: 10.1134/S0012266113120112.
8. **Gordievskikh D.M., Fedorov V.E.** Solutions for initial boundary value problems for some degenerate equations systems of fractional order with respect to the time // The Bulletin of Irkutsk State University. Series Mathematics, 2015. Vol. 12. P. 12–22.
9. **Plekhanova M.V.** Quasilinear equations that are not solved the higher-order time derivative // Siberian Mathematical Journal, 2015. Vol. 56. P. 725–735.
10. **Plekhanova M.V.** Nonlinear equations with degenerate operator at fractional Caputo derivative // Mathematical Methods in the Applied Sciences. 2016. In press. DOI: 10.1002/mma.3830
11. **Kostic M.** Abstract time-fractional equations: Existence and growth of solutions // Fractional Calculus and Applied Analysis. 2011. Vol. 14, no. 2. P. 301–316.
12. **Glushak A.V.** Correctness of Cauchy-type problems for abstract differential equations with fractional derivatives // Russian Mathematics. 2009. Vol. 53, no. 9. P. 10–19.
13. **Vorob'eva S.A., Glushak A.V.** An abstract EulerPoissonDarboux equation containing powers of an unbounded operator // Differential Equations. 2001. Vol. 37, no. 5. P. 743–746.
14. **Fursikov, A.V.** Optimalnoe upravlenie raspredelennymi sistemami. Teoriya i prilozheniya. Optimal Control of Distributed Systems. Theory and Applications, Novosibirsk, 1999. [In Russian]
15. **Fedorov V.E., Plekhanova M.V.** Optimal control of Sobolev type linear equations // Differential equations, 2004. Vol. 40. P. 1548–1556.
16. **Fedorov V.E., Plekhanova M.V.** The problem of start control for a class of semilinear distributed systems of Sobolev type // Proceeding of the Steklov institute if mathematics. 2011. Vol. 275. P. 40–48.

17. **Plekhanova M.V.** Distributed control problems for a class of degenerate semilinear evolution equations // J. of Computational and Applied Mathematics, 2017. Vol. 312. P. 39–46. http://dx.doi.org/10.1016/j.cam.2015.09.034
18. **Kochubei N.** Fractional-parabolic systems // Potential Anal. 2012. Vol. 37. P. 1–30.
19. **Kamocki R.** On the existence of optimal solutions to fractional optimal control problems // Applied Mathematics and Computation. 2014. Vol. 235. P. 94–104.
20. **Kerboua M., Debbouche A., Baleanu D.** Approximate controllability of Sobolev type fractional stochastic nonlocal nonlinear differential equations in Hilbert spaces // Electronic J. of Qualitative Theory of Differential Equations, 2014. Vol. 58. P. 1–16.
21. **Showalter R.E.** Nonlinear degenerate evolution equations and partial differential equations of mixed type // SIAM J. Math. Anal., 1975. Vol. 6. P. 25–42. DOI: 10.1137/0506004
22. **Sidorov N.A.** A class of degenerate differential equations with convergence // Math. Notes., 1984. Vol. 35. P. 300–305.
23. **Ladyzhenskaya O.A.** The Mathematical Theory of Viscous Incompressible Flow. Gordon and Breach, Science Publishers, New York, London, Paris, 1969.

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

REGULARIZATION OF PONTRYAGIN MAXIMUM PRINCIPLE IN OPTIMAL CONTROL OF DISTRIBUTED SYSTEMS[1]

Mikhail I. Sumin

Nizhnii Novgorod State University,
Nizhnii Novgorod, Russia, m.sumin@mail.ru

Abstract: This article is devoted to studying dual regularization method applied to parametric convex optimal control problem of controlled third boundary-value problem for parabolic equation with boundary control and with equality and inequality pointwise state constraints. This dual regularization method yields the corresponding necessary and sufficient conditions for minimizing sequences, namely, the stable, with respect to perturbation of input data, sequential or, in other words, regularized Lagrange principle in nondifferential form and Pontryagin maximum principle for the original problem. Regardless of the fact that the stability or instability of the original optimal control problem, they stably generate a minimizing approximate solutions in the sense of J. Warga for it. For this reason, we can interpret these regularized Lagrange principle and Pontryagin maximum principle as tools for direct solving unstable optimal control problems and reducing to them unstable inverse problems.

Key words: Optimal boundary control, Parabolic equation, Minimizing sequence, Dual regularization, Stability, Lagrange principle, Pontryagin maximum principle

Introduction

Pontryagin maximum principle is the central result of all optimal control theory, including optimal control for differential equations with partial derivatives. Its statement and proof assume, first of all, that the optimal control problem is considered in an ideal situation, when its input data are known exactly. However, in the vast number of important practical problems of optimal control, as well as numerous problems reducing to optimal control problems, the requirement of exact defining input data is very unnatural, and in many undoubtedly interest cases is simply impracticable. In similar problems, we can not, strictly speaking, take as an approximation to the solution of the initial (unperturbed) problem with the exact input data, a control formally satisfying the maximum principle in the perturbed problem. The reason of such situation lies in the natural instability of optimization problems with respect to perturbation of its input data. As a typical property of optimization problems in general, including constrained ones, instability fully manifests itself in optimal control problems (see., e.g., [10]). As a consequence, the above mentioned instability implies "instability" of the classical optimality conditions, including the conditions in the form of Pontryagin maximum principle. This instability manifests itself in selecting arbitrarily distant "perturbed" optimal elements from their unperturbed counterparts in the case of an arbitrarily small perturbations of the input data. The above applies, in full measure, both to discussed below optimal control problem with pointwise state constraints for linear parabolic equation in divergent form, and to the classical optimality conditions in the form of the Lagrange principle and the Pontryagin maximum principle for this problem.

[1]This work was supported by the Russian Foundation for Basic Research (project no. 15–47–02294–r_povolzh'e_ – a), by the Ministry of Education and Science of the Russian Federation within the framework of project part of state tasks in 2014–2016 (code no. 1727) and by the grant within the agreement of August 27, 2013 No. 02.B.49.21.0003 between the Ministry of Education and Science of the Russian Federation and Lobachevskii State University of Nizhnii Novgorod.

In this paper we discuss how to overcome the problem of instability of the classical optimality conditions in optimal control problems applying dual regularization method (see., e.g., [11–13]) and simultaneous transition to the concept of minimizing sequence of admissible elements as the main concept of optimization theory. The latter role acts the concept of the minimizing approximate solution in the sense of J. Warga [23]. The main attention in the paper is given to the discussion of the so–called regularized or, in other words, stable, with respect to perturbation of input data, sequential Lagrange principle in the nondifferential form and Pontryagin maximum principle. Regardless of the stability or instability of the original optimal control problem, they stably generate minimizing approximate solutions for it. For this reason, we can interpret the regularized Lagrange principle and Pontryagin maximum principle that are obtained in the article as tools for direct solving unstable optimal control problems and reducing to them unstable inverse problems [10,14–16]. Thus, they contribute to a significant expansion of the range of applicability of the theory of optimal control in which a central role belongs to classic constructions of the Lagrange and Hamilton–Pontryagin functions. Finally, we note that discussed in this paper regularized Lagrange principle in the nondifferential form and Pontryagin maximum principle may have another kind, more convenient for applications [4,9,15]. Justification of these alternative forms of the regularized Lagrange principle and Pontryagin maximum principle is based on the so-called method of iterative dual regularization [11,12]. In this case, they take the form of iterative processes with the corresponding stopping rules when the error of input data is fixed and finite. Here these alternative forms are not considered.

1. Statement of optimal control problem

We consider the fixed-time parametric optimal control problem

$$g_0^\delta(\pi) \to \min, \quad \pi \equiv (u,w) \in \mathcal{D} \subset L_2(Q_T) \times L_2(S_T), \tag{$P_{p,r}^\delta$}$$

$$g_1^\delta(\pi)(x,t) \equiv \varphi_1^\delta(x,t) z^\delta[\pi](x,t) = h^\delta(x,t) + p(x,t) \quad \text{for a.e. } (x,t) \in Q,$$

$$g_2^\delta(\pi)(x,t) \equiv \varphi_2^\delta(x,t,z^\delta[\pi](x,t)) \le r(x,t) \quad \text{for a.e. } (x,t) \in Q$$

with equality and inequality pointwise state constraints understood as ones in the Hilbert space $\mathcal{H} \equiv L_2(Q)$;

$$\mathcal{D} \equiv \{u \in L_2(Q_T) : u(x,t) \in U \text{ for a.e. } (x,t) \in Q_T\} \times \{w \in L_2(S_T) : w(x,t) \in W \text{ for a.e. } (x,t) \in S_T\};$$

$U, W \subset \mathbb{R}^1$ are convex compact sets. In this problem, $p \in \mathcal{H}$ and $r \in \mathcal{H}$ are parameters; $g_0^\delta : L_2(Q_T) \times L_2(S_T)$ is a continuous convex functional, $Q \subset \overline{Q}_{\iota,T}$ is a compact set without isolated points with a nonempty interior, $\iota \in (0,T)$, $Q = \operatorname{cl}\operatorname{int} Q$; and $z^\delta[\pi] \in V_2^{1,0}(Q_T) \cap C(\overline{Q}_T)$ is a weak solution [6] to the third boundary-value problem[2]

$$z_t - \frac{\partial}{\partial x_i}(a_{i,j}(x,t) z_{x_j}) + a^\delta(x,t) z + u(x,t) = 0, \tag{1.1}$$

$$z(x,0) = v_0(x), \quad x \in \Omega, \qquad \frac{\partial z}{\partial \mathcal{N}} + \sigma^\delta(x,t) z = w(x,t), \quad (x,t) \in S_T.$$

The superscript δ in the input data of Problem ($P_{p,r}^\delta$) indicates that these data are exact ($\delta = 0$) or perturbed ($\delta > 0$), i.e., they are specified with an error, $\delta \in [0, \delta_0]$, where $\delta_0 > 0$ is a fixed number.

[2]Here and below, we use the notations for the sets Q_T, S_T, $Q_{i,T}$ and also for functional spaces and norms of their elements adopted in monograph [6].

For definiteness, as a target functional we take the terminal one

$$g_0^\delta(\pi) \equiv \int_\Omega G^\delta(x, z^\delta[\pi](x,T))dx.$$

The input data for Problem $(P_{p,r}^0)$ are assumed to meet the following conditions:

a) It is true that $a_{i,j} \in L_\infty(Q_T)$, $i,j = 1,\dots,n$, $a^\delta \in L_\infty(Q_T)$, $\sigma^\delta \in L_\infty(S_T)$, $v_0^\delta \in C(\overline{\Omega})$,

$$\nu|\xi|^2 \le a_{i,j}(x,t)\xi_i\xi_j \le \mu|\xi|^2 \quad \forall (x,t) \in Q_T, \quad \nu, \mu > 0,$$

$$a^\delta(x,t) \ge C_0 \text{ for a.e. } (x,t) \in Q_T, \quad \sigma^\delta(x,t) \ge C_0 \quad \text{for a.e. } (x,t) \in S_T;$$

b) It is true that ϕ_1^δ, $h^\delta \in L_\infty(Q)$; $\phi_2^\delta : Q \times \mathbb{R}^1 \to \mathbb{R}^1$ is Lebesgue measurable function that is continuous and convex with respect to z for a.e. $(x,t) \in Q$, $\varphi_2^\delta(\cdot,\cdot,z(\cdot,\cdot)) \in L_\infty(Q)$ $\forall z \in C(Q)$; $G^\delta : \Omega \times \mathbb{R}^1 \to \mathbb{R}^1$ is Lebesgue measurable function that is continuous and convex with respect to z for a.e. $x \in \Omega$, $G^\delta(\cdot, z(\cdot,T)) \in L_\infty(\Omega)$ $\forall z(\cdot,T) \in C(Q)$;

c) $\Omega \subset \mathbb{R}^n$ be a bounded domain with piece–wise smooth boundary S.

Assume that the following estimates hold:

$$\begin{gathered}
|G^\delta(x,z) - G^0(x,z)| \le C_M\delta \quad \forall\, (x,z) \in \Omega \times S_M^1, \\
\|\varphi_1^\delta - \varphi_1^0\|_{\infty,Q} \le C\delta, \quad \|h^\delta - h^0\|_{\infty,Q} \le C\delta, \\
|\varphi_2^\delta(x,t,z) - \varphi_2^0(x,t,z)| \le C_M\delta \quad \forall\, (x,t,z) \in Q \times S_M^1, \\
\|a^\delta - a^0\|_{\infty,Q_T} \le C\delta, \quad |v_0^\delta - v_0^0|_{\overline{\Omega}}^{(0)} \le C\delta, \quad \|\sigma^\delta - \sigma^0\|_{\infty,S_T} \le C\delta,
\end{gathered} \tag{1.2}$$

where C, $C_M > 0$ are independent of δ; $S_M^n \equiv \{x \in \mathbb{R}^n : |x| < M\}$. Let's note, that the conditions on the input data of Problem $(P_{p,r}^\delta)$, and also the estimates of deviations of the perturbed input data from the exact ones can be weakened.

2. Basic concepts and auxiliary propositions

In this paper we use for discussing the main results, related to the stable sequential Lagrange principle and Pontryagin maximum principle in Problem $(P_{p,r}^0)$, a scheme of studying the similar optimization problems in the papers [17,19] for a system of controlled ordinary differential equations (see also [20,21] for the case of distributed systems). In these works, both spaces of admissible controls and spaces, containing lie images of the operators that define the pointwise state constraints, were presented as Hilbert spaces of square-integrable functions. For this reason, we put the set $\mathcal{D}$ of admissible controls π into a Hilbert space also, i.e., assume that

$$\mathcal{D} \subset Z \equiv L_2(Q_T) \times L_2(S_T), \quad \|\pi\| \equiv (\|u\|_{2,Q_T}^2 + \|w\|_{2,S_T}^2)^{1/2}.$$

At the same time, we note that the conditions on the input data of Problem $(P_{p,r}^\delta)$ allow formally to consider that the operators g_1^δ, g_2^δ, specifying the state constraints of the problem, act into space $L_p(Q)$ with any index $p \in [1, +\infty]$. However, in this paper, taking into account the above remark, we will put images of these functional operators in the Hilbert space $L_2(Q) \equiv \mathcal{H}$. We note here that the imbedding the images of the operators g_1^δ, g_2^δ, specifying the state constraints, into reflexive space $L_p(Q)$ with $1 < p < 2$, in general, permits significantly to weaken the conditions on the input data and to get, strictly speaking, a stronger result in Problem $(P_{p,r}^0)$.

If Problem $(P^0_{p,r}\ )$ is solvable (it has a unique solution if g^0_0 is strictly (strongly) convex), then its solutions are denoted by $\pi^0_{p,r} \equiv (u^0_{p,r}, w^0_{p,r})$, and the set of all such solutions is designated as $U^0_{p,r}$. Define the Lagrange functional, a set of its minimizers and the concave dual problem

$$L^\delta_{p,r}(\pi,\lambda,\mu) \equiv g^\delta_0(\pi) + \langle \lambda, g^\delta_1(\pi) - h^\delta - p\rangle + \langle \mu, g^\delta_2(\pi) - r\rangle, \quad \pi \in \mathcal{D},$$

$$U^\delta[\lambda,\mu] \equiv \operatorname{Argmin}\{L^\delta_{p,r}(\pi,\lambda,\mu) : \pi \in \mathcal{D}\} \quad \forall (\lambda,\mu) \in \mathcal{H}\times\mathcal{H}_+,$$

$$\mathcal{H}_+ \equiv \{z \in \mathcal{H} :\ z(x,t) \geq 0 \ \text{ for a.e.} \quad (x,t) \in Q\},$$

$$V^\delta_{p,r}(\lambda,\mu) \to \sup, \quad (\lambda,\mu) \in \mathcal{H}\times\mathcal{H}_+, \quad V^\delta_{p,r}(\lambda,\mu) \equiv \inf_{\pi\in\mathcal{D}} L^\delta_{p,r}(\pi,\lambda,\mu).$$

Since the Lagrange functional is continuous and convex for any pair $(\lambda,\mu) \in \mathcal{H}\times\mathcal{H}_+$, and the set $\mathcal{D}$ is bounded, the dual functional $V^\delta_{p,r}$, is obviously defined and finite for any $(\lambda,\mu) \in \mathcal{H}\times\mathcal{H}_+$.

The concept of a minimizing approximate solution in the sense of J. Warga [23] is of great importance for the design of a dual regularizing algorithm for Problem $(P^0_{p,r})$. Recall that a minimizing approximate solution is a sequence $\pi^i \equiv (u^i, w^i)$, $i = 1, 2, \ldots$ such that $g^0_0(\pi^i) \leq \beta(p,r) + \delta^i$, $\pi^i \in \mathcal{D}^{0,\epsilon^i}_{p,r}$ for some nonnegative number sequences δ^i and ϵ^i, $i = 1, 2, \ldots$, that converge to zero. Here, $\beta(p,r)$ is the generalized infimum, i.e., S-function:

$$\beta(p,r) \equiv \lim_{\epsilon\to+0} \beta_\epsilon(p,r), \quad \beta_\epsilon(p,r) \equiv \inf_{\pi\in\mathcal{D}^{0,\epsilon}_{p,r}} g^0_0(\pi), \qquad \beta_\epsilon(p,r) \equiv +\infty \text{ if } \mathcal{D}^{0,\epsilon}_{p,r} = \emptyset,$$

$$\mathcal{D}^{\delta,\epsilon}_{p,r} \equiv \{\pi \in \mathcal{D} :\ \|g^\delta_1(\pi) - h^\delta - p\|_{2,Q} \leq \epsilon,\ \min_{z\in\mathcal{H}_-} \|g^\delta_2(\pi) - r - z\|_{2,Q} \leq \epsilon\}, \quad \epsilon \geq 0,$$

$$\mathcal{D}^{00}_{p,r} \equiv \mathcal{D}^0_{p,r}, \quad \mathcal{H}_- \equiv \{z \in \mathcal{H} :\ z(x,t) \leq 0 \ \text{ for a.e. } (x,t) \in Q\}, \quad \mathcal{H}_+ \equiv -\mathcal{H}_-.$$

Obviously, in the general situation, $\beta(p,r) \leq \beta_0(p,r)$, where $\beta_0(p,r)$ is the classical value of the problem. However, in the case of Problem $(P^0_{p,r})$, we have $\beta(p,r) = \beta_0(p,r)$. Simultaneously, we may asset that $\beta :\ \mathcal{H}\times\mathcal{H} \to \mathbb{R}^1 \cup \{+\infty\}$ is a convex and lower semicontinuous function. Note here that the existence of a minimizing approximate solution in Problem $(P^0_{p,r})$ obviously implies its solvability.

From the conditions a) – c) and from the theorem on the existence of a weak solution of the third boundary-value problem for a linear parabolic equation of the divergent type [6, ch. III, section 5] (see also [5, 7]), it follows that the direct boundary-value problem (1.1) and the corresponding adjoint problem are uniquely solvable in $V^{1,0}_2(Q_T)$.

Proposition 1. *For any pair $(u,w) \in L_2(Q_T)\times L_2(S_T)$ and for any $T > 0$ the direct boundary-value problem (1.1) is uniquely solvable in $V^{1,0}_2(Q_T)$ and we have the estimate*

$$|z^\delta[\pi]|_{Q_T} + \|z^\delta[\pi]\|_{2,S_T} \leq C_T(\|u\|_{2,Q_T} + \|v_0\|_{2,\Omega} + \|w\|_{2,S_T})$$

where the constant C_T is independent of $\delta \geq 0$ and pair $\pi \equiv (u,w) \in L_2(Q_T)\times L_2(S_T)$. Also the adjoint problem

$$-\eta_t - \frac{\partial}{\partial x_j} a_{i,j}(x,t)\eta_{x_i} + a^\delta(x,t)\eta = \chi(x,t),$$

$$\eta(x,T) = \psi(x), \quad x \in \Omega, \qquad \frac{\partial\eta}{\partial\mathcal{N}} + \sigma^\delta(x,t)\eta = \omega(x,t), \quad (x,t) \in S_T$$

is uniquely solvable in $V^{1,0}_2(Q_T)$ for any $\chi \in L_2(Q_T)$, $\psi \in L_2(\Omega)$, $\omega \in L_2(S_T)$ and any $T > 0$. Its solution is denoted as $\eta[\chi,\psi,\omega]$. Simultaneously, the estimate

$$|\eta^\delta[\chi,\psi,\omega]|_{Q_T} + \|\eta^\delta[\chi,\psi,\omega]\|_{2,S_T} \leq C^1_T(\|\chi\|_{2,Q_T} + \|\psi\|_{2,\Omega} + \|\omega\|_{2,S_T}),$$

is true where the constant C^1_T is independent of $\delta \geq 0$ and a triple (χ,ψ,ω).

Simultaneously, from conditions a) – c) and the theorems on the existence of a weak (generalized) solution of the third boundary-value problem for a linear parabolic equation of the divergent type (see, e.g., [3, 8]), it follows that the direct boundary-value problem is uniquely solvable in $V_2^{1,0}(Q_T) \cap C(\overline{Q}_T)$.

Proposition 2. *Let us $l > n+1$. For any pair $(u, w) \in L_l(Q_T) \times L_l(S_T)$ and any $T > 0$, $\delta \in [0, \delta_0]$ the direct boundary-value problem (1.1) is uniquely solvable in $V_2^{1,0}(Q_T) \cap C(\overline{Q}_T)$ and the estimate*

$$|z^\delta[\pi]|^{(0)}_{\overline{Q}_T} \le C_T(\|u\|_{l,Q_T} + |v_0|^{(0)}_{\overline{\Omega}} + \|w\|_{l,S_T}),$$

is true where the constant C_T is independent of pair $\pi \equiv (u, w)$ and δ.

Further, the minimization problem for Lagrange functional

$$L^\delta_{p,r}(\pi, \lambda, \mu) \to \min, \quad \pi \in \mathcal{D}, \quad \text{when} \quad (\lambda, \mu) \in \mathcal{H} \times \mathcal{H}_+ \tag{2.1}$$

plays the central role in all subsequent constructions. It is usual problem without equality and inequality constraints. It is solvable as a minimization problem for weakly semicontinuous functional on the weak compact set $\mathcal{D} \subset L_2(Q_T) \times L_2(S_T)$. Here, the weak semicontinuity is a consequence of the convexity and continuity with respect to π of the Lagrange functional. Minimizers $\pi^\delta[\lambda, \mu] \in U^\delta[\lambda, \mu]$ for this optimal control problem satisfy the Pontryagin maximum principle under supplementary assumption of the existence of Lebesgue measurable with respect to $(x, t) \in Q$ for all $z \in \mathbb{R}^1$ and continuous with respect to z for a.e. x, t gradients $\nabla_z \varphi^\delta_2(x, t, z)$, $\nabla_z G^\delta(x, z)$ with the estimates

$$|\nabla_z \varphi^\delta_2(x, t, z)| \le C_M, \quad |\nabla_z G^\delta(x, z)| \le C_M, \quad \forall z \in S^1_M,$$

where $C_M > 0$ is independent of δ. The following lemma is true due to the estimates of the propositions 1, 2 and to the so called two-parameter variation [22] of the pair $\pi^\delta[\lambda, \mu]$ that is needle-shaped with respect to control u and classical with respect to control w.

Lemma 1. *Let $H(y, \eta) \equiv -\eta y$ and the additional condition that specified above is fulfilled. Any pair $\pi^\delta[\lambda, \mu] = (u^\delta[\lambda, \mu], w^\delta[\lambda, \mu]) \in U^\delta[\lambda, \mu]$, $(\lambda, \mu) \in \mathcal{H} \times \mathcal{H}_+$ satisfies the (usual) Pontryagin maximum principle in the problem* (2.1): *for $\pi = \pi^\delta[\lambda, \mu]$ the following maximum relations*

$$H(u(x, t), \eta^\delta(x, t)) = \max_{u \in U} H(u, \eta^\delta(x, t)) \text{ for a.e. } Q_T, \tag{2.2}$$

$$H(w(s, t), \eta^\delta(s, t)) = \max_{w \in W} H(w, \eta^\delta(s, t)) \text{ for a.e. } S_T$$

hold, where $\eta^\delta(x, t)$, $(x, t) \in Q_T$ is a solution for $\pi = \pi^\delta[\lambda, \mu]$ of the adjoint problem

$$-\eta_t - \frac{\partial}{\partial x_j}(a_{i,j}(x, t)\eta_{x_i}) + a^\delta(x, t)\eta = \varphi^\delta_1(x, t)\lambda(x, t) + \nabla_z \varphi^\delta_2(x, t, z^\delta[\pi](x, t))\mu(x, t), \quad (x, t) \in Q_T,$$

$$\eta(x, T) = \nabla_z G^\delta(x, z^\delta[\pi](x, T)), \quad x \in \Omega,$$

$$\frac{\partial \eta(x, t)}{\partial \mathcal{N}} + \sigma^\delta(x, t)\eta = 0, \quad (x, t) \in S_T.$$

Remark 1. Note that here and below, if the functions φ^δ_1, $\nabla_z \varphi^\delta_2(\cdot, \cdot, z(\cdot, \cdot))$, $\lambda, \mu \in \mathcal{H}$ are considered on the entire cylinder Q_T, we set that the equalities $\varphi^\delta_1(x, t) = \nabla_z \varphi^\delta_2(x, t, z(x, t)) = \lambda(x, t) = \mu(x, t) = 0$ take place for $(x, t) \in Q_T \setminus Q$; the same notation is preserved if these functions are taken on the entire cylinder.

An important result for the subsequent presentation is the following lemma, which is a consequence of the classical asymmetric minimax theorem [2, Chapter 6, Section 2, Theorem 7].

Lemma 2. *The minimax equality*

$$\inf_{\pi\in\mathcal{D}}\sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+} L^0_{p,r}(\pi,\lambda,\mu) = \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+}\inf_{\pi\in\mathcal{D}} L^0_{p,r}(\pi,\lambda,\mu),$$

is true. It can be rewritten as the duality relation

$$g_0^0(\pi^0_{p,r}) = \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+} V^0_{p,r}(\lambda,\mu). \tag{2.3}$$

In the next section we construct minimizing approximate solutions for Problem $(P^0_{p,r})$ from the elements $\pi^\delta[\lambda,\mu]$, $(\lambda,\mu)\in\mathcal{H})\times\mathcal{H}_+$. As consequence, this construction leads us to various versions of the stable sequential Lagrange principle and Pontragin maximum principle. In the case of strong convexity and subdifferentiability of the target functional g_0^0, these versions are statements about stable approximations of the solutions of Problem $(P^0_{p,r})$ in the metric of $Z\equiv L_2(Q_T)\times L_2(S_T)$ by the points $\pi^\delta[\lambda,\mu]$. Due to the estimates (1.2) and the propositions 1, 2 we may assert that the estimates

$$|g_0^\delta(\pi)-g_0^0(\pi)|\le C_1\delta \quad \forall\pi\in\mathcal{D}, \quad \|g_1^\delta(\pi)-g_1^0(\pi)\|_{2,Q}\le C_2\delta(1+\|\pi\|) \quad \forall\pi\in Z, \tag{2.4}$$

$$\|h^\delta-h^0\|_{2,Q}\le C\delta, \quad \|g_2^\delta(\pi)-g_2^0(\pi)\|_{2,Q}\le C_3\delta \quad \forall\pi\in\mathcal{D},$$

hold, in which the constants C_1, C_2, $C_3>0$ are independent of $\delta\in(0,\delta_0]$, π.

Since the set $\mathcal{D}$ is bounded, the dual functional is obviously defined and finite for any element $(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}$. Moreover, it is also obvious that the value $V^\delta_{p,r}(\lambda,\mu)$ is reached at elements $\pi^\delta[\lambda,\mu]$ of the set $U^\delta[\lambda,\mu]\equiv \mathrm{Argmin}\,\{L^\delta_{p,r}(\pi,\lambda,\mu),\ \pi\in\mathcal{D}\}$ for $(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+$,

$$\mathcal{H}_+\equiv\{z\in\mathcal{H}:\ z(x,t)\ge 0 \text{ for a.e. } (x,t)\in Q\}.$$

Note also that, by virtue of estimates (2.4) and since $\mathcal{D}$ is bounded, we have the estimate

$$|V^\delta_{p,r}(\lambda,\mu)-V^0_{p,r}(\lambda,\mu)|\le C\delta(1+\|\lambda\|+\|\mu\|), \tag{2.5}$$

where $C>0$ is a constant independent of λ, μ, δ.

3. Stable sequential Pontryagin maximum principle

In this section we discuss the so-called regularized or, in other words, stable, with respect to errors of input data, sequential Pontryagin maximum principle for Problem $(P^0_{p,r})$ as necessary and sufficient condition for elements of minimizing approximate solutions. Simultaneously, we may treat this condition as one for existence of a minimizing approximate solutions in Problem $(P^0_{p,r})$ with perturbed input data or as condition of stable construction of a minimizing sequence in this problem. The proof of the necessity of this condition is based on the dual regularization method [11–13] that is a stable algorithm of constructing a minimizing approximate solutions in Problem $(P^0_{p,r})$.

3.1. Dual regularization for optimal control problem with pointwise state constraints

The estimates (2.4) give a possibility to organize the procedure of the dual regularization in accordance with a scheme of the paper [19] for constructing a minimizing approximate solution in Problem $(P^0_{p,r})$. In accordance with this scheme the dual regularization consists in the direct solving dual of Problem $(P^0_{p,r})$ and Tikhonov stabilized problem

$$R^{\delta,\alpha(\delta)}_{p,r}(\lambda,\mu)\equiv V^\delta_{p,r}(\lambda,\mu)-\alpha(\delta)\|(\lambda,\mu)\|^2\to\max, \quad (\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+$$

under consistency condition

$$\frac{\delta}{\alpha(\delta)} \to 0, \quad \alpha(\delta) \to 0, \quad \delta \to 0. \tag{3.1}$$

Let us denote $(\lambda_{p,r}^{\delta,\alpha}, \mu_{p,r}^{\delta,\alpha}) \equiv \operatorname{argmax}\{R_{p,r}^{\delta,\alpha}(\lambda,\mu) : (\lambda,\mu) \in \mathcal{H} \times \mathcal{H}_+\}$. The above dual regularization leads to constructing minimizing approximate solution in Problem $(P_{p,r}^0)$ from the elements $\pi^\delta[\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}] \in \operatorname{Argmin}\{L_{p,r}^\delta(\pi,\lambda,\mu) : \pi \in \mathcal{D}\}$, when $\delta \to 0$.

In this section, we extend the algorithm of the dual regularization [12, 18] to the case of Problem $(P_{p,r}^0)$ in which the objective functional is only convex. Below we prove convergence theorem for dual regularization method in exact accordance with a scheme of proving the similar theorem in [19]. We note only that, as in [19], this proving uses a weak continuity of the operators g_1^δ, g_2^δ that is consequence of the conditions on the input data of Problem $(P_{p,r}^0)$ and a regularity of the bounded solutions of the boundary-value problem (1.1) (see Proposition 2) inside of the cylinder Q_T [6, ch.III, theorem 10.1].

Let Problem $(P_{p,r}^0)$ be solvable. To prove the convergence theorem for dual regularization method, first of all, we give a formula for the superdifferential (in the sense of a convex analysis) of the concave value functional $V_{p,r}^\delta$. The proof of this formula can be found in [12].

Lemma 3. *The superdifferential of the concave value functional $V_{p,r}^\delta(\lambda,\mu)$ at the point $(\lambda,\mu) \in \mathcal{H} \times \mathcal{H}$ is equal*

$$\partial V_{p,r}^\delta(\lambda,\mu) = \partial_C V_{p,r}^\delta(\lambda,\mu) = \operatorname{cl}\operatorname{conv}\{w - \lim_{i\to\infty}(g_1^\delta(u^i) - h^\delta - p, g_2^\delta(u^i) - r) : \pi^i \in \mathcal{D},$$

$$L_{p,r}^\delta(\pi^i,\lambda,\mu) \to \inf_{\pi\in\mathcal{D}} L_{p,r}^\delta(\pi,\lambda,\mu),\ i \to \infty\},$$

where $\partial_C V_{p,r}^\delta(\lambda,\mu)$ is Clarke's generalized gradient of the functional $V_{p,r}^\delta(\lambda,\mu)$ at the point (λ,μ) and the limit $w-\lim$ is understood in the sense of weak convergence in the space $\mathcal{H} \times \mathcal{H}$.

Further, to substantiate the dual regularization method in the case under consideration, we write the inequality $\forall\,(\lambda',\mu') \in \mathcal{H} \times \mathcal{H}_+$

$$\langle (I_1, I_2) - 2\alpha(\delta)(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}), (\lambda',\mu') - (\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)})\rangle \le 0$$

for some element $(I_1, I_2) \in \partial V_{p,r}^\delta(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)})$.

By Lemma 3 and the classical properties of closed convex hulls (see [1, p. 210, 217]), we obtain

$$\langle \lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(w - \lim_{j\to\infty}(g_1^\delta(\pi_{s,i}^j) - h^\delta - p, g_2^\delta(\pi_{s,i}^j)) - r) - 2\alpha(\delta)(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}), \tag{3.2}$$

$$(\lambda',\mu') - (\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)})\rangle \le 0 \quad \forall\,(\lambda',\mu') \in \mathcal{H} \times \mathcal{H}_+,$$

where $\sum\limits_{i=1}^{l(s,\delta)} \gamma_i(s,\delta) = 1$, $\gamma_i(s,\delta) \ge 0$, $i = 1,\dots,l(s,\delta)$, and $\pi_{s,i}^j \in \mathcal{D}$, $j = 1,2,\dots$ is a sequence such that

$$L_{p,r}^\delta(\pi_{s,i}^j, \lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}) \to \min_{\pi\in\mathcal{D}} L_{p,r}^\delta(\pi, \lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}),\ j \to \infty.$$

Assume without loss of generality that the sequence $\pi_{s,i}^j \in \mathcal{D}$, $j = 1,2,\dots$, converges weakly as $j \to \infty$ to an element $\pi_{s,i} \in \mathcal{D}$, which obviously belongs to the set $U^\delta[\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}]$. Due to weak continuity of the operators g_i^δ, $i = 1,2$, and boundedness of $\mathcal{D}$, from (3.2) the inequality follows

$$\langle \lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_1^\delta(\pi_{s,i}) - h^\delta - p, g_2^\delta(\pi_{s,i}) - r) - 2\alpha(\delta)(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}),$$

$$(\lambda', \mu') - (\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)})\rangle \le 0 \quad \forall\, (\lambda', \mu') \in \mathcal{H} \times \mathcal{H}_+.$$

The above inequality implies the limit relations

$$\lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_1^\delta(\pi_{s,i}) - h^\delta - p) = 2\alpha(\delta)\lambda_{p,r}^{\delta,\alpha(\delta)}, \tag{3.3}$$

$$\lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_2^\delta(\pi_{s,i})(x,t) - r(x,t)) = 2\alpha(\delta)\mu_{p,r}^{\delta,\alpha(\delta)}(x,t) \tag{3.4}$$

$$\text{for a.e. } (x,t) \in \{(x,t) \in Q :\, \mu_{p,r}^{\delta,\alpha(\delta)}(x,t) > 0\},$$

$$\lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_2^\delta(\pi_{s,i})(x,t) - r(x,t)) \le 0 \quad \text{for a.e. } (x,t) \in \{(x,t) \in Q :\, \mu_{p,r}^{\delta,\alpha(\delta)}(x,t) = 0\}. \tag{3.5}$$

In turn, the limit relations (3.3)–(3.5) imply the limit equalities

$$\lim_{s\to\infty} \langle \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_1^\delta(\pi_{s,i}) - h^\delta - p), \lambda_{p,r}^{\delta,\alpha(\delta)}\rangle = 2\alpha(\delta)\|\lambda_{p,r}^{\delta,\alpha(\delta)}\|^2 \ge 0, \tag{3.6}$$

$$\lim_{s\to\infty} \langle \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta) g_2^\delta(\pi_{s,i}) - r, \mu_{p,r}^{\delta,\alpha(\delta)}\rangle = 2\alpha(\delta)\|\mu_{p,r}^{\delta,\alpha(\delta)}\|^2 \ge 0.$$

From (3.4) we obtain also: if $\mu_{p,r}^{\delta,\alpha(\delta)}(x,t) > 0$ for some (x,t) belonging to a set of full measure in $\{(x,t) \in Q :\, \mu_{p,r}^{\delta,\alpha(\delta)}(x,t) > 0\}$, then

$$\lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_2^\delta(\pi_{s,i})(x,t) - r(x,t)) - 2\alpha(\delta)\mu_{p,r}^{\delta,\alpha(\delta)}(x,t) = 0, \tag{3.7}$$

$$\lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_2^\delta(\pi_{s,i})(x,t) - r(x,t))\mu_{p,r}^{\delta,\alpha(\delta)}(t) > 0.$$

This implies that for a.e. $(x,t) \in Q$ such that $\lim\limits_{s\to\infty} \sum\limits_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_2^\delta(\pi_{s,i})(x,t) - r(x,t)) < 0$, the equality $\mu_{p,r}^{\delta,\alpha(\delta)}(x,t) = 0$ holds. From (3.4) and (3.7) we obtain simultaneously that

$$\mu_{p,r}^{\delta,\alpha(\delta)}(x,t) \lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_2^\delta(\pi_{s,i})(x,t) - r(x,t)) \ge 0 \quad \text{for a.e. } (x,t) \in Q.$$

Besides, from (3.6) we get the inequality

$$\begin{gathered}\lim_{s\to\infty} \langle \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g_1^\delta(\pi_{s,i}) - h^\delta - p, g_2^\delta(\pi_{s,i}) - r), (\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)})\rangle = \\ 2\alpha(\delta)(\|\lambda_{p,r}^{\delta,\alpha(\delta)}\|^2 + \|\mu_{p,r}^{\delta,\alpha(\delta)}\|^2) \ge 0.\end{gathered} \tag{3.8}$$

Further, since for any $\pi_{p,r}^0 \in U_{p,r}^0$

$$L_{p,r}^\delta(\pi_{s,i}, \lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}) \equiv g_0^\delta(\pi_{s,i}) + \langle \lambda_{p,r}^{\delta,\alpha(\delta)}, g_1^\delta(\pi_{s,i}) - h^\delta - p\rangle + \langle \mu_{p,r}^{\delta,\alpha(\delta)}, g_2^\delta(\pi_{s,i}) - r\rangle \le$$

$$L^{\delta}_{p,r}(\pi^0_{p,r},\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) \equiv g^{\delta}_0(\pi^0_{p,r}) + \langle\lambda^{\delta,\alpha(\delta)}_{p,r}, g^{\delta}_1(\pi^0_{p,r}) - h^{\delta} - p\rangle + \langle\mu^{\delta,\alpha(\delta)}_{p,r}, g^{\delta}_2(\pi^0_{p,r}) - r\rangle \le$$

$$g^0_0(\pi^0_{p,r}) + [g^{\delta}_0(\pi^0_{p,r}) - g^0_0(\pi^0_{p,r})] + \|\lambda^{\delta,\alpha(\delta)}_{p,r}\|\|g^{\delta}_1(\pi^0_{p,r}) - h^{\delta} - p\| + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|\|g^{\delta}_2(\pi^0_{p,r}) - g^0_2(\pi^0_{p,r})\|,$$

due to the estimates (2.4) and the limit equality (3.8) and doing some elementary transformation, we obtain the estimate

$$2\alpha(\delta)(\|\lambda^{\delta,\alpha(\delta)}_{p,r}\|^2 + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|^2) \le$$

$$C_1\delta\|\lambda^{\delta,\alpha(\delta)}_{p,r}\| + C_1\delta\|\mu^{\delta,\alpha(\delta)}_{p,r}\| + g^0_0(\pi^0_{p,r}) + C_1\delta - \min_{\pi\in\mathcal{D}} g^{\delta}_0(\pi) \le$$

$$\sqrt{2}C_1\delta\sqrt{\|\lambda^{\delta,\alpha(\delta)}_{p,r}\|^2 + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|^2} + g^0_0(\pi^0_{p,r}) + C_1\delta - \min_{\pi\in\mathcal{D}} g^{\delta}_0(\pi)$$

or

$$\alpha(\delta)(\|\lambda^{\delta,\alpha(\delta)}_{p,r}\|^2 + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|^2) - C_2\delta\sqrt{\|\lambda^{\delta,\alpha(\delta)}_{p,r}\|^2 + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|^2} - g^0_0(\pi^0_{p,r}) - C_1\delta + \min_{\pi\in\mathcal{D}} g^{\delta}_0(\pi) \le 0,$$

where C_1, $C_2 > 0$ are independent of constant δ. From here, the estimate follows

$$\sqrt{\|\lambda^{\delta,\alpha(\delta)}_{p,r}\|^2 + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|^2} \le \frac{C_2\delta + \sqrt{(C_2\delta)^2 - 4\alpha(\delta)K(\delta)}}{2\alpha(\delta)},$$

where $K(\delta) \equiv \min\limits_{\pi\in\mathcal{D}} g^{\delta}_0(\pi) - g^0_0(\pi^0_{p,r}) - C\delta$. In turn, this estimate implies the limit realtions

$$\alpha(\delta)\|\lambda^{\delta,\alpha(\delta)}_{p,r}\| \to 0, \quad \alpha(\delta)\|\mu^{\delta,\alpha(\delta)}_{p,r}\| \to 0, \quad \delta \to 0. \tag{3.9}$$

Further, the limit relations (3.3)–(3.5), (3.9) imply

$$\lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g^{\delta}_1(\pi_{s,i}) - h^{\delta} - p) \to 0, \quad \delta \to 0,$$

$$\lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g^{\delta}_2(\pi_{s,i}) - r) \le \phi(\delta), \ \ \|\phi(\delta)\| \to 0, \quad \delta \to 0,$$

where the inequality $\lim\limits_{s\to\infty} \sum\limits_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g^{\delta}_2(\pi_{s,i}) - r) \le \phi(\delta)$ is understood in the sense of ordering on the cone of nonpositive functions $\mathcal{H}_-$.

Denoting by $\pi_\delta \in U^{\delta}[\lambda^{\delta,\alpha(\delta)}_{p,r}, \mu^{\delta,\alpha(\delta)}_{p,r}]$ any weak limit point of the sequence $\sum\limits_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)\pi_{s,i}$, $s = 1, 2, \ldots$ and taking into account the inequality

$$g^{\delta}_2(\sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)\pi_{s,i}) \le \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta) g^{\delta}_2(\pi_{s,i}),$$

which is understood also in the sense of ordering on the cone of nonpositive functions, we obtain the limit relations

$$g^{\delta}_1(\pi_\delta) - h^{\delta} - p \to 0, \quad g^{\delta}_2(\pi_\delta) - r \le \lim_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)(g^{\delta}_2(\pi_{s,i}) - r) \le \phi(\delta), \quad \delta \to 0,$$

and, as a consequence, due to the boundedness of $\mathcal{D}$, the limit relations

$$g^0_1(\pi_\delta) - h^0 - p \to 0, \quad g^0_2(\pi_\delta) - r \le \bar{\phi}(\delta), \quad \|\bar{\phi}(\delta)\| \to 0, \quad \delta \to 0. \tag{3.10}$$

Simultaneously, due to the inclusion $\pi_{s,i} \in U^{\delta}[\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}]$ we have the inequality

$$g_0^{\delta}(\pi_{s,i}) + \langle \lambda_{p,r}^{\delta,\alpha(\delta)}, g_1^{\delta}(\pi_{s,i}) - h^{\delta} - p\rangle + \langle \mu_{p,r}^{\delta,\alpha(\delta)}, g_2^{\delta}(\pi_{s,i}) - r\rangle \leq$$

$$g_0^{\delta}(\pi) + \langle \lambda_{p,r}^{\delta,\alpha(\delta)}, g_1^{\delta}(\pi) - h^{\delta} - p\rangle + \langle \mu_{p,r}^{\delta,\alpha(\delta)}, g_2^{\delta}(\pi) - r\rangle \quad \forall\, \pi \in \mathcal{D}.$$

Hence, due to the limit relation (3.8), we can write for any $u_{p,r}^0 \in U_{p,r}^0$

$$\liminf_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta) g_0^{\delta}(\pi_{s,i}) \leq g_0^{\delta}(\pi_{p,r}^0) + \langle \lambda_{p,r}^{\delta,\alpha(\delta)}, g_1^{\delta}(\pi_{p,r}^0) - h^{\delta} - p\rangle + \langle \mu_{p,r}^{\delta,\alpha(\delta)}, g_2^{\delta}(\pi_{p,r}^0) - r\rangle.$$

In turn, from here, due to the consistency condition (3.1), the estimates (1.2) and the boundedness of $\mathcal{D}$ we derive

$$\liminf_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta) g_0^0(\pi_{s,i}) \leq g_0^0(\pi_{p,r}^0) + \tilde{\phi}(\delta), \quad \tilde{\phi}(\delta) \to 0, \quad \delta \to 0$$

or

$$g_0^0(\pi_{\delta}) \leq \liminf_{s\to\infty} g_0^0\Big(\sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta)\pi_{s,i}\Big) \leq \liminf_{s\to\infty} \sum_{i=1}^{l(s,\delta)} \gamma_i(s,\delta) g_0^0(\pi_{s,i}) \leq$$

$$g_0^0(\pi_{p,r}^0) + \tilde{\phi}(\delta), \quad \tilde{\phi}(\delta) \to 0, \quad \delta \to 0.$$

Thus, by virtue of the boundedness of $\mathcal{D}$, weak lower semicontinuity of g_0^0 and weak continuity of g_i^0, $i = 1, 2$, we constructed the family of elements $\pi_{\delta} \in U^{\delta}[\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}]$, depending on δ, such that the limit relations (3.10) hold and simultaneously

$$g_0^0(\pi_{\delta}) \to \min_{\pi\in\mathcal{D}_{p,r}^0} g_0^0(\pi), \quad \delta \to 0.$$

Moreover, weak limit point $\bar{\pi}$ of any weakly converging sequence $\pi_{\delta^k},\ k = 1, 2, \ldots,\ \delta^k \to 0,\ k \to \infty$, is obviously a solution of Problem $(P_{p,r}^0)$.

We can assert that simultaneously the family of elements $(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)})$, in view of the estimates (1.2), (2.5) and the consistency condition (3.1), satisfies the limit relation (see [12,13,15,18])

$$\lim_{\delta\to+0} V_{p,r}^0(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}) = \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+} V_{p,r}^0(\lambda,\mu), \tag{3.11}$$

which, combined with the estimate (2.5), the consistency condition (3.1), and the duality relation (2.3) yields the limit relation (see [12, 13, 15, 18])

$$\big\langle (\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}), (g_1^{\delta}(\pi_{\delta}) - h^{\delta} - p, g_2^{\delta}(\pi_{\delta})) - r\big\rangle \to 0, \quad \delta \to 0.$$

Let us prove the limit relation (3.11). Since

$$V_{p,r}^{\delta}(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}) - \alpha(\delta)\|\lambda_{p,r}^{\delta,\alpha(\delta)}\|^2 - \alpha(\delta)\|\mu_{p,r}^{\delta,\alpha(\delta)}\|^2 \geq$$

$$V_{p,r}^{\delta}(\lambda,\mu) - \alpha(\delta)\|\lambda\|^2 - \alpha(\delta)\|\mu\|^2 \quad \forall\, (\lambda,\mu) \in \mathcal{H}\times\mathcal{H}_+,$$

we can write, thanks to (2.5), the estimates

$$V_{p,r}^{\delta}(\lambda_{p,r}^{\delta,\alpha(\delta)}, \mu_{p,r}^{\delta,\alpha(\delta)}) \geq V_{p,r}^0(\lambda,\mu) + \alpha(\delta)\|\lambda_{p,r}^{\delta,\alpha(\delta)}\|^2 + \alpha(\delta)\|\mu_{p,r}^{\delta,\alpha(\delta)}\|^2 -$$

$$C\delta(1 + \|\lambda\| + \|\mu\|) - \alpha(\delta)\|\lambda\|^2 - \alpha(\delta)\|\mu\|^2,$$

$$V^0_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) = V^\delta_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) + [V^0_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) - V^\delta_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r})] \geq$$
$$V^\delta_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) - C\delta(1 + \|\lambda^{\delta,\alpha(\delta)}_{p,r}\| + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|),$$

whence we obtain

$$V^0_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) \geq V^0_{p,r}(\lambda,\mu) + \alpha(\delta)\|\lambda^{\delta,\alpha(\delta)}_{p,r}\|^2 + \alpha(\delta)|\mu^{\delta,\alpha(\delta)}_{p,r}|^2 - C\delta(1 + \|\lambda^{\delta,\alpha(\delta)}_{p,r}\| + \|\mu^{\delta,\alpha(\delta)}_{p,r}\|) -$$
$$C\delta(1 + \|\lambda\| + \|\mu\|) \quad \forall\,(\lambda,\mu) \in \mathcal{H} \times \mathcal{H}_+.$$

From here, we deduce, due to the consistency condition (3.1) and limit relations (3.9), that for any fixed $M > 0$ and for any fixed $\epsilon > 0$ there exists such $\delta(\epsilon) > 0$ for which the estimate

$$V^0_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) \geq \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+:\,\|\lambda\|\leq M,\,\|\mu\|\leq M} V^0_{p,r}(\lambda,\mu) - \epsilon \tag{3.12}$$
$$\forall\,\delta \leq \delta(\epsilon)\ \ \forall\,(\lambda,\mu) \in \{(\lambda,\mu) \in \mathcal{H} \times \mathcal{H}_+ :\, \|\lambda\| \leq M,\, \|\mu\| \leq M\}$$

holds.

Suppose now that the limit relation (3.11) is not true. Then there exists such a sequence δ_s, $s = 1, 2, \ldots$ convergent to zero that the inequality

$$V^0_{p,r}(\lambda^{\delta_s,\alpha(\delta_s)}_{p,r},\mu^{\delta_s,\alpha(\delta_s)}_{p,r}) \leq \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+} V^0_{p,r}(\lambda,\mu) - l, \quad s = 1, 2, \ldots$$

is fulfilled for some $l > 0$.

Since

$$\sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+} V^0_{p,r}(\lambda,\mu) - \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+:\,\|\lambda\|\leq M,\,\|\mu\|\leq M} V^0_{p,r}(\lambda,\mu) \to 0,$$

for $M \to +\infty$, we deduce from the last estimate that for all sufficiently large positive M the inequality

$$V^0_{p,r}(\lambda^{\delta_s,\alpha(\delta_s)}_{p,r},\mu^{\delta_s,\alpha(\delta_s)}_{p,r}) \leq \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+:\,\|\lambda\|\leq M,\,\|\mu\|\leq M} V^0_{p,r}(\lambda,\mu) - l/2$$

is true. This estimate contradicts to the estimate obtained above (3.12). The last contradiction proves correctness of the limit relation (3.11).

Summarizing the above arguments, we assert that the following "convergence" theorem for the dual regularization method in Problem $(P^0_{p,r})$ is valid.

Theorem 1. *Let Problem $(P^0_{p,r})$ be solvable. Regardless of the properties of the solvability of the dual problem to Problem $(P^0_{p,r})$ or, in other words, regardless of the properties of the subdifferential $\partial\beta(p,r)$ (it is empty or not empty), it is true that exist elements $\pi^\delta \in U^\delta[\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}]$ such that the relations*

$$g^0_0(\pi^\delta) \to g^0_0(\pi^0_{p,r}), \quad g^0_1(\pi^\delta) - h^0 - p \to 0, \quad g^0_2(\pi^\delta) - r \leq \kappa(\delta), \quad \|\kappa(\delta)\| \to 0, \quad \delta \to 0, \tag{3.13}$$
$$\left\langle(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}),(g^\delta_1(\pi^\delta) - h^\delta - p, g^\delta_2(\pi^\delta) - r)\right\rangle \to 0, \quad \delta \to 0$$

hold, in which the inequality $g^0_2(\pi^\delta) - r \leq \kappa(\delta)$ is understood in the sense of ordering on a cone of nonpositive functions in $\mathcal{H}$. Simultaneously, the equality

$$\lim_{\delta\to+0} V^0_{p,r}(\lambda^{\delta,\alpha(\delta)}_{p,r},\mu^{\delta,\alpha(\delta)}_{p,r}) = \sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+} V^0_{p,r}(\lambda,\mu)$$

is valid. In addition, the duality relation (2.3) holds. If the dual of Problem $(P^0_{p,r})$ is solvable, then the limit relation $(\lambda^{\delta,\alpha(\delta)}_{p,r},\ \mu^{\delta,\alpha(\delta)}_{p,r}) \to (\lambda^0_{p,r},\mu^0_{p,r})$, $\delta \to 0$ is valid also, where $(\lambda^0_{p,r},\mu^0_{p,r})$ denotes the minimum-norm solution of the dual problem.

3.2. Stable sequential Lagrange principle for optimal control problem with pointwise state constraints

We formulate in this subsection the necessary and sufficient condition for existence of a minimizing approximate solution in Problem $(P^0_{p,r})$. Also, for this problem it can be called by stable sequential Lagrange principle in nondifferential form. Simultaneously, as we deal only with regular Lagrange function, the formulated theorem may be called by Kuhn–Tucker theorem in nondifferential form. Note that the necessity of the conditions of the theorem formulated below follows from the theorem 1. At the same time, their sufficiency is a simple consequence of the convexity of Problem $(P^0_{p,r})$ and the conditions on its input data. A verification of these propositions for similar situation of the convex programming problem in a Hilbert space may be found in [10, 15, 16].

Theorem 2. *Regardless of the properties of the subdifferential $\partial\beta(p,r)$ (it is empty or not empty) or, in other words, regardless of the properties of the solvability of the dual problem to Problem $(P^0_{p,r})$, necessary and sufficient conditions for Problem $(P^0_{p,r})$ to have a minimizing approximate solution is that there is a sequence of dual variables $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$, $k=1,2,\dots$, such that $\delta^k\|(\lambda^k,\mu^k)\|\to 0$, $k\to\infty$, and relations*

$$\pi^{\delta^k}[\lambda^k,\mu^k]\in\mathcal{D}^{\delta^k,\epsilon^k}_{p,r},\quad \epsilon^k\to 0,\quad k\to\infty, \tag{3.14}$$

$$\Big\langle(\lambda^k,\mu^k),(g_1^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])-h^{\delta^k}-p,g_2^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])-r)\Big\rangle\to 0,\quad k\to\infty \tag{3.15}$$

hold for some elements $\pi^{\delta^k}[\lambda^k,\mu^k]\in U^{\delta^k}[\lambda^k,\mu^k]$. The sequence $\pi^{\delta^k}[\lambda^k,\mu^k]$, $k=1,2,\dots$, is the desired minimizing approximate solution and each of its weak limit points is a solution of Problem $(P^0_{p,r})$. As $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$, $k=1,2,\dots$, we can use the sequence of the points $(\lambda^{\delta^k,\alpha(\delta^k)}_{p,r},\mu^{\delta^k,\alpha(\delta^k)}_{p,r})$, $k=1,2,\dots$, generated by the dual regularization method of the theorem 1. If the dual of Problem $(P^0_{p,r})$ is solvable, the sequence $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$, $k=1,2,\dots$, should be assumed to be bounded. The limit relation

$$V^0_{p,r}(\lambda^k,\mu^k)\to\sup_{(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+}V^0_{p,r}(\lambda,\mu) \tag{3.16}$$

holds as a consequence of the relations (3.14), (3.15). *Furthermore, each weak limit point (if such points exist) of the sequence $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$, $k=1,2,\dots$ is a solution of the dual problem $V^0_{p,r}(\lambda,\mu)\to\max$, $(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+$.*

P r o o f. To prove the necessity, we first note that problem $(P^0_{p,r})$ is solvable (i.e., $U^0_{p,r}\neq\emptyset$) due to the conditions on the initial data and to the existence of a minimizing approximate solution. Now the existence of the indicated sequence $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$, $k=1,2,\dots$ and the limit relations (3.14) and (3.15) follow from Theorem 1 if the points (λ^k,μ^k) and $\pi^{\delta^k}[\lambda^k,\mu^k]$ are defined as $(\lambda^{\delta^k,\alpha(\delta^k)}_{p,r},\mu^{\delta^k,\alpha(\delta^k)}_{p,r})$, and π_{δ^k}, $k=1,2,\dots$, respectively. These limit relations imply that (3.16) holds as well. Really, combining estimates (2.4) with the limit relation $\delta^k\|(\lambda^k,\mu^k)\|\to 0$, $k\to\infty$, we conclude (see the estimate (2.5)) that $V^{\delta^k}_{p,r}(\lambda^k,\mu^k)-V^0_{p,r}(\lambda^k,\mu^k)\to 0$, $k\to\infty$. Then, in view of (2.3), (3.15), and the limit relation $f^0(z^{\delta^k}[\lambda^k,\mu^k])\to f^0(z^0_{p,r})$, $k\to\infty$ (see (3.13)), we have

$$V^{\delta^k}_{p,r}(\lambda^k,\mu^k)=f^{\delta^k}(z^{\delta^k}[\lambda^k,\mu^k])+\big\langle(\lambda^k,\mu^k),(A^{\delta^k}z^{\delta^k}[\lambda^k,\mu^k]-h^{\delta^k}-p,g^{\delta^k}(z^{\delta^k}[\lambda^k,\mu^k])-r)\big\rangle\to f^0(z^0_{p,r}),$$

therefore, the limit relation (3.16) holds true.

To prove the sufficiency, we first note also that the set $U^0_{p,r}\subset\mathcal{D}^{0\epsilon^k}_{p,r}$ is not empty. This follows from the inclusion $\pi^{\delta^k}[\lambda^k,\mu^k]\in\mathcal{D}^{\delta^k,\epsilon^k}_{p,r}$, from the fact that the sequence $\pi^{\delta^k}[\lambda^k,\mu^k]$, $k=1,2,\dots$

is bounded, and from the conditions on the initial data in Problem $(P^0_{p,r})$. Furthermore, since the point $\pi^{\delta^k}[\lambda^k,\mu^k]$ minimizes on $\mathcal{D}$ the functional $L^{\delta^k}_{p,r}(\cdot,\lambda^k,\mu^k)$, we have

$$g_0^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])+\big\langle(\lambda^k,\mu^k),(g_1^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])-h^{\delta^k}-p,g_2^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])-r)\big\rangle\le$$

$$g_0^{\delta^k}(\pi)+\big\langle(\lambda^k,\mu^k),(g_1^{\delta^k}(\pi)-h^{\delta^k}-p,g_2^{\delta^k}(z)-r)\big\rangle\quad\forall\,\pi\in\mathcal{D}.$$

By the assumptions of the theorem, it follows that

$$g_0^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])\le g_0^{\delta^k}(\pi)+\big\langle(\lambda^k,\mu^k),(g_1^{\delta^k}(\pi)-h^{\delta^k}-p,g_2^{\delta^k}(\pi)-r)\big\rangle+\psi^k\quad\forall\pi\in\mathcal{D},\quad\psi^k\to0,\quad k\to\infty.$$

Setting $\pi=\pi^0_{p,r}\in U^0_{p,r}$ and using the consistency condition $\delta^k\|(\lambda^k,\mu^k)\|\to0$, $k\to\infty$, we obtain $g_0^0(\pi^{\delta^k}[\lambda^k,\mu^k])\le g_0^0(\pi^0_{p,r})+\widetilde{\psi}^k$, $\widetilde{\psi}^k\to0$, $k\to\infty$. Since we also have the inclusion $\pi^{\delta^k}[\lambda^k,\mu^k]\in\mathcal{D}^{\delta^k,\epsilon^k}_{p,r}$, using the classical weak compactness properties of a bounded convex closed set and the weak lower semicontinuity of a continuous convex functional in a Hilbert space, we easily derive $g_0^0(\pi^{\delta^k}[\lambda^k,\mu^k])\to g_0^0(\pi^0_{p,r})$, $k\to\infty$; i.e., the sequence $\pi^{\delta^k}[\lambda^k,\mu^k]$, $k=1,2,\dots$ is a minimizing approximate solution in Problem $(P^0_{p,r})$. In view of (3.15) and the obtained limit relation $g_0^0(\pi^{\delta^k}[\lambda^k,\mu^k])\to g_0^0(\pi^0_{p,r})$, $k\to\infty$, we can write

$$V^{\delta^k}_{p,r}(\lambda^k,\mu^k)=g_0^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])+\big\langle(\lambda^k,\mu^k),(g_1^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])-h^{\delta^k}-p,g_2^{\delta^k}(\pi^{\delta^k}[\lambda^k,\mu^k])-r)\big\rangle\to g_0^0(\pi^0_{p,r}),$$

therefore, limit relation (3.16) holds by virtue of estimate (2.5), equality (2.3), and the limit relation $\delta^k\|(\lambda^k,\mu^k)\|\to0$, $k\to\infty$. To conclude, we note that, it is easy to show that each weak limit point of the sequence $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$, $k=1,2,\dots$ (if such points exist) is a solution of the dual problem $V^0_{p,r}(\lambda,\mu)\to\max$, $(\lambda,\mu)\in\mathcal{H}\times\mathcal{H}_+$.

Remark 1. *If the functional g_0^0 is strongly convex and subdifferentiable on $\mathcal{D}$ then from the weak convergence of the unique in this case elements $\pi^{\delta^k}[\lambda^k,\mu^k]$ to unique element $\pi^0_{p,r}$ as $k\to\infty$, and numerical convergence $g_0^0(\pi^{\delta^k}[\lambda^k,\mu^k])\to g_0^0(\pi^0_{p,r})$, $k\to\infty$ follows the strong convergence $\pi^{\delta^k}[\lambda^k,\mu^k]\to\pi^0_{p,r}$, $k\to\infty$. Problem $(P^0_{p,r})$ with the strongly convex g_0^0 for linear system of ordinary differential equations but with exact input data is studied in* [17].

3.3. Stable sequential Pontryagin maximum principle for optimal control problem with pointwise state constraints

Denote by $U^\delta_{max}[\lambda,\mu]$ a set of elements $\pi^\delta_{\max}[\lambda,\mu]\in\mathcal{D}$ that satisfy all relations of the maximum principle (2.2) of the lemma 1. Under the supplementary condition of existence of continuous with respect to z gradients $\nabla_z\varphi^\delta_2(x,t,z)$, $\nabla_z G^\delta(x,z)$ with corresponding estimates, it follows that the proposition of the Theorem 2 may be rewritten in the form of the stable sequential Pontryagin maximum principle. It is obviously that the equality $U^\delta_{max}[\lambda,\mu]=U^\delta[\lambda,\mu]$ takes place under mentioned supplementary condition.

Theorem 3. *Regardless of the properties of the subdifferential $\partial\beta(p,r)$ (it is empty or not empty) or, in other words, regardless of the properties of the solvability of the dual problem to Problem $(P^0_{p,r})$, necessary and sufficient conditions for Problem $(P^0_{p,r})$ to have a minimizing approximate solution is that there is a sequence of dual variables $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$, $k=1,2,\dots$, such that $\delta^k\|(\lambda^k,\mu^k)\|\to0$, $k\to\infty$, and relations* (3.14), (3.15) *hold for some elements $\pi^{\delta^k}[\lambda^k,\mu^k]\in U^{\delta^k}_{max}[\lambda^k,\mu^k]$. Moreover, the sequence $\pi^{\delta^k}[\lambda^k,\mu^k]$, $k=1,2,\dots$, is the desired minimizing approximate solution and each of its weak limit points is a solution of Problem $(P^0_{p,r})$. As $(\lambda^k,\mu^k)\in\mathcal{H}\times\mathcal{H}_+$,*

$k = 1, 2, \ldots$, we can use the sequence of the points $(\lambda_{p,r}^{\delta^k,\alpha(\delta^k)}, \mu_{p,r}^{\delta^k,\alpha(\delta^k)})$, $k = 1, 2, \ldots$, generated by the dual regularization method of the theorem 1. If the dual of Problem $(P_{p,r}^0)$ is solvable, the sequence $(\lambda^k, \mu^k) \in \mathcal{H} \times \mathcal{H}_+$, $k = 1, 2, \ldots$, should be assumed to be bounded. The limit relation (3.16) *holds as a consequence of the relations* (3.14), (3.15).

Remark 2. *When the inequality constraint in Problem $(P_{p,r}^0)$ is absent, i.e., $(P_{p,r}^0) = (P_p^0)$, and $\phi_1(x,t) \equiv 1$, the target functional g_0^0 is taken, for example, in the form $g_0^0(\pi) \equiv \|\pi\|^2 \equiv \|u\|^2 + \|w\|^2$ then Problem (P_p^0) acquires the typical form of unstable inverse problem. In this case the stable sequential Pontryagin maximum principle of the Theorem 3 becomes a tool for the direct solving such unstable inverse problem.*

Remark 3. *In important partial case of Problem $(P_{p,r}^0) = (P_r^0)$, when it has only the inequality constraint $(\varphi_1^\delta(x,t) = h^\delta(x,t) = p(x,t) = 0,\ (x,t) \in Q)$, "weak" passage to the limit in the relations of the Theorem* 3 *leads to usual for similar optimal control problems Pontryagin maximum principle (see, e.g.,* [3, 8]*) with nonnegative Radon measures in the input data of the adjoint equation.*

REFERENCES

1. **Alekseev V.M., Tikhomirov V.M., Fomin S.V.** Optimal Control. Moscow: Nauka, 1979. 432 p. [in Russian]
2. **Aubin J.–P., Ekeland I.** Applied Nonlinear Analysis. New York: Wiley, 1984. 518 p.
3. **Casas E., Raymond J.–P., Zidani H.** Pontryagin's principle for local solutions of control problems with mixed control-state constraints // SIAM J. Control Optim. 2000. Vol. 39, no. 4. P. 1182–1203.
4. **Kalinin A.V., Sumin M.I., Tyukhtina A.A.** Stable sequential Lagrange principles in the inverse final observation problem for the system of Maxwell equations in the quasistationary magnetic approximation // Differential Equations. 2016. Vol. 52, no. 5. P. 587–603.
5. **Kuzenkov O.A., Plotnikov V.I.** Existence and uniqueness of a generalized solution to a linear vector equation of parabolic type in the third boundary value problem // Mathematical Modeling and Optimization Methods (Gorky State University). 1989. P. 132–144. [in Russian].
6. **Ladyzhenskaya O.A., Solonnikov V.A., Ural'tseva N.N.** Linear and quasilinear equations of parabolic type. Providence, R.I.: Am. Math. Soc., 1968. 648 p.
7. **Plotnikov V.I.** Existence and uniqueness theorems and a priori properties of weak solutions // Dokl. Akad. Nauk SSSR. 1965. Vol. 165, no. 1. 33–35. [in Russian]
8. **Raymond J.–P., Zidani H.** Pontryagin's principle for state-constrained control problems governed by parabolic equations with unbounded controls // SIAM J. Control Optim. 1998. Vol. 36. no. 6. P. 1853–1879.
9. **Gaikovich K.P., Gaikovich P.K., Sumin M.I.** Stable sequential Kuhn–Tucker theorem in one–dimensional inverse problems of dielectric reflectometry // Proc. of the 16th International Conference on Transparent Optical Networks: ICTON–2014. 2014. P. Th.A4.6. P. 1–4.
10. **Sumin M.I.** Stable sequential convex programming in a Hilbert space and its application for solving unstable problems // Comput. Math. Math. Phys. 2014. Vol. 54, no. 1. P. 22–44.
11. **Sumin M.I.** A regularized gradient dual method for the inverse problem of a final observation for a parabolic equation // Comput. Math. Math. Phys. 2004. Vol. 44, no. 11. P. 1903–1921.
12. **Sumin M.I.** Duality-based regularization in a linear convex mathematical programming problem // Comput. Math. Math. Phys. 2007. Vol. 47, no. 4. P. 579–600.
13. **Sumin M.I.** Regularized parametric Kuhn–Tucker theorem in a Hilbert space // Comput. Math. Math. Phys. 2011. Vol. 51, no. 9. P. 1489–1509.
14. **Sumin M.I.** Dual regularization and Pontryagin's maximum principle in a problem of optimal boundary control for a parabolic equation with nondifferentiable functionals // Proc. Steklov Inst. Math. 2011. Suppl. 1. P. S161–S177.
15. **Sumin M.I.** On the stable sequential Kuhn–Tucker theorem and its applications // Appl. Math. 2012. Vol. 3, no. 10A. P. 1334–1350.

16. **Sumin M.I.** On the stable sequential Lagrange principle in the convex programming and its applications for solving unstable problems // Proc. of the Inst. of Math. and Mech., Ural Branch of the RAS. 2013. Vol. 19, no. 4. P. 231–240. [in Russian]
17. **Sumin M.I.** Parametric dual regularization for an optimal control problem with pointwise state constraints // Comput. Math. Math. Phys. 2009. Vol. 49, no. 12. P. 1987–2005.
18. **Sumin M.I.** Nekorrektnye zadachi i metody ikh resheniya. Materialy k lektsiyam dlya studentov starshikh kursov (Ill-Posed Problems and Solution Methods). Nizhnii Novgorod: Nizhnii Novgorod State University, 2009. 289 p. [in Russian]
19. **Sumin M.I.** Stable sequential Pontryagin maximum principle in optimal control problem with state constraints // Proc. of the XIIth All-Russia Conference on Control Problems, Inst. of Control Sci. of RAS, Moscow. 2014. P. 796–808. [in Russian]
20. **Sumin M.I.** Stable sequential Pontryagin maximum principle in optimal control for distributed systems // Proc. of Intern. conf. "Systems Dynamics and Control Processes" dedicated to the 90-th anniversary of academician N.N. Krasovskii (Ekaterinburg, Russia, Sept. 15-20, 2014). Ekaterinburg: Ural Federal University, 2015. P. 301–308. [in Russian]
21. **Sumin M.I.** Subdifferentiability of value functions and regularization of Pontryagin maximum principle in optimal control for distributed systems // Tambov State University Reports. Series: Natural and Tech. Sci. 2015. Vol. 20, no. 5. P. 1461–1477. [in Russian]
22. **Sumin M.I.** The first variation and Pontryagin's maximum principle in optimal control for partial differential equations // Comput. Math. Math. Phys. 2009. Vol. 49, no. 6. P. 958–978.
23. **Warga J.** Optimal control of differential and functional equations. New York: Academic Press, 1972. 531 p.

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

OPTIMAL MULTIATTRIBUTE SCREENING[1]

Thomas A. Weber

Chair of Operations, Economics and Strategy
Swiss Federal Institute of Technology, Lausanne, Switzerland
thomas.weber@epfl.ch

Abstract: We provide a technique for constructing optimal multiattribute screening contracts in a general setting with one-dimensional types based on necessary optimality conditions. Our approach allows for type-dependent participation constraints and arbitrary risk profiles. As an example we discuss optimal insurance contracts.

Key words: Asymmetric Information, Incentive Contracting, Maximum Principle, Nonlinear Pricing.

1. Introduction

Starting with the seminal contribution by Mirrlees [1], optimal screening contracts have found many applications in economics, including taxation, nonlinear pricing, and the regulation of monopolies [1–4]. The underlying game of asymmetric information contains two periods. In the first period, the principal announces a menu of attribute bundles to an agent who possesses private information about his utility function, and who can select an attribute bundle by sending a message to the principal. In the second period, allocations are made according to a publicly known enforceable mapping from the message space to both attribute bundles and monetary transfers from agent to principal. Subject to the agent's participation constraint (individual rationality) and the agent's self-interested choice behavior (incentive compatibility), an optimal screening menu maximizes the principal's expected payoff. In this paper, which is related to [5,6], we provide an explicit method to construct optimal screening contracts with continuously distributed, one-dimensional "types" representing the agents' private information and multiple instruments (or attributes), based on necessary optimality conditions. The problem of finding optimal screening contracts can be formulated as an optimal control problem, for which we derive necessary optimality conditions using the maximum principle by Pontryagin et al. [7] and the technique of successive approximation [8]. In this vein, the method approximates an optimal schedule by directly computable solutions to a sequence of relaxed screening problems. We allow for type-dependent participation constraints and payoff functions that are nonlinear in each contract instrument. This is useful when considering the effects of variable outside options and/or risk-aversion on the optimal contracts. The set of participating types is also subject to optimization.

The paper is organized as follows. In Section 2, we introduce a general model for multiattribute screening and formulate the principal's optimization problem in an optimal-control setting. In Section 3, we provide a full set of necessary optimality conditions that can be used to construct a solution to the screening problem, including an iterative approximation method. In Section 4, we discuss the design of a menu of optimal insurance contracts for an agent with unknown risk aversion as a practical example. Section 5 concludes.

[1]This paper was presented as a plenary lecture in October 2016 in Ekaterinburg, at the International Conference on "Systems Analysis: Modeling and Control" in memory of Academician Arkady Kryazhimskiy.

2. Model

2.1. Screening Problem

We consider a standard screening setup with a principal ("she") and an agent ("he"). The agent's possible type θ, known only to him, lies in the type space $\Theta = [0,1]$.[2] The type θ summarizes all the private information the agent has. From the principal's viewpoint it is distributed with the continuous probability density $f = \dot{F}$ on the support Θ.[3] The principal designs a schedule of instruments (or attributes) $x = (x_0, x_1, \ldots, x_n) : \Theta \to \mathbb{R}_+^{n+1}$ so as to maximize her expected payoff,

$$\bar{V}(x;\theta_0) = \int_{\theta_0}^{1} V(x(\theta),\theta) f(\theta)\, d\theta, \tag{2.1}$$

where $n \geq 1$ is a given integer which denotes the number of attributes characterizing the bundle, and $\Theta_0 = [\theta_0, 1] \subseteq \Theta$ is the set of participating types, with the marginal type $\theta_0 \in \Theta$ subject to optimization.[4] In order to ensure implementability of a favored attribute schedule, the principal's optimization problem is subject to the agent's self-interested behavior, which manifests itself in the form of two constraints. First, given that an agent of type θ has net utility $U(\xi,\theta)$ for a bundle $\xi \in \mathbb{R}_+^{n+1}$, the optimal type announcement ϑ satisfies the *incentive-compatibility constraint*

$$U(x(\vartheta),\theta) \geq U(x(\hat{\theta}),\theta), \quad \forall\, \vartheta, \hat{\theta} \in \Theta_0. \tag{2.2}$$

Second, since the agent is free to walk away from the principal's proposed menu of contracts (attribute schedule), the participation set Θ_0 is defined by the agent's *participation constraint*

$$\vartheta, \theta \in \Theta_0 \quad \Leftrightarrow \quad U(x(\vartheta),\theta) \geq U(0,\theta). \tag{2.3}$$

That is, an agent of type θ participates if by pretending to be any other participating agent type ϑ (including himself) he achieves a utility that is at least equal to the utility derived from a zero bundle (corresponding to nonparticipation). The principal's screening problem is to find an (absolutely continuous) attribute schedule $x(\cdot)$ which maximizes the objective in (2.1) subject to the implementability constraints (2.2) and (2.3).

2.2. Key Assumptions

The attribute x_0 corresponds to a *numéraire* good such as a monetary transfer from the agent to the principal.[5] The principal's payoff function $V : \mathbb{R}^{n+1} \times \Theta \to \mathbb{R}$ is continuously differentiable in the attribute vector x and continuous in the type θ. In addition, we make the following two assumptions on V.

P1. $V_{x_0} > 0 > V_{x_i}$ for all $i \in \{1, \ldots, n\}$ (P-Monotonicity).[6]

P2. $V(0,\theta) \geq 0$ for all $\theta \in \Theta$ (Possibility of Inaction).

[2] Using a simple affine transformation this does in fact allow for *any* compact interval on the real line.

[3] Throughout we use the dot-notation for total derivatives with respect to θ, e.g., $\dot{F} = \frac{dF}{d\theta}$.

[4] The problem of maximizing the principal's expected utility can be viewed as a "dynamic optimization problem" if $\bar{V}(x;\theta_0)$ represents an average payoff generated by a trajectory $x(\theta)$ on the time interval $[\theta_0, 1]$, where the starting time θ_0 is subject to optimization. The methods presented in this paper extend to intervals of the form $[\theta_0, \theta_1]$ where both boundaries are subject to optimization (cf. footnote 23).

[5] We sometimes use x to denote bundles (i.e., points) in the attribute space $\mathbb{R}_+^{n+1} = \{\hat{x} \in \mathbb{R}^{n+1} : \hat{x} \geq 0\}$ rather than full schedules (i.e., functions); the same 'notational flexibility' is used for components x_i of x.

[6] Throughout the text, we denote partial derivatives using subscripts, e.g., $V_{x_0} = \partial V / \partial x_0$.

Agent type θ's net utility from the bundle x is $U(x,\theta)$, where $U : \mathbb{R}^{n+2} \to \mathbb{R}$ is a twice continuously differentiable function. His outside option (which is realized after not agreeing to any of the principals' proposed contracts) yields the *reservation utility* $r(\theta) = U(0,\theta)$, where $r : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function. We make the following three assumptions on U.

A1. For all $(x,\theta) \in \mathbb{R}^{n+1}_+ \times \Theta$: (i) $U_\theta(0,x_1,\ldots,x_n,\theta) \geq \dot{r}(\theta)$, and (ii) $U_{x_i}(x,\theta) > 0 > U_{x_0}(x,\theta)$ for all $i \in \{1,\ldots,n\}$ (A-Monotonicity).

A2. For all $(x_1,\ldots,x_n,\theta) \in \mathbb{R}^n_+ \times \Theta$: (i) $U(0,x_1,\ldots,x_n,\theta) \geq r(\theta)$ (Attribute Desirability), and (ii) $r(\theta) > \lim_{x_0\to\infty} U(x_0,x_1,\ldots,x_n,\theta)$ (Transfer Sensitivity).

A3. For all $(x,\theta) \in \mathbb{R}^{n+1}_{++} \times \Theta$: $|U_{x_{i'}\theta}(x,\theta)| > 0$ for some $i' \in \{1,\ldots,n\}$ (Incentive Regularity).[7]

Finally, to exclude unbounded screening contracts, the gains from trade between the principal and the agent need to be bounded. To formalize this notion, we first introduce the principal's *equivalent variation* $E(x_1,\ldots,x_n,\theta)$, which corresponds to the minimum payment in terms of the numéraire good x_0 she would be willing to accept to provide the attribute bundle $(x_1,\ldots,x_n)$ to the agent, i.e.,

$$E(x_1,\ldots,x_n,\theta) = \inf\{x_0 \in \mathbb{R}_+ : V(x_0,x_1,\ldots,x_n,\theta) \geq 0\}, \tag{2.4}$$

where we adopt the convention that $\inf \emptyset = \infty$. On the other hand, agent type θ's *compensating variation* $C(x_1,\ldots,x_n,\theta)$ is his maximum willingness to pay (in terms of x_0) for the attribute bundle $(x_1,\ldots,x_n)$ provided by the principal, i.e.,

$$C(x_1,\ldots,x_n,\theta) = \sup\{x_0 \in \mathbb{R}_+ : U(x_0,x_1,\ldots,x_n,\theta) \geq r(\theta)\}, \tag{2.5}$$

where we adopt the convention that $\sup \emptyset = -\infty$. The following assumption guarantees that gains from trade between principal and agent exist and the contracts between the two involve only bounded attribute bundles. Let

$$\mathcal{X} = \{(x_0,x_1,\ldots,x_n) \in \mathbb{R}^{n+1}_+ : C(x_1,\ldots,x_n,\theta) \geq x_0 \geq E(x_1,\ldots,x_n,\theta) \quad \text{and} \quad \theta \in \Theta\}$$

be the (feasible) contract space.[8]

T1. $\mathcal{X}$ is bounded (Bounded Contract Space).

We now comment on the six assumptions P1, P2, A1–A3, and T1. Assumption P2 bounds the principal's value of the screening problem from below by zero, since the zero schedule $x = 0$ is always feasible. Assumption P1 means that the principal's preferences are nonsatiated in the numéraire good. The monotonicity of the principal's payoff with respect to the non-numéraire attributes $(x_1,\ldots,x_n)$ is unimportant and can be relaxed in situations which involve cooperation between the principal and the agent. Correspondingly, we assume in A1(ii) that the agent dislikes paying the numéraire to the principal, but finds all other attributes desirable. The first inequality in P1 and A1(ii) can be always satisfied by relabelling and simple sign-transformations, as long as there is one attribute that both the principal and the agent like, so it is possible for the agent to compensate the principal for her actions regarding the other attributes. Assumption A1(i) guarantees that the set of participating types is convex (of the form $\Theta_0 = [\theta_0, 1]$ for some $\theta_0 \in$

[7] Without any loss of generality, we assume in what follows that $i' = 1$; furthermore, $\mathbb{R}^{n+1}_{++} = \text{int}(\mathbb{R}^{n+1}_+)$.

[8] The equivalent and compensating variations are classical welfare measures [9]. A normative relationship between them, in particular a lack of exchange due to the endowment effect (when $E > C$) for identical individuals, is discussed by [10]. Here, assumptions A1 and P1 imply that the contracting parties' preferences differ, and assumptions A2 and P2 that gains from trade exist.

$[0,1]$) and includes the highest type $\theta = 1$,[9] as long as it is nonempty. The latter is guaranteed by P2, since the principal can always provide the zero-attribute bundle at no disbenefit. The first inequality in A2 means that any agent always accepts the bundle $(0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1}_+$ when he obtains the non-numéraire attributes for free. The second inequality means that for any attribute vector $(x_1, \ldots, x_n) \in \mathbb{R}^n_+$ there is a price (in terms of x_0) that is too high. The incentive-regularity condition A3 requires that the agent perceives increasing (or decreasing) differences with respect to one non-numéraire attribute and his type. This condition is needed only in a neighborhood of points where the agents' incentive compatibility becomes a binding constraint. Finally, the assumption T1 stipulating that the gains from trade be bounded is naturally satisfied in any practical situation, e.g., when both U and V are bounded and there exists an ε such that $U_{x_0} < -\varepsilon < 0 < \varepsilon < V_{x_0}$. Intuitively it is enough when the principal's marginal costs $-V_{x_i}$ of providing a bundle (to a given agent type) increase fast enough in the attributes, and at the same time the agent's marginal utilities U_{x_i} for these attributes decrease. Assumption T1 guarantees that the solution to the principal's screening problem behaves as if it were constrained to attribute schedules with values in the set $\mathcal{X}$ without the need for an explicit consideration of this constraint. Note that assumptions P2 and A2 imply that the contract space is nonempty, as it must contain the zero bundle, i.e., $0 \in \mathcal{X}$.

3. Optimal Screening Contracts

We treat the screening problem in an optimal-control framework. Accordingly, the *admissible schedules* $x : \Theta \to \mathbb{R}^{n+1}_+$ are in the Sobolev space $\mathbf{W}_{1,\infty}$ of absolutely continuous functions with essentially bounded derivatives, and the corresponding class of *admissible controls* $u : \Theta \to \mathbb{R}^n$ is the Lebesgue space $\mathbf{L}_\infty$ of all essentially bounded functions. An admissible marginal type is any $\theta_0 \in \Theta$. Correspondingly, let $\mathcal{D} = \mathbf{L}_\infty(\Theta, \mathbb{R}^{n+1}_+) \times \mathbf{W}_{1,\infty}(\Theta, \mathbb{R}^n) \times \Theta$ be the *domain of admissibility* for solutions (x^*, u^*, θ_0^*) to the screening problem.[10]

Theorem 1. *Under assumptions P1 and A1–A2, the principal's screening problem can be written in the form:*

$$\sup_{(x,u,\theta_0)\in\mathcal{D}} \bar{V}(x;\theta_0), \tag{P}$$

subject to

$$\dot{x} = \Phi(x,\theta)u,\ U(x(\theta_0),\theta_0) = r(\theta_0), \tag{3.1}$$

and

$$\min_{\hat{\theta}\in\Theta_0} \left\{ (\hat{\theta} - \theta) \Big(U_\theta(x(\hat{\theta}),\theta) - U_\theta(x(\theta),\theta) \Big) \right\} \geq 0, \tag{3.2}$$

for all $\theta \in \Theta_0 = [\theta_0, 1]$, *where*[11]

$$\Phi(x,\theta) = \begin{bmatrix} \varphi(x,\theta) \\ \mathbf{I}_n \end{bmatrix} \in \mathbb{R}^{(n+1)\times n}, \quad \varphi(x,\theta) = -\frac{(U_{x_1}(x,\theta),\ldots,U_{x_n}(x,\theta))}{U_{x_0}(x,\theta)} \geq 0, \tag{3.3}$$

and $\mathbf{I}_n$ *denotes an* $(n \times n)$*-identity matrix.*

P r o o f. Given an admissible schedule $x : \Theta \to \mathbb{R}^{n+1}$, we conclude from A1 that $U(x(\hat{\theta}),\theta_1) - r(\theta_1) \geq U(x(\hat{\theta}),\theta_0) - r(\theta_0)$ for any $\hat{\theta}, \theta_1, \theta_0 \in \Theta$ with $\theta_1 \geq \theta_0$. Thus, type θ_0's participation

[9] Convexity of Θ_0 can be achieved also when $U - r$ is only quasiconcave in θ [11], in which case the upper marginal type becomes subject to optimization also.

[10] At an optimal solution (x^*, u^*, θ_0^*), the schedule x^* and control u^* need only be defined on the optimal participation set $\Theta_0^* = [\theta_0^*, 1] \subseteq \Theta$.

[11] A vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ satisfies $y \geq 0$ if and only if $y_i \geq 0$ for all $i \in \{1, \ldots, n\}$.

implies participation for all types $\theta \in \Theta_0 = [\theta_0, 1]$. Since by P1 and A1 it is $V_{x_0} > 0 > U_{x_0}$, the lowest participating type θ_0 cannot get any surplus, so necessarily $U(x(\theta_0), \theta_0) = r(\theta_0)$. The lowest type θ_0 is itself subject to optimization. Equation (3.1) states that $d(x_1, \dots, x_n)/d\theta = u$, where u is subject to optimization. When searching for schedules $x : \Theta \to \mathbb{R}^{n+1}$ that maximize expected profits $\bar{V}(x; \theta_0)$, by the revelation principle [12, 13] the principal can restrict attention to schedules under which all types report truthfully,[12] so

$$U(x(\theta), \theta) \geq U(x(\hat{\theta}), \theta) \tag{3.4}$$

for all $\theta, \hat{\theta} \in \Theta$. The incentive-compatibility condition (3.4) is equivalent to (3.1)–(3.3). To prove this, we first show that (3.4) implies (3.1)–(3.3). Indeed, by subtracting $U(x(\hat{\theta}), \hat{\theta})$ from (3.4) and switching the labels for θ and $\hat{\theta}$ we obtain that

$$U(x(\theta), \theta) - U(x(\hat{\theta}), \hat{\theta}) \geq U(x(\hat{\theta}), \theta) - U(x(\hat{\theta}), \hat{\theta})$$

and

$$U(x(\hat{\theta}), \hat{\theta}) - U(x(\theta), \theta) \geq U(x(\theta), \hat{\theta}) - U(x(\theta), \theta).$$

Combining the last two inequalities yields

$$U(x(\theta), \theta) - U(x(\theta), \hat{\theta}) \geq U(x(\theta), \theta) - U(x(\hat{\theta}), \hat{\theta}) \geq U(x(\hat{\theta}), \theta) - U(x(\hat{\theta}), \hat{\theta}) \tag{3.5}$$

for all $\theta, \hat{\theta} \in \Theta$. Selecting any $\hat{\theta}, \theta \in \operatorname{int} \Theta$ and taking the limit for $\hat{\theta} \to \theta^+$ and $\theta \to \theta^-$ we get

$$\frac{dU(x(\theta), \theta)}{d\theta} = U_\theta(x(\theta), \theta)$$

almost everywhere (a.e.) on Θ, which is equivalent to

$$U_x(x(\theta), \theta)\, \dot{x}(\theta) = 0. \tag{3.6}$$

The last equation defines the system dynamics $\dot{x}(\theta) = \Phi(x(\theta), \theta)\, u(\theta)$ on Θ_0, as specified in (3.1) and (3.3). Note that $\dot{x}_0(\theta) = \varphi(x(\theta), \theta)$ by solving (3.6) for $\dot{x}_0$. The fact that

$$\varphi(x, \theta) = (\varphi_1(x, \theta), \dots, \varphi_n(x, \theta)) \geq 0$$

(componentwise) follows directly from A1. Since by admissibility of the menu x and smoothness of U, the function $U(x(\theta), \theta)$ is absolutely continuous in θ, by the fundamental theorem of calculus [14, p. 134] it is

$$U(x(\hat{\theta}), \hat{\theta}) - U(x(\theta), \theta) = \int_\theta^{\hat{\theta}} U_\theta(x(\vartheta), \vartheta)\, d\vartheta. \tag{3.7}$$

The fundamental theorem of calculus also implies that

$$U(x(\hat{\theta}), \hat{\theta}) - U(x(\hat{\theta}), \theta) = \int_\theta^{\hat{\theta}} U_\theta(x(\hat{\theta}), \vartheta)\, d\vartheta, \tag{3.8}$$

so by subtracting (3.7) from (3.8) one can rewrite the first inequality in (3.5) in the form

$$\int_\theta^{\hat{\theta}} \Big[U_\theta(x(\hat{\theta}), \vartheta) - U_\theta(x(\vartheta), \vartheta) \Big]\, d\vartheta \geq 0$$

[12] The revelation principle describes the (almost trivial) fact that when the principal is able to commit to a mechanism (i.e., an allocation function $\hat{x} : \Theta \to \mathcal{X}$) by solving the agent's problem (2.2) of announcing a type $\vartheta = \vartheta^*(\theta) \in \Theta_0$ that maximizes his expected utility, she can simply choose x, with $x(\theta) \mapsto \hat{x}(\vartheta^*(\theta))$, instead of $\hat{x}$ to obtain a mechanism that reveals the agent's type truthfully.

for all $\theta, \hat{\theta} \in \Theta$. This last inequality holds for all $\theta, \hat{\theta} \in \Theta$ if and only if

$$\begin{cases} U_\theta(x(\hat{\theta}), \theta) \geq U_\theta(x(\theta), \theta), & \text{if } \hat{\theta} > \theta, \\ U_\theta(x(\hat{\theta}), \theta) \leq U_\theta(x(\theta), \theta), & \text{if } \hat{\theta} < \theta, \end{cases}$$

i.e., if and only if the single-crossing condition (3.2) is satisfied for all $\theta \in \Theta_0$. Hence, condition (3.4) is necessary for conditions (3.1)–(3.3) in Theorem 1. In order to show that conditions (3.1)–(3.3) are also sufficient for the incentive compatibility constraint (3.4), consider any $\hat{\theta}, \theta \in \Theta$ and rewrite (3.4) using the integral representation (3.7) and the fundamental theorem of calculus as

$$U(x(\theta), \theta) - U(x(\hat{\theta}), \hat{\theta}) = \int_{\hat{\theta}}^{\theta} U_\theta(x(\vartheta), \vartheta)\, d\vartheta \geq \int_{\hat{\theta}}^{\theta} U_\theta(x(\hat{\theta}), \vartheta)\, d\vartheta = U(x(\hat{\theta}), \theta) - U(x(\hat{\theta}), \hat{\theta}),$$

or in the more compact form

$$\int_{\theta}^{\hat{\theta}} \Big[U_\theta(x(\hat{\theta}), \vartheta) - U_\theta(x(\vartheta), \vartheta) \Big]\, d\vartheta \geq 0.$$

This last inequality holds for all $\hat{\theta}, \theta \in \Theta$ if and only if $U_\theta(x, \theta)$ exhibits the single-crossing property (3.2) for all $\theta \in \Theta$, which yields the desired sufficiency of (3.1)–(3.3) for (3.4). As alluded to earlier, the initial condition in (3.1) is due to the agent's participation constraint, $U(x(\theta), \theta) \geq r(\theta)$ for all $\theta \in \Theta_0$. By A2, the set of participating types Θ_0 is nonempty (given that $\mathcal{X} \neq \emptyset$). □

The intuition for the system dynamics in (3.1) is that, except for the component x_0 (which contains the agent's payment in the numéraire good), the rate of change of the principal's schedule x as a function of the type θ is governed by the control variable u. The dynamics of the system are constrained by the fact that a participating agent $\theta \in [\theta_0, 1]$ needs to find it optimal to report his type truthfully, such that $U(x(\theta), \theta) \geq U(x(\hat{\theta}), \theta)$. In other words, by consuming a bundle $x(\hat{\theta})$ possibly different from the bundle $x(\theta)$ designed for him, agent θ cannot be better off. This incentive compatibility (or "implementability") is responsible for the dynamics of x_0 and the constraint (3.2). The latter constraint renders problem (P) a nonstandard optimal control problem,[13] for it involves the schedule x generically at different points of the type space. Hence, the corresponding necessary optimality conditions, summarized by the following result, differ from standard versions of the maximum principle. The existence of a solution to (P) is implied by [11, Thm. 2]. We now provide a set of necessary optimality conditions.

Theorem 2. *Given that assumption T1 is satisfied, let (x^*, u^*, θ_0^*) be an optimal solution to the screening problem* (P) *and let $\Theta_0^* = [\theta_0^*, 1]$ be the corresponding set of participating types. Then there exists a function $\psi = (\psi_0, \ldots, \psi_n) : \Theta_0^* \to \mathbb{R}^{n+1}$ of bounded variation, a constant $\lambda_0 \geq 0$, and a nonnegative Borel measure ν, such that the following optimality conditions are satisfied.*

C1. Adjoint Equation: $\psi(\theta) \in \mathbb{R}^{n+1}$ satisfies

$$\psi(\theta) = \int_{\theta}^{1} (\lambda_0 V_x(x^*, \vartheta) f(\vartheta) + \psi_0 \varphi_x(x^*, \vartheta) u^*)\, d\vartheta - \int_{\theta}^{1} (\rho(x^*, \vartheta) - \vartheta) U_{x\theta}(x^*, \vartheta)\, d\nu,$$

for all $\theta \in \Theta_0^*$, where the measurable selection ρ satisfies[14]

$$\rho(\bar{x}, \theta) \in \arg\min_{\hat{\theta} \in \Theta_0^*} \Big\{ (\hat{\theta} - \theta) \Big(U_\theta(x^*(\hat{\theta}), \theta) - U_\theta(\bar{x}, \theta) \Big) \Big\},$$

[13] In the remainder of the text, references to problem (P) (or to its relaxation (P') introduced below) include the complete problem setting with assumptions and relevant constraints.

[14] A set-valued minimizer of a continuous function over a compact set is nonempty, compact-valued, and upper semi-continuous in θ. It therefore has a measurable selection [15, p. 44].

for any $(\bar{x}, \theta) \in \mathbb{R}^{n+1} \times \Theta_0^*$.

C2. Maximality:

$$u_i^*(\theta) \neq 0 \quad \Rightarrow \quad \psi_0(\theta)\varphi_i(x^*(\theta), \theta) + \psi_i(\theta) = 0,$$

a.e. on Θ_0^*, *for all* $i \in \{1, \ldots, n\}$.

C3. Transversality:

$$\textit{(i)}\ \ \theta_0^* = 0 \quad \Rightarrow \quad \exists\, \lambda_1 \in \mathbb{R} :\ \psi^*(0) + \lambda_1 U_x(x^*(0), 0) = 0;$$

$$\textit{(ii)}\ \ \theta_0^* > 0 \quad \Rightarrow \quad \psi(\theta_0^*)\left(U_\theta(x^*(\theta_0^*), \theta_0^*) - \dot{r}(\theta_0^*)\right) + \lambda_0 V(x^*(\theta_0^*), \theta_0^*) U_x(x^*(\theta_0^*), \theta_0^*) = 0.$$

C4. Complementary Slackness:

$$\textit{supp}(\nu) \subseteq \left\{ \theta \in \Theta : \min_{\hat{\theta} \in \Theta_0^*} \left\{ (\hat{\theta} - \theta) \left(U_\theta(x^*(\hat{\theta}), \theta) - U_\theta(x^*(\theta), \theta) \right) \right\} = 0 \right\}.$$

C5. Nontriviality: $\lambda_0 + \sup_{\theta \in \Theta_0^*} \|\psi(\theta)\| > 0$.

P r o o f. See appendix. □

The adjoint variable $\psi(\theta)$ corresponds to the shadow value of the optimal attribute schedule at type θ, given that the evolution of the optimal schedule satisfies the ordinary differential equation (ODE) in (3.1) as well as the incentive-compatibility constraint (3.2). The adjoint equation C1 describes the evolution of ψ on the set of participating types. In particular, the shadow value of the attribute schedule vanishes for the highest type, $\theta = 1$. The maximality condition C2 requires that the optimal control u^* can be essentially bounded only if, while maximizing the principal's expected payoff, the gradient of Φu with respect to u (corresponding to the right-hand side of (3.1)) vanishes. The transversality condition C3 is implied by the optimality of the marginal agent of type θ_0^*, who is indifferent between participating or not. If θ_0^* is not a boundary solution (i.e., when $\theta_0^* > 0$), condition C3 (ii) means that the total change in value, as measured by the agent's marginal surplus through the change in his type movement (evaluated at the shadow value $\psi(\theta_0^*)$) plus his marginal utility for the attribute bundle (evaluated at the principal's net payoff $V(x^*(\theta_0^*), \theta_0^*)$) must vanish for the indifferent type θ_0^*. If it is optimal to serve all agents (i.e., when $\theta_0^* = 0$), one obtains a distortion of the shadow values described by the transversality condition C3 (i). The complementary slackness condition C4 shows that the support of the measure ν must be inside the set of types for which the incentive-compatibility constraint (3.2) is binding. Thus, if (3.2) is never binding, then ν vanishes. Condition C5 ensures that the necessary optimality conditions are nontrivial in the sense that λ_0 and ν cannot vanish together. This can imply an important simplification: if $\nu = 0$, then necessarily $\lambda_0 > 0$, so that, without any loss in generality, we can set $\lambda_0 = 1$, since all other optimality conditions are positively homogeneous in λ_0.

It is generally difficult to construct a measure ν without the precise knowledge of x^*, since ρ is defined using x^* and θ_0^*. When taking the limit $\hat{\theta} \to \theta$ in the maximand of the left-hand side in (3.2), the following constraint is implied:

$$\gamma \cdot u \geq 0, \tag{3.2'}$$

where $\gamma = (1, \gamma_2, \ldots, \gamma_n)$ and $\gamma_j = U_{x_j\theta}/U_{x_1\theta}$ corresponds to the agent's marginal rate of substitution (with respect to U_θ) between x_1 and x_j for $j \in \{2, \ldots, n\}$. We refer to (P) with constraint (3.2) replaced by (3.2') therefore as the *relaxed screening problem* (P'). Contrary to the screening problem (P), the relaxed screening problem (P') can be solved explicitly.

Theorem 3. *Under assumptions P1, P2, and A1–A3, if* (x^*, u^*, θ_0^*) *is an optimal solution to the relaxed screening problem* (P')*, then there exists an absolutely continuous function* $\psi : \Theta \to \mathbb{R}^{n+1}$ *such that the transversality condition C3 is satisfied, and*[15]

$$-\dot{\psi} = V_x(x^*, \theta) f(\theta) + (\psi_0 \varphi_x(x^*, \theta) + \mu \gamma_x(x^*, \theta))\, u^*, \quad \psi(1) = 0, \tag{3.9}$$

and

$$[\psi_0 \varphi_x + \mu \gamma_x, \Phi] u^* = \psi_0 \varphi_\theta + \mu \gamma_\theta - f V_x \Phi, \tag{3.10}$$

where $\mu = \psi_0 \varphi_1(x^*, \theta) + \psi_1$*, and* $\mu \gamma \cdot u^* = 0$*.*

P r o o f. Apply the optimality conditions in Theorem 2 to the optimal control problem (P'). For this, consider the Hamiltonian $H = \lambda_0 V f + \psi \cdot \Phi u$, and first examine the case where the constraint (3.2') is not binding, i.e., where $\gamma \cdot u^* > 0$. If the optimal control u^* is "proper," i.e., independent of $\bar{u}$, then the maximality condition with respect to the relaxed screening problem (P') implies that $H_u = \psi \Phi = 0$ on an optimal state-control trajectory. Differentiating both sides with respect to θ, taking into account the adjoint equation, yields

$$-\dot{\psi}_i = -\left(V_{x_0} f + \psi_0 \varphi_{x_0} \cdot u^*\right) \varphi_i + \psi_0 \varphi_{i,x} \Phi u^* + \psi_0 \varphi_{i,\theta} = V_{x_i} f + \psi_0 \varphi_{x_i} \cdot u^* \tag{3.11}$$

for any $i \in \{1, \ldots, n\}$. These n equations can be rewritten more compactly in the form

$$[\psi_0 \varphi_x, \Phi]\, u^* = \psi_0 \varphi_\theta - f V_x \Phi. \tag{3.12}$$

In the case where $\gamma \cdot u^* = 0$, maximization of the Hamiltonian subject to (3.2') yields the optimality condition $\psi \Phi = \mu \gamma$, where $\mu \geq 0$ is the corresponding Lagrange multiplier. The first component of this condition yields that $\mu = \psi_0 \varphi_1(x^*, \theta) + \psi_1$ as claimed. □

The Lagrange multiplier μ is associated with the relaxed incentive-compatibility constraint (3.2'). By eliminating u^* from the above relations and from (3.1), a solution to the relaxed screening problem can therefore be obtained by solving a system of $n + 3$ ODEs.

Corollary 1. *Let assumptions P1, P2, and A1–A3 be satisfied, and let* (x^*, u^*, θ_0^*) *be an optimal solution to the relaxed screening problem* (P')*. (i) If the matrix* $R = [\varphi_x, \Phi]$ *is nonsingular at the optimal solution, then there exist absolutely continuous functions* $\psi_0, \psi_1 : \Theta_0^* \to \mathbb{R}$ *such that*

$$\left\{ \begin{array}{rcl} \dot{x}^* & = & \Phi R^{-1} \left(\varphi_\theta - \frac{f V_x \Phi}{\psi_0} \right), \\ -\dot{\psi}_0 & = & V_{x_0} f + \varphi_{x_0} R^{-1} \left(\psi_0 \varphi_\theta - f V_x \Phi \right), \end{array} \right. \tag{3.13}$$

provided that $\gamma \cdot R^{-1} \left(\varphi_\theta - \frac{f V_x \Phi}{\psi_0} \right) > 0$*. (ii) Otherwise, setting* $\hat{\gamma} = (\gamma_2, \ldots, \gamma_n)$ *and* $\hat{u} = (u_2, \ldots, u_n)$*), there exist absolutely continuous functions* $\psi_0, \psi_1 : \Theta_0^* \to \mathbb{R}$ *such that*

$$\left\{ \begin{array}{rcl} \dot{x}^* & = & \hat{\Phi} \left[\psi_0 \hat{\varphi}_x + \psi_1 \hat{\gamma}_x, \hat{\Phi} \right]^{-1} \left(\psi_0 \hat{\varphi}_\theta + \psi_1 \hat{\gamma}_\theta - f V_x \hat{\Phi} \right), \\ -\dot{\psi}_i & = & V_{x_i} f + \left(\psi_0 \hat{\varphi}_{x_i} + \psi_1 \hat{\gamma}_{x_i} \right) \left[\psi_0 \hat{\varphi}_x + \psi_1 \hat{\gamma}_x, \hat{\Phi} \right]^{-1} \left(\psi_0 \hat{\varphi}_\theta + \psi_1 \hat{\gamma}_\theta - f V_x \hat{\Phi} \right), \end{array} \right. \tag{3.14}$$

for $i \in \{0, 1\}$*. (iii) The boundary conditions for both (3.13) and (3.14) are* $U(x^*(\theta_0^*), \theta_0^*) = r(\theta_0^*)$ *and* $\psi(1) = 0$*.*

[15] Given two matrices $A \in \mathbb{R}^{m \times l}$ and $B \in \mathbb{R}^{l \times m}$, where m, l are positive integers, the *Lie bracket* of A and B is given by $[A, B] = AB - (AB)^T = AB - B^T A^T \in \mathbb{R}^{m \times m}$, where $(\cdot)^T$ denotes the transpose of $(\cdot)$.

To obtain the $n+2$ initial values $x(\theta_0^*)$ and θ_0^* for the system (3.13), one can use the $n+1$ transversality conditions C3 in conjunction with the initial condition in (3.1).

Remark 1. In the case where all payoff functions are quasilinear in the numéraire, it is $\varphi_{x_0} = 0$ and $V_{x_0} = 1$, which implies that $\psi_0(\theta) = 1 - F(\theta)$ by virtue of (3.13), provided that

$$\left(1, \frac{U_{x_2\theta}}{U_{x_1\theta}}, \ldots, \frac{U_{x_n\theta}}{U_{x_1\theta}}\right) \cdot [\varphi_x, \Phi]^{-1} \left(\varphi_\theta - \frac{f}{1-F} V_x \Phi\right) \geq 0. \tag{3.15}$$

The last inequality can be checked *ex ante*. It generalizes the standard Spence—Mirrlees sorting condition [16, Eq. (7), p. 155]. Note that the last inequality (3.15) features all primitives of the problem, and it is satisfied if the relaxed incentive-compatibility constraint (3.2') is not binding.

Remark 2. In the case where $R = 0$, as in the optimal insurance example discussed in Section 4, relation (3.10) immediately implies that

$$\psi_0 \varphi_\theta + \mu \gamma_\theta = f V_x \Phi. \tag{3.16}$$

Solving this n-dimensional optimality condition, one can determine real-valued functions q_i, such that

$$x_i = q_i(x_0, \theta, \psi_0), \quad i \in \{1, \ldots, n\}, \tag{3.17}$$

which then allows for the solution of the Hamiltonian system consisting of the state equation in (3.1) and the adjoint equation (3.9).

If the solution to the relaxed screening problem (P') is feasible in the screening problem (P), then it is also a solution to the principal's screening problem. Failing that, we can approximate an optimal solution to (P) by solutions to appropriate relaxed problems. For this we introduce a sequence of relaxed screening problems $\{(P'_{kl})\}_{k,l\geq 1}$. For any $k, l \geq 1$, problem (P'_{kl}) is identical to problem (P') with V replaced by

$$V^{kl}(x, \theta) = V(x, \theta) - k \left(g_-(x, \theta; x^{k-1,l-1})\right)^2 - l \left\|x - x^{k-1,l-1}\right\|^2,$$

and

$$g_-(\bar{x}, \theta; x^{k-1,l-1}) = \min_{\hat{\theta} \in [\theta_0^k, 1]} \left\{0, (\hat{\theta} - \theta) \left(U_\theta(x^{k-1,l-1}(\hat{\theta}), \theta) - U_\theta(\bar{x}, \theta)\right)\right\}$$

for any $(\bar{x}, \theta) \in \mathbb{R}^{n+1} \times \Theta$. We denote a solution to problem (P'_{kl}) by $(x^{kl}, u^{kl}, \theta_0^{kl})$. To initialize the sequence of relaxed screening problem, we set $V^{00} = V$, thus adding a problem (P'_{00}) which is identical to problem (P'); its solution $(x^{00}, u^{00}, \theta_0^{00})$ is therefore described by our earlier results.

Theorem 4. *(i) For any given $k \geq 1$, the sequence $\left\{(x^{kl}, u^{kl}, \theta_0^{kl})\right\}_{l\geq 1}$ of solutions to (P'_{kl}) converges to $(x^k, u^k, \theta_0^k) \in \mathbf{W}_{1,\infty} \times \mathbf{L}_\infty \times \Theta$. (ii) The sequence $\left\{(x^k, u^k, \theta_0^k)\right\}_{k\geq 1}$ converges to a solution (x^*, u^*, θ_0^*) of* (P).

P r o o f. (i) Fix any $k \geq 1$. The sequence $\{x^{kl}\}_{l\geq 1}$ is by construction a Cauchy sequence in the Banach space $\mathbf{L}_2(\Theta)$, and therefore converges strongly. This implies weak convergence of the sequence $\{u^{kl}\}_{l\geq 1}$. Lastly, by the Bolzano—Weierstrass theorem the sequence $\{\theta_0^{kl}\}_{l\geq 1} \subset \Theta$ contains a convergent subsequence with limit in Θ. Consider the limits $\theta_0^k \leq \hat{\theta}_0^k$ of any two such convergent subsequences. Then $U(x^k(\theta_0^k), \theta_0^k) = r(\theta_0^k)$ and $U(x^k(\hat{\theta}_0^k), \hat{\theta}_0^k) = r(\hat{\theta}_0^k)$. On the other hand, as already noted in the proof of Theorem 1, by A1 for any $\theta \in [\theta_0^k, 1]$ it is $U(x(\theta), \theta_0^k) - r(\theta_0^k) \leq$

$U(x(\theta),\hat{\theta}_0^k)-r(\hat{\theta}_0^k)$, so necessarily $U(x(\theta),\theta)-r(\theta)=0$ for all $\theta\in[\theta_0^k,\hat{\theta}_0^k]$. Without loss of generality, we can therefore take $\theta_0^k=\liminf_{l\to\infty}\theta_0^{kl}$. The limit (x^k,u^k,θ_0^k) solves the relaxed screening problem (P') with V replaced by $V^k(x,\theta)=V(x,\theta)-k\left(g_-(x,\theta;x^k)\right)^2$. (ii) The convergence of the sequence $\{\theta_0^k\}_{k\geq 1}\subset\Theta$ to $\theta_0^*\in\Theta$ obtains as in part (i). Furthermore, any limit (x^*,u^*,θ_0^*) of the sequence $\left\{(x^k,u^k,\theta_0^k)\right\}_{k\geq 1}$ satisfies $g_-(x^*(\theta),\theta;x^*)=0$ a.e. on Θ_0^*, and is thus a feasible solution to the screening problem (P), provided that $(x^*,u^*)\in\mathbf{W}_{1,\infty}\times\mathbf{L}_\infty$. The latter follows from the existence of a solution to (P). □

4. Application: Optimal Insurance

Consider the problem of designing a nonlinear insurance contract with multiple contingencies, which dates back at least to Stiglitz [17]. An agent has constant absolute risk aversion $\theta\in\Theta=[0,1]$.[16] The type parameter θ belongs to the agent's private information, and—from the principal's point of view—it is distributed with the differentiable probability density $f>0$ on the type space Θ.[17] The agent's utility for any real-valued monetary payoff ξ is

$$v(\xi,\theta)=-\exp\left(-\theta\xi\right),$$

and his current wealth is zero. The agent faces n possible, mutually exclusive loss events $L_1,\ldots,L_n$, which are ordered by magnitude such that $0<L_1<\cdots<L_n<\infty$. The probability of loss event L_i is $p_i>0$, so $p_0=1-(p_1+\cdots+p_n)\in(0,1)$ is the probability that no loss event occurs.[18] An insurance contract subsidizes the agent by an amount x_i when loss event L_i occurs, and asks the agent for a net payment of x_0 in the absence of a loss. Given an insurance contract $x=(x_0,\ldots,x_n)$, the agent's expected utility is

$$U(x,\theta)=p_0v(-x_0,\theta)+\sum_{i=1}^n p_iv(x_i-L_i,\theta).$$

The agent's reservation utility without contract is therefore

$$r(\theta)=U(0,\theta)=-p_0+\sum_{i=1}^n p_iv(-L_i,\theta).$$

On the other hand, the principal's expected payoff is

$$V(x,\theta)=p_0x_0-\sum_{i=1}^n p_ix_i,$$

independent of θ. It is straightforward to verify that both the agent and the principal have utility functions which satisfy assumptions A1–A3 and P1–P2, respectively. The principal's equivalent variation in (2.4) is

$$E(x_1,\ldots,x_n,\theta)=\sum_{i=1}^n\left(\frac{p_i}{p_0}\right)x_i,$$

while the agent's compensating variation in (2.5) takes the form

$$C(x_1,\ldots,x_n,\theta)=\left(\frac{1}{\theta}\right)\ln\left[1+\sum_{i=1}^n\left(\frac{p_i}{p_0}\right)e^{\theta L_i}\left(1-e^{-\theta x_i}\right)\right].$$

[16] By a change of units (i.e., renormalization) this is without loss of generality; cf. also footnote 2.

[17] To satisfy assumption T1, the agent's risk aversion needs to be strictly positive. By considering $\theta/\bar{\theta}$ instead of θ, the analysis generalizes to positive risk aversions in $[0,\bar{\theta}]$, for any $\bar{\theta}>0$.

[18] The vector $(p_0,\ldots,p_n)$ lies in the interior of the n-simplex $\Delta_n=\{(\hat{p}_0,\ldots,\hat{p}_n)\in\mathbb{R}_{++}^{n+1}:\hat{p}_0+\cdots+\hat{p}_n=1\}$.

Thus, assumption T1 is satisfied as long as $f(\theta) = 0$ on $[0, \varepsilon]$ for some minimum risk-aversion $\varepsilon \in (0, 1)$. When the agent is risk-neutral, then the principal would be willing to offer an actuarially fair contract to the agent. By continuous completion for $\varepsilon \to 0^+$ (or, alternately, by imposing a very small capital cost on the principal), we can include risk-neutral agents who are then offered a zero contract. Using the optimality conditions in Theorem 2 and Corollary 1, we find that the Lie product $R = [\varphi_x, \Phi]$ vanishes identically, so by Remark 3:

$$(x_0 - y_i)\,\psi_0 = \frac{f(p_0\varphi_i - p_i) - \mu\gamma_{i\theta}}{-\varphi_i} = \frac{\big(1 - e^{-\theta(x_0 - y_i)}\big)p_0 f + (p_0/p_i)\mu\gamma_{i\theta}}{e^{-\theta(x_0-y_i)}}, \quad i \in \{1, \dots, n\}, \quad (4.18)$$

where $\varphi_i = -(p_i/p_0)\,v(x_0 - y_i, \theta)$ and $\gamma_{i\theta} = -(p_i/p_1)(y_i/y_1)(y_1 - y_i)v(y_1 - y_i, \theta)$ using the abbreviation $y_i \equiv L_i - x_i \geq 0$.

Full Coverage. One immediate solution of (4.18) is $y_i = L_i - x_i = x_0$, for all $i \in \{1, \dots, n\}$, which yields *full coverage* for the participating agent types.[19] The lowest participating type θ_0 in the full-coverage scenario is determined by setting that agent's insurance premium equal to his "certainty equivalent." The latter corresponds to the agent's compensating variation for the insurance contract, so necessarily $x_0 = C(L_1 - x_0, \dots, L_n - x_0, \theta_0)$, which in turn implies that

$$x_0 = g(\theta_0) \equiv \begin{cases} \ln(-r(\theta_0))/\theta_0, & \text{if } \theta_0 \in (0, 1], \\ \bar{L}, & \text{if } \theta_0 = 0, \end{cases}$$

where $\bar{L} = \sum_{i=1}^n p_i L_i$ denotes the agent's expected loss.[20] The principal's expected payoff under full coverage is $\bar{V}(x; \theta_0) = (x_0 - \bar{L})(1 - F(\theta_0))$, so that the optimal participation threshold becomes the global solution of a scalar maximization problem on an interval (for details, see [18]):

$$\theta_0^* \in \arg\max_{\theta_0 \in [0,1]} \left\{(g(\theta_0) - \bar{L})(1 - F(\theta_0))\right\}.$$

As a result, the optimal (constant) schedule is $x^* = (x_0^*, L_1 - x_1^*, \dots, L_n - x_n^*)$, where $x_0^* = g(\theta_0^*)$ and $x_i^* = L_i - x_0^*$ for $i \in \{1, \dots, n\}$. The full-coverage solution leads to no information revelation at all, as all the agent types are offered the same contract. This is also referred to as "bunching" [4].

Partial Coverage. Based on the available optimality conditions it may be possible to construct another solution to the optimal insurance problem, which involves at least partial information revelation. Indeed, for a given $\theta \in (0, 1]$, provided that $\mu = 0$ and $\phi(\theta) \equiv p_0 f(\theta)/\psi_0(\theta) > 1/\theta$, there is a negative solution to (4.18), i.e., there exists a $\zeta = \zeta(\theta) < 0$ such that

$$\zeta = \left(e^{\theta\zeta} - 1\right)\phi. \quad (4.19)$$

In this case, the solution $\zeta = x_0 - y_i = x_0 + x_i - L_i < 0$ is independent of $i \in \{1, \dots, n\}$. For $\phi(\theta) \in [0, 1/\theta]$, the only solution to (4.19) is $\zeta = 0$, reverting back to the full-insurance regime (for that agent type θ). Because by the transversality condition (C1) it is $\psi_0(1) = 0$, this implies that for large enough agent types θ the principal may find it optimal to use partial coverage.

We now continue to follow the solution algorithm outlined in Remark 3, via (3.1) and (3.9). Indeed, the law of motion in (3.1), together with $\dot{x}_i = \dot{\zeta} - \dot{x}_0$ for all $i \in \{1, \dots, n\}$, implies that

$$\dot{x}_0 = \varphi(x, \theta) \cdot (\dot{x}_1, \dots, \dot{x}_n) = \left(\dot{\zeta} - \dot{x}_0\right)\sum_{i=1}^n \varphi_i(x, \theta) = -\frac{\frac{1-p_0}{p_0}\,v(\zeta, \theta)\,\dot{\zeta}}{1 - \frac{1-p_0}{p_0}\,v(\zeta, \theta)} = \frac{\frac{1-p_0}{p_0}\,\dot{\zeta}}{e^{\theta\zeta} + \frac{1-p_0}{p_0}}. \quad (4.20)$$

[19] By the adjoint equation (C1), it is $\psi_0 = (1 - F)p_0$ and $\psi_i = -(1 - F)p_i$ on $[\theta_0^*, 1]$ for all $i \in \{1, \dots, n\}$. (Thus, nontriviality (C5) holds.) Theorem 3 yields $\mu(\theta) = \psi_0(\theta)\varphi_1(x^*(\theta), \theta) + \psi_1(\theta) = 0$ for all $\theta \in [\theta_0^*, 1]$.

[20] By l'Hôpital's rule and the definition of r, it is $\lim_{\theta_0 \to 0^+} \ln(-r(\theta_0))/\theta_0 = r'(0)/r(0) = \bar{L}$.

Using again the law of motion, the first component of the adjoint equation (3.9) becomes

$$-\frac{\dot{\psi}_0}{\psi_0} = \frac{p_0 f}{\psi_0} + (\varphi_x(x,\theta)u)_0 = \phi + \frac{\theta v(\zeta,\theta)}{p_0}\sum_{i=1}^{n} p_i u_i = \phi - \theta\dot{x}_0. \tag{4.21}$$

Since on the one hand $\dot{\psi}_0/\psi_0 = (\dot{f}/f) - (\dot{\phi}/\phi)$, by the definition of ϕ, and since on the other hand

$$\frac{\dot{\phi}}{\phi} = \left(1 - \theta\phi e^{\theta\zeta}\right)\frac{\dot{\zeta}}{\zeta} + \frac{\zeta e^{\theta\zeta}}{1-e^{\theta\zeta}},$$

by virtue of (4.19), one obtains—taking account of (4.20)—that relation (4.21) is equivalent to

$$-\frac{\dot{f}}{f} = -\frac{\dot{\phi}}{\phi} + \phi - \theta\dot{x}_0 = -\left(1 + \frac{\theta\zeta\, e^{\theta\zeta}}{1-e^{\theta\zeta}}\right)\frac{\dot{\zeta}}{\zeta} - \frac{\zeta}{1-e^{\theta\zeta}} - \frac{\theta\dot{\zeta}}{1+\frac{p_0}{1-p_0}e^{\theta\zeta}}.$$

But the last equation implies an initial-value problem,

$$\dot{\zeta} = \left[\frac{1}{\zeta} + \frac{\theta e^{\theta\zeta}}{1-e^{\theta\zeta}} + \frac{(1-p_0)\theta}{(1-p_0)+p_0 e^{\theta\zeta}}\right]^{-1}\left(\frac{\dot{f}}{f} - \frac{\zeta}{1-e^{\theta\zeta}}\right), \quad \zeta(\theta_0) = \zeta_0, \tag{4.22}$$

where the initial value is equal to the certainty equivalent of the marginal agent's exposure conditional on a loss,[21]

$$\zeta_0 = -\frac{\ln\left(\sum_{i=1}^{n}\frac{p_i e^{\theta L_i}}{1-p_0}\right)}{\theta} < 0,$$

thus rendering the contract worthless for the type θ_0, and consequently: $x_0(\theta_0) = 0$.

5. Conclusion

The solution to multiattribute screening problems with one-dimensional types and type-dependent participation constraints can be obtained using optimality conditions derived from a nonstandard version of Pontryagin's maximum principle.[22] Contrary to the extant literature on screening, we do not assume representations of preferences that are quasilinear in the numéraire attribute, thus allowing for arbitrary risk profiles. We also do not require payoff functions to be supermodular in all nonmonetary attributes but impose incentive compatibility as a nonlocal constraint. We have shown that a solution to the multiattribute screening problem, in the case where the incentive-compatibility constraint is binding, can be obtained by solving a sequence of relaxed screening problems, in which constraint violations are increasingly penalized. The results depend in essence only on the convexity of the participation set[23] and on (local) incentive regularity (where needed). Thus, even if A1–A2 and P1–P2 are not satisfied everywhere, one may use our results to construct solutions and then verify the assumptions in a neighborhood of the obtained solutions *ex post*. We further show that the dimensionality of the $2(n+1)$-dimensional Hamiltonian system can often be reduced significantly in concrete problems (e.g., to a 1-dimensional differential equation in our optimal-insurance application in Section 4).

[21] The agent's expected utility is $U(x,\theta) = p_0 v(-x_0,\theta) + (1-p_0)v(\zeta - x_0,\theta) = \left(p_0 + (1-p_0)e^{-\rho\zeta}\right)v(-x_0,\theta)$.

[22] Type-dependent participation constraints arise when agents have heterogeneous outside options, as illustrated in Section 4.

[23] If $U - r$ is quasiconcave in θ (instead of nondecreasing) the participation set is of the form $[\theta_0,\theta_1]$ and θ_1 becomes subject to optimization, leading to an additional transversality condition analogous to C3.

Acknowledgements

The author would like to thank Sergey Aseev and Arkardy Kryazhimskiy, as well as participants of the 2015 INFORMS Annual Meeting in Philadelphia and the 5th International Conference on Continuous Optimization (ICCOPT 2016) in Tokyo, for helpful discussions on related research. A number of very useful comments by an anonymous referee, Ksenia Melnikova, and participants of the 2016 International Conference on "System Analysis: Modeling and Control" in Yekaterinburg are gratefully acknowledged. This paper is dedicated to the memory of Academician Arkady Kryazhimskiy, an outstanding researcher and very kind individual, who will be missed.

REFERENCES

1. **Mirrlees, J.A.** An exploration in the theory of optimal income taxation // Rev. Econ. Stud. 1971. Vol. 38, no. 2, P. 175–208.
2. **Baron, D.P., Myerson, R.B.** Regulating a monopolist with unknown costs // Econometrica. 1982. Vol. 50, no. 4, P. 911–930.
3. **Mirman, L.J., Sibley, D.** Optimal nonlinear prices for multiproduct monopolies // Bell J. Econ. 1980. Vol. 11, no. 2, P. 659–670.
4. **Mussa, M., Rosen, S.** Monopoly and product quality // J. Econ. Theory. 1978. Vol. 18, no. 2, P. 301–317.
5. **Guesnerie, R., Laffont, J.-J.** A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm // J. Public Econ. 1984. Vol. 25, no. 3, P. 329–369.
6. **Matthews, S., Moore, J.** Monopoly provision of quality and warranties: an exploration in the theory of multidimensional screening // Econometrica. 1987. Vol. 55, no. 2, P. 441–467.
7. **Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., Mishchenko, E.F.** The mathematical theory of optimal processes, New York: Wiley Interscience, 1962. 360 p.
8. **Aseev, S.** Methods of regularization in nonsmooth problems of dynamic optimization // J. Math. Sci. 1999. Vol. 94, no. 3, P. 1366–1393.
9. **Hicks, J.R.** Value and capital, Oxford: Clarendon Press, 1939. 331 p.
10. **Weber, T.A.** Hicksian welfare measures and the normative endowment effect // Am. Econ. J.: Microecon. 2010. Vol. 2, no. 4, P. 171–194.
11. **Weber, T.A.** Screening with externalities. Working Paper, Stanford University, 2005.
12. **Gibbard, A.** Manipulation of voting schemes: a general result // Econometrica. 1973. Vol. 41, no. 4, P. 587–601.
13. **Myerson, R.B.** Incentive compatibility and the bargaining problem // Econometrica. 1979. Vol. 47, no. 1, P. 61–74.
14. **Rudin, W.** Principles of mathematical analysis (3rd edition). New York: McGraw-Hill, 1976. 342 p.
15. **Jayne, J.E., Rogers, C.A.** Selectors. Princeton: Princeton University Press, 2002. 167 p.
16. **Laffont, J.-J.** The economics of uncertainty and information. Cambridge: MIT Press, 1989. 289 p.
17. **Stiglitz, J.E.** Monopoly, non-linear pricing and imperfect information: the insurance market // Rev. Econ. Stud. 1977. Vol. 44, no. 3, P. 407–430.
18. **Weber, T.A.** Global optimization on an interval // J. Optimiz. Theory App. 2016. Forthcoming. [doi:10.1007/s10957-016-1006-y]
19. **Arutyunov, A.V.** Optimality conditions: abnormal and degenerate problems. Dordrecht: Kluwer, 2000. 299 p.
20. **Weber, T.A.** Optimal control theory with applications in economics. Cambridge: MIT Press, 2011. 360 p. (Preface by A.V. Kryazhimskiy)
21. **Gelfand, I.M., Fomin, S.V.** Calculus of variations. Englewood-Cliffs: Prentice-Hall, 1963. 240 p.
22. **Kolmogorov, A.N., Fomin, S.V.** Elements of the theory of functions and functional analysis, parts I & II. Rochester: Graylock Press, 1957. 288 p.
23. **Megginson, R.E.** An introduction to Banach space theory. New York: Springer, 1998. 596 p.
24. **Dunford, N., Schwartz, J.T.** Linear operators, part I: general theory. New York: Wiley Interscience, 1958. 858 p.

25. **Milyutin, A.A., Osmolovskii, N.P.** Calculus of variations and optimal control. Providence: American Mathematical Society, 1998. 372 p.
26. **Zorich, V.A.** Mathematical analysis, vol. I. New York: Springer, 2004. 574 p.
27. **Kirillov, A.A., Gvishiani, A.D.** Theorems and problems in functional analysis. New York: Springer, 1982. 347 p.
28. **Taylor, A.E.** General theory of functions and integration. New York: Blaisdell Publishing, 1965. 437 p.
29. **Giaquinta, M., Modica, G., Souček, J.** Cartesian currents in the calculus of variations, vol. I. New York: Springer, 1998. 711 p.
30. **Riesz, F., Sz.-Nagy, B.** Functional analysis, New York: Ungar Publishing, 1955. 491 p.
31. **Schechter, M.** Principles of functional analysis (2nd edition). Providence: American Mathematical Society, 2002. 425 p.

Appendix: Proof of Theorem 2

P r o o f. The argument proceeds in six steps (cf. [7, 8, 19, 20]). For any $(\bar{x}, \theta) \in \mathbb{R}^{n+1} \times \Theta$, let

$$g(\bar{x}, \theta) = \min_{\hat{\theta} \in \Theta_0^*} \left\{ (\hat{\theta} - \theta) \left(U_\theta(x^*(\hat{\theta}), \theta) - U_\theta(\bar{x}, \theta) \right) \right\}.$$

Step 1: Approximate the problem (P) *by a sequence of relaxed problems* $\left\{(\bar{\mathrm{P}}_k)\right\}_{k \geq 1}$.

We approximate the principal's screening problem (P) by a sequence of problems $(\bar{\mathrm{P}}_k)$, $k = 1, 2, \ldots$, in each of which the constraints are relaxed. The sequence of relaxed problems approximates the original problem, since deviations from the constraints and the optimal control u^* are penalized using successively increasing weights. We first fix the positive numbers ε, δ, and

$$\bar{u} \geq 2 + \operatorname*{ess\,sup}_{\theta \in \Theta_0^*} \|u^*(\theta)\|,$$

relative to which the sequence $\{(\bar{\mathrm{P}}_k)\}_{k=1}^{\infty}$ of relaxed problems will be defined. For any $k \geq 1$, let

$$V^k(x, u, \theta) = V(x, \theta) - \delta \|u(\theta) - u^*(\theta)\|^2 - k g_-(x, \theta)^2, \tag{5.1}$$

where $g_- = \min\{g, 0\}$ takes on nonzero (negative) values whenever the relevant constraint (3.2) of the original problem (P) is violated. Similarly, for any $\theta_0 \in \Theta$ and $\underline{x} \in \mathbb{R}_+^{n+1}$ we set

$$K(\underline{x}, \theta_0) = (\min\{\theta_0, 0\})^2 + \left(U(\underline{x}, \theta_0) - r(\theta_0)\right)^2 \tag{5.2}$$

to penalize deviations from the endpoint constraints (including $\theta_0 \geq 0$) imposed in the original problem. We are now ready to formulate the relaxed problem $(\bar{\mathrm{P}}_k)$ for any $k \geq 1$, given an optimal solution $(x^*, u^*, \theta_0^*) \in \mathcal{D}$ to the original problem (P):

$$\left\{ \begin{array}{l} \sup_{(x,u,\theta_0,\underline{x}) \in \hat{\mathcal{D}}} \left\{ \int_{\theta_0}^1 V^k(x(\theta), u(\theta), \theta)\, dF(\theta) - (\underline{x} - \underline{x}^*)^2 - (\theta_0 - \theta_0^*)^2 - k K(\underline{x}, \theta_0) \right\} \\ \text{s.t.} \\ \quad \dot{x} = \Phi u, \quad x(\theta_0) = \underline{x}, \\ \quad \varepsilon \geq \|x - x^*\|_\infty + (\underline{x} - \underline{x}^*)^2 + (\theta_0 - \theta_0^*)^2, \\ \quad u \in \mathcal{U}, \end{array} \right\} \tag{$\bar{\mathrm{P}}_k$}$$

where $\hat{\mathcal{D}} = \mathcal{D} \times \mathbb{R}_+^{n+1}$ is an augmented domain of admissibility, $\mathcal{U} = \{\hat{u} \in \mathbb{R}^n : \|\hat{u}\| \leq \bar{u}\}$ is a control-constraint set, and $\underline{x}^* = x^*(\theta_0^*)$ is the bundle offered to the optimal marginal agent type. We denote a solution to the relaxed problem $(\bar{\mathrm{P}}_k)$ by $(x^k, u^k, \theta_0^k, \underline{x}^k)$. By setting u^k to zero for all $\theta \notin [\theta_0^k, 1]$ we extend any such solution to all types in Θ (containing the optimal interval $[\theta_0^k, 1]$), so the state trajectory x^k is constant on $[0, \theta_0^k]$.

Step 2: Show that each problem $(\bar{\mathrm{P}}_k)$ *has a solution.*
For any $k \geq 1$, there exists a solution $(x^k, u^k, \theta_0^k, \underline{x}^k)$ to the relaxed problem $(\bar{\mathrm{P}}_k)$. Let

$$\{(x^{k,j}, u^{k,j}, \theta_0^{k,j}, \underline{x}^{k,j})\}_{j=1}^{\infty}$$

be an admissible *maximizing sequence* [21, p. 193] for the problem $(\bar{\mathrm{P}}_k)$. Since $u^{k,j}$ takes values in the closed ball of $\mathbb{R}^n$ at 0 of radius $\bar{u}$, and *a forteriori* the contract space (which contains all bundles that can actually be transacted between principal and agent) is bounded by T1,[24] this maximizing sequence is uniformly bounded, which allows the following three conclusions for an appropriate subsequence (and for simplicity we identify our original maximizing sequence with this subsequence by relabelling indices if necessary). First, from the definition of an admissible sequence $\{x^{k,j}\}_{j=1}^{\infty} \subset \mathbf{W}_{1,\infty}$ and the uniform boundedness of $\{\dot{x}^{k,j}\}_{j=1}^{\infty}$ this sequence of state trajectories is equicontinuous, so by the Arzelà—Ascoli theorem [22, Part I, p. 54] it converges uniformly to $\hat{x}^k$. Second, we obtain pointwise convergence of $(\underline{x}^{k,j}, \theta_0^{k,j})$ to $(\hat{\underline{x}}^k, \hat{\theta}_0^k)$ as $j \to \infty$. Third, since the space of admissible controls $\mathbf{L}_\infty$ is a subset of the Hilbert space $\mathbf{L}_2$ (which can be identified with its dual), $u^{k,j}$ converges weakly to $\hat{u}^k$ as $j \to \infty$.[25]

We now show that in fact the above limits coincide with the solution to $(\bar{\mathrm{P}}_k)$, i.e.,

$$(x^k, u^k, \theta_0^k, \underline{x}^k) = (\hat{x}^k, \hat{u}^k, \hat{\theta}_0^k, \hat{\underline{x}}^k). \tag{5.3}$$

For any $\theta \in [\theta_0^k, 1]$ it is

$$x^{k,j}(\theta) = x^{k,j}(\theta_0^{k,j}) + \int_{\theta_0^{k,j}}^{\theta} \Phi(x^{k,j}(\vartheta), \vartheta) u^{k,j}(\vartheta)\, d\vartheta,$$

so by taking the limit for $j \to \infty$:

$$\hat{x}^k(\theta) = \hat{x}^k(\hat{\theta}_0^k) + \int_{\hat{\theta}_0^k}^{\theta} \Phi(\hat{x}^k(\vartheta), \vartheta) \hat{u}^k(\vartheta)\, d\vartheta.$$

Hence, the limiting tuple $(\hat{x}^k, \hat{u}^k, \hat{\theta}_0^k, \hat{\underline{x}}^k)$ is consistent with the Cauchy problem for the state evolution, i.e., it satisfies

$$\dot{\hat{x}}^k = \Phi \hat{u}^k, \quad \hat{x}^k(\hat{\theta}_0^k) = \hat{\underline{x}}^k.$$

The state constraint $\varepsilon \geq \|\hat{x}^k - x^*\|_\infty + \left\|\hat{\underline{x}}^k - \underline{x}^*\right\|^2 + \left(\hat{\theta}_0^k - \theta_0^*\right)^2$ is satisfied by uniform convergence of the maximizing sequence. Lastly, the control constraint $\hat{u}^k \in \mathcal{U}$ is satisfied, since each $u^{k,j}$, $j = 1, 2, \ldots$, is feasible ($\bar{u}$ has been chosen appropriately large). The weak convergence $u^{k,j} \overset{w}{\to} \hat{u}^k$ as $j \to \infty$ implies, by Mazur's compactness theorem [23, p. 254], that there exists a sequence $\{v^{k,j}\}_{j=1}^{\infty}$ with elements in the convex hull $\operatorname{co}\{u^{k,j}\}_{j=1}^{\infty}$ which converges strongly to $\hat{u}^k$ in $\mathbf{L}_2^n(\Theta)$. We therefore obtain that equation (5.3) holds, i.e., the limit point $(\hat{x}^k, \hat{u}^k, \hat{\theta}_0^k, \hat{\underline{x}}^k)$ of the maximizing sequence describes an admissible solution to the relaxed problem $(\bar{\mathrm{P}}_k)$.

Step 3: Show that the solutions of $(\bar{\mathrm{P}}_k)_{k\geq 1}$ *converge to the solution of* (P).
As before, there exists an admissible tuple $(\hat{x}, \hat{u}, \hat{\theta}_0) \in \mathcal{D}$ such that $x^k \rightrightarrows \hat{x}^k$, $u^k \overset{w}{\to} \hat{u}$, and $\theta_0^k \to \hat{\theta}_0^k$. We now show that

$$(\hat{x}, \hat{u}, \hat{\theta}_0) = (x^*, u^*, \theta_0^*), \tag{5.4}$$

i.e., in particular that $x^k \rightrightarrows x^*$, $u^k \overset{w}{\to} u^*$, and $\theta_0^k \to \theta_0^*$.

[24] The assumption T1 is used here to guarantee the boundedness of the first (numéraire) component of the attribute schedule, as its dynamics in (3.1) are not directly governed by the control variable.

[25] By the Banach—Alaoglu theorem the unit ball in L_2 is weakly* (and therefore weakly) compact, so by the Eberlein—Šmulian theorem it is also weakly sequentially compact [23, p. 229/248]. This property of reflexive Banach spaces can also be deduced from the uniform boundedness principle [24, Ch. 2].

Let $\pi^k(x,u,\theta_0,\underline{x}) = \int_{\theta_0}^1 V^k(x(\theta),u(\theta),\theta)\,dF(\theta) - \|\underline{x}-\underline{x}^*\|^2 - (\theta_0-\theta_0^*)^2 - kK(\underline{x},\theta_0)$. As a consequence of the uniform boundedness of the state-control trajectories, there exists a constant $M>0$ such that $M \geq \pi^k(x^k,u^k,\theta_0^k,\underline{x}^k)$ for all k. Since $\pi^k(x^k,u^k,\theta_0^k,\underline{x}^k) - \pi^k(x^*,u^*,\theta_0^*,\underline{x}^*) \geq 0$, it is

$$\frac{M}{k} \geq \int_{\theta_0^k}^1 \left(\left(\frac{\delta}{k}\right)\|u^k-u^*\|^2 + g_-^2\right)\,dF(\theta) + \left(\frac{1}{k}\right)\left[\|\underline{x}^k-\underline{x}^*\|^2 + (\theta_0^k-\theta_0^*)^2\right] + K(\underline{x}^k,\theta_0^k) \geq 0.$$

Taking the limit for $k\to\infty$, by continuity of K it is $K(\hat{\underline{x}},\hat{\theta}_0)=0$, i.e., $(\hat{\underline{x}},\hat{\theta}_0)$ satisfies the endpoint constraints $U(\hat{\underline{x}},\hat{\theta}_0) = r(\hat{\theta}_0)$ and $\hat{\theta}_0 \geq 0$. Moreover,

$$\lim_{k\to\infty}\int_{\theta_0^k}^1 \left(g_-(x^k,\theta)\right)^2\,dF(\theta) = 0.$$

Since the type distribution F has full support Θ (no type can be excluded for sure), it is

$$g_-(x^k,\theta) = 0.$$

Hence, $g(x^k,\theta)=0$, which is equivalent to

$$\begin{cases} \hat{\theta}>\theta & \Rightarrow \quad U_\theta(x^*(\hat{\theta}),\theta) \geq U_\theta(x^k(\theta),\theta), \\ \hat{\theta}\leq\theta & \Rightarrow \quad U_\theta(x^*(\hat{\theta}),\theta) \leq U_\theta(x^k(\theta),\theta), \end{cases}$$

for all $\theta,\hat{\theta}\in\Theta$. Because U is by assumption twice differentiable, and $U_\theta(x,\theta)$ is Lipschitz-continuous in x, there exists a constant $L>0$ such that $|U_\theta(\bar{x},\theta) - U_\theta(\bar{y},\theta)| \leq L\|\bar{x}-\bar{y}\|$ for all $\bar{x},\bar{y}$ in the relevant (by assumption T1 bounded) contract space $\mathcal{X}$ and all $\theta\in\Theta$. For $\hat{\theta}>\theta$ this means that

$$U_\theta(x^k(\hat{\theta}),\theta) + L\|x^*-x^k\| \geq U_\theta(x^k(\hat{\theta}),\theta) + \Big(U_\theta(x^*(\hat{\theta}),\theta) - U_\theta(x^k(\hat{\theta}),\theta)\Big) \geq U_\theta(x^k(\theta),\theta),$$

and, for $\hat{\theta}<\theta$, that

$$U_\theta(x^k(\hat{\theta}),\theta) - L\|x^*-x^k\| \geq U_\theta(x^k(\hat{\theta}),\theta) + \Big(U_\theta(x^*(\hat{\theta}),\theta) - U_\theta(x^k(\hat{\theta}),\theta)\Big) \leq U_\theta(x^k(\theta),\theta).$$

Hence, it is

$$\min_{\hat{\theta}\in\Theta_0^*}\left\{(\hat{\theta}-\theta)\left(U_\theta(x^*(\hat{\theta}),\theta) - U_\theta(x^*(\theta),\theta)\right)\right\} \geq -L\varepsilon,$$

and for $\varepsilon\to 0^+$ the limit $(\hat{x},\hat{u},\hat{\theta}_0,\hat{\underline{x}})$ becomes admissible in problem (P). This implies

$$\pi(x^*,u^*,\theta_0^*,\underline{x}^*) \geq \pi(\hat{x},\hat{u},\hat{\theta}_0,\hat{\underline{x}}). \tag{5.5}$$

On the other hand, $\pi^k(x^k,u^k,\theta_0^k,\underline{x}^k) \geq \pi^k(x^*,u^*,\theta_0^*,\underline{x}^*) =: \pi(x^*,u^*,\theta_0^*,\underline{x}^*) = \bar{V}(x^*;\theta_0^*)$, whence

$$\pi(x^k,u^k,\theta_0^k,\underline{x}^k) - \|\underline{x}^k-\underline{x}^*\|^2 - (\theta_0^k-\theta_0^*)^2 - \delta\int_{\theta_0^k}^1 \|u^k-u^*\|^2\,dF(\theta) \geq \pi(x^*,u^*,\theta_0^*,\underline{x}^*)$$

for all $k\geq 1$. Hence, taking the limit for $k\to\infty$ yields

$$\pi(\hat{x},\hat{u},\hat{\underline{x}},\hat{\theta}_0) - \left(\hat{\theta}_0-\theta_0^*\right)^2 - \delta\lim_{k\to\infty}\int_{\Theta_0}\|u^k-u^*\|^2\,dF(\theta) \geq \pi(x^*,u^*,\theta_0^*,\underline{x}^*),$$

which together with (5.5) implies that $\hat{\theta}_0=\theta_0^*$, $\hat{u}=u^*$, and

$$\lim_{k\to\infty}\int_\Theta \|u^k-u^*\|^2\,dF(\theta) = 0,$$

so the sequence $\{u^k\}_{k=1}^{\infty}$ converges to u^* a.e. on Θ.

Step 4: Show that the problem $(\bar{\mathrm{P}}_k)$ *becomes a standard OCP* $(\bar{\mathrm{P}}'_k)$ *for large* k.
Because of the uniform convergence of the optimal state-trajectories x^k and the pointwise convergence of the boundary points θ_0^k (as $k \to \infty$) to the corresponding optimal state trajectory x^* and optimal boundary point θ_0^* of the original problem (P), the state constraint in the relaxed problem $(\bar{\mathrm{P}}_k)$ is strictly not binding for large enough k, i.e.,

$$\varepsilon > \|x^k - x^*\|_\infty + \left(\theta_0^k - \theta_0^*\right)^2,$$

as long as k is sufficiently large. Hence, for fixed constants ε, δ, and $\bar{u}$ there exists a $k_0 = k_0(\varepsilon, \delta, \bar{u}) \geq 1$ such that for all $k \geq k_0$ the problem $(\bar{\mathrm{P}}_k)$ can be rewritten equivalently in the form

$$\left\{ \begin{array}{l} \sup_{(x,u,\theta_0,\underline{x})\in\hat{\mathcal{D}}} \left\{ \int_{\theta_0}^1 V^k(x(\theta),u(\theta),\theta)\, dF(\theta) - \|\underline{x} - \underline{x}^*\|^2 - (\theta_0 - \theta_0^*)^2 - kK(\underline{x},\theta_0) \right\} \\ \text{s.t.} \\ \quad \dot{x} \;=\; \Phi u, \quad x(\theta_0) \;=\; \underline{x}. \end{array} \right\} \qquad (\bar{\mathrm{P}}'_k)$$

Necessary optimality conditions based on the maximum principle [7] for problem $(\bar{\mathrm{P}}'_k)$ are readily available.

Step 5: Obtain necessary optimality conditions for $(\bar{\mathrm{P}}'_k)$.
We provide here a version of the maximum principle by Milyutin and Osmolovskii [25, P. 24–25].[26] Let

$$H^k(x,u,\theta,\lambda_0^k,\psi^k) = \lambda_0^k V^k f + \psi^k \cdot \Phi u$$

be the Hamiltonian function associated with problem $(\bar{\mathrm{P}}'_k)$, where $\lambda_0^k \in \mathbb{R}$ is a constant multiplier, $\psi^k \in \mathbb{R}^{n+1}$ is an adjoint variable, and x is the state of the system. The Hamiltonian corresponds to the instantaneous payoff to the principal including the current benefit of state velocities. The shadow prices of these are measured by the adjoint variables λ_0^k and ψ^k respectively.

Maximum Principle for Problem $(\bar{\mathrm{P}}'_k)$. *If* $(x^k, u^k, \theta_0^k, \underline{x}^k)$ *is an optimal solution for the problem* $(\bar{\mathrm{P}}'_k)$*, then there exists an absolutely continuous function* $\psi^k : [\theta_0^k, 1] \to \mathbb{R}^{n+1}$*, and a constant* $\lambda_0^k > 0$*, such that the following relations hold.*

1. *Adjoint Equation:*

$$-\dot{\psi}^k(\theta) = H_x^k(x^k(\theta), u^k(\theta), \theta, \lambda_0^k, \psi^k(\theta)) \qquad (5.6)$$

2. *State Transversality:*

$$\psi^k(\theta_0^k) \;=\; 2k\left(U(\underline{x}^k,\theta_0^k) - r(\theta_0^k)\right) U_x(\underline{x}^k,\theta_0^k) + 2\left(\underline{x}^k - \underline{x}^*\right) \qquad (5.7)$$

$$\psi^k(1) \;=\; 0 \qquad (5.8)$$

3. *Maximality:*

$$u^k(\theta) \in \arg\max_{u\in\mathcal{U}} H^k(x^k(\theta), u, \theta, \lambda_0^k, \psi^k(\theta)) \quad \textit{a.e. on } [\theta_0^k, 1] \qquad (5.9)$$

One can extend the standard maximum principle comprised by the above conditions, and obtain the following type-transversality condition.

4. *Type Transversality:*

$$\left\{ \begin{array}{c} 2\left(k\min\{\theta_0^k, 0\} + k\left(U(\underline{x}^k,\theta_0^k) - r(\theta_0^k)\right)\left(U_\theta(\underline{x}^k,\theta_0^k) - \dot{r}(\theta_0^k)\right) + (\theta_0^k - \theta_0^*)\right) \\ + \sup_{u\in\mathcal{U}} H^k(\underline{x}^k, u, \theta_0^k, \lambda_0^k, \psi^k(\theta_0^k)) = 0. \end{array} \right\} \qquad (5.10)$$

[26] They consider a slightly more general Mayer problem. Problem $(\bar{\mathrm{P}}'_k)$ is a Bolza problem on a variable interval, which can be reduced to a Mayer problem (in which the objective function depends only on the endpoints of the state trajectory) by introducing an additional real-valued state variable.

The type-transversality condition is crucial for determining the missing multiplier associated with the optimization of the participation set in the principal's screening problem. To prove type transversality, consider some real τ (with $|\tau|$ sufficiently small), and let

$$\hat{\omega}_0^k(\tau) = (x^k(\theta_0^k + \tau), \theta_0^k + \tau),$$

and

$$\hat{\Pi}^k(\tau) = \int_{\theta_0^k+\tau}^1 \lambda_0^k V^k(x^k(\theta), u^k(\theta), \theta)\, dF(\theta) - \|\hat{\omega}_0^k(\tau) - (\underline{x}^*, \theta_0^*)\|^2 - kK(\hat{\omega}_0^k(\tau)),$$

where if necessary we extend the state-control trajectory beyond the lowest participating type. Then by optimality of θ_0^k,

$$\begin{aligned}
0 &\leq \hat{\Pi}_2^k(0) - \hat{\Pi}_2^k(\tau) \\
&= \int_{\theta_0^k}^{\theta_0^k+\tau} \lambda_0^k V^k(x^k, u^k, \theta)\, dF(\theta) + \left(\|\hat{\omega}_0^k(\tau) - (\underline{x}^*, \theta_0^*)\|^2 - \|(\underline{x}^k - \underline{x}^*, \theta_0^k - \theta_0^*)\|^2\right) \\
&\qquad\qquad + k\left(K(\hat{\omega}_0^k(\tau)) - K(\underline{x}^k, \theta_0^k)\right) \\
&= \int_{\theta_0^k}^{\theta_0^k+\tau} \left(\lambda_0^k V^k f + \left(2\left(x^k(\theta) - \underline{x}^*\right) + kK_{\underline{x}}(x^k(\theta), \theta)\right) \cdot \dot{x}^k(\theta) + 2(\theta - \theta_0^*) + kK_{\theta_0}(x^k(\theta), \theta)\right) d\theta \\
&= \int_{\theta_0^k}^{\theta_0^k+\tau} \left(H^k(x^k, u^k, \theta, \lambda_0^k, \psi^k)\big|_{\psi^k = 2(x^k(\theta) - \underline{x}^*) + kK_{\underline{x}}(x^k(\theta), \theta)} + kK_{\theta_0}(x^k(\theta), \theta)\right) d\theta + 2(\theta_0^k - \theta_0^*)\tau + \tau^2 \\
&\leq \int_{\theta_0^k}^{\theta_0^k+\tau} \left(\sup_{u\in\mathcal{U}} H^k(x^k, u, \theta, \lambda_0^k, 2\left(x^k - \underline{x}^*\right) + kK_{\underline{x}}(x^k, \theta)) + kK_{\theta_0}(x^k(\theta), \theta)\right) d\theta + 2(\theta_0^k - \theta_0^*)\tau + \tau^2.
\end{aligned}$$

By the first mean-value theorem for the integral [26, p. 352], the last integral can be written in the form

$$\left(\sup_{u\in\mathcal{U}} H^k\left(x(\check{\theta}_0^k), u, \check{\theta}_0^k, \lambda_0^k, 2\left(x^k(\check{\theta}_0^k) - \underline{x}^*\right) + kK_{\underline{x}}(x^k(\check{\theta}_0^k), \check{\theta}_0^k)\right) + kK_{\theta_0}(x^k(\check{\theta}_0^k), \check{\theta}_0^k)\right)\tau$$

for some appropriate $\check{\theta}_0^k \in [\theta_0^k, \theta_0^k + \tau]$. Hence, dividing the previous inequality by $\tau > 0$ and subsequently taking the limit for $\tau \to 0^+$ yields

$$0 \leq \sup_{u\in\mathcal{U}} H^k(x^k, u, \theta, \lambda_0^k, 2\left(\underline{x}^k - \underline{x}^*\right) + kK_{\underline{x}}(\underline{x}^k, \theta_0^k)) + kK_{\theta_0}(\underline{x}^k, \theta_0^k) + 2(\theta_0^k - \theta_0^*). \tag{5.11}$$

Consider now the case where we extend the type interval by an increment $\hat{\tau} > 0$ to the left beyond the optimal lowest type θ_0^k, which corresponds to $\tau < 0$ in the above relations. If we set $\hat{\tau} = -\tau > 0$, then it is (using our earlier computations)

$$0 \leq -\int_{\theta_0^k-\hat{\tau}}^{\theta_0^k} \left(H^k(x^k, u^k, \theta, \lambda_0^k, \psi^k)\Big|_{\psi^k = 2\left(x^k(\theta) - \underline{x}^*\right) + kK_{\underline{x}}(x^k(\theta), \theta)} + kK_{\theta_0}(x^k(\theta), \theta)\right) d\theta - 2(\theta_0^k - \theta_0^*)\hat{\tau} + \hat{\tau}^2,$$

where we can extend $u^k(\theta)$, for all $\theta \in [\theta_0^k - \hat{\tau}, \theta_0^k]$ to the left. In particular, the latter extension can be performed such that

$$\rho \geq \left[\sup_{u\in\mathcal{U}} H^k\left(x^k, u, \theta, \lambda_0^k, kK_{\underline{x}}(x^k(\theta), \theta)\right) - H^k\left(x^k, u^k, \theta, \lambda_0^k, kK_{\underline{x}}(x^k(\theta), \theta)\right)\right]$$

on the interval $[\theta_0^k - \hat{\tau}, \theta_0^k]$ for some arbitrary $\rho > 0$. Hence,

$$0 \leq -\left(\sup_{u\in\mathcal{U}} H^k(x^k(\check{\theta}_0^k), u, \check{\theta}_0^k, \lambda_0^k, 2\left(x^k(\check{\theta}_0^k) - \underline{x}^*\right) + kK_{\underline{x}}(x^k(\check{\theta}_0^k), \check{\theta}_0^k)) - \rho + kK_{\theta_0}(x^k(\check{\theta}_0^k), \check{\theta}_0^k) + 2(\theta_0^k - \theta_0^*)\right)\hat{\tau} + \hat{\tau}^2,$$

for some $\check{\theta}_0^k \in [\theta_0^k - \hat{\tau}, \theta_0^k]$, so by dividing through $\hat{\tau} > 0$ and taking the limits for $\hat{\tau} \to 0^+$ and $\rho \to 0^+$, we get

$$0 \geq \sup_{u \in \mathcal{U}} H^k(\underline{x}^k, u, \theta_0^k, \lambda_0^k, 2\left(\underline{x}^k - \underline{x}^*\right) + kK_{\underline{x}}(\underline{x}^k, \theta_0^k)) + kK_{\theta_0}(\underline{x}^k, \theta_0^k) + 2(\theta_0^k - \theta_0^*). \tag{5.12}$$

Relations (5.11) and (5.12), together with (5.7), are equivalent to (5.10).

Using the above extended maximum principle (5.6)–(5.10) for the relaxed problem $(\bar{\mathrm{P}}_k')$, one obtains the adjoint equation

$$-\dot{\psi}^k = \lambda_0^k V_x f + \nu^k g_x + \psi^k \cdot \left(\Phi_x u^k\right) = \lambda_0^k V_x f + \nu^k g_x + \psi_0^k \varphi_x u^k \tag{5.13}$$

on Θ, where

$$\nu^k(\theta) = \begin{cases} -2k\lambda_0^k g_-(x^k, \theta), & \theta \in [\theta_0^k, 1], \\ 0, & \text{otherwise.} \end{cases} \tag{5.14}$$

The maximality condition (5.9) constitutes a constrained optimization problem, for which there exists a Lagrange multiplier $\varsigma^k = \varsigma^k(\theta)$ such that

$$\psi_0 \varphi_i + \psi_i - 2\lambda_0^k \delta(u_i^k - u_i^*) + 2\varsigma^k u_i^k = 0$$

for all $i \in \{1, \ldots, n\}$, with complementary slackness $\varsigma^k \left(\|u^k\| - \bar{u}\right) = 0$. By Step 3 we know that $u^k \to u^*$ a.e. on Θ. Hence, by Egorov's theorem [27, p. 24] for any $\bar{\delta} > 0$ there is a subset $\Theta_{\bar{\delta}}$ of Θ, such that $\int_{\Theta \setminus \Theta_{\bar{\delta}}} d\theta < \bar{\delta}$ and $u^k \rightrightarrows u^*$ uniformly on $\Theta_{\bar{\delta}}$. Since u^* is feasible, this uniform convergence implies that $u^k \notin \partial\mathcal{U}$ on $\Theta_{\bar{\delta}}$ for k large enough. By virtue of complementary slackness, the corresponding Lagrange multipliers ς^k and ρ^k therefore vanish on $\Theta_{\bar{\delta}}$, as long as k is large enough. In other words,

$$\psi_0 \varphi_i + \psi_i - 2\lambda_0^k \delta(u^k - u^*) = 0 \tag{5.15}$$

a.e. on Θ for all $i \in \{1, \ldots, n\}$, as long as k is large enough.

Step 6: Derive necessary optimality conditions for (P).
The sequence $\{\lambda_0^k\}_{k=1}^{\infty}$ is uniformly bounded and $\{\psi_k\}_{k=1}^{\infty}$ is also equicontinuous. Hence we conclude, as before (in Step 2), that there exist $\psi_{\varepsilon,\bar{u}}$ and $\lambda_{\delta,\bar{u}}$, such that

$$\psi^k \rightrightarrows \psi_{\delta,\bar{u}}, \quad \lambda_0^k \to \lambda_{\delta,\bar{u}}.$$

As already indicated through the notation, the limits $\psi_{\delta,\bar{u}}$ and $\lambda_{\delta,\bar{u}}$ generically depend on the constants δ and $\bar{u}$. More specifically, these limits correspond to the optimal solution to the screening problem (P) if we replace V by $V - \delta\|u - u^*\|^2$ and introduce the additional constraint $\|u\| \leq \bar{u}$.

Adjoint Equation. Since by the maximum principle for problem $(\bar{\mathrm{P}}_k')$ (cf. Step 5) it is $\lambda_0^k > 0$, relations (5.6)–(5.8) are positively homogeneous of degree one in ψ^k/λ_0^k, and relation (5.9) is positively homogeneous of degree zero, it is possible to multiply equations (5.6)–(5.8) with positive numbers (and relabel the variables λ_0^k and ψ^k back), such that

$$0 < \lambda_0^k + \max_{\theta \in [\theta_0^k, 1]} \|\psi^k(\theta)\|^2 + \int_{\theta_1^k}^1 \left(\nu^k(\theta)\right)^2 d\theta \leq 1. \tag{5.16}$$

Integrating the components of the adjoint equation (5.13) yields (using the state-transversality condition (5.8))

$$\psi^k(\theta) \;=\; \int_\theta^1 \left(\lambda_0^k V_x f + \psi_0^k \varphi_x u^k\right) d\theta + \int_\theta^1 \nu^k g_x d\theta \tag{5.17}$$

for all $\theta \in \Theta$ (using the standard extension from $[\theta_0^k, 1]$ to Θ explained in Step 1).

Consider now a Borel measure with density given ν^k by (5.14).[27] By abuse of notation we denote this measure also by ν^k. The Borel measure ν^k is nonnegative by construction and bounded as a consequence of (5.16). Thus, there exists a Borel measure ν, such that (passing to a subsequence if necessary) $\nu^k \xrightarrow{w} \nu$ as $k \to \infty$.

Since the total variation of ψ^k on Θ is uniformly bounded for all k by (5.16) (for every absolutely continuous function is of bounded variation on a compact interval [28, p. 412]), and the sequence $\{\psi^k\}$ is also uniformly bounded as a consequence of (5.17), by Helly's selection principle [28, p. 398] there exists a function $\hat{\psi}$, so (a subsequence of) the sequence $\{\psi^k\}$ converges to $\hat{\psi}$. By taking the limit in (5.17) for $k \to \infty$ we thus obtain

$$\hat{\psi}_x(\theta) = \int_\theta^1 \left(\lambda_0 V_x f + \hat{\psi} \cdot \Phi_x\right) d\theta + \int_\theta^1 g_x d\nu, \tag{5.18}$$

for almost all $\theta \in \Theta_0^*$, where $\hat{\psi} = \psi_{\delta,\bar{u}}$. The adjoint equation C1 then follows by noting that (using the envelope theorem) $g_x(x^*(\theta),\theta) = -\left(\rho(x^*(\theta),\theta) - \theta\right) U_{x\theta}(x^*(\theta),\theta)$.

Maximality. Consider the maximality condition (5.15) for Problem $(\bar{\mathrm{P}}_k)$, which holds for k large enough. Since $x^k \rightrightarrows x^*$ and $u^k \to u^*$ (a.e. on Θ) for $k \to \infty$, we obtain the maximality condition C2.

Transversality. Since $\theta_0^k \to \theta_0^*$ as k tends to infinity (cf. Step 3), by setting

$$\lambda_1 = -\lim_{k\to\infty} 2k\left(U(\underline{x}^k,\theta_0^k) - r(\theta_0^k)\right) \text{ and } \lambda_2 = -\lim_{k\to\infty} 2k \min\{\theta_0^k, 0\}$$

we obtain from the state-transversality condition (5.7) for $k \to \infty$ (taking into account the maximality condition C2) that

$$\psi(\theta_0^*) = -\lambda_1 U_x(\underline{x}^*,\theta_0^*), \tag{5.19}$$

and from the type transversality condition (5.10) that

$$-\lambda_1\left(U_\theta(\underline{x}^*,\theta_0^*) - \dot{r}(\theta_0^*)\right) - \lambda_2 + \lambda_0 V(\underline{x}^*,\theta_0^*) = 0. \tag{5.20}$$

If $\theta_0^* > 0$, then by the definition of λ_2 it is $\lambda_2 = 0$. Combining relations (5.19) and (5.20) for $\lambda_2 = 0$ yields the transversality condition C3 (ii). If on the other hand $\theta_0^* = 0$, then we obtain the transversality condition C3 (i) from relation (5.19) directly.

[27] Recall [29, Ch. 1] that all Borel measures ν on Θ form a linear normed space with norm

$$\|\nu\| = \sup_{h\in C[0,1]:\|h\|_\infty=1} \int_0^1 h(\vartheta)\, d\nu(\vartheta).$$

Interpreting the above Stieltjes integral as a scalar product, we can think of any Borel measure ν as an element of the dual space $(C[0,1])^*$. A sequence of Borel measures $\nu^1, \nu^2, \ldots$ weak*-converges to ν if for all $h \in C[0,1]$,

$$\int_0^1 h(\vartheta)\, d\nu^k(\vartheta) \to \int_0^1 h(\vartheta)\, d\nu(\vartheta),$$

as $k \to \infty$. Since $C[0,1]$ is separable, by the theorem of choice every (bounded) sequence of measures has a weak*-convergent subsequence [30, p. 64; 31, p. 189]. Since for any $h \in C[0,1]$ the function $\int_0^\theta h(\vartheta)\, d\nu(\vartheta)$ is continuous a.e. on Θ, for any sequence of points $\{\theta_k\}_{k=1}^\infty$ with limit θ, for almost every such $\theta \in \Theta$ it is:

$$\lim_{k\to\infty} \int_0^{\theta_k} h(\vartheta)\, d\nu(\vartheta) = \int_0^\theta h(\vartheta)\, d\nu(\vartheta).$$

Moreover, if in addition the sequence $\{\nu^k\}_{k=0}^\infty$ of Borel measures has the Borel measure ν as its limit, then

$$\lim_{k\to\infty} \int_0^{\theta_k} h(\vartheta)\, d\nu^k(\vartheta) = \int_0^\theta h(\vartheta)\, d\nu(\vartheta)$$

for almost every $\theta \in \Theta$.

Complementary Slackness. The definition (5.14) of the measure ν^k, and the fact that $\nu^k \xrightarrow{w} \nu$ as $k \to \infty$ (cf. also footnote 27) yields the complementary slackness condition C4.

Nontriviality. If λ_0 and ψ are trivial, then (λ_0, ψ) must vanish identically on Θ_0^*. From the adjoint equation C1 it follows that

$$\int_\theta^1 g_x(x^*(\vartheta), \vartheta)\, d\nu(\vartheta) = 0$$

for all $\theta \in \Theta_0^*$. By the complementary slackness condition C4, whenever $\nu(\theta) \neq 0$ it is also $g(x^*(\theta), \theta) = 0$. This implies that

$$\int_{\theta_0^*}^1 d\nu = 0.$$

On the other hand, it is possible to renormalize the middle term in (5.16) in the solution of the approximate problem $(\bar{\mathrm{P}}_k)$ such that

$$\lambda_0^k + \max_{\theta \in [\theta_0^k, 1]} \|\psi^k(\theta)\|^2 + \int_{\theta_0^k}^1 + \left(\nu^k(\theta)\right)^2 d\theta = 1$$

for any $k \geq 1$. By taking the limit for $k \to \infty$,

$$\int_{\theta_0^*}^1 \left(\nu(\theta)\right)^2 d\theta = 1.$$

By the nonnegativity of the measure ν, this yields a contradiction. Hence, the incentive regularity condition A3 ensures that (λ_0, ψ) does not vanish identically on the optimal participation set Θ_0^*, which implies the nontriviality condition C5. □

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

A NUMERICAL METHOD FOR SOLVING LINEAR–QUADRATIC CONTROL PROBLEMS WITH CONSTRAINTS[1]

Mikhail I. Gusev

N.N. Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia,
gmi@imm.uran.ru

Igor V. Zykov

N.N. Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russia,
zykoviustu@mail.ru

Abstract: The paper is devoted to the optimal control problem for a linear system with integrally constrained control function. We study the problem of minimization of a linear terminal cost with terminal constraints given by a set of linear inequalities. For the solution of this problem we propose two-stage numerical algorithm, which is based on construction of the reachable set of the system. At the first stage we find a solution to finite-dimensional optimization problem with a linear objective function and linear and quadratic constraints. At the second stage we solve a standard linear-quadratic control problem, which admits a simple and effective solution.

Key words: Optimal control, Reachable set, Integral constraints, Convex programming, Semi-infinite linear programming.

Introduction

The optimal control problems under integrally constrained controls were studied in many papers (see, for example, [1, 5, 7–9, 12, 13]). In [2, 3] the authors considered the linear control system with integrally constrained control. They studied the problem of minimization of a linear terminal cost under linear terminal constraints and proposed a saddle-point method to solve it.

We propose here a two-stage numerical algorithm for the solution of above optimal control problem. It is based on constructing the reachable set of the control system, and we use here the well-known result that this set for a linear control system with integral quadratic constraints on controls is an ellipsoid in the state space. Then, at the first stage, we find a minimum of a linear function on the intersection of the polyhedron and ellipsoid. This problem may be solved numerically in different ways. At the second stage, we solve the standard linear-quadratic control problem with fixed endpoints of the trajectory, this problem has a simple and effective solution in the linear case. The typical unpleasant feature of the optimal control problem with terminal cost is a nonuniqueness of solutions, which always takes place if the endpoint of the optimal trajectory belongs to the interior of the reachable set. This leads to additional difficulties in the construction and implementation of numerical algorithms. The method proposed here avoids these problems.

[1]The research is supported by Russian Science Foundation, project no. 16–11–10146.

1. Notation and problem statement

We use the following notations. By $A^\top$ we denote the transpose of a real matrix A, 0 stands for a zero vector of appropriate dimension, I is an identity matrix. For $x, y \in \mathbb{R}^n$ let $(x, y) = x^\top y$ be the inner product, $x^\top = (x_1, \ldots, x_n)$, $\|x\| = (x, x)^{\frac{1}{2}}$ be the Euclidean norm, and $B_r(\bar{x})$: $B_r(\bar{x}) = \{x \in \mathbb{R}^n : \|x - \bar{x}\| \le r\}$ be a ball of radius $r > 0$ centered at $\bar{x}$. For a set $S \subset \mathbb{R}^n$ let ∂S, $\mathrm{int}S$, $\mathrm{cl}S$, $\mathrm{co}S$ be a boundary, an interior, a closure, and a convex hull of S respectively; $\nabla g(x)$ is the gradient of a function $g(x)$ at the point x, $\frac{\partial f}{\partial x}(x)$ is the Jacobi matrix of a vector-valued function $f(x)$. For a real $k \times m$ matrix A a matrix norm is denoted as $\|A\|_{k\times m}$. The symbols $\mathbb{L}_1$, $\mathbb{L}_2$ and $\mathbb{C}$ stand for the spaces of summable, square summable and continuous functions respectively. The norms in these spaces are denoted as $\|\cdot\|_{\mathbb{L}_1}$, $\|\cdot\|_{\mathbb{L}_2}$, $\|\cdot\|_{\mathbb{C}}$.

We consider a linear control system with integral constraints on a controls

$$\dot{x}(t) = A(t)x(t) + B(t)u(t), \quad t \in [t_0, t_1], \quad x(t_0) = x^0, \tag{1.1}$$

where $x \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^r$, $A(t), B(t)$ are integrable on $[t_0, t_1]$ matrix functions. Let the control constraints are defined by the integral quadratic inequality

$$u(\cdot) \in U = \left\{ u(\cdot) \in \mathbb{L}_2 : \ J(u(\cdot)) = \|u(\cdot)\|^2_{\mathbb{L}_2} = \int_{t_0}^{t_1} u^\top(t)u(t)dt \le \mu^2 \right\}, \tag{1.2}$$

where $\mu > 0$ is a given number. For any $u(\cdot) \in \mathbb{L}_2$ there exists a unique absolutely continuous solution $x(t)$ of system (1.1), which is defined on interval $[t_0, t_1]$.

Assume that $m \times n$–matrix D, vectors $d \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ are given. We consider here the following optimal control problem for system (1.1):

$$I(u(\cdot)) = c^\top x(t_1) \to \min, \tag{1.3}$$

under constraints

$$u(\cdot) \in U, \quad Dx(t_1) \le d. \tag{1.4}$$

Definition 1. *The reachable set $G(t_1)$ of system* (1.1) *at time instant t_1 is a set of all states $x(t_1)$ that can be reached by the trajectories of* (1.1) *corresponding to controls $u(\cdot) \in U$:*

$$G(t_1) = \{x \in \mathbb{R}^n : \exists u(\cdot) \in U, \ x = x(t_1)\},$$

where $x(t)$ is a solution of (1.1).

The considered optimization problem may be split into two following subproblems.

Problem 1 (the first stage): to find $x^* \in \mathbb{R}^n$ that solves the finite-dimensional optimization problem

$$c^\top x \to \min,$$

under constraints

$$x \in G(t_1), \quad Dx \le d.$$

Here the reachable set $G(t_1)$ is an ellipsoid in $\mathbb{R}^n$ those parameters are calculated effectively. Thus, the optimization Problem 1 may be solved by employing the methods of linear or convex programming.

Problem 2 (the second stage): to find a control $u(\cdot) \in U$ that steers the trajectory $x(t)$ of (1.1) to point $x(t_1) = x^*$ and minimizes functional $J(u(\cdot))$.

2. Description of reachable sets

Let $X(t,\tau) = \Phi(t)\Phi^{-1}(\tau)$, where $\Phi(t)$ is a Cauchy matrix, satisfying the equation

$$\dot{\Phi}(t) = A(t)\Phi(t), \quad \Phi(t_0) = I.$$

A solution of (1.1) at time instant t_1 has the form

$$x(t_1) = \hat{x} + \int_{t_0}^{t_1} X(t_1,\tau)B(\tau)u(\tau)d\tau, \tag{2.1}$$

where $\hat{x} = X(t_1,t_0)x^0$. Take an arbitrary vector $l \in \mathbb{R}^n$, $l \neq 0$ and calculate the maximum of inner product $(l, x(t_1))$ over all $x(t_1) \in G(t_1)$:

$$\max_{x(t_1)\in G(t_1)}(l,x(t_1)) = l^\top\hat{x} + \max_{u(\cdot)\in U}\int_{t_0}^{t_1} l^\top X(t_1,\tau)B(\tau)u(\tau)d\tau = l^\top\hat{x} + \max_{\langle u(\cdot),u(\cdot)\rangle\le\mu^2}\langle v(\cdot),u(\cdot)\rangle =$$
$$= l^\top\hat{x} + \mu\|v(\cdot)\|_{\mathbb{L}_2} = \mu\sqrt{l^\top W(t_1)l} + l^\top\hat{x}.$$

Here $v(t) = B^\top(t)X^\top(t_1,t)l$,

$$\langle v(\cdot),u(\cdot)\rangle = \int_{t_0}^{t_1} v^\top(t)u(t)dt$$

is an inner product of functions $v(\cdot), u(\cdot)$ in the Hilbert space $\mathbb{L}_2$, and symmetric matrix $W(t)$ is defined by the equality

$$W(t) = \int_{t_0}^{t} X(t,\tau)B(\tau)B^\top(\tau)X^\top(t,\tau)d\tau.$$

Differentiating the last equality in t we get the following matrix differential equation for W:

$$\dot{W}(t) = A(t)W(t) + W(t)A^\top(t) + B(t)B^\top(t), \quad W(t_0) = 0. \tag{2.2}$$

It is known (see, for example, [10]), that system (1.1) is completely controllable on $[t_0,t_1]$, if and only if $W(t_1)$ is positive definite. In this case $G(t_1)$ is a nondegenerate ellipsoid

$$G(t_1) = \{(x-\hat{x})^\top W^{-1}(t_1)(x-\hat{x}) \le \mu^2\}.$$

If the system is not completely controllable, the reachable set is a degenerate ellipsoid (ellipsoid, lying in a subspace of dimension less than n). It is obvious, that $x \in \partial G(t_1)$ if and only if there exists a vector $l \neq 0$ such that $(l,x) = \max_{y\in G(t_1)}(l,y)$, and hence, a control $u(\cdot)$ steering the system trajectory to point x satisfies the relation

$$\int_{t_0}^{t_1} l^\top X(t,\tau)B(\tau)u(\tau)d\tau = \max_{\langle u(\cdot),u(\cdot)\rangle\le\mu^2}\langle v(\cdot),u(\cdot)\rangle, \tag{2.3}$$

where $v(t)$ is defined above. If the system is completely controllable, $v(\cdot) \neq 0$ and the equality (2.3) uniquely determines $u(t) = \alpha v(t)$, where $\alpha = \mu/\|v(\cdot)\|_{\mathbb{L}_2}$. Denote $p(t) = \alpha X^\top(t_1,t)l$, assuming that $p(t)$ is a nontrivial solution of the adjoint differential equation $\dot{p}(t) = -A^\top(t)p(t)$ and the relations

$$u(t) = B^\top(t)p(t), \quad J(u(\cdot)) = \mu^2$$

hold. Define a Hamiltonian $H(p,t,x,u)$ by the expression

$$H(p,t,x,u) = -\frac{1}{2}u^\top u + p^\top(A(t)x + B(t)u).$$

Equating to zero the gradient of $H(p(t), t, x(t), u)$ in u, from the concavity of the Hamiltonian in u we get that control $u(t) = B^\top(t)p(t)$ satisfies the maximum principle

$$H(p(t), t, x(t), u(t)) = \max_{v \in \mathbb{R}^r} H(p(t), t, x(t), v),$$

where $p(t)$ is a nontrivial solution of the adjoint differential equation

$$\dot{p}(t) = -\frac{\partial H}{\partial x}(p(t), t, x(t), u(t)).$$

This form of the maximum principle corresponds to the problem of minimization of a convex functional $J(u(\cdot))$ on solutions of linear system (1.1)

$$J(u(\cdot)) \to \min, \quad x(t_0) = x^0, \quad x(t_1) = x. \tag{2.4}$$

Since for a linear-convex optimal control problem the maximum principle provides necessary and sufficient optimality conditions, a control $u(t) = B^\top(t)p(t)$, found from the maximum principle, solves problem (2.4). Inversely, let $u(\cdot)$ be the solution to problem (2.4) and let $J(u(\cdot)) = \mu^2$. Then we will have $x \in G(t_1)$. Indeed, if we assume that $x \notin \partial G(t_1)$, then

$$(x - \hat{x})^\top W^{-1}(t_1)(x - \hat{x}) = \nu^2 < \mu^2.$$

Hence, the control system can be transferred to the point x by the control $v(\cdot)$, for this control we have $J(v(\cdot)) \le \nu^2 < \mu^2$ which contradicts to the optimality of $u(t)$. Thus, we come to the following statement.

Assertion 1. *Let system* (1.1)4 *be completely controllable. In order to control* $u(\cdot)$ *steers a trajectory of system* (1.1) *to point* x*, lying on the boundary of the reachable set* $G(t_1)$*, it is necessary and sufficient that this control solves the extremal problem* (2.4) *and the minimum of functional* J *equals to* μ^2*.*

3. The algorithm of solving the optimal control problem

Having the parameters of the reachable set given, let us describe the algorithm of the solution of the problem. Consider system (1.1) and assume that it is completely controllable. Then, at the first stage, we should solve the Problem 1

$$\begin{gathered} c^\top x \to \min, \\ Dx \le d, \\ (x - \hat{x})^\top W^{-1}(t_1)(x - \hat{x}) \le \mu^2. \end{gathered} \tag{3.1}$$

The problem (3.1) is a linear programming problem with an additional constraint, which is defined by an inequality with a positive definite quadratic form. This problem may be solved by various algorithms. For example, in [11] the authors proposed a finite convergent algorithm for solution of the problem of type (3.1).

The second way of finding the solution concerns with description of the ellipsoidal reachable set via its support function:

$$G(t_1) = \{x \in \mathbb{R}^n : (l, x) \le \psi(l), \ \forall l \in S\},$$

where the support function has the form

$$\psi(l) = \mu\sqrt{l^\top W(t_1) l} + (\hat{x}, l),$$

with $S = \{l : \|l\| = 1\}$. Then, the problem (3.1) may be written as follows

$$\begin{gathered} c^\top x \to \min, \\ Dx \le d, \quad l^\top x - \psi(l) \le 0, \quad l \in S. \end{gathered} \tag{3.2}$$

The problem (3.2) is a semi-infinite linear programming problem [6], which can be solved by a number of effective numerical algorithms.

Choosing a finite grid of N vectors $l_i \in S$, we can approximate the problem (3.2) by the following linear programming problem with a finite number of constraints

$$\begin{gathered} c^\top x \to \min, \\ Dx \le d, \quad {l_i}^\top x - \psi(l_i) \le 0, \quad i = 1, ..., N. \end{gathered} \tag{3.3}$$

Let a solution x^* of problem (3.1) ((3.2), (3.3)) be obtained (this solution is unique as a rule). Then we come to the next stage: to find the solution of the next problem

$$J(u(\cdot)) \to \min, \quad x(t_0) = x^0, \quad x(t_1) = x^*. \tag{3.4}$$

For completely controllable system this solution does exist and is unique. It may be obtained from the maximum principle:

$$u(t) = B^\top(t)p(t), \quad \dot{p} = -A^\top(t)p(t), \quad p^1 = p(t_1). \tag{3.5}$$

Represent $p(t)$ as follows

$$p(t) = X^\top(t_1, t)p^1,$$

then we have

$$x(t_1) = \hat{x} + \int_{t_0}^{t_1} X(t_1, \tau)B(\tau)B^\top(\tau)X^\top(t_1, \tau)p^1 d\tau = x^*. \tag{3.6}$$

Thus, to find an optimal control it is sufficient:
1) to find a vector p^1 from the linear equation

$$W(t_1)p^1 + \hat{x} = x^*, \tag{3.7}$$

2) to integrate adjoint equation (3.5) with boundary condition $p(t_1) = p^1$ and substitute $p(t)$ into the formula for $u(t)$.

We can put from the very beginning $x(t_1) = \hat{x} + W(t_1)p^1$ and solve a semi-infinite linear programming problem in the dual variables p^1

$$\begin{gathered} c^\top W(t_1)p^1 \to \min \\ DW(t_1)p^1 \le d - D\hat{x}, \quad l^\top W(t_1)p^1 - \mu\sqrt{l^\top W(t_1)l} \le 0, \quad l \in S. \end{gathered} \tag{3.8}$$

The last form is more convenient, it may be used even in the case of degenerate matrix $W(t_1)$. Really, it is known [8] that if there exists a control from $\mathbb{L}_2$ that steers the control system from x^0 to x^*, than control $u(\cdot)$ solving the problem (3.4) is a linear combination of columns of matrix $B^\top(\tau)X^\top(t_1, \tau)$. That is

$$u(\tau) = B^\top(\tau)X^\top(t_1, \tau)p^1$$

for some $p^1 \in \mathbb{R}^n$. Substituting the control $u(t)$ into equality (3.6) we get (3.7). Thus, the following theorem holds.

Theorem 1. *Assume that the system of constraints* (1.1), (1.4) *is consistent. Then the optimal control in the problem* (1.3)–(1.4) *is given by formulas* (3.5), *where* p^1 *is a solution to the linear semi-infinite optimization problem* (3.8).

4. Example

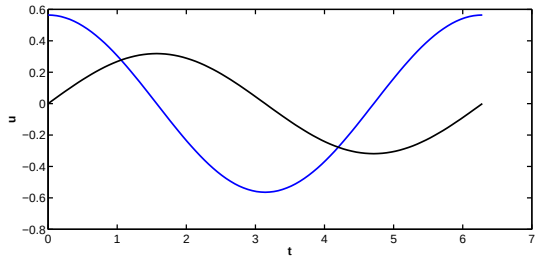


Figure 1. Graphs of optimal controls

Consider an illustrative example of optimal control problem for a linear control system described by the equations

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 + x_3, \\ \dot{x}_3 &= u \end{aligned} \tag{4.1}$$

with $t_0 = 0,\ t_1 = 2\pi$ and $x^0 = (1, 0, 0)^\top$.

We consider the integral quadratic constraints on controls given by the inequality (1.2) with $\mu^2 = 1$.

Assume a matrix D and a vector d that determine the terminal constraints are

$$D = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad d = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}.$$

Integrating a differential equations (2.2) for matrix W and calculating vector $\hat{x}$ we get

$$W(2\pi) = \begin{pmatrix} 9.4264 & 0.0005 & 6.2831 \\ 0.0005 & 3.1399 & 0.0000 \\ 6.2831 & 0.0000 & 6.2832 \end{pmatrix}, \quad \hat{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Let the terminal cost is determined by vector $c = (0, 2, 2)^\top$. Solving the problem (3.1) we get the solution

$$x^* = (0.9993, -1.0000, 0.0000)^\top, \quad p^1 = (-0.0002, -0.3185, 0.0002)^\top.$$

The graph of the optimal control is shown in Figure 1 by the black line. Here x^* is the interior point of the reachable set, so there are infinitely many admissible control inputs that steer the trajectories to point x^*. Among these inputs the considered control input has a minimal value of the integral functional $J(u(\cdot))$ which equals to 0.3185.

Consider another case and put $c = (1, 0, 0)^\top$. In this case we have the following solution

$$x^* = (-0.7729, -0.0003, 0.0000)^\top, \quad p^1 = (-0.5640, -0.0000, 0.5640)^\top.$$

Here $(x^* - \hat{x})^\top W^{-1}(2\pi)(x^* - \hat{x}) = 1$, hence x^* belongs to the boundary of the reachable set. In this case there exists a unique optimal control input which is shown in Figure 1 by the blue line.

5. Further generalizations

Consider here the generalized statement for the previous problem. We assume that integral constraints restrict simultaneously a control and a trajectory of system (1.1) as follows

$$J(u(\cdot)) = \frac{1}{2}\int_{t_0}^{t_1} \Big(x^\top(t)Q(t)x(t) + u^\top(t)R(t)u(t) \Big)\, dt \le \frac{\mu^2}{2}, \tag{5.1}$$

where $Q(t)$ is a nonnegative definite and $R(t)$ is a positive definite matrix for every $t \in [t_0, t_1]$. Suppose $Q(t)$ and $R(t)$ to be the measurable and bounded matrix functions.

Assume that a pair $(A(t), B(t))$ is completely controllable on $[t_0, t_1]$. Let $x(t_0) = x^0$ be fixed and let $x \in G(t_1)$. The set of trajectories satisfying inequality (5.1) is a compact set in $\mathbb{C}$. Hence, there exists the solution to the problem

$$J(u(\cdot)) \to \min, \quad x(t_0) = x^0, \quad x(t_1) = x,$$

this solution is unique due to strict convexity of the functional $J(u(\cdot))$.

Consider a Hamiltonian

$$H(t, x, u, (p_0, p)) = -p_0 \frac{1}{2} \left(x^\top Q(t) x + u^\top R(t) u \right) + p^\top (A(t)x + B(t)u).$$

According to the maximum principle (see, for example, [4, 10]) there exist $(p_0, p(\cdot)) \neq 0$ such that $u(t)$ maximizes a Hamiltonian, hence we have $\frac{\partial H}{\partial u} = 0$ and

$$\dot{p}(t) = -\frac{\partial H}{\partial u} = -A^\top(t) p(t) + p_0 Q(t) x.$$

Assuming $p_0 = 0$, we get $p(\cdot) \neq 0$. Then $p(t)$ is a nonzero solution of the equation

$$\dot{p}(t) = -A^\top(t) p(t),$$

and the condition $\partial H / \partial u = 0$ implies the equality $p^\top(t) B \equiv 0$, this contradict to the controllability conditions. Thus, $p_0 \neq 0$, so we can take $p_0 = 1$. In this case we have

$$u(t) = R^{-1}(t) B^\top(t) p(t). \tag{5.2}$$

Substituting control (5.2) into equations of the control system we get

$$\begin{aligned} \dot{x} &= A(t)x + B(t) R^{-1}(t) B^\top(t) p, \\ \dot{p} &= -A^\top(t) p + Q(t) x. \end{aligned} \tag{5.3}$$

Thus, (5.3) is a linear homogeneous system of differential equations

$$\begin{pmatrix} \dot{x} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} A(t) & B(t) R^{-1}(t) B(t) \\ Q(t) & -A^\top(t) \end{pmatrix} \begin{pmatrix} x \\ p \end{pmatrix}. \tag{5.4}$$

The solution of (5.4) with the initial state $x(t_0) = x^0$, $p(t_0) = p^0$ has the form

$$\begin{pmatrix} x(t) \\ p(t) \end{pmatrix} = Y(t) \begin{pmatrix} x^0 \\ p^0 \end{pmatrix},$$

where $Y(t)$ is the fundamental matrix of system (5.4) satisfying the initial condition $Y(t_0) = I$. Representing $Y(t)$ as a block matrix

$$Y(t) = \begin{pmatrix} Y_{11}(t) & Y_{12}(t) \\ Y_{21}(t) & Y_{22}(t) \end{pmatrix},$$

we will have

$$\begin{aligned} x(t) &= Y_{11}(t) x^0 + Y_{12}(t) p^0, \\ p(t) &= Y_{12}(t) x^0 + Y_{22}(t) p^0, \\ x = x(t_1) &= Y_{11}(t_1) x^0 + Y_{12}(t_1) p^0, \\ u(t) &= R^{-1}(t) B^\top(t) \left(Y_{12}(t) x^0 + Y_{22}(t) p^0 \right). \end{aligned} \tag{5.5}$$

Substituting $x(t)$ and $u(t)$ into $J(u)$ we get

$$J(u(\cdot)) = \frac{1}{2}\int_{t_0}^{t_1} \Big[\Big({x^0}^\top {Y_{11}}^\top(t) + {p^0}^\top {Y_{12}}^\top(t)\Big) Q(t) \big(Y_{11}(t)x^0 + Y_{12}(t)p^0\big)$$

$$+ \Big({x^0}^\top {Y_{12}}^\top(t) + {p^0}^\top {Y_{22}}^\top(t)\Big) B(t){R^{-1}}^\top(t)B^\top(t) \big(Y_{12}(t)x^0 + Y_{22}(t)p^0\big) \Big] dt \le \frac{\mu^2}{2},$$

or

$$ {x^0}^\top S_{11}x^0 + {x^0}^\top S_{12}p^0 + {p^0}^\top S_{22}p^0 \le \mu^2, \tag{5.6}$$

where

$$S_{11} = \int_{t_0}^{t_1} \Big({Y_{11}}^\top(t)Q(t)Y_{11}(t) + {Y_{12}}^\top(t)B(t)R^{-1}(t)B^\top(t){Y_{12}}^\top(t)\Big)\, dt,$$
$$S_{12} = 2\int_{t_0}^{t_1} \Big({Y_{11}}^\top(t)Q(t)Y_{12}(t) + {Y_{12}}^\top(t)B(t)R^{-1}(t)B^\top(t){Y_{22}}^\top(t)\Big)\, dt,$$
$$S_{22} = \int_{t_0}^{t_1} \Big({Y_{12}}^\top(t)Q(t)Y_{12}(t) + {Y_{22}}^\top(t)B(t)R^{-1}(t)B^\top(t){Y_{22}}^\top(t)\Big)\, dt.$$

Matrices S_{11}, S_{22} are, obviously, nonnegative definite.

Assertion 2. *If the pair $(A(t), B(t))$ is completely controllable on $[t_0, t_1]$ then S_{22} is a positive definite matrix.*

P r o o f. Suppose to the contrary that there exists $p^0 \neq 0 : S_{22}p^0 = 0$. Let us take $x^0 = 0$ and $\alpha \in \mathbb{R}$. Denote

$$\bar p(t) = Y_{22}(t)p^0, \quad \bar u(t) = R^{-1}B^\top(t)\bar p(t), \quad \bar x(t) = Y_{12}(t)p^0.$$

Multiply p^0 on α, then $\alpha\bar u(t)$ satisfies (5.1), that is

$$\alpha^2 \int_{t_0}^{t_1} \bar u^\top(t)R(t)\bar u(t)dt \le \frac{\mu^2}{2}$$

for any α, this implies $\bar u(t) \equiv 0$, that is $B^\top(t)\bar p(t) \equiv 0$. To zero controller $\bar u(t)$ and state $x^0 = 0$ there corresponds the trajectory $\bar x(t) \equiv 0$. Hence, from the equations of adjoint system we get

$$\dot{\bar p}(t) = -A^\top(t)\bar p(t) + Q(t)\bar x(t) = -A^\top(t)\bar p(t).$$

Thus, $\bar p(t)$ is a nonzero solution of the adjoint homogenous system such that $\bar p^\top(t)B(t) \equiv 0$. This contradicts to the controllability of the system.

In the last case the problem (3.1) may be written as follows

$$\begin{gathered} \bar c^\top p^0 \to \min \\ \bar D p^0 \le \bar d, \quad {p^0}^\top S_{22}p^0 + {x^0}^\top S_{12}p^0 \le \mu^2 - {x^0}^\top S_{11}x^0, \end{gathered} \tag{5.7}$$

where

$$\bar c^\top = c^\top Y_{12}(t_1), \quad \bar D = DY_{12}(t_1), \quad \bar d = d - DY_{11}(t_1)x^0.$$

This is also a problem of the above type with the linear constraints and one quadratic constraint. Solving this problem we can obtain an optimal control by the explicit formulas (5.5).

The reduction of the integral constraints to quadratic constraint (5.6) allows us to easily generalize the considered problem to the case, when x^0 is not fixed but belongs to some polyhedron.

6. Conclusion

We consider an optimal control problem for a linear system with integrally constrained control, with a linear terminal cost and with terminal constraints given by a set of linear inequalities. This problem is, in fact, the ill-posed problem because of nonuniqueness of optimal control, which always takes place if the end point of the optimal trajectory belongs to the interior of the reachable set of the control system. We propose here a simple numerical algorithm for solving the optimal control problem, which uses a known explicit description of the reachable sets for linear systems with integral quadratic constraints on control functions. The algorithm is based on the reduction of considered problem to the solution of a finite-dimensional convex programming problem in primal or dual variables. This method allows to avoid difficulties related to nonuniqueness of optimal control.

REFERENCES

1. **Anan'ev B.I.** Motion correction of a statistically uncertain system under communication constraints // Automation and Remote Control, 2010. Vol. 71, no. 3. P. 367–378.
2. **Antipin A.S., Khoroshilova E.V.** Linear programming and dynamics // Trudy Inst. Mat. i Mekh. UrO RAN, 2013. Vol. 19, no. 2. P. 7–25.
3. **Antipin A.S., Khoroshilova E.V.** Linear programming and dynamics // Ural Mathematical Journal, 2015. Vol. 1, no. 1. P. 3–19. DOI: http://dx.doi.org/10.15826/umj.2015.1.001
4. **Arutyunov A.V., Magaril-Il'yaev G.G. and Tikhomirov V.M.** Pontryagin maximum principle. Proof and applications. Moscow: Factorial press,(in Russian), 2006. 124 p.
5. **Dar'in A.N., Kurzhanski A.B.** Control under indeterminacy and double constraints // Differential Equations, 2003. Vol. 39, no. 11. P. 1554–1567.
6. **Goberna M.A.,Lopez M.A.**Linear semi-infinite programming theory: An updated survey //European Journal of Operational Research, 2002. Vol. 143, Issue 2. P. 390–405.
7. **Gusev M.I.** On optimal control problem for the bundle of trajectories of uncertain system // Lecture Notes in Computer Sciences. Springer,2010. Vol. 5910. P. 286–293.
8. **Krasovskii N.N.** Theory of Control of Motion. Moscow: Nauka, 1968. 476 p. [in Russian]
9. **Kurzhanski A.B.** Control and Observation under Conditions of Uncertainty. Moscow: Nauka, 1977. 392 p. [in Russian]
10. **Lee E.B., Marcus L.** Foundations of Optimal Control Theory. Jhon Willey 'S' Sons, Inc., 1967. 124 p.
11. **Martein L., Schaible S.** On solving a linear program with one quadratic constraint // Rivista di matematica per le scienze economiche e sociali, 1988. P. 75–90.
12. **Ushakov V.N.** Extremal strategies in differential games with integral constraints // J. of Applied Mathematics and Mechanics, 1972. Vol. 36, no. 1. P. 15–23.
13. **Ukhobotov V.I.** On a class of differential games with an integral constraint // J. of Applied Mathematics and Mechanics, 1977. Vol. 41, no. 5. P. 838–844.

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

ON PARAMETERIZED COMPLEXITY OF THE HITTING SET PROBLEM FOR AXIS–PARALLEL SQUARES INTERSECTING A STRAIGHT LINE

Daniel M. Khachay

N.N. Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences and
Ural Federal University, Ekaterinburg, Russia,
dmx@imm.uran.ru

Michael Yu. Khachay

N.N. Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences and
Ural Federal University, Ekaterinburg, Russia,
mkhachay@imm.uran.ru

Abstract: The Hitting Set Problem (HSP) is the well known extremal problem adopting research interest in the fields of combinatorial optimization, computational geometry, and statistical learning theory for decades. In the general setting, the problem is NP-hard and hardly approximable. Also, the HSP remains intractable even in very specific geometric settings, e.g. for axis-parallel rectangles intersecting a given straight line. Recently, for the special case of the problem, where all the rectangles are unit squares, a polynomial but very time consuming optimal algorithm was proposed. We improve this algorithm to decrease its complexity bound more than 100 degrees of magnitude. Also, we extend it to the more general case of the problem and show that the geometric HSP for axis-parallel (not necessarily unit) squares intersected by a line is polynomially solvable for any fixed range of squares to hit.

Key words: Hitting set problem, Dynamic programming, Computational geometry, Parameterized complexity.

Introduction

We consider the parameterized complexity of a geometric statement of the well-known Hitting Set Problem (HSP), engaging researchers in combinatorial optimization, computational geometry and statistical learning from early 1980-th.

To the best of our knowledge, HSP gains theoretical interest because it was the first intractable combinatorial optimization problem, whose approximation algorithms were dramatically improved [11] on the basis of Vapnik and Chervonenkis's [15] results in statistical learning theory. The development of randomized algorithms for HSP and related combinatorial problems defined on range spaces of finite VC-dimension, initiated by seminal papers [1] and [6] established a new field in modern computational geometry.

On the other hand, the concepts of hitting set and classifier ensemble, making decisions by some voting logic, seem to be related very closely. Consequently, approximation techniques developed for HSP and its dual Set Cover problem are closely related to the well-known boosting learning technique [14], especially in the context of the minimal committee problem looking for minimum VC-dimension correct majoritary classifier ensemble (see, e.g., [8–10]).

In addition, new efficient optimal and approximation algorithms for Hitting Set and Set Cover problems have a practical importance, e.g. in design of reliable wireless networks [13].

The Hitting Set Problem for Axis-Parallel Rectangles (HSP–APR) is a well-studied geometric setting of the HSP. This setting is also NP-hard [5] and remains intractable even for unit squares.

In papers [2,7], first polynomial time approximation schemes (PTAS) are proposed for axis-parallel squares. Paper [3] introduces 6-approximation polynomial time algorithm for the case of rectangles intersecting some axis-monotone curve. In [4], this particular case of HSP–APR is proved to be NP-hard even for a straight line and the first 4-approximation algorithm is constructed.

In this paper, we improve one of the recent results describing a polynomial time solvable subclass of this problem. Recently, Mudgal and Pandit [12] introduced an optimal polynomial time algorithm for the Hitting Set Problem for Axis Parallel Unit Squares Intersecting a given Straight Line (HSP–APUS–ISL). The theoretical importance of this result can hardly be overestimated, since almost all known geometric settings of the HSP, including extremely specific ones, are intractable. Unfortunately, this algorithm is impractical due to its incredibly high time consumption of $O(n^{145})$. In Section 2, we propose the improved version of the algorithm, whose complexity bound $O(n^{37})$ is still high but by more than 100 degrees of magnitude better. Further, in Section 3, we extend this algorithm on a case of squares of different sizes (HSP–APS–ISL) and show that this problem can be solved to optimal in polynomial time for any fixed range of square sizes.

1. Problem statement

We consider the following geometric setting of the well-known Hitting Set Problem, which is called the Hitting Set Problem for Axis-Parallel Squares Intersecting a Straight Line (HSP–APS–ISL) (see Fig. 1). In the Euclidean plane, a finite collection $S = \{Q_1, \ldots, Q_n\}$ of axis-parallel (closed) squares intersecting some straight line d is given. For the collection S, it is required to find a hitting set P^* of the minimum size, i.e.

$$P^* = \arg\min\{|P| \colon P \subset \mathbb{R}^2,\ P \cap Q_j \neq \varnothing, j = 1, \ldots, n\}.$$

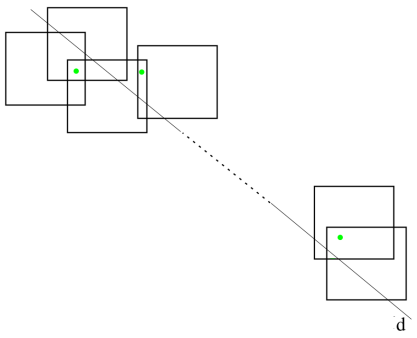


Figure 1. Problem statement

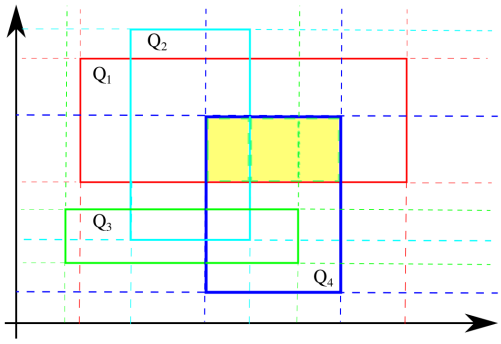


Figure 2. K does not exceed the number of rectangular cells induced by the lines defining borders of $Q_1, \ldots, Q_n$

Without loss of generality we assume that the line d is defined by the equation $kx + y = 0$ for some $k \geq 1$.

The collection S partitions the plane onto mutually disjunctive regions $\theta_1, \ldots, \theta_K$ such that, any points p_1 and p_2 belong to the same region θ_k if and only if

$$(\forall Q_j \in S)\ ((p_1 \in Q_j) \iff (p_2 \in Q_j)).$$

Since each minimal hitting set contains at most one point p_k taken from any region θ_k, the initial continuous problem is polynomially equivalent to the corresponding combinatorial one, which is of

finding a minimal hitting set among subsets of the finite set

$$\mathcal{P} = \{p_1, \ldots, p_K\}, \quad p_k \in \theta_k \setminus \bigcup_{l \neq k} \theta_l.$$

Indeed, for any collection of n axis-parallel squares (and even rectangles), the corresponding set $\mathcal{P}$ contains at most $O(n^2)$ elements (see Fig. 2) and can be constructed in polynomial time.

2. Improved algorithm for unit squares

In this section we describe parameterized optimal algorithm for HSP–APS–ISL and discuss its application to solving the special case of this problem, HSP–APUS–ISL, where collection S consists of equal squares (without loss of generality, which are assumed to be unit). We start with the similar (but not the same) notation to introduced in [12].

First, we partition the plane by straight lines $l_0, \ldots, l_{r+2}$ orthogonal to d with distance of $\sqrt{2}/2$ between each neighboring lines such that, for each square $Q_j \in S$, its center C_j is located between l_1 and l_{r+1} (hereinafter all tights are broken arbitrarily). For any $i = 0, \ldots, r+1$, we denote by R_i the stripe located between l_i and l_{i+1}. Next, we introduce the notation $S_i = \{Q_j \colon Q_j \cap R_i \neq \varnothing\}$, $S_i^{in} = \{Q_j \in S_i \colon C_j \in R_i\}$, and $S_i^{out} = S_i \setminus S_i^{in}$. By construction, $S_i^{out} \subset S_{i-1}^{in} \cup S_{i+1}^{in}$.

As in [12], we assume that any stripe R_i is intersected at least by a single square Q_j. Further, we find an optimal hitting set recursively, by the dynamic programming procedure presented in Algorithm 1.

Algorithm 1 Parameterized exact DP based algorithm

Input: a collection $S = \{Q_1, \ldots, Q_n\}$ of axis-parallel squares intersecting a given straight line d
Outer parameter: an upper bound q of the size of subsets to search for
Output: the minimum size hitting set P for S.

1: Construct a set $\mathcal{P}$ induced by the collection S; let $\mathcal{P}_i = \mathcal{P} \cap R_i$;
2: **for all** $U \subset \mathcal{P}_{r-1}$ and $V \subset \mathcal{P}_r$, s.t. $|U|, |V| \leq q$ **do**
3: define $\mathcal{W}_r = \{W \subset \mathcal{P}_{r+1} \colon |W| \leq q,\ U \cup V \cup W \cap Q_j \neq \varnothing\ (Q_j \in S_r)\}$ and

$$T(r, U, V) = \begin{cases} \min\{|U \cup V \cup W| : W \in \mathcal{W}_r\}, & \text{if } \mathcal{W}_r \neq \varnothing, \\ +\infty, & \text{otherwise} \end{cases}$$

4: **end for**
5: **for all** $1 \leq i \leq r-1$ **do**
6: **for all** $U \subset \mathcal{P}_{i-1}$ and $V \subset \mathcal{P}_i$, s.t. $|U|, |V| \leq q$ **do**
7: define $\mathcal{W}_i = \{W \subset \mathcal{P}_{i+1} \colon |W| \leq q, U \cup V \cup W \cap Q_j \neq \varnothing\ (Q_j \in \bigcup_{l \geq i} S_l)\}$ and

$$T(i, U, V) = \begin{cases} |U| + \min\{T(i+1, V, W) : W \in \mathcal{W}_i\}, & \text{if } \mathcal{W}_i \neq \varnothing, \\ +\infty, & \text{otherwise} \end{cases}$$

8: **end for**
9: **end for**
10: Output

$$\arg\min\{T(1, U, V) \colon U \subset \mathcal{P}_0, V \subset \mathcal{P}_1, |U|, |V| \leq q\}.$$

Indeed, for any $i \in 1, \ldots, r$, denote $\mathcal{P}_i = \mathcal{P} \cap R_i$. Let, for $U \subset \mathcal{P}_{i-1}$ and $V \subset \mathcal{P}_i$, $T(i, U, V)$ be the size of a smallest hitting set P for $\bigcup_{l \geq i} S_l$ such that $P \cap \mathcal{P}_{i-1} = U$ and $P \cap \mathcal{P}_i = V$. Similarly to [12], we express $T(i, U, V)$ in terms of $T(i+1, U', V')$ but for a substantially smaller subsets U' and V'.

Algorithm 1 has an outer parameter q, which meaning is twofold. On the first hand, q depends on size-length of the squares to hit and provides a uniform upper bound for the smallest size of a hitting set for an arbitrary S_i. On the other hand, q bounds the number of subset enumerated at each iteration of Algorithm 1. Therefore, its complexity bound can be defined in terms of q again.

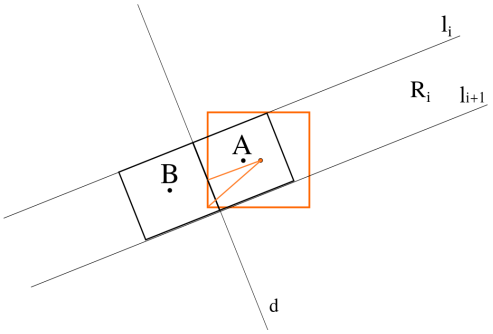


Figure 3. Any unit square $Q_j \in S_i^{in}$ is hitted by one of the centers A and B of $\sqrt{2}/2$-squares

The following Theorem summarizes the properties of Algorithm 1.

Theorem 1. *For $q = 6$, Algorithm 1 finds an optimal hitting set for the collection S in time of $O(n^{37})$.*

P r o o f. We start with the following simple fact. By construction, for any $i \in \{1, \ldots, r\}$ and any $j \in S_i^{in}$, $Q_j \cap \{A, B\} \neq \varnothing$ (see Fig. 3). As a consequence, for any optimal hitting set P and any $i \in \{1, \ldots, r\}$, $|P_i| \leq 6$, where $P_i = P \cap R_i$. Indeed, assume by contradiction that, for some i, $|P_i| > 6$. Since $S_i \subset S_{i-1}^{in} \cup S_i^{in} \cup S_{i+1}^{in}$ and $P_i \cap Q_j = \varnothing$ for any $Q_j \notin S_i$, we can substitute P_i by an appropriate 6-point subset P_i' such that $P \cup P_i' \setminus P_i$ remains a hitting set for S and $|P'| < |P|$. The contradiction obtained with optimality of P finalizes our argument. Hence, Algorithm 1 realizing classic dinamic programming technique finds an optimal hitting set for the given collection S.

Let us obtain an upper bound for its running time. Obviously, the loop 5-9 having $r - 1 = O(n)$ iterations is the most time consuming part of Algorithm 1. In each iteration, $O(|\mathcal{P}_{i-1}|^6) \times O(|\mathcal{P}_i|^6) = O(n^{24})$ subproblems each having time complexity of $O(n^{12})$ should be solved. Therefore, the overall running time is $O(n^{37})$. □

3. General case of HSP–APS–ISL

By scaling, we can easily show that the result of Section 2 remains valid in the case of equal squares of any side-length. In this section, we extend this result to the more general case. Let a and b be the minimum and the maximum values of side-lengths of the given squares. By the same reason, assume that $a = 1$.

3.1. Case of $k = 1$

We proceed with the following observation. For $k = 1$, as in Section 2, any square Q of size at least 1, whose center belongs to some stripe R' of width $\sqrt{2}/2$ orthogonal to the line d, is hitted by the points A and B (like in Fig. 3). Therefore, in this case, we can adapt Algorithm 1 to take into account the squares, whose side-lengths are greater then 1.

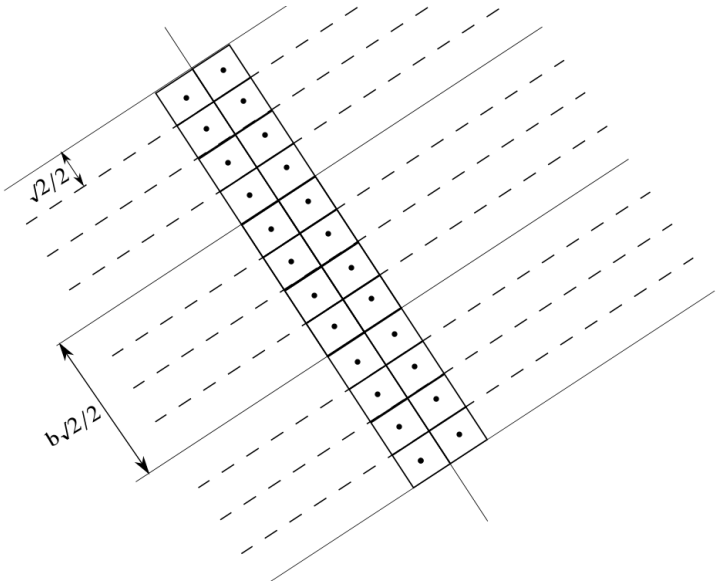


Figure 4. Partition of the plane for $b = 4$

Indeed, as above, consider stripes R_i of width $b\sqrt{2}/2$ consisting all the squares. Then, partition each of them onto $\lceil b \rceil$ substripes of width $\sqrt{2}/2$ (see Fig. 4) and use all other notation introduced in Section 2 as is. The following assertion is valid.

Theorem 2. *Let the given collection S consists of squares with side-lengths from $[1, b]$. Algorithm 1 with $q = 6\lceil b \rceil$ finds an optimal hitting set for this collection in time of $O(n^{6q+1}) = O(n^{36\lceil b \rceil+1})$.*

The argument proving Theorem 2 is similar to the proof of the Theorem 1. For the sake of brevity, we skip the proof.

3.2. What if $k > 1$

In this section, we show that to find an optimal solution for HSP–APS–ISL we can use Algorithm 1 again with an adjusted value of the parameter q. As above, this value is defined by the number of points needed to hit any square intersecting the line d, whose center belong to some stripe of the width $\sqrt{2}/2$. Although, for $k > 1$, points A and B (as in Fig. 3) do not hit all such squares, we can still provide a finite point collection that does.

Without loss of generality, assume that the strip R (of width $\sqrt{2}/2$)) orthogonal to the line d is located symmetrically with respect to the origin. An arbitrary square Q intersecting the line d, whose center C belongs to the stripe R is called *R-centered*.

Consider finite point sequences $\{A_t\}$ and $\{B_t\}$ defined by the following equation

$$A_t = -B_t = \left[\frac{k + 2t}{2\sqrt{2(1 + k^2)}}, \frac{1 - 2tk}{2\sqrt{2(1 + k^2)}} \right] \quad (t \in \{-1, \ldots, p\}). \tag{3.1}$$

Theorem 3. *For any $k > 1$, any R-centered square Q of size belonging to the range $[1, p\sqrt{2}]$ is hitted by the points $A_0, \ldots A_p, B_0, B_1, \ldots B_p$.*

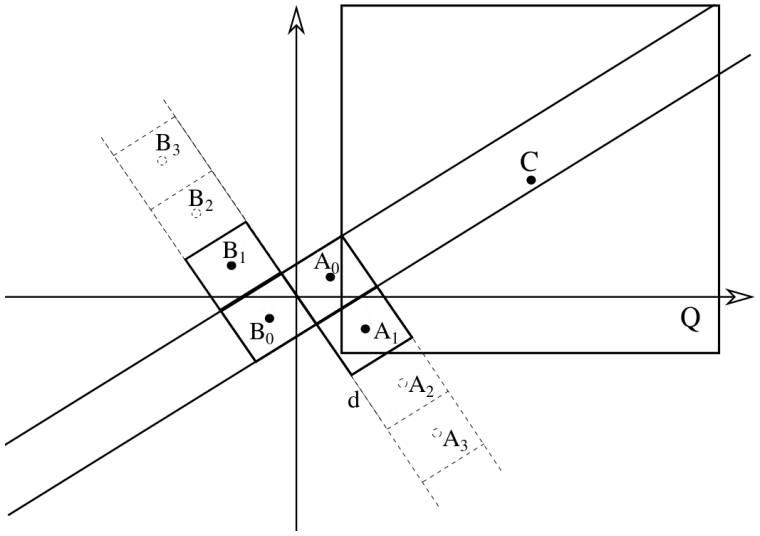


Figure 5. Hitting of large squares by the centers of neighboring $\sqrt{2}/2$-squares

P r o o f. 1. Consider an arbitrary R-centered square Q. Theorem 3 is evidently valid if the center C of this square belongs to one of $\sqrt{2}/2$-squares centered at A_0 or B_0. Consider the other option. Without loss of generality, assume that C belongs to right-upper part of the stripe R (as in Fig. 4). The square Q coincides with an intersection of four closed halfplanes bordering it from the left, top, right, and bottom sides. We denote them by H_L, H_T, H_R, and H_B, respectively. To proceed with the argument, it is sufficient to prove that there exists a point $A_t \in Q = H_L \cap H_T \cap H_R \cap H_B$.

The inclusion $A_t \in H_T$ is valid for any $t = 0, 1, \ldots, p$, since $y_{A_t} \leq y_C$ by the location assumption for the square Q. Furthermore, this assumption implies that A_{-1} can not be located to the right of the border of H_L. Suppose, $A_{t-1} \notin H_L$ and $A_i \in H_L$ for any $i \geq t$. Now, we show that A_t is the desired point hitting the square Q. Indeed, consider the intersection point D of the line d with the vertical line visiting the point A_{i-1}. Since

$$x_D = \frac{k + 2(t-1)}{2\sqrt{2(1+k^2)}}$$

and

$$kx_D + y_D = 0,$$

we obtain

$$y_{A_t} - y_D = \frac{1 - 2tk + k(k + 2(t-1))}{2\sqrt{2(1+k^2)}} = \frac{(k-1)^2}{2\sqrt{2(1+k^2)}} \geq 0.$$

Therefore, $A_t \in H_B$ (see Fig. 6).

Inclusion $A_t \in H_R$ follows easily from equation (3.1). Indeed, for any $k > 1$

$$x_{A_t} - x_{A_{t-1}} = \frac{1}{2\sqrt{2(1+k^2)}} < 1/2 \leq x_C - x_{A_{t-1}},$$

since a size of the square Q is at least 1. Thus, $A_t \in H_L \cap H_T \cap\ H_R \cap H_B = Q$.

2. To obtain the upper side-length bound of the fittable squares, it is sufficient to calculate the minimum side-length of the R-centered square touching the point A_p by its left side (Fig. 7). It

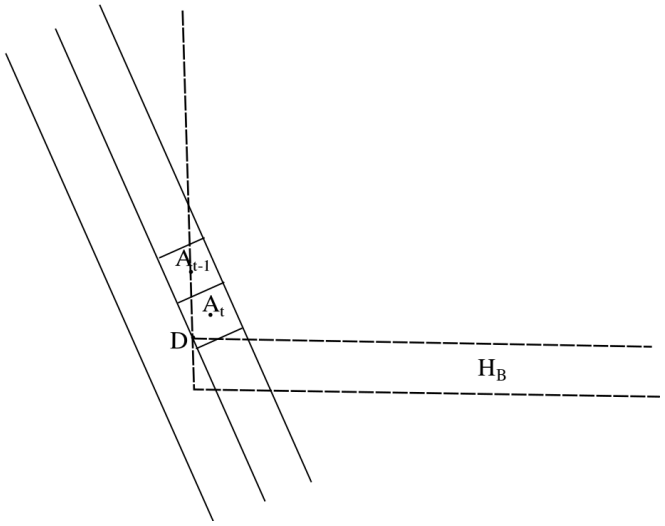


Figure 6. A_t belongs to H_B.

is easy to show that this length coincides with $s = 2(x_F - x_{A_p})$, where X_F can be found from the following system

$$\begin{cases} x_E = x_{A_p} = \dfrac{k+2p}{2\sqrt{2(1+k^2)}}, \\ kx_E + y_E = 0, \\ -x_E + y_E = z, \\ -x_F + ky_F = -\dfrac{\sqrt{1+k^2}}{2\sqrt{2}}, \\ -x_F + y_F = z, \end{cases}$$

i.e.

$$x_F = \frac{k^3 + 2pk^2 + 2pk - 1}{2\sqrt{2}(k-1)\sqrt{1+k^2}}$$

and

$$s = \frac{k^3 + 2pk^2 + 2pk - 1}{(k-1)\sqrt{2(1+k^2)}} - \frac{k+2p}{\sqrt{2(1+k^2)}} = \frac{\sqrt{2(1+k^2)}}{2} + \frac{p\sqrt{2(1+k^2)}}{k-1}.$$

To complete our proof, we should minimize $s = s(k)$ for $k > 1$.

The derivative

$$s'(k) = \frac{\sqrt{2}}{2}\frac{k(k-1)^2 - 2p(k+1)}{(k-1)^2\sqrt{1+k^2}}$$

is vanishing if and only if

$$k^3 - 2k^2 + k = 2p(k+1). \tag{3.2}$$

For $p = 0$, the function $s(k)$ has no minimizers in $(1, \infty)$. The right limit

$$\lim_{k \to +0} s(k) = \inf\{s(k) \colon k > 1\} = 1,$$

although $s(1) = +\infty$, as it follows from Subsection 3.1.

Given that $p \geq 1$, it is sufficient to consider a few cases. If $p = 1$ we have a single root (in the feasible domain $\{k\colon k > 1\}$) and it is easy to see that this root is a minimizer of $s(k)$, since $s'(k)$ changes its sign at this point. Further, it can be verified that, for any $p > 1$, we also have the unique extremal point.

Denote by $\bar{k} = \bar{k}(p)$ this extremum for the given p. Using equation (3.2), we obtain

$$s(\bar{k}) = \frac{\sqrt{2}(1+\bar{k}^2)^{3/2}}{2(1+\bar{k})}.$$

Therefore, since $\bar{k} > 1$,

$$\frac{s(\bar{k}(p))}{p} = \frac{\sqrt{2}(1+\bar{k}^2)^{3/2}}{\bar{k}(1-\bar{k})^2} \geq \frac{\sqrt{2}(3/2+\bar{k}^2)}{(\bar{k}-1)^2} > \sqrt{2}.$$

Theorem is proved. □

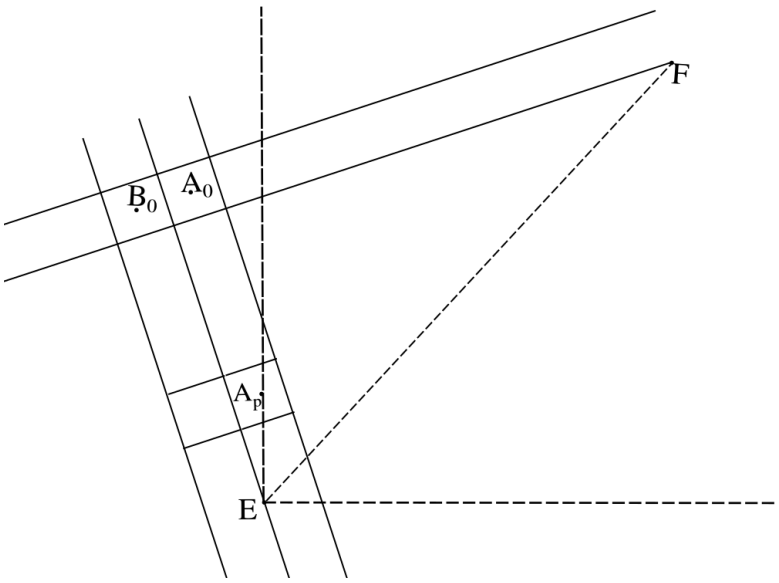


Figure 7. Estimation of $s(\bar{k})$.

Remark 1. *It is easily to verify that $\bar{k} = \bar{k}(p)$ is a monotonically increasing function and tends to $+\infty$ as $p \to +\infty$. Therefore,*

$$\lim_{p\to+\infty} \frac{s(\bar{k}(p))}{p} = \lim_{\bar{k}\to+\infty} \frac{\sqrt{2}(1+\bar{k}^2)^{3/2}}{2(1+\bar{k})} = \sqrt{2}.$$

Applying the approach proposed in Subsection 3.1, we obtain our final result. Indeed, let we should find the minimum hitting set for n squares intersecting the line d; sizes of the squares belong to $[a', b']$. First, by scaling, transform their sizes to the range $[1, b]$, where $b = b'/a'$.

Further, partition the plane onto d-orthogonal stripes of width $b\sqrt{2}/2$; we call these stripes *wide*. Finally, we partition each wide stripe onto $\lceil b \rceil$ $\sqrt{2}/2$-width *narrow* substripes.

By construction, any square intersecting a wide stripe is centered at this or two neighboring wide stripes. Therefore, by Theorem 3, it can be hit by $q = 6\lceil b \rceil + 2\lceil b/\sqrt{2} \rceil$, and the optimal hitting set can be found by Algorithm 1 using this value of q. Hence, we proved the following theorem.

Theorem 4. *For any constant c and any square collection with size-range $[a, c\grave{a}]$, the problem HSP–ASP–ISL can be solved to optimality in time $O(n^{6q+1})$, where $q = 6\lceil c\rceil + 2\lceil c/\sqrt{2}\rceil$.*

Remark 2. *Results of Theorem 2 and 3 shows that HSP–APS–ISL is polynomial solvable for any fixed range of squares, since the running time bound of Algorithm 1 in this case is*

$$O(n^{6(6\lceil c\rceil+2\lceil c/\sqrt{2}\rceil)+1}).$$

Unfortunately, the question of constructing for this problem an FPT algorithm having parameterized complexity bound like $f(c) \cdot n^{O(1)}$ still remains open.

4. Conclusion

In the paper, the improved version of the optimal polynomial time hitting set construction algorithm for axis-parallel squares intersecting the given straight line introduced in [12] is proposed. Our modification has better upper time complexity bound by 100 orders of magnitude.

Also, we propose an extension of this algorithm to the case of non-unit squares and show that the problem can be solved to optimality in polynomial time for any fixed range of squares.

As for the future work, it would be interesting to establish the complexity status of the considered problem in the case, where this parameter is unbounded. Also, it is interesting to answer the question, does the Hitting Set Problem for Axis-Parallel Squares belong to the class of Fixed Parameter Tractable (FPT) problems.

Acknowledgements

This research was supported by Russian Foundation for Basic Research, grant no. 16-07-00266 and Complex Program of Ural Branch of RAS, grant no. 15-7-1-23.

REFERENCES

1. **Brönnimann H. and Goodrich M. T.** Almost optimal set covers in finite vc-dimension // Discrete & Computational Geometry, 1995. Vol. 14, no. 4. P. 463–479. DOI: 10.1007/BF02570718
2. **Chan T. M.** Polynomial-time approximation schemes for packing and piercing fat objects // J. of Algorithms, 2003. Vol. 46, no. 2. P.178–189. DOI: 10.1016/S0196-6774(02)00294-8
3. **Chepoi V. and Felsner S.** Approximating hitting sets of axis-parallel rectangles intersecting a monotone curve. Computational Geometry, 2013. Vol. 46, no. 9. P. 1036–1041. DOI: 10.1016/j.comgeo.2013.05.008
4. **Correa J., Feuilloley L., Perez-Lantero P. and Soto J. A.**Independent and hitting sets of rectangles intersecting a diagonal line // Algorithms and complexity. Discrete & Computational Geometry, 2015. Vol. 53, no. 2. P. 344–365. DOI: 10.1007/s00454-014-9661-y
5. **Fowler R. J., Paterson M. S. and. Tanimoto S. L.** Optimal packing and covering in the plane are np-complete. Information Processing Letters, 1981. Vol. 12, no. 3. P. 133–137. DOI: 10.1016/0020-0190(81)90111-3
6. **Haussler D. and Welzl E.** Epsilon-nets and simplex range queries // Discrete & Computational Geometry, 1987. Vol. 2, no. 2. P. 127–151. DOI: 10.1007/BF02187876
7. **Hochbaum D.and Maass W.** Approximation schemes for covering and packing problems in image processing and vlsi // J. ACM, 1985. Vol. 32, no. 1. P. 130–136. DOI: 10.1145/2455.214106
8. **Khachay M.** Committee polyhedral separability: complexity and polynomial approximation // Machine Learning, 2015. Vol. 101, no. 1. P. 231–251. DOI: 10.1007/s10994-015-5505-0
9. **Khachay M. and Poberii M.** Complexity and approximability of committee polyhedral separability of sets in general position. Informatica, 2009. Vol. 20, no. 2. P. 217–234.

10. **Khachay M., Pobery M. and Khachay D.** Integer partition problem: Theoretical approach to improving accuracy of classifier ensembles // International J. of Artificial Intelligence, 2015. Vol. 13, no. 1. P. 135–146.
11. **Matoušek J.** Lectures on Discrete Geometry. Springer: New York, 2002. DOI: 10.1007/978-1-4613-0039-7
12. **Mudgal A. and Pandit S.** Covering, hitting, piercing and packing rectangles intersecting an inclined line // Proceedings of the Combinatorial Optimization and Applications: 9th International Conference, (COCOA 2015, Houston, TX, USA, December 18–20, 2015), Zaixin Lu, Donghyun Kim, Weili Wu, Wei Li, and Ding-Zhu Du (Ed.). LNCS, Springer International Publishing: Cham, 2015. Vol. 9486. P. 126–137. DOI: 10.1007/978-3-319-26626-8_10
13. **Ramakrishnan S. and Emary I. M. M. El.** Wireless sensor networks: from theory to applications. CRCPress, Taylor & Francis, 2014.
14. **Schapire R. and Freund Y.** Boosting: Foundations and algorithms. MIT Press, 2012.
15. **Vapnik V. and Chervonenkis A.** On the uniform convergence of relative frequencies of events to their probabilities // Theory Probab. Appl., 1971. Vol. 16. P. 264–280. DOI: 10.1137/1116025

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

ON SOME NUMERICAL INTEGRATION CURVES FOR PDE IN NEIGHBORHOOD OF "BUTTERFLY" CATASTROPHE POINT[1]

Oleg Y. Khachay

Ural Federal University, Ekaterinburg, Russia,
khachay@yandex.ru

Pavel A. Nosov

Ural Federal University, Ekaterinburg, Russia,
pavel.nosov@urfu.ru

Abstract: We consider a three-dimensional nonlinear wave equation with the source term smoothly changing over time and space due to a small parameter. The behavior of solutions of this PDE near the typical "butterfly" catastrophe point is studied. In the framework of matched asymptotic expansions method we derive a nonlinear ODE of the second order depending on three parameters to search for the special solution describing the rapid restructuring of the solution of the PDE in a small neighborhood of the catastrophe point, matching with expansion in a more outer layer. Numerical integration curves of the equation for the leading term of the inner asymptotic expansion are obtained.

Key words: Matched asymptotic expansions, Numerical integration, Butterfly catastrophe, Nonlinear ODE and PDE.

Introduction

This paper is devoted to the study of specific behavior of a solution of the nonlinear wave PDE

$$-\frac{\partial^2 U}{\partial^2 T}+\frac{\partial^2 U}{\partial X^2}+\frac{\partial^2 U}{\partial Y^2}+\frac{\partial^2 U}{\partial Z^2}=f(\varepsilon T,\varepsilon X,\varepsilon Y,\varepsilon Z,U) \tag{0.1}$$

as a smoothed shock wave, the so called step-like contrast structure. Since this equation (0.1) contains 4 independent variables in every open domain of the arguments of function in the right-hand side of this equation there typically exists a point of catastrophe related to degeneration of f up to 5-th order with respect to the unknown function, that corresponds to "butterfly" type catastrophe [2].

The purpose of this paper is by applying the matched asymptotic expansions method [3] and catastrophe theory [2] to deduce the nonlinear ODE of the second order, which depends on three parameters, the ODE, which would be satisfied by the special solution related to a step-like contrast structure. We explore also the variants of such special solutions behavior depending on the parameter settings.

The similar equation with two independent variables and the corresponding typical point of "cusp" catastrophe was considered in [12]. The detailed study of special solutions with obtaining a uniform asymptotic expansion was carried out in [5], [6], this paper mainly follows the approach taken in these works.

[1]This work was supported by RFBR, research project No 16–31–00222.

1. Preliminary constructions

We consider first a more general than in equation (0.1) form of the differential operator in the left-hand side of PDE the resultant ODE in some cases will have the first order while in others it will be the second order. We suppose to study the constructed ODE in our following works in the framework of RFBR Research Project mentioned above.

Consider a nonlinear PDE of such form

$$\sum_{|\boldsymbol{\alpha}|=1}^{2} A_{\boldsymbol{\alpha}}(\varepsilon \boldsymbol{R}) \partial^{\boldsymbol{\alpha}} U(\varepsilon, \boldsymbol{R}) = f(\varepsilon \boldsymbol{R}, U(\varepsilon, \boldsymbol{R})). \tag{1.1}$$

Following [13, p. 15] we denote by $\partial^{\boldsymbol{\alpha}}$ the operator of differentiation with respect to independent variables and assume that the multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)$ corresponds to the independent variables, in particular, $(R_1, \ldots, R_4) = \boldsymbol{R}$ in equality (1.1).

The relation (0.1) is a particular case of the equation (1.1), which corresponds to the following values of coefficients: $A_{\boldsymbol{\alpha}} = 0$ for $0 \le \alpha_n < 2, n = 1, \ldots, 4$ and $1 = -A_{2,0,0,0} = A_{0,2,0,0} = A_{0,0,2,0} = A_{0,0,0,2}$.

In order to reduce the equation (1.1) to the standard form of singular equations with a small parameter multiplying the derivative, we make the change of variables $\boldsymbol{s} = \varepsilon \boldsymbol{R}$, $V(\varepsilon, \boldsymbol{s}) = U(\varepsilon, \varepsilon \boldsymbol{R})$, resulting equation (1.1) takes the form

$$\sum_{|\boldsymbol{\alpha}|=1}^{2} \varepsilon^{|\boldsymbol{\alpha}|} A_{\boldsymbol{\alpha}}(\boldsymbol{s}) \partial^{\boldsymbol{\alpha}} V(\varepsilon, \boldsymbol{s}) = f(\boldsymbol{s}, V(\varepsilon, \boldsymbol{s})). \tag{1.2}$$

Let the function $f(\boldsymbol{s}, V)$ be smooth and satisfy the inequality $f_V(\boldsymbol{s}, V) \neq 0$, $(\boldsymbol{s}, V) \in \Omega_{\boldsymbol{s}} \times \Omega_V$. Let some conditions relevant to the equation (1.2) specify a solution on the domain $\Omega_{\boldsymbol{s}}$ asymptotically approximated by a series of the following form

$$V(\varepsilon, \boldsymbol{s}) = V_0(\boldsymbol{s}) + \sum_{n=1}^{\infty} \varepsilon^n V_n(\boldsymbol{s}), \quad \varepsilon \to 0, \tag{1.3}$$

on the domain $\Omega_{\boldsymbol{s}}$ except for a neighborhood of its borders, and let the principal term $V_0(\boldsymbol{s})$ of the series (1.3) satisfy the equation

$$f(\boldsymbol{s}, V_0(\boldsymbol{s})) = 0. \tag{1.4}$$

Examples of such statements of problems are presented, for instance, in the monograph [11] and in the paper [5].

At the boundary of $\Omega_{\boldsymbol{s}}$ could be located points of degeneracy of the function $f(\boldsymbol{s}, V)$ with respect to the unknown function, i.e., zeros of the function $f_V(\boldsymbol{s}, V)$ on the boundary of the set $\Omega_{\boldsymbol{s}}$, which may contain the manifolds of points of "fold" catastrophe and "cusp" catastrophe, smooth lines of "swallowtail" catastrophe and isolated points of "butterfly" catastrophe [2].

2. Derivation of special equation

Let a point $(\boldsymbol{s}^\circ, V^\circ)$ on the boundary of the domain $\Omega_{\boldsymbol{s}} \times \Omega_V$ be an isolated point of "butterfly" $(A_{\pm 5})$ type catastrophe [2, p. 5,11] of the function $f(\boldsymbol{s}, V)$. Then Taylor asymptotic expansion of f in the neighborhood of this point is

$$f(\boldsymbol{s}, V) = \sum_{m=0}^{\infty} \left(b_m + G_m(\boldsymbol{s}) + \widetilde{G_m}(\boldsymbol{s}) \right) (V - V^\circ)^m,$$

where

$$G_m(\boldsymbol{s}) = \sum_{j=1}^{4} b_{m,j}(s_j - s_j^\circ),$$

$$\widetilde{G_m}(\boldsymbol{s}) = \sum_{|\boldsymbol{\alpha}|=2}^{\infty} b_{m,\boldsymbol{\alpha}}(\boldsymbol{s} - \boldsymbol{s}^\circ)^{\boldsymbol{\alpha}},$$

$$0 = b_0 = b_1 = b_2 = b_3 = b_4, \tag{2.1}$$

$$b_5 = \varkappa \neq 0. \tag{2.2}$$

We introduce the following notation

$$\boldsymbol{B} = \left(b_{n-1,j}\right)_{n,j=1,\ldots,4}.$$

The vanishing of the five coefficients (2.1) is achieved by the choice of the values of the five coordinates of a point $(\boldsymbol{s}^\circ, V^\circ)$; the existence of a solution of a system of five equations

$$\frac{\partial^n}{\partial V^n} f(\boldsymbol{s}, V) = 0, \quad n = 0, \ldots, 4$$

with respect to five unknowns is a typical situation, while satisfying any additional independent relations would not be the typical for functions $f(\boldsymbol{s}, V)$, and this, in particular, validates the assumption (2.2) and the inequality

$$\det \boldsymbol{B} \neq 0. \tag{2.3}$$

Performing the change of variables

$$\sigma_n = -G_{n-1}(\boldsymbol{s}),\ n = 1, \ldots, 4;\ \boldsymbol{\sigma} = \boldsymbol{B}(\boldsymbol{s} - \boldsymbol{s}^\circ); \quad V(\varepsilon, \boldsymbol{s}) = V(\varepsilon, \boldsymbol{s}^\circ + \boldsymbol{B}^{-1}\boldsymbol{\sigma}) = \mathcal{W}(\varepsilon, \boldsymbol{\sigma}) \tag{2.4}$$

and using the approach of catastrophe theory [2, p. 37,38,43], we find that there is a diffeomorphism of a neighborhood of $(\boldsymbol{s}^\circ, V^\circ)$ onto a neighborhood of the origin of variables $(\boldsymbol{\sigma}, \mathcal{W})$ satisfying the asymptotic relation

$$V - V^\circ = H_0(\boldsymbol{\sigma}) + \mathcal{W}(1 + H_1(\boldsymbol{\sigma})) + \sum_{m=2}^{\infty} \mathcal{W}^m (C_m + H_m(\boldsymbol{\sigma})), \tag{2.5}$$

$$H_m(\boldsymbol{\sigma}) = \sum_{|\boldsymbol{\alpha}|=1}^{\infty} C_{m,\boldsymbol{\alpha}} \boldsymbol{\sigma}^{\boldsymbol{\alpha}}$$

with a set of numerical coefficients $C_m, C_{m,\boldsymbol{\alpha}}$. Additionally, in the neighborhood of $(\boldsymbol{\sigma}^\circ, V^\circ)$ this diffeomorphism satisfies the identity

$$f(\boldsymbol{\sigma}, V) \equiv -\varphi_1 - \varphi_2 \mathcal{W} - \varphi_3 \mathcal{W}^2 - \varphi_4 \mathcal{W}^3 + \varkappa \mathcal{W}^5, \tag{2.6}$$

where the number $\varkappa$ is the same as in (2.2), and the coefficients $\varphi_n = \varphi_n(\boldsymbol{\sigma})$ have the asymptotic representation

$$\varphi_n(\boldsymbol{\sigma}) = \sigma_n + \sum_{|\boldsymbol{\alpha}|=2}^{\infty} c_{n,\boldsymbol{\alpha}} \boldsymbol{\sigma}^{\boldsymbol{\alpha}}. \tag{2.7}$$

To study in details the behavior of the solution near the point of catastrophe $\boldsymbol{s}^\circ$ corresponding to the value $\boldsymbol{\sigma} = 0$ we perform the coordinate stretching

$$\sigma_n = \varepsilon^{\beta_n} S_n,\ \beta_n > 0,\ n = 1, \ldots, 4; \quad \mathcal{W}(\varepsilon, \boldsymbol{\sigma}) = \varepsilon^{\gamma} W(\varepsilon, \boldsymbol{S}),\ \gamma > 0 \tag{2.8}$$

with some exponents $(\beta_1, \ldots, \beta_4) = \boldsymbol{\beta}$ and γ, the exact values of which will be defined below.

Since the leading term $\mathcal{W}$ of the series (2.5) has to satisfy the limiting equation (1.4) with expression (2.6) substituted into it instead of function f, then it turns out, that it depends on the all $\varphi_n(\boldsymbol{\sigma})$, $n = 1, \ldots, 4$, and hence, by virtue of the asymptotics (2.7), it is dependent on all σ_n, $n = 1, \ldots, 4$. To achieve this effect, when making the change of variables (2.8), we need to balance all the exponents of powers of ε arising from the terms in equation (1.4). Setting

$$\beta_n = (n+1)\gamma_n, \;\; n = 1, \ldots, 4, \tag{2.9}$$

we obtain the asymptotic approximation

$$f(\boldsymbol{s}, V(\varepsilon, \boldsymbol{s})) = \varepsilon^{5\gamma}\left(-S_1 - S_2 W - S_3 W^2 - S_4 W^3 + \varkappa W^5\right) + \sum_{m=6}^{\infty} \varepsilon^{m\gamma} P_m(\boldsymbol{S}, W), \;\; \varepsilon \to 0,$$

where $W = W(\varepsilon, \boldsymbol{S})$. Below, we will analyze the behavior of the principal term of the asymptotic expansion

$$W(\varepsilon, \boldsymbol{S}) = w(\boldsymbol{S}) + \sum_{n=1}^{\infty} \varepsilon^n w_n(\boldsymbol{S}), \quad \varepsilon \to 0. \tag{2.10}$$

To derive the principal term of the asymptotics of the left-hand side part of the equation (1.2) we transform the operators of differentiation with the use of changes of variables written in equalities (2.4), (2.8):

$$\varepsilon^{\beta_n} dS_n = d\sigma_n = \sum_{j=1}^{4} b_{n-1,j}\, ds_j.$$

To simplify the calculations we denote $\varepsilon^{\gamma} W(\varepsilon, \boldsymbol{S}) = \widetilde{\mathcal{W}}(\varepsilon, \boldsymbol{S})$, then, as

$$\sum_{n=1}^{4} \frac{\partial}{\partial S_n}(\widetilde{\mathcal{W}}(\varepsilon, \boldsymbol{S}))\, dS_n = d\widetilde{\mathcal{W}}(\varepsilon, \boldsymbol{S})) = dV(\varepsilon, \boldsymbol{s}) = \sum_{j=1}^{4} \frac{\partial}{\partial s_j} V(\varepsilon, \boldsymbol{s})\, ds_j,$$

we have

$$\sum_{n=1}^{4} \varepsilon^{-\beta_n} \frac{\partial}{\partial S_n}(\widetilde{\mathcal{W}}(\varepsilon, \boldsymbol{S})) \sum_{j=1}^{4} b_{n-1,j}\, ds_j = \sum_{j=1}^{4} \frac{\partial}{\partial s_j} V(\varepsilon, \boldsymbol{s})\, ds_j,$$

whence

$$\frac{\partial}{\partial s_j} = \sum_{n=1}^{4} \varepsilon^{-\beta_n} b_{n-1,j} \frac{\partial}{\partial S_n},$$

$$\sum_{|\boldsymbol{\alpha}|=1}^{2} \varepsilon^{|\boldsymbol{\alpha}|} A_{\boldsymbol{\alpha}}(\boldsymbol{s}) \partial^{\boldsymbol{\alpha}} V(\varepsilon, \boldsymbol{s}) = \varepsilon^{\gamma} \sum_{|\boldsymbol{\alpha}|=1}^{2} \varepsilon^{|\boldsymbol{\alpha}|} \widehat{A}_{\boldsymbol{\alpha}}(\boldsymbol{S}) \prod_{j=1}^{4} \left(\sum_{n=1}^{4} \varepsilon^{-\beta_n} b_{n-1,j} \frac{\partial}{\partial S_n} \right)^{\alpha_j} W(\varepsilon, \boldsymbol{S}),$$

where

$$\widehat{A}_{\boldsymbol{\alpha}}(\boldsymbol{S}) = A_{\boldsymbol{\alpha}}(\boldsymbol{s}^{\circ} + \boldsymbol{B}^{-1}\boldsymbol{\sigma}(\boldsymbol{S})), \quad \boldsymbol{\sigma}(\boldsymbol{S}) = (\varepsilon^{\beta_n} S_n)_{n=1,\ldots,4}\,.$$

Consequently, under the condition of smoothness of functions $A_{\boldsymbol{\alpha}}(\boldsymbol{s})$ and taking into account (2.8)–(2.10) at a fixed value of $\boldsymbol{S}$ and $\varepsilon \to 0$ the equation (1.2) can be written as

$$\varepsilon^{1-\beta_1+\gamma}\left(M \frac{\partial}{\partial S_1} W(\varepsilon, \boldsymbol{S}) + O(\varepsilon^{\gamma})\right) + \varepsilon^{2-2\beta_1+\gamma}\left(N \frac{\partial^2}{\partial (S_1)^2} W(\varepsilon, \boldsymbol{S}) + O(\varepsilon^{\gamma})\right) = \varepsilon^{5\gamma}\left(-S_1 - S_2 w - S_3 w^2 - S_4 w^3 + \varkappa w^5 + O(\varepsilon^{\gamma})\right), \tag{2.11}$$

if we collect on the left-hand side in one bracket only the terms with derivatives of the first order and in another bracket – only with second-order derivatives; here

$$M = \sum_{|\boldsymbol{\alpha}|=1} Q(\boldsymbol{\alpha}), \quad N = \sum_{|\boldsymbol{\alpha}|=2} Q(\boldsymbol{\alpha}),$$

$$Q(\boldsymbol{\alpha}) = A_{\boldsymbol{\alpha}}(\boldsymbol{s}^{\circ})\boldsymbol{b}^{\boldsymbol{\alpha}}, \quad \boldsymbol{b} = (b_{0,1}, \ldots, b_{0,4}).$$

In particular, the constants $M = 0$ and $N = -(b_{0,1})^2 + (b_{0,2})^2 + (b_{0,3})^2 + (b_{0,4})^2$ correspond to the equation (0.1) and the constants $M = -b_{0,1}$ and $N = (b_{0,2})^2 + (b_{0,3})^2 + (b_{0,4})^2$ are matched with the diffusion equation

$$-\frac{\partial U}{\partial T} + \frac{\partial^2 U}{\partial X^2} + \frac{\partial^2 U}{\partial Y^2} + \frac{\partial^2 U}{\partial Z^2} = f(\varepsilon T, \varepsilon X, \varepsilon Y, \varepsilon Z, U).$$

In accordance with the practice of matched asymptotic expansions method [3], when transiting to an internal scale, we need to choose γ such, that the exponents of powers of ε at the main terms of the asymptotics of the left-hand and right-hand sides of the equation coincide with each other after the transition to the new variables.

Let us consider the following two situations separately:

1) when the constant $M \neq 0$;

2) when in the original equation (1.1) all the coefficients of the first derivatives are identical to zero and the constant $N \neq 0$.

Note that implementation of these inequalities for the constant M and N, the value of which, except for special cases (for example, the lack of the first order in the original equation (1.1)), depends on the choice of f on the right-hand part of the original equation, is a typical situation similar to inequality (2.3), as noted above.

In the first case, by virtue of (2.9) and (2.11) we come to the conclusion, that γ has to satisfy the relation

$$\min\{1 - 4\gamma, 2 - 9\gamma\} = 5\gamma.$$

Solving it and taking into account the inequality $\gamma > 0$, we obtain the value $\gamma = 1/9$. Substituting it into the equality (2.11), we receive the estimate

$$M\frac{\partial}{\partial S_1} w(\boldsymbol{S}) + S_1 + S_2 w(\boldsymbol{S}) + S_3 w^2(\boldsymbol{S}) + S_4 w^3(\boldsymbol{S}) - \varkappa w^5(\boldsymbol{S}) = O(\varepsilon^{1/9}). \tag{2.12}$$

The limiting equation to 2.14 is an ODE of the first order with respect to $w(\boldsymbol{S})$ as a function of one variable S_1 and three parameters S_2, S_3, S_4. By making in the ODE the linear change of variables

$$S_1 = (M^5\varkappa^{-1})^{1/9}\, x, \quad S_2 = (M^4\varkappa)^{1/9}\, y, \quad S_3 = (M\varkappa)^{1/3}\, z, \quad S_4 = (M^2\varkappa^5)^{1/9}\, t,$$
$$u(x) = (M\varkappa^{-2})^{1/9}\, w(S_1(x), S_2(y), S_3(z), S_4(t)),$$

we obtain the first order nonlinear ODE

$$u_x = u^5 - tu^3 - zu^2 - yu - x, \tag{2.13}$$

which depends on three parameters y, z, t. We plan the study of the behavior of solutions of ODE (2.13) to be hold in subsequent papers in the framework of the above mentioned RFBR research project.

In the second case, in the relation (2.11) the first term of the equation is completely missing and the equation for γ becomes simpler $2-9\gamma=5\gamma$, whence $\gamma=1/7$ and therefore equality (2.11) takes the form:

$$N\frac{\partial^2}{\partial(S_1)^2}w(\boldsymbol{S})+S_1+S_2w(\boldsymbol{S})+S_3w^2(\boldsymbol{S})+S_4w^3(\boldsymbol{S})-\varkappa w^5(\boldsymbol{S})=O(\varepsilon^{1/7}). \tag{2.14}$$

We turn to the equation obtained from (2.14) as a result of passing to the limit as $\varepsilon\to 0$ and at the same time producing a linear change of variables

$$\begin{gathered}S_1=-\operatorname{sgn}(\varkappa)\left|N^5\varkappa^{-1}\right|^{1/14}x,\quad S_2=-\operatorname{sgn}(\varkappa)\left|N^2\varkappa\right|^{1/7}y,\\ S_3=-\operatorname{sgn}(\varkappa)\left|N^3\varkappa^5\right|^{1/14}z,\quad S_4=-\operatorname{sgn}(\varkappa)\left|N\varkappa^4\right|^{1/7}t,\\ u(x)=\left|N\varkappa^{-3}\right|^{1/14}w(S_1(x),S_2(y),S_3(z),S_4(t)).\end{gathered}$$

Thus, we obtain the desired nonlinear second-order ODE

$$\operatorname{sgn}(N)\,u_{xx}=u^5-tu^3-zu^2-yu-x, \tag{2.15}$$

depending on three parameters y,z,t. This ODE is held for a special solution describing the rapid reconstruction of the original PDE solution in a small neighborhood of the catastrophe point $\boldsymbol{s}^\circ$.

3. Matching with more outer layer condition

Consider the multi-valued relation

$$u=h(x;y,z,t), \tag{3.1}$$

each value of which is a root of the equation

$$0=u^5-tu^3-zu^2-yu-x.$$

In order to apply the matched asymptotic expansions method, it is necessary to construct a function $u(x)=u(x;y,z,t)$, which is a solution of the equation (2.15), i.e., it is the principal term of the asymptotic expansion of the solution of the original problem for PDE in the inner layer (the layer, projection of which onto the axis Ox is the set $|x|<\varepsilon^{-2/7+\delta_1}$) in such way that it would be matched with the solution of the original problem in the more outer layer (projection of which onto the axis Ox is the set $|x|>\varepsilon^{-2/7+\delta_2}$), where $0<\delta_1<\delta_2<2/7$ are some numbers. Therefore it is necessary to match $u(x;y,z,t)$ and $V_0(\boldsymbol{s})$, the leading term of the series (1.3), which satisfy the equation (1.4). As a consequence of that we will look for those solutions $u(x)=u(x;y,z,t)$ of equation (2.15), which satisfies the limiting relation

$$\lim_{x\to-\infty}|u(x;y,z,t)+H(x;y,z,t)|=0, \tag{3.2}$$

where $u=H(x)=H(x;y,z,t)$ stands for the maximal extension over axis Ox in the right-hand direction of a smooth branch of the root (3.1) defined in a neighborhood of $x=-\infty$. Thus, we will configure the function $u(x;y,z,t)$ to match with the $V_0(\boldsymbol{s})$ on the left-hand side along axis Ox and then we will study the behavior of such a curve with an increase in the variable x.

4. Numerical search for special integral curves

Integral curves of solutions given in this paper shown in bold and green on the figures below were calculated using explicit Runge–Kutta–Felberg (4,5) method [1] with variable step and accuracy control. In all figures a red thick line represents the graph of the root (3.1).

Note that the behavior of solutions of equations (2.15) significantly depends on the $\operatorname{sgn}(\varkappa)$ in the left-hand side. For minus sign it is typical to appear rapid fluctuations of solutions after passing the branching point of a multi-valued relation (3.1). Figure 1 illustrates the occurrence of such fluctuations of the integral curve of the equation

$$-u_{xx} = u^5 + 9u^3 - 8u - x, \tag{4.1}$$

computed under condition (4.6). In this work we do not perform a detailed study of such solutions

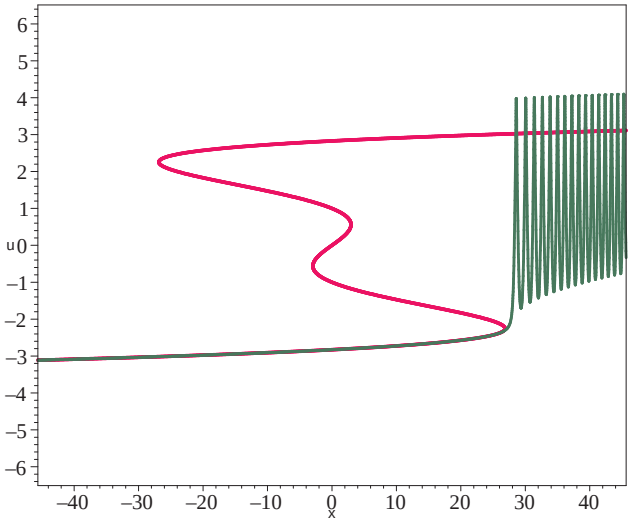


Figure 1. Integral curve for equation (4.1) under condition (4.6) and graph of root line (3.1)

with fluctuations; we focus on so-called step-like contrast structures. Therefore, we will consider only the equation with plus sign:

$$u_{xx} = u^5 - tu^3 - zu^2 - yu - x. \tag{4.2}$$

The purpose of construction and further study of the behavior of the integral curves was to test the hypothesis about the possibility of proving the existence of such special solutions within the framework of the approach of [5], [6]. Therefore, at this stage, we have considered the case when the parameter z in the right-hand side of equation (4.2) is zero, since this option, on the one hand, includes the very point of "butterfly" catastrophe (at the origin of all variables and parameters). On the other hand, it is very close to "cusp" catastrophe type of discussed in these papers, since in this case, the root line (3.1) is odd (symmetric with respect to the origin), and therefore the solution can be constructed as the extension based on oddness onto the negative half-axis ($x < 0$) of a solution $u(x)$ of the Cauchy problem for the equation

$$u_{xx} = u^5 - tu^3 - yu - x \tag{4.3}$$

with the initial conditions

$$u(0) = 0, \tag{4.4}$$

$$u_x(0) = \alpha. \tag{4.5}$$

In this case, the limiting relation (3.2) takes the form

$$\lim_{x\to+\infty} |u(x;y,0,t) + H(x;y,0,t)| = 0. \tag{4.6}$$

Thus, below, we consider integral curves only for the Cauchy problem (4.3)–(4.5). As a problem of obtaining the principal term of the internal expansion of the original problem for PDE it presents the extension of study conducted by one of the authors of this paper of bisingular initial problems with one small parameter for the systems of one, two or more of ODEs, which also has the property of degeneration of high-order of right-hand part of the equation in respect to the unknown function [4, 7–10].

According to the theorem proved in [6], one can find the initial value of α_0 of the derivative $u_x(x)$ corresponding to the solution that satisfies the asymptotics (4.6) by calculating the exact lower boundary for a certain set of M. The numerical set M depending on the parameters of the differential equation was determined in [6] as follows: the number of α belong to the set M, if and only if the solution of the Cauchy $u_\alpha(x)$ corresponding to initial condition (4.5) with this value of α has the property, that there exists a point x_α such that

- for $0 < x < x_\alpha$ function $u_\alpha(x)$ is less than the function $H(x)$;
- at the point x_α functions u_α and $H(x)$ coincide.

It is not difficult to see that the most top branch of the root (3.1) defined by the formula $u = H(-x)$, to which the desired solution has to approach (see, eg., Fig. 4.a), consists by virtue of the equation (4.3) (except, perhaps, for its leftmost point) of points of repulsion, which damps on approaching the line $u = H(-x)$ and rapidly increases with increasing the distance from this line in any vertical direction. Therefore, most of the curves produced with the initial conditions (4.4), (4.5), either pass through a line of $u = H(-x)$, and quickly grow up, tending to $+\infty$, or, if the initial velocity (4.5) is not sufficient, can not come close to this line and are broken down to $-\infty$. Search for the fine line of the balance between these two states of computed integral curves is consistent with the idea of the works [5], [6] of the initial value of $\alpha_0 = \inf M$ in (4.5).

We have implemented a simple binary search algorithm of this balance: in the transition from the range $[a_m, b_m]$ of possible values $u_x(0)$ at the current step to the next step range $[a_{m+1}, b_{m+1}]$ we always choose one of the intervals $[a_m, c_m]$, $[c_m, b_m]$ (where c_m is the middle of the segment $[a_m, b_m]$), for which the integral curve started with an initial rate of a_{m+1} breaks down to $-\infty$, and one with initial rate of b_{m+1} grows up to $+\infty$. Clearly, the value b_m pretend to be an element α of the set M described above and the limit of the decreasing sequence b_m could play the role of $\alpha_0 = \inf M$.

5. Illustrations of section at $z = 0$ of separatrix of "butterfly" catastrophe ($A_{\pm5}$)

This section of the paper is complementary, although it still has an indirect relationship to the purpose of this paper. Here, for the first time, as far as we know, the three-dimensional illustration of the cross section for $z = 0$ separatrix [2, p. 51] "butterfly" catastrophe ($A_{\pm5}$) are given. Separatrix corresponding to such a point of catastrophe is the set of points $(x^\circ, y^\circ, z^\circ, t^\circ)$ included into the parameter space of the parametric family of functions

$$\psi(u;x,y,z,t) = u^5 - tu^3 - zu^2 - yu - x, \tag{5.1}$$

with the following property: when such a point $(x^\circ, y^\circ, z^\circ, t^\circ)$ is substituted into (5.1) the resultant function $\psi^\circ(u)$ has at least one non-Morse point, i.e., the following relation holds

$$\exists u^\circ\colon \big(\psi^\circ(u^\circ) = 0, \frac{\partial\psi^\circ}{\partial u}(u^\circ) = 0\big).$$

The separatrix [2, p. 57–58] divides the parameter space into open domains corresponding to which subfamilies of the general family of functions (5.1) are structurally stable.

To construct a three-dimensional surface data we implemented the approach applied in [2, p. 62–63] to the function of "swallowtail" catastrophe (A_4). Knowledge about the configuration of the

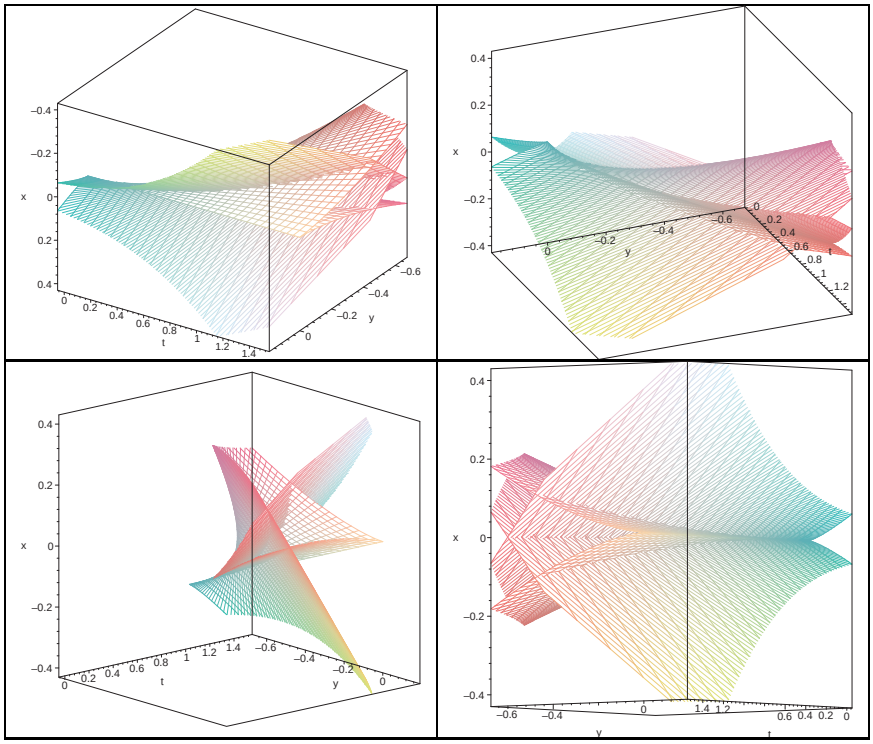


Figure 2. Section at $z = 0$ of separatrix of family of functions (5.1)

section of the separatrix allows us to study not all possible options for the behavior of curves at $z = 0$ through taking only one representative from each of the open domains of the parameters (x, y, t) corresponding to Morse functions from the family (5.1) and additional representatives from the border separatrix itself, detailed analysis of which can be done by ranking the degree of degeneracy of the function (5.1) with respect to u. Some of the chosen within the framework of this concept values of the parameters (x, y, t) are given in Table 1. The figures with the corresponding integral curves of the Cauchy problem (4.3)–(4.5) illustrating the results of the described above algorithm of the binary search for the solution of the asymptotic problem (4.3), (4.4), (4.6) are placed in the next section of the paper.

Table 1. Parameter values and corresponding numbers of figures

No	y	t	Corresponding figures
1	-8	9	Fig. 3, 4
2	-10	5	Fig. 5
3	8	-9	Fig. 6
4	8	8	Fig. 7
5	-8	-8	Fig. 8
6	8	0	Fig. 9
7	0	0	Fig. 10

6. Illustrations of numerical calculations

This section is devoted to illustrations of the integral curves for the Cauchy problem (4.3)–(4.5) obtained using the binary search for the initial speed as described above. We remind that in the figures bold green line represents the integral curves and the red bold line corresponds to the graph of the root function (3.1). In figures with contour distributions the increase in the intensity of green shade in the color of lines corresponds to the increase of positive values of the function, and increase in the intensity of blue shade in the color of lines corresponds to the increase in the absolute value of the negative values. The contour distributions are given in the paper to show the rate of the function (5.1) changing with increasing distance from the line of the root (3.1) in different parts of the line.

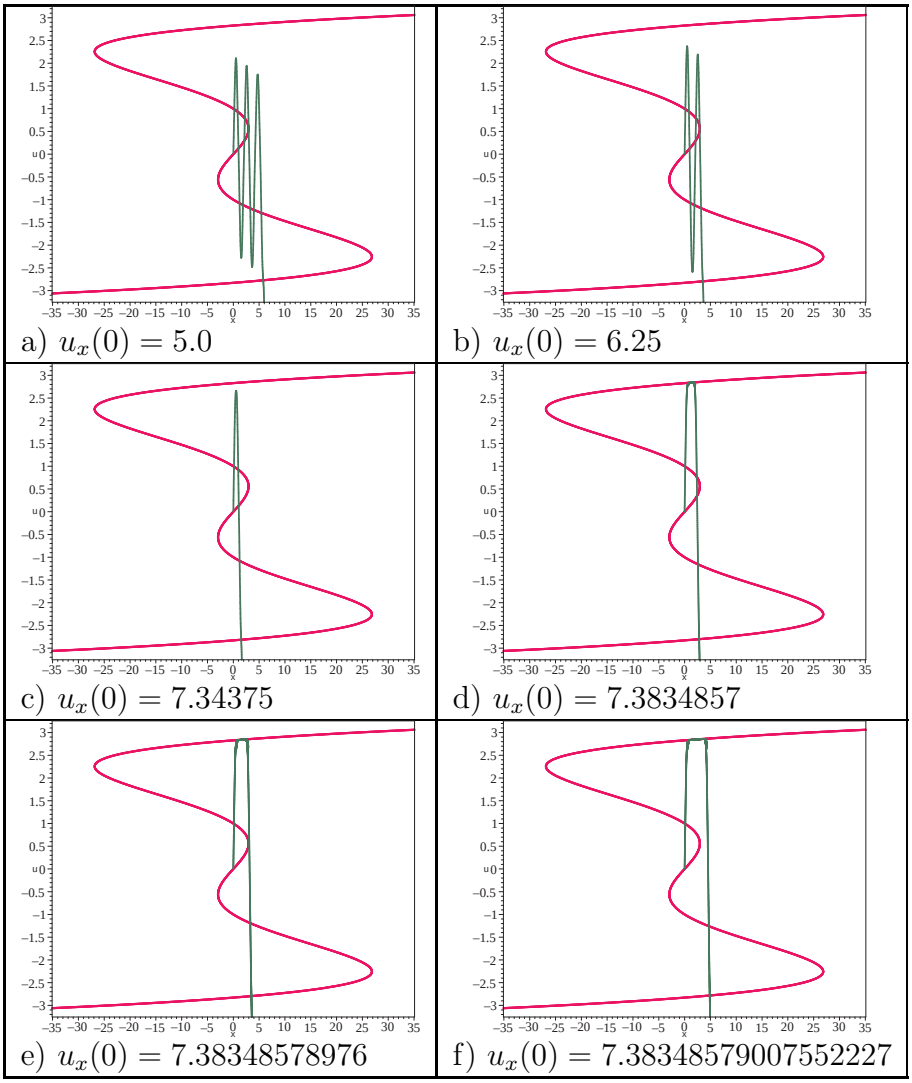


Figure 3. Some intermediate stages of search for optimal initial rate $u_x(0)$ for the Cauchy problem (4.3)–(4.5) for $y = -8$, $t = 9$

The root line (3.1) for $y = -8$, $t = 9$ and $z = 0$ shown in Fig. 3 and 4 has an explicit double system of bends, in virtue of which the integral curves of the Cauchy problem (4.3)–(4.5) obtained for different values of the initial velocity demonstrate a variety of processes including fluctuations. Partial figures of the Fig. 3 marked with letters from a) to f) correspond to selected stages of the

binary search, which embodied such variants of the curve behavior as the gradual disappearance of the interval of oscillation and the slow emergence and expand of the interval, on which the curve passes very close to the line of the root (3.1).

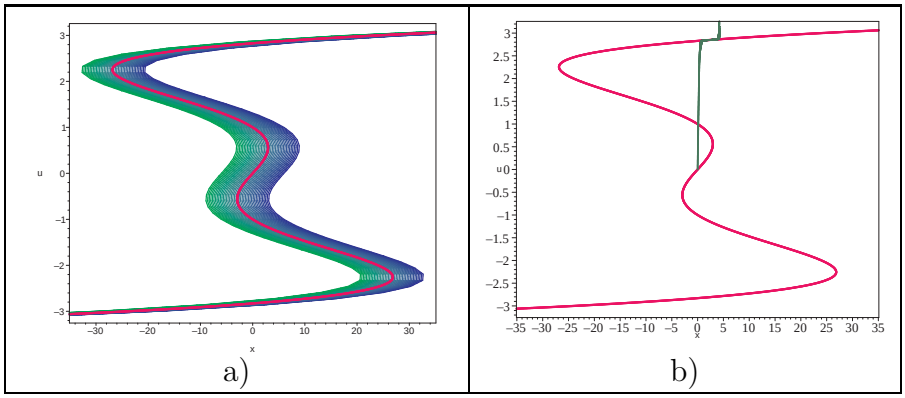


Figure 4. Results obtained for $y = -8$, $t = 9$ $z = 0$: a) contour distributions of function (5.1); b) integral curves for the Cauchy problem (4.3)–(4.5) for $u_x(0) = 7.38348579007552236$

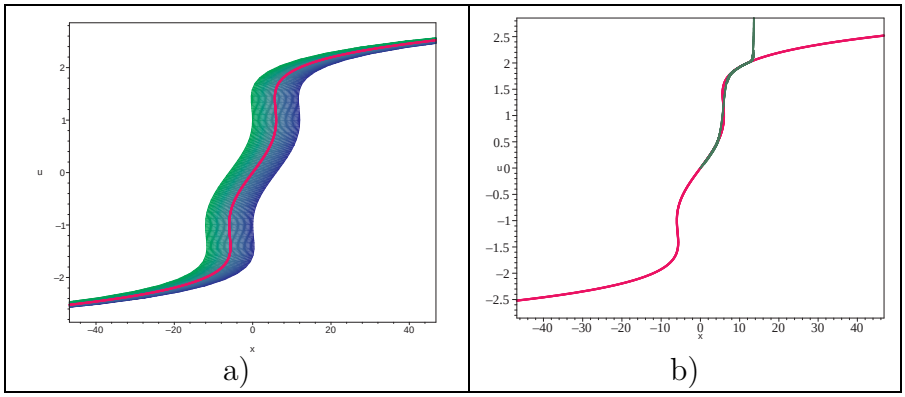


Figure 5. Results obtained for $y = -10$, $t = 5$ $z = 0$: a) contour distributions of function (5.1); b) integral curves for the Cauchy problem (4.3)–(4.5) for $u_x(0) = 0.10031199880359876075$

The situation that occurs when $y = -10$, $t = 5$ and $z = 0$ is shown in Fig. 5. It is characterized by the presence of two deflections of green curve, which is the approximation of the desired solution of the asymptotic problem (4.3), (4.4), (4.6), from the root line (3.1) in various ways: in the beginning integral curve follows the line of the root, then exceeds it, then crosses and passes below the root line and later again starts to follow it.

Situation in Fig. 6 that occurs when $y = 8$, $t = -9$ and $z = 0$, is very similar to the pattern for case of "cusp" catastrophe studied in the works [5], [6], the resemblance in the shape of the root lines (3.1) and of integral curves is explicit.

If $y = 8$, $t = 8$ and $z = 0$, as shown in Fig. 7, the root line (3.1) has a larger number inflection points, than one in the case of "cusp" catastrophe or than in Fig. 6, but the behavior of the integral curve remains broadly similar: a sharp increase is replaced by following the root line.

The graph the function (3.1) in Fig. 8 for $y = -8$, $t = -8$ and $z = 0$ increases gradually and remains single-valued for all points of positive half-axis ($x > 0$). By virtue of this fact, immediately from the very point $x = u = 0$, the root line (3.1) becomes an attractor for the approximation curve for the solution of asymptotic problem (4.3), (4.4), (4.6) and the integral curve following it until the final rapid movement to the vertical infinity.

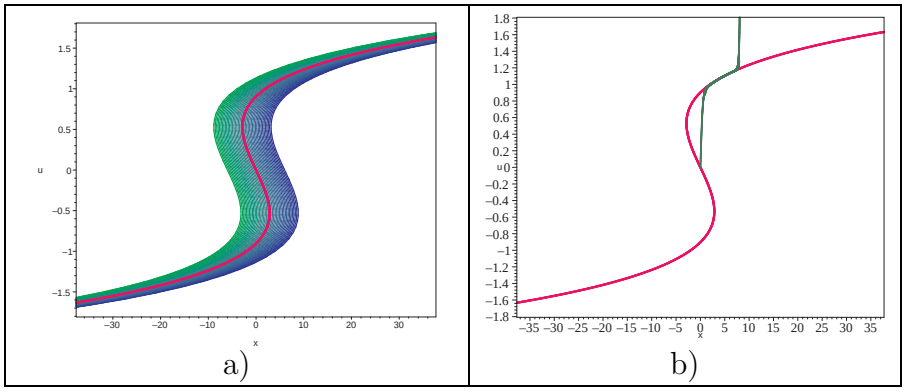


Figure 6. Results obtained for $y = 8$, $t = -9$ $z = 0$: a) contour distributions of function (5.1); b) integral curves for the Cauchy problem (4.3)–(4.5) for $u_x(0) = 1.982684405999750261$

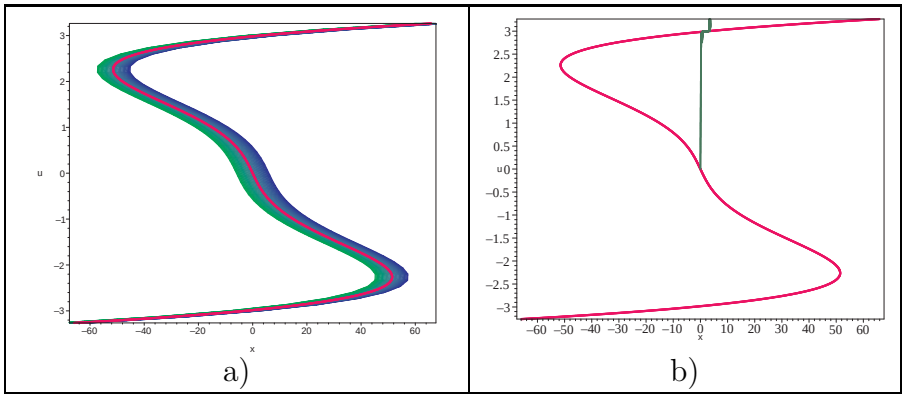


Figure 7. Results obtained for $y = 8$, $t = 8$ $z = 0$: a) contour distributions of function (5.1); b) integral curves for the Cauchy problem (4.3)–(4.5) for $u_x(0) = 12.406010731733428955$

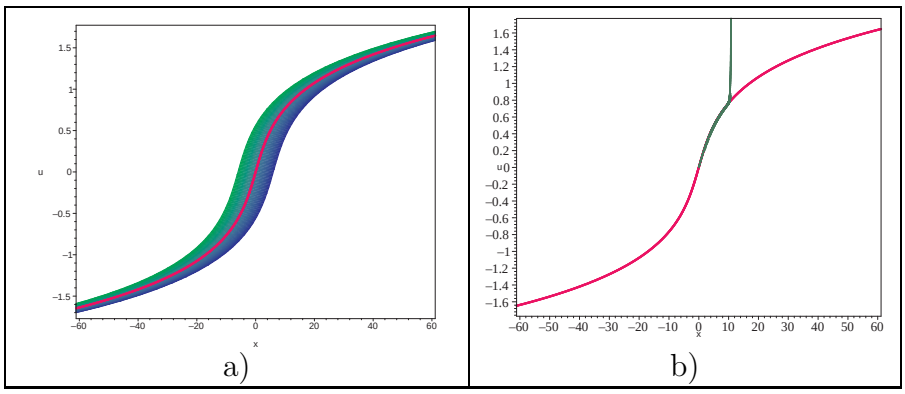


Figure 8. Results obtained for $y = -8$, $t = -8$ $z = 0$: a) contour distributions of function (5.1); b) integral curves for the Cauchy problem (4.3)–(4.5) for $u_x(0) = 0.1236970813110499762$

The situation in Fig. 9, corresponding to the values $y = 8$, $t = 0$ and $z = 0$ of the parameters and therefore to the absence of not only a quadratic but also a cubic term in the formula (5.1), the shape of the root line (3.1) has a more extended than in Fig. 6, the interval of domination of the linear term near the origin, but the behavior of the integral curve remains broadly the same as in the Fig. 6.

The values of the parameters for which the curves in Fig. 10 are constructed, correspond to the

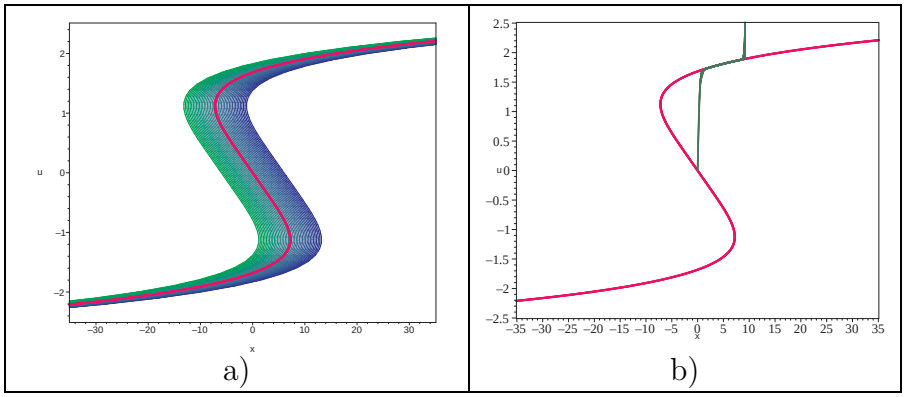


Figure 9. Results obtained for $y = 8$, $t = 0$ $z = 0$: a) contour distributions of function (5.1); b) integral curves for the Cauchy problem (4.3)–(4.5) for $u_x(0) = 4.00170961880393095411279 89$

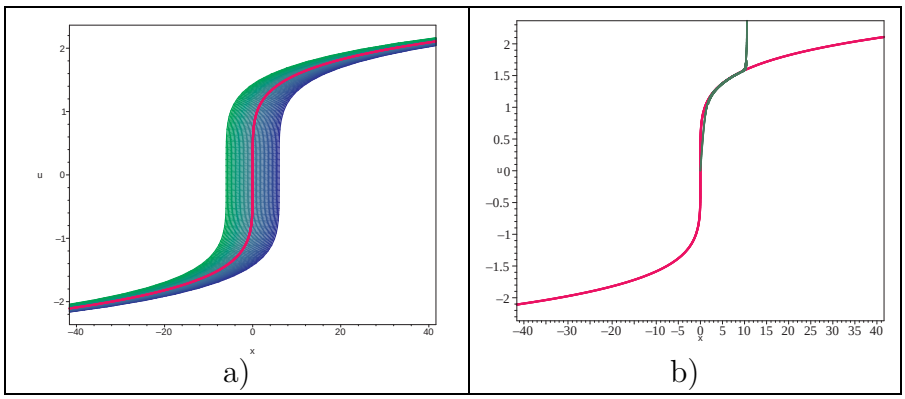


Figure 10. Results obtained for $y = 0$, $t = 0$ $z = 0$: a) contour distributions of function (5.1); b) integral curves for the Cauchy problem (4.3)–(4.5) for $u_x(0) = 0.9725146994563902$

origin of the parameter space of the function (5.1). In Fig. 10 the point $x = u = 0$ is the very point of "butterfly" catastrophe. One can note that in a neighborhood of this point the integral curve slightly deviates from almost vertically growing root line.

Thus, the results of calculations for the values of the parameters in Table 1 show that, gradually increasing the accuracy of the curve fitting method [1] and using a binary search algorithm, it is possible to obtain curves, which for more and more long time remain close to the line of balance noted above, that is consistent with hypothesis about the construction of the desired curve satisfying the asymptotic condition (4.6) as the solution of the Cauchy problem (4.3)–(4.5) with the value $\alpha_0 = \inf M$ of the initial rate.

7. Conclusion

We considered a PDE with an arbitrary linear combination of differentiation operators with smooth coefficients of the first and second orders on the left-hand side of the equation and a nonlinear function in the right-hand side, which depends on the desired function and contains a small parameter. We consider and analyze also the three-dimensional nonlinear wave equation with the source term smoothly changing over time and space.

The primary study of the behavior of solutions of this PDE near the typical point of "butterfly" catastrophe was held. We deduce two nonlinear ODE of the first and the second orders, respectively, depending on three parameters. The order of equation depends on the configuration

of the coefficients of the linear combination on the left-hand side of the original PDE and of the properties of the nonlinear function on its right-hand side.

We use the resulting second-order ODE to search for a special solution describing the rapid reconstruction of the solution of the wave PDE in a small neighborhood of the catastrophe point matched with expansion in a more outer layer. We have done a primary study: we produced integral curves that allow one to analyze the behavior of such a special solution. The results revealed no contradictions with the possibility to prove the existence of special solution within the framework of the approach given in [5, 6].

Acknowledgements

This work was supported by RFBR, research project No. 16-31-00222.

REFERENCES

1. **Fehlberg E.** Low-order classical Runge-Kutta formulas with stepsize control// NASA Technical Report R–315, 1969.
2. **Gilmore R.** Catastrophe theory for scientists and engineers// New York: Dover Publications, 1993. 666 p.
3. **Il'in A.M.** Matching of asymptotic expansions of solutions of boundary value problems// Transl. Math. Monogr., Vol.102, Providence: Amer. Math. Soc., 1992. 281 p.
4. **Il'in A.M., Leonychev Y.A., Khachay, O.Y.** The asymptotic behaviour of the solution to a system of differential equations with a small parameter and singular initial point// Sbornik Mathematics, 2010. Vol. 201, no. 1, P. 79–101.
5. **Il'in A.M., Suleimanov B.I.** On two special functions related to fold singularities// Doklady Mathematics, 2002. Vol. 66, no. 3, P. 327–329.
6. **Il'in A.M., Suleimanov B.I.** Birth of step-like contrast structures connected with a cusp catastrophe// Sbornik Mathematics, 2004. Vol. 195, no 12. P. 1727–1746.
7. **Khachay O.Y.** Asymptotic expansion of the solution of the initial value problem for a singularly perturbed ordinary differential equation// Differential Equations, 2008. Vol. 44, no. 2, P. 282–285.
8. **Khachay O.Y.** Asymptotics of the solution of a system of nonlinear differential equations with a small parameter and with a higher-order extinction effect// Differential Equations, 2011. Vol. 47, no. 4, P. 604–607.
9. **Khachay O.Y.** On the matching of powerlogarithmic asymptotic expansions of a solution to a singular Cauchy problem for a system of ordinary differential equations// Trudy Instituta Matematiki i Mekhaniki, 2013. Vol. 19, no. 1, P. 300–315. [in Russian]
10. **Khachai O.Y.** On the application of the method of matching asymptotic expansions to a singular system of ordinary differential equations with a small parameter// Differential Equations, 2014. Vol. 50, no. 5, P. 608–622.
11. **Maslov V.P., Danilov V.G., Volosov K.A.** Mathematical modeling of heat and mass transfer processes. Evolution of dissipative structures// Nauka. Moscow, 1987. 352 p. [in Russian]
12. **Suleimanov B.I.** Cusp catastrophe in slowly changing equilibrium positions// JETPh, 2002. Vol. 122, no. 5 (11). P. 1093–1106. [in Russian]
13. **Vladimirov V.S., Jarinov V.V.** Equations of mathematical physics// Fizmatlit, Moscow, 2004. 400 p., ISBN: 5-9221-0310-5. [in Russian]

URAL MATHEMATICAL JOURNAL, Vol. 2, No. 2, 2016

IMPULSE–SLIDING REGIMES IN SYSTEMS WITH DELAY[1]

Alexander N. Sesekin

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of Russian Academy of Sciences, Ekaterinburg, Russia,
sesekin@imm.uran.ru

Natalya I. Zhelonkina

Krasovskii Institute of Mathematics and Mechanics,
Ural Branch of Russian Academy of Sciences, Ekaterinburg, Russia,
312115@mail.ru

Abstract: The paper is devoted to the formalization of a concept of impulse-sliding regimes generated by positional impulse controls for systems with delay. We define the notion of impulse-sliding trajectory as a limit of a sequence of Euler polygonal lines generated by a discrete approximation of the impulse position control. The equations describing the trajectory of impulse-sliding regime are received.

Key words: Impulse position control, Systems with delay, Impulse-sliding regime, Euler polygonal lines.

Introduction

Usually, the positional control algorithms are introduced by substitution in the program control models the initial time and the initial model position to an arbitrary time moment and to an arbitrary state. Such replacement may result in that we will need to realize an impulse control action at each time moment. This fact leads to the appearance of a moving or a so called sliding impulse. Such phenomenon from the point of view of the theory of differential equations requires an appropriate formalization. In addition, this motion type in the space of positions creates the motion sliding on some functional manifold. Impulse-sliding regimes in systems without delay were considered in [1–3]. Impulse-sliding regimes for linear systems with delay were studied in [4]. The reaction of nonlinear systems with delay to impulse actions is understood here as in the paper [5]. The definition of a solution of nonlinear systems with delay given in [5] is a generalization for the notion of solution for systems without delay in [6, 7].

1. Formalization of impulse-sliding regime

Consider a dynamic system with impulse control

$$\dot{x}(t) = f\big(t, x(t), x(t-\tau)\big) + B(t, x(t))u, \quad t \in [t_0, \vartheta], \tag{1.1}$$

with the initial condition

$$x(t) = \varphi(t), \quad t \in [t_0 - \tau, t_0], \tag{1.2}$$

where $f(\cdot,\cdot,\cdot)$ is a function with value in R^n, $B(\cdot,\cdot)$ is a $m \times n$-matrix function. Elements of f and B are continuous functions and satisfy the conditions, which guarantee the existence and

[1]The research was supported by Russian Science Foundation (RSF) (project No.16–11–10146).

uniqueness of a solution for any summable function $u(t)$. Let $x_t(\cdot)$ be a function-prehistory $x_t(\cdot) = \{x(t+s);\ -\tau \le s < 0\}$. The function $\varphi(t)$ here is a function of bounded variation for $t \in [t_0-\tau, t_0]$.

We will assume that the function $B(t,x)$ satisfies the well-known Frobenius condition [8],

$$\sum_{\nu=1}^{n} \frac{\partial b_{ij}(t,x)}{\partial x_\nu} b_{\nu l}(t,x) = \sum_{\nu=1}^{n} \frac{\partial b_{il}(t,x)}{\partial x_\nu} b_{\nu j}(t,x). \tag{1.3}$$

According to [5,6] this condition ensures the uniqueness of the system response to the impact of a generalized control $u(t)$ (the generalized derivative of a bounded variation function). We note that there are various ways of defining a solution for the equation (1.1) which lead generally to various implementations of the trajectories [6]. We will use the definition that is based on the closure of the set of smooth trajectories in the space of functions of bounded variation [6]. This definition is the most natural from the point of view of control theory. This is due to the fact that impulse controls are often some control idealizations operating in short time intervals and with great intensity.

By an impulse positional control we will mean an operator $t,\ x_t(\cdot) \longrightarrow U(t,x(t))$ mapping the space of extended states $\{t,x(t)\}$ into the space of m-vector-valued distributions

$$U(t,x(t)) = r(t,x(t))\,\delta_t. \tag{1.4}$$

In this paper we assume that a delay is only in $f(t,x(t),x(t-\tau))$ and a control function does not contain a delay.

Here $r(t,x(t))$ is m-dimensional vector function, δ_t is the Dirac impulse function concentrated at t. The system reaction to the impulse position control $U(t,x(t))$ (which we call an impulse-sliding regime) is defined as the set of Euler polygonal functions $x^h(\cdot)$, $h = \max(t_{k+1} - t_k)$ corresponding to all decompositions $t_0 < t_1 < ... < t_p = \vartheta$ of the interval $[t_0, \vartheta]$. The Euler polygonal function (Euler line) $x^h(\cdot)$ is constructed as a left continuous function of bounded variation such that the equation holds

$$\dot{x}^h(t) = f\big(t, x^h(t), x^h(t-\tau)\big) + \sum_{i=1}^{p} B(t,x^h(t)) r(t_i, x(t_i)) \delta_{t_i} \tag{1.5}$$

with the initial condition $x(t) = \varphi(t),\ \ t \in [t_0-\tau, t_0]$.

The Euler line satisfies the equation

$$x^h(t) = \varphi(t_0) + \int_{t_0}^{t} f\big(\xi, x^h(\xi), x^h(\xi-\tau)\big)\,d\xi + \sum_{t_i<t} S\big(t_i, x^h(t_i), r(t_i, x(t_i))\big) \tag{1.6}$$

and the jump functions are defined by the equations

$$S\big(t_i, x^h(t_i), r(t_i, x^h(t_i))\big) = z(1) - z(0), \tag{1.7}$$

$$\dot{z}(\xi) = B(t, z(\xi))\, r(t_i, x^h(t_i)), \quad z(0) = x^h(t_i). \tag{1.8}$$

Here the jump function $S(t,x,\mu)$ is the solution of the equation

$$\frac{\partial y}{\partial \mu} = B(t,y). \tag{1.9}$$

We will assume that the equality

$$r\Big(t,\ x(t) + S\big(t, x(t), r(t,x(t))\big)\Big) = 0. \tag{1.10}$$

is true.

This equality means that after the action of impulse at the system at time t, the state $\{t,x(t)\}$ will belong to the manifold $r(t,x(t)) = 0$.

2. Properties of the impulse-sliding regime

Lemma 1. *Assume that for all admissible values t_1, t_2, x_1, x_2, y_1 and y_2 the following inequalities are true*

$$\|f(t,x,y)\| \le C\big(1 + \sup_{[t_0-\tau]} \|x(\cdot)\|\big), \tag{2.1}$$

$$\|S(t_1,x_1,r(t_1,x_1)) - S(t_2,x_2,r(t_2,x_2))\| \le L\big(|t_1-t_2| + \|x_1-x_2\|\big). \tag{2.2}$$

Then for all decompositions h and all $t \in [t_0,\vartheta]$ the set of Euler polygonal functions $x^h(\cdot)$ is bounded, what means that there exists a constant M such that

$$\|x^h(t)\| \le M. \tag{2.3}$$

P r o o f. From (1.6) and (2.1) the following inequality follows

$$\|x^h(t)\| \le \|\varphi(t_0)\| + C\int_{t_0}^{t} \big(1 + \sup_{[t_0-\tau,\xi]} \|x^h(\cdot)\|\big)d\xi + \sum_{t_i<t} \big\|S\big(t_i,\, x^h(t_i),\, r(t_i,x^h(t_i))\big)\big\|. \tag{2.4}$$

Due to the fact that

$$S\big(t_{i-1},\, x^h(t_{i-1}+0),\, r(t_{i-1},x^h(t_{i-1}+0))\big) = 0,$$

in view of (2.2), we have the inequalities

$$\begin{gathered}\|S(t_i,x^h(t_i),r(t_i,x^h(t_i))\| = \|S(t_i,\, x^h(t_i),\, r(t_i,x^h(t_i))\| - \\ -S\big(t_{i-1},\, x^h(t_{i-1}+0),\, r(t_{i-1},x^h(t_{i-1}+0))\big) \le L\big(t_i - t_{i-1} + \|x^h(t_i) - x^h(t_{i-1}+0)\|\big).\end{gathered} \tag{2.5}$$

At the same time, in view of (2.1),

$$\begin{aligned}\|x^h(t_i) - x^h(t_{i-1}+0)\| &\le \int_{t_{i-1}}^{t_i} \|f(\xi,x^h(\xi),x^h(\xi-\tau))\|d\xi \\ &\le C\Big(t_i - t_{i-1} + L\int_{t_{i-1}}^{t_i} (1 + \sup_{[t_0-\tau,\xi]} \|x^h(\cdot)\|)d\xi\Big).\end{aligned} \tag{2.6}$$

In consequence, from (2.4) in view of (2.5) and (2.6) we get the following inequality

$$\|x^h(t)\| \le \|\varphi(t_0)\| + (L+C)(t-t_0) + L(1+C)\int_{t_0}^{t} \sup_{[t_0-\tau,\xi]} \|x^h(\cdot)\|d\xi. \tag{2.7}$$

As in [9], from the last inequality we get

$$\sup_{[t_0-\tau,t]} \|x^h(\cdot)\| \le R + (L+C)(t-t_0) + L(1+C)\int_{t_0}^{t} \sup_{[t_0-\tau,\xi]} \|x^h(\cdot)\|d\xi, \tag{2.8}$$

where

$$R = \sup_{[t_0-\tau,t_0]} \|\varphi(\cdot)\|.$$

Applying the result of [10] we get from (2.8) the estimate

$$\sup \|x^h(\cdot)\| \le \big(R + (L+C)(\vartheta - t_0)\big)e^{L(1+C)(\vartheta-t_0)},$$

which completes the proof of the lemma. □

Note that as a constant M we can take the following number

$$M = \big(R + (L + C)(\vartheta - t_0)\big)e^{L(1+C)(\vartheta - t_0)}.$$

Let D be a bounded closed set which contains all $x^h(\cdot)$. By continuity we may assume that all functions $f(t, x, y)$, $B(t, x)$ and $r(t, x)$ are bounded.

Denote

$$M_1 = \max_{[t_0,\vartheta]\times D\times D} \|f(t, x, y)\|, \quad M_2 = \max_{[t_0,\vartheta]\times D} \|B(t, x)\|, \quad M_3 = \max_{[t_0,\vartheta]\times D} \|r(t, x)\|. \tag{2.9}$$

Lemma 2. *Under the above assumptions from each confinal sequence of Euler functions $\{x^h(\cdot)\}$ we can select a subsequence $\{x^{h_p}(\cdot)\}$ uniformly at $(t_0, \vartheta]$ converging to absolutely continuous function $x(\cdot)$. Moreover for all $t \in (t_0, \vartheta]$ we have $r(t, x(t)) = 0$ $\big(x(t) = \varphi(t)$ for $t \in [t_0 - \tau, t_0]\big)$, in other words the limit element of the impulse-sliding regime moves over the manifold which is described by the equation $r(t, x(t)) = 0$.*

P r o o f. The proof of convergence of $x^h(\cdot)$ uses the generalization of Arzela's lemma from [11]. Let $x^{h_i}(\cdot)$ be a confinal sequence. Then according to (1.6) we have

$$\|x^{h_i}(t'') - x^{h_i}(t')\| \le \int_{t'}^{t''} \|f(t, x^h(t), x^h(t - \tau)\|ds + \sum_{k=m(t')+1}^{m(t'')} \|S(t_k, x^{h_i}(t_k), r(t_k, x^{h_i}(t_k)))\|, \tag{2.10}$$

where $m(t)$ is the nearest point on the left in the decomposition which generates the polygonal line $x^{h_i}(\cdot)$. In accordance with (1.6) we have

$$\|S(t_k, x^{h_i}(t_k), r(t_k, x^h(t_k)))\| = \|S(t_k, x^h(t_k), r(t_k, x^h(t_k)))\| -$$

$$-\big\|S\big(t_{k-1}, x^{h_i}(t_{k-1} + 0), r(t_{k-1}, x^{h_i}(t_{k-1} + 0))\big)\big\|.$$

Considering (2.2) we get

$$\|S(t_k, x^{h_i}(t_k), r(t_k, x^h(t_k)))\| \le L\big(t_k - t_{k-1} + \|x^{h_i}(t_k) - x^{h_i}(t_{k-1} + 0)\|\big).$$

At the same time

$$x^{h_i}(t_k) - x^{h_i}(t_{k-1} + 0) = \int_{t_{k-1}}^{t_k} f(\xi, x^h(\xi))d\xi.$$

By taking into account (2.8), we obtain

$$\big\|S\big(t_k, x^{h_i}(t_k), r(t_k, x^h(t_k))\big)\big\| \le L\big(t_k - t_{k-1} + M_1(t_k - t_{k-1})\big) = L(1 + M_1)(t_k - t_{k-1}). \tag{2.11}$$

From (2.10) and (2.11) it follows that

$$\|x^{h_i}(t'') - x^{h_i}(t')\| \le \big(M_1 + L(1 + M_1)\big)(t'' - t') + L(2 + M)(t' - t_{t_i h_i}), \tag{2.12}$$

where $t_{t_i h_i}$ is the nearest point at the left in partition h_i to the point t'. The last inequality allows to apply the generalization of Arzela's lemma from [11] and ensures the existence of a subsequence $x^{h_i}(\cdot)$ which uniformly converges to the function $x(\cdot)$.

Now we pass to the limit in the inequality (2.12) as $i \to \infty$. As a result we have

$$\|x(t'') - x(t')\| \le (M_1 + L(1 + M_1))(t'' - t').$$

This means that $x(t)$ is an absolutely continuous function at $(t_0, \vartheta]$.

Now let us show that the limit element $x^h(\cdot)$ belongs to the manifold $r(t,x)=0$. Let $t_{m_t h_i}$ be the nearest point from the left in partition h_i by the time t. The following inequality holds

$$\begin{aligned}\|r(t,x(t)\| &\le \|r(t,x(t)) - r(t,x^{h_i}(t)) + r(t,h^i(t))\| \\ &\le \|r(t,x(t)) - r(t,x^{h_i}(t))\| + \|r(t_{m_t h_i} x^{h_i}(t_{m_t h_i}+0)) - r(t,x^{h_i}(t))\| \\ &\le L\big[\|x(t)-x^{h_i}(t)\| + (t-t_{m_t h_i}) + \|x^{h_i}(t_{m_t h_i}+0) - x^{h_i}(t)\|\big] \\ &\le L\big[\|x(t)-x^{h_i}(t)\| + (L+M)(t-t_{m_i h_i})\big].\end{aligned}$$

By the uniform convergence of the sequence $x^{h_i}(\cdot)$ the first term at the right hand part at the last inequality tends to zero. The second one tends to zero because $h_i \to 0$ when $i \to \infty$. Therefore $r(t,x(t)) \equiv 0$ when $t \in (t_0, \vartheta]$, this completes the proof of lemma. □

Lemma 3. *Let $r(t,x)$ be a vector function continuously differentiable in all variables. Then the following equality holds*

$$\begin{aligned}&S(t_k, x^h(t_k), r(t_k, x^h(t_k))) - S(t_{k-1}, x^h(t_{k-1}+0), r(t_{k-1}, x^h(t_{k-1}+0))) \\ &= \int_{t_{k-1}}^{t_k} \left[\frac{\partial S(\xi, x^h(\xi), r(t, x^h(\xi)))}{\partial \xi} + \frac{\partial S(\xi, x^h(\xi), r(\xi, x^h(\xi)))}{\partial x} f(\xi, x^h(\xi), x^h(\xi-\tau)) + \right. \\ &\left. + \frac{\partial S(\xi, x^h(\xi), r(\xi, x^h(\xi)))}{\partial r}\Big(\frac{\partial r(\xi, x^h(\xi))}{\partial \xi} + \frac{\partial r(\xi, x^h(\xi))}{\partial x} \dot f(\xi, x^h(\xi), x^h(\xi-\tau))\Big)\right] d\xi.\end{aligned} \tag{2.13}$$

The lemma follows from the the formula for differentiating a composite function.

Theorem 1. *Let all assumptions given above hold. Then an impulse-sliding regime on $(t_0, \vartheta]$ is described by the equation*

$$\begin{aligned}\dot x(t) &= \frac{\partial S(t, x(t), r(t, x(t)))}{\partial t} + \frac{\partial S(t, x(t), r(t, x(t)))}{\partial r}\frac{\partial r(t, x(t))}{\partial t} + \\ &+ \left[E + \frac{\partial S(t, x(t), r(t, x(t)))}{\partial x} + \frac{\partial S(t, x(t), r(t, x(t)))}{\partial r} \times \frac{\partial r(t, x(t)))}{\partial x}\right] f(t, x(t), x(t-\tau)), \\ x(t_0+0) &= x(t_0) + S\big(t_0, x(t_0), r(t_0, x(t_0))\big).\end{aligned} \tag{2.14}$$

P r o o f. According to (1.6) and Lemma 3 $x^h(t)$ satisfies the equation

$$\begin{aligned}x^{h_i}(t) &= \varphi(t_0) + \int_{t_0}^{t} f(\xi, x^{h_i}(\xi), x^{h_i}(\xi-\tau))d\xi + \int_{t_0}^{t_{mh_i}} \left[\frac{\partial S(\xi, x^{h_i}(\xi), r(t, x^{h_i}(\xi)))}{\partial \xi} + \right. \\ &+ \left(\frac{\partial S(\xi, x^{h_i}(\xi), r(\xi, x^{h_i}(\xi)))}{\partial x} + \frac{\partial S(\xi, x^{h_i}(\xi), r(\xi, x^{h_i}(\xi)))}{\partial r} \times \frac{\partial r(\xi, x^{h_i}(\xi))}{\partial x}\right) f(\xi, x^{h_i}(\xi), x^{h_i}(\xi-\tau)) + \\ &\left. + \frac{\partial S(\xi, x^{h_i}(\xi), r(\xi, x^{h_i}(\xi)))}{\partial r} \cdot \frac{\partial r(\xi, x^{h_i}(\xi))}{\partial \xi}\right] d\xi.\end{aligned}$$

Passing to the limit at the last equation and bearing in mind that $x(t)$ is an absolutely continuous function, we can see that the theorem is true. □

3. Conclusion

The formalization of the impulse-sliding regime for a nonlinear system with time delay is made. The equation to describe the limiting element of impulse-sliding regime is obtained.

REFERENCES

1. **Zavalishchin S.T., Sesekin A.N.** Impulse-sliding regimes of nonlinear dynamic systems // Differ. Equations, 1983. Vol. 19, no. 5. P. 562–571.
2. **Finogenko I.A., Ponomarev D.V.** On differential inclusions with positional discontinuous and pulse controls // Trudy Inst. Mat. i Mekh. UrO RAN, 2013. Vol. 19, no. 1. P. 284–299.
3. **Sesekin A.N., Nepp A.N.** Impulse position control algorithms for nonlinear systems // AIP Conference Proceeding, 2015. Vol. 1690, 040002. P. 1–6. http://dx.doi.org/10.1063/1.4936709
4. **Andreeva I.Yu., Sesekin A.N.** Degenerate linear-quadratic optimization with time delay // Autom. Remote Control, 1997. Vol. 58, no. 7. Part 1. P. 1101–1109.
5. **Fetisova Yu.V., Sesekin A.N.** Discontinuous solutions of differential equations with time delay // WSEAS transactions on systems, 2005. Vol. 4, no. 5. P. 487–492.
6. **Zavalishchin S.T, Sesekin A.N.** Dynamic impulse systems: theory and applications. Kluwer Academic Publishers. Dolrbrecht, 1997.
7. **Sesekin A.N.** Dynamic systems with nonlinear impulse structure // Proceedings of the Steklov Institute of Mathematics (Supplementary issues), 2000. Suppl. 2, P. S158–S172.
8. **Cartan H.** Formes differentielles. Calcul differentiel. Hermann Paris, 1997.
9. **Lukojanov N.Yu.** Functional equations of Hamilton-Jacobi and control tasks with hereditary information. Ural Federal University. Ekaterinburg, 2011.
10. **Bellman R., Cooke K.** Differential-difference equations. Academik Press. New Yourk, London, 1963.
11. **Filippov A.F.** The existence of solutions of generalized differential equations // Mathematical Notes. Vol. 10, no. 3. P. 307–313.

Editor: Tatiana F. Filippova

Managing Editor: Oksana G. Matviychuk

Design: Alexander R. Matvichuk

Contact Information

16 S. Kovalevskaya str., Ekaterinburg, Russia, 620990

Phone: +7 (343) 375-34-73

Fax: +7 (343) 374-25-81

Email: secretary@umjuran.ru

Web-site: https://umjuran.ru

N.N.Krasovskii Institute of Mathematics and Mechanics
of the Ural Branch of Russian Academy of Sciences

Ural Federal University named after the first President of Russia B.N.Yeltsin

Distributed for free