National Research University
Higher School of Economics
Faculty of Mathematics

As a manuscript

Tikhon V. Bogachev

EXTREMAL PROBLEMS IN SOME PROBABILISTIC RESOURCE ALLOCATION MODELS

Summary of the thesis
for the purpose of obtaining academic degree
Doctor of Philosophy in Mathematics

Supervisor:
Doctor of Science, Professor
Alexandr V. Kolesnikov

Moscow-2023

Introduction

Field of study

The problems of the optimal distribution of resources in the system are being studied by mathematicians for a relatively long time. Many examples in finance, physics, information technology and queuing theory show that intuitive approaches to optimisation may give results that are far from the real optimum. In particular, there are situations in which the addition of an extra resource into the system leads to deterioration of the overall performance. As an example, let us take the famous Braess paradox [1], describing the theoretical (in the form of a graph) road configuration, in which the construction of a new connecting road may slow down the average motion speed, even if the number of cars remains constant. And vice versa, closing one road in the Braes network will allow all vehicles to travel faster on average. Other situations of this kind from the field of computing systems [2] are also highlighted.

In classical optimisation problems, deterministic systems are considered, which leads to optimal control problems (see, for example, [3], [4]). Meanwhile, in systems describing processes in time, such as the exchange of information in a communication network, establishing connections in a social network, movement of particles and exchange of energy, interaction of players in the market, probabilistic models of processes are of great importance. If an aggregate measure of the system's performance over a certain period is taken, then we get a function that does not contain time as a parameter. One of perhaps the least studied problems of this kind in terms of rigorous mathematical results is the optimal taxation. The impact of taxation on the economy was considered in practice rather intuitively. More rigorous research in this area has been initiated by specialists in mathematical economics in two main branches: the impact of tax changes on the distribution of various goods in the economy (see [5]) and problems of the optimal income taxation. On the second topic we mention a very well-known article by the Nobel laureate D.A. Mirrlis [6], his review [7], as well as [8], [9]. Also, in mathematical economics commodity taxation is distinguished (see [10]), but specific mathematical models are not strictly tied to a particular economic issue and appear in different senses in a variety of fields. Let us give the formulation of the problem investigated by the author in the papers [11] (co-authored with S. N. Popova) and [12]. We will consider the income tax model and the optimisation problem in this model, motivated by [13] and [14]. This problem consists in maximising the integral functional on the space of increasing functions in the presence of non-linear constraint leading to rather singular objects. Therefore, in contrast to many works from applied economics using heuristic methods to differentiate all the necessary functions and assuming that they have zero derivatives at the extremum points, a rigorous mathematical analysis of the problem leads to interesting questions in the theory of functions.

We will assume that an economic agent is an abstract object of some class, fully characterised by the performance type $\theta \in \Theta \subset \mathbb{R}^n_+$. Agents are distributed among the types according to the probability measure P on Θ (in general, measure should be probabilistic up to normalisation, but for the sake of convenience we will further assume it is probabilistic, which will not affect the analysis). Denote by $l \in \mathbb{R}^n_+$ the labour effort of particular subject. As a result of the scalar product we have a numerical income $y = (\theta, l)$ and utility $U(\theta, l) = y - T(y) - f(l)$, where $T(y)$ is a tax collected from the income y, determined by the regulator, and $f(l)$ is a function of labour effort, describing the financial costs arising from the agent's activities. Utility can be understood as the net profit of the agent after all costs have been incurred. It is also specified that $f(l)$ is twice continuously differentiable, increasing, strictly convex function. Based on common sense considerations, $T(y)$ and $y - T(y)$ are increasing non-negative continuous functions. Normally, we will also assume that the function T is convex, which makes sense, given the fact that the tax is assumed to grow faster for higher incomes. For the particular performance type θ and fixed taxation function T we solve the optimisation problem

$$\max_{l_i > 0 \, \forall i} U(\theta, l) = \max_{l \in \mathbb{R}^n_+} ((\theta, l) - T(\theta, l) - f(l)). \tag{1}$$

Having found the points of maxima $l_{max}(\theta)$ for each θ and, henceforth, $y_{max}(\theta) = (\theta, l_{max}(\theta))$, we define the government revenue as

$$R(T) = \int_\Theta T(y_{max}(\theta))P(d\theta). \tag{2}$$

By analogy, the overall utility in the economy can be defined as

$$\int_\Theta U(\theta, l_{max}(\theta))P(d\theta).$$

The monopoly problem

In the study of interaction of agents (customers) in economic models one encounters problems of a monopolist, multidimensional screening and auctions [15], [16], [17], [18]. The subject of study is a Dirichlet type functional

$$\Phi(u) = \int_X \Big(\langle x, \nabla u \rangle - u - \varphi(\nabla u)\Big)\rho dx,$$

where $X = [0, 1]^n$, φ is a convex function on X, ρ is a probability density on X. The problem is to maximise the functional Φ.

In the paper [15], an equivalent formulation of the monopolist problem was obtained in the form of finding the maximum of the functional Φ on the set $\mathcal{U}_0$ of convex and coordinate-wise increasing functions on X, equal to 0 at zero (maximisation of the total income).

In the paper [17], another representation of the functional Φ was introduced. Under the assumption of a sufficient smoothness of ρ we integrate by parts in the term

$$\int_X \langle x, \nabla u \rangle \rho dx.$$

This gives the following representation on the set of all Lipschitz functions:

$$\int_X \Big(\langle x, \nabla u \rangle - u \Big) \rho dx + u(0) = \int_X u dm,$$

where m is a measure of bounded variation with the property $m(X) = 0$. We note that m can contain singular components, including the Dirac measure at zero and a nontrivial measure on ∂X. Thus, the auction problem for a singe customer reduces to the problem of finding

$$\int_X u dm \to \max$$

on the set $\mathcal{U}(X) \cap \mathrm{Lip}_1(X)$. As shown in the paper [17], this representation enables us to find a connection between the original problem and the transport problem with the cost function $c(x, y) = |x - y|$ and the corresponding distance W_1. Namely, the auction problem is dual for the transport problem in the following sense:

$$\max_{u \in \mathcal{U}(X) \cap \mathrm{Lip}_1(x)} \int_X u dm = \min_{m_+ \preceq \gamma_1, m_- \succeq \gamma_2} W_1(\gamma_1, \gamma_2). \tag{3}$$

Here $m = m_+ - m_-$ is the decomposition of m into positive and negative parts, γ_i are nonnegative measures with the property $m_+(X) = m_-(X) =$

$\gamma_1(X) = \gamma_2(X)$; $\mu_1 \preceq \mu_2$ means that for each function $u \in \mathcal{U}$ we have $\int_X u d\mu_1 \leq \int_X u d\mu_2$.

In the paper [18] the following equality was proved generalising (3):

$$\sup_{u \in \mathcal{U}(X)} \Phi(u) = \inf_{c \in \mathcal{C}} \int_X \varphi^*(c) \rho dx, \tag{4}$$

where $\mathcal{C}$ is the set of vector fields with the property

$$\mathcal{C} = \Big\{ c : \int_X u dm \leq \int_X \langle \nabla u, c \rangle \rho dx, \ \forall u \in \mathcal{U}(X) \Big\}.$$

For sufficiently regular fields this relationship can be written as

$$m \preceq -\mathrm{div}(c \cdot \rho),$$

so equality (4) takes the form

$$\sup_{u \in \mathcal{U}(X)} \Phi(u) = \inf_{m \preceq \pi} \mathrm{Beck}_{\rho, \varphi^*}(\pi), \tag{5}$$

where

$$\mathrm{Beck}_{\rho, \varphi^*}(m) = \inf_{c : \mathrm{div}(c \cdot \rho) = -m} \int_X \varphi^*(c) \rho dx$$

is the Beckman functional introduced in the paper [19] for modelling transport flows. It is shown in Chapter 4 of this dissertation that the functional in the right side of (5) attains a minimum.

Queueing system

In addition to representing a resource-constrained system as a probabilistic distribution of elements of this system, a powerful tool such as network representation emerges. Namely, in such representation we consider a network (graph) of vertices and links (edges) between them, where each vertex and each link has its own qualities. Within the network, processes of origination and processing of some nominal queries (*jobs*) take place. Such processes are modelled by stochastic processes. These processes require the network to allocate a resource to optimise its performance. Examples of systems can be observed in nature, such that they optimise their own activities on the basis of natural laws in a distributed way. The behaviour of liquid or gas can be approximated to a microscopic level by considering process in terms of the mechanics of each molecule, with the edges of the virtual graph describing physical interaction between particles. At this level of detail, the velocity norm of a molecule looks like a random process with the stationary distribution found by Maxwell and later discussed by Erlang ([24]). Moving to a larger scale, we get an aggregated description, which operates with values such as temperature and pressure. Similarly, the behaviour of electrons in a power grid can be described in terms of random walks, and this simple way of modelling leads to very complex behaviour at the macroscopic level: the structure of the potentials in the resistor network is such that it minimises heat dissipation at a given level of current flow ([25]). The local, random behaviour of electrons forces the network as a whole to solve a rather difficult optimisation problem.

When considering communication networks, one can imagine a situation in which "smart" control system redirects connections established on blocked lines. This, in turn, causes the next connections to be redirected. As a result, it starts a chain reaction leading to a fatal outcome for the performance of the entire network. In this way, when a system is too efficient, it can overdo it. These examples show how important are routing algorithms ([26]). We also mention issues of random network growth and other concepts from the area of random graphs (an overview can be found in [27] or [28]).

Apart from the mentioned approaches to the extremely broad problem of optimisation of the structure of networks with a load, we should consider optimisation of processes taking place in a given network and subject to a certain probabilistic model (a systematic review can be found in [29]). We will give a brief introduction to this topic to state the problem explored in Chapter 3 of this work. At the micro level, the analysis of a basic service system is queuing theory ([30]). The first thing to consider is a single queue with an incoming flow of jobs and a processor of a given capacity that processes these jobs. The resulting process will have Markov properties. When switching to a network of queues, the concept of Poisson flows in the network is arising. For the first time these properties were discovered for telephone networks by Erlang ([31]). The concept of loss networks is also considered, the essence of which is that the connection between two vertices requires simultaneously holding a line on each intermediate edge in a route between vertices (see [32] for more on this). The main difference between a loss network and a queuing network is that in the first, network congestion causes a loss of connection, while in the second, congestion leads to an increase of delays. In both types of networks, problems can be formulated at different scales of network size and time. When moving from a single node to the network as a whole, we get flows of jobs that use resources throughout their paths in the network. Then problems arise related to the concepts of fairness with respect to processing of different flows (see [33], [34]). Among other things, in [34] fluid scheduling models were investigated, which in many ways became a source of inspiration for problem statements in this dissertation. However, if we move to a larger time scale, then the job flows are being considered as objects that appear in the system and run out. At this level, the entire network as a whole is considered as a processor-sharing system. There is a connection between the policies of the scheduler at different scales of consideration, it is partially studied, for example, in [35].

This dissertation explores a problem at the junction of different scales. The model we are about to present is explored in [36] by the author of the thesis. In this model, a single network node can be considered, or an entire network serving several job flows in parallel. Consider a system of N queues and one processor of power 1 (see fig. 1). There is a constraint M describing

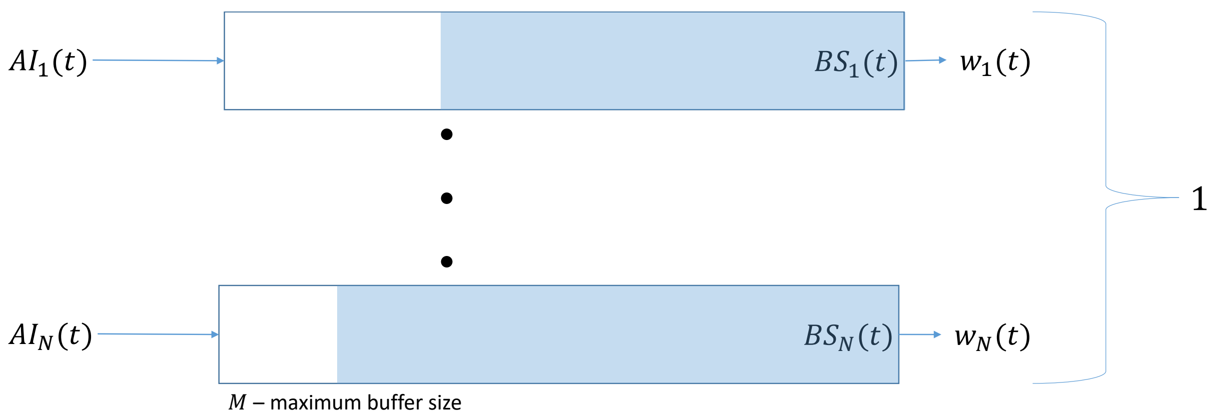

Figure 1: Job scheduler model

the maximum size of a single queue. In other words, M is the size of the buffer allocated to a single queue. If the buffer size is exceeded, the job is lost. At any given time moment, the processor determines how to allocate its processing power between the queues. The processor is assumed to have a high number of logical threads, so that it can perform several jobs, i.e. serve several queues, at the same time. But within a single queue, jobs are carried out strictly in sequence. The queues are formed by jobs entering them from outside the system. Jobs coming in from a particular flow j (the flow of jobs coming into the queue j) represent a continuous stochastic process $\mathcal{A}_j(t)$. Suppose that all jobs are relatively small in size and arrive in large numbers. Then we can look at each jobs stream as a flow of a fluid pouring at a speed of $AI_j(t)$, with speed changes corresponding to a particular trajectory of the process $\mathcal{A}_j(t)$.

Specific vector of computational distribution resource $w = (w_1, \ldots, w_N)$, where $w_1 + \cdots + w_N = 1$, describes the behaviour of the system at a particular point in time, and the vector field $w(\cdot)$ describes the whole scheduling policy. Researchers usually consider the arrival flows of jobs with stationary properties, e.g. distributed according to Poisson's law or to a more complex

Poisson's law with Markov intensities (the so-called Markov-modulated Poisson process), and try to provide a scheduler policy that is «good» for such flows. In this study we make *local* assumptions for the forthcoming time period of duration T_{upd}, i.e. T_{upd} is the update frequency of the scheduling policy. The reason for this approach is long-term sustainability. Indeed, by acting in the described way, we will not be as dependent on whether the arrivals of the jobs are really well estimated by stationary processes in the long term. Moreover, in recent years, researchers have learned to detect moments of Markov "jumps" with fairly good accuracy ([37]). We can then assume that in the near future, after the decision has been made, the arrival of the jobs in the queue with number i will be close to a Poisson process with intensity a_i. More specifically, we will assume that each $\mathcal{A}_i(t)$ is a Gaussian stochastic process defined as follows:

1. $\mathcal{A}_i(0) = 0$.
2. increments of the process $\mathcal{A}_i(t)$ are independent.
3. for $t_1 < t_2$ is true that $\mathcal{A}_i(t_2) - \mathcal{A}_i(t_1)$ is a Gaussian random variable with the expectation $a_i(t_2 - t_1)$ and the same variance.

This makes sense in terms of real-world applications, where incoming jobs during the upcoming period of time can be modelled by Poisson processes with intensities a_i, but the total capacity and intensity values are large. Then the discrete arrival process is approximated by a smooth Gaussian process, and, up to scale, the total capacity is 1.

Denote by b_i the initial queue size $\mathcal{Q}_i(0)$. The queue size at time t is then described by the random variable

$$\xi_i(t) = \mathcal{A}_i(t) + b_i - w_i t. \tag{6}$$

If we cut the value of $\xi_i(t)$ inside of the segment $[0, M]$, i.e. put 0 where it is less than zero, and M where $\xi_i(t) > M$, then we get a random value $\mathcal{Q}_i(t)$, which is the actual queue size.

For an event $\omega \in \Omega$ and a time moment $u \in [0, T_{upd}]$ let's define a set of $\tau_i(\omega, u) = \{t \in [0, u] : \ \xi_i(\omega, t) \geq M\}$. Data from the queue will be lost at points in time $t \in \tau_i(\omega, u)$. The amount of data that did not fit in queue for the whole set of intervals $\tau_i(\omega, u)$ and have been removed, is assessed by the penalty function

$$L_i(\omega, u) = \int_{\tau_i(\omega,u)} (\xi_i(\omega, t) - M)dt. \tag{7}$$

By analogy, define the set $\beta_i(\omega, u) = \{t \in [0, u] : \ \ \xi_i(\omega, t) \leq 0\}$ and the bonus function

$$B_i(\omega, u) = \int_{\beta_i(\omega,u)} (-\xi_i(\omega, t))dt. \tag{8}$$

The resulting random processes L_i and B_i with the time on $[0, T_{upd}]$ describe the amount of loss and the total downtime bonus for the time up to u.

The mathematical expectation of the reference value $\xi_i(t)$ is

$$s_i(t) = \mathbb{E}\xi_i(t) = b_i + a_i t - w_i t \tag{9}$$

Definition 0.0.1. *The queue size prediction* i at time t is the value $q_i(t)$, given by restricting the value of $s_i(t)$ into the range $[0, M]$.

Note that this value is not equivalent to the calculation of the mathematical expectation of the true queue size $\mathcal{Q}_i(t)$.

Definition 0.0.2. *Expected delay* $D_i(t)$ is the time required to process the predicted queue $q_i(t)$ using the allocated bandwidth. It is based on a function

$$d_i(t) = \frac{a_i - w_i}{w_i} t + \frac{b_i}{w_i} \tag{10}$$

after limiting its values inside the boundaries of $[0, M/w_i]$.

Note that if $\mathcal{Q}_i(t) = 0$, then $D_i(t) = 0$ independent of w_i.

Remark. Note that this method of estimating the delay is not equivalent to the calculation of the mathematical expectation of the delay process $\mathcal{Q}_i(t)/w_i$.

Definition 0.0.3. For each flow we define the *mean local delay*:

$$\mathcal{D}(w_i, a_i, b_i) = \frac{\int_0^{T_{upd}} D_i(t)\ dt}{T_{upd}} \tag{11}$$

In this approach, the values of (a_i, b_i) are assumed to be known, the question of how best to predict the value of a_i is left out of consideration. The control variable is the vector w with the constraint $w \in \mathcal{W}$, where $\mathcal{W}$ — set of points $w = (w_1, \ldots, w_N)$ such that $w_i \geq 0$ and $w_1 + \cdots + w_N = 1$, i.e. the standard N-symplex.

Let us introduce various performance metrics for the entire system, generating the corresponding optimisation problems.

Problem (Minimising the sum of the mean delays)**.**
The problem of minimising the arithmetic mean of mean local delays of each stream (or, equally, their sums, since the summation goes on a fixed set) for a given state of the system:

$$\min_{w \in \mathcal{W}} \left(\sum_1^N \mathcal{D}(w_i, a_i, b_i). \right) \tag{12}$$

Note that this problem is different from the problem of minimisation of the overall mean delay over the entire system.

Problem (Minimax mean delay)**.**
The problem of minimising the maximum of all mean local delays:

$$\min_{w \in \mathcal{W}} \left[\max_i \left(\mathcal{D}(w_i, a_i, b_i) \right) \right]. \tag{13}$$

Further, we narrow down the requirements for the resource allocation vector and the initial state of the entire system, based on some conceptual considerations from the realm of practical requirements. Namely, the concept of *steadiness* of a state is introduced, which has the meaning that for this state, there is at least one resource allocation such that none of the N queues will not qualitatively change their status over the next period of time. We will call such allocations *uniform*. The classification of queue and system statuses is collectively a set of simple, but cumbersome relations on a_i, b_i, M, T_{upd}, so we won't present them here. We will only point out that as a result these conceptual considerations in this model, they acquire a specific form of constraints in the form of inequalities. So in the end the resource allocation vector, in addition to being on the standard $\mathcal{W}$ simplex, must be inside the parallelepiped, which we denote by $\mathcal{R}$. The resulting constraint area is the $(N-1)$-dimensional polytope $\mathcal{P}$ inside the simplex.

Research objectives and results

The aim of the study is to optimise the policy of a complex system in the various models presented above, or finding analytical properties of optimal solutions. Here are the main results of the work.

1 (Taxation, smooth univariate type distribution)*.* In the described taxation model, associated with a probability distribution of economic agents among types θ, consider the following clarifications. First, we will consider θ, l in one-dimensional form, i.e. as numbers instead of vectors. In Chapter 2 it is shown that this does not limit the generality in this model, namely the consideration of vector parameters is reduced to the consideration of the norms of these vectors. Make a substitution and consider the problem in the variables y, θ, rather than θ, l, which translates $f(l)$ into $f(y, \theta)$. Recall that agents are distributed among the types according to the probability measure P on Θ. Let's impose the following conditions on the functions listed:

- $f(y, \theta)$ is a non-negative continuous function where $y \geq 0$, $\theta > 0$ (or θ lies in some interval $[\theta_{min}, \theta_{max}] \subset (0, +\infty)$).
- $f(0, \theta) = 0$ for every θ, function $f(y, \theta)$ increases with y.
- derivative $\partial_\theta f(y, \theta)$ exists and is continuous with respect to θ and decreasing with respect to y.
- there is a derivative $\partial_y f(y, \theta) > 0$.
- $T\colon [0, +\infty) \to [0, +\infty)$ is continuous increasing, $T(0) = 0$ and the function $y \mapsto y - T(y)$ increases. Such functions form the class $\mathcal{T}$.
- For each $\theta > 0$ for sufficiently large values of y we have $y - f(y, \theta) < 0$. Concretely, there is P-integrable locally bounded function $\gamma > 0$, for which
$$y - f(y, \theta) \leq 0 \quad \text{for } y \geq \gamma(\theta).$$

The class $\mathcal{T}$ can be described as the set of all 1-lipschitz functions, which are increasing and equal to zero at zero. Let $y_T(\theta)$ be the minimum point at which the maximum utility (1) is reached. Such a point exists as a consequence of the conditions on f and T. Assume $y_T(0) = 0$.

Let the Borel probability measure of the distribution of agents be given by the density: $P = p\,dx$ on $(0, +\infty)$. We will assume that either the density p is positive on $(0, +\infty)$ or the measure P is concentrated on a segment and the density of p is positive on that segment. Put

$$F(\theta) = P([\theta, +\infty)) = \int_\theta^{+\infty} p(t)\,dt.$$

We investigate the problem of finding the value of

$$J(f, P) = \sup_{T \in \mathcal{T}} R(T), \tag{14}$$

i.e. the tax collected is maximised. Note that it follows from what is written above that $J(f,P) \le \|R\|_{L^1(P)}$.

Denote by $\mathcal{Y}$ the set of all increasing continuous functions $y\colon [0,+\infty) \to [0,+\infty)$, for which $y(0)=0$ and for all $\theta > 0$ hold the inequalities:

$$y(\theta) \le \gamma(\theta), \quad \partial_y f(y(\theta),\theta) \le 1, \quad f(y(\theta),\theta) \le y(\theta).$$

For each function $y \in \mathcal{Y}$ we introduce the right inverse function with the formula

$$v(s) = \sup\{t\colon y(t) \le s\}.$$

The function v also increases and is left-continuous. Moreover, $y(v(s)) = s$ at points from the set of values of y, for such a point $v(s)$ — the maximal point in $y^{-1}(s)$.

The main result of this part of the work is as follows.

Theorem 0.0.4. *The following equality is true*

$$\begin{aligned} J(f,P) = \sup_{y\in\mathcal{Y}} \int \left(y(\theta) - f(y(\theta),\theta) + \int_0^\theta \partial_\theta f(y(\tau),\tau)\, d\tau \right) P(d\theta) = \\ = \sup_{y\in\mathcal{Y}} \int [y(\theta) - f(y(\theta),\theta)]\, P(d\theta) + \int \partial_\theta f(y(\theta),\theta) F(\theta)\, d\theta, \quad (15) \end{aligned}$$

and the supremum can be taken over increasing infinitely differentiable functions y with $y' > 0$. Approximations to the supremum by T can be obtained using mappings of the form

$$T(s) = s - f(s, v(s)) + \int_0^{v(s)} \partial_\theta f(y(\tau),\tau)\, d\tau,$$

where

$$y \in \mathcal{Y}$$

is an infinitely differentiable function with inverse function v.

Using the theorem 0.0.4 a description of the solution to the optimisation problem is obtained in the case

$$f(y, \theta) = \frac{y^2}{2\theta^2} \tag{16}$$

Namely:

$$J(f, P) = \frac{1}{2} \int \frac{p^2(\theta)\theta^3}{p(\theta)\theta + 2F(\theta)} d\theta. \tag{17}$$

2 (Piecewise linear taxes). Consider a piecewise linear taxation function, represented by N linear parts. The corresponding segments are described by the split points $m_1 \leq \ldots \leq m_{N-1}$ and the coefficients $k_1 \leq \ldots \leq k_N$, so that

$$T(y) = \begin{cases} k_1 y, \text{ if } y \leq m_1 \\ k_1 m_1 + k_2(y - m_1), \text{ if } m_1 \leq y \leq m_2 \\ \vdots \\ k_1 m_1 + k_2(m_2 - m_1) + \cdots + k_N(y - m_{N-1}) \text{ at } m_{N-1} \leq y. \end{cases} \tag{18}$$

Then the following statements are true.

Theorem 0.0.5 (On the form of optimal income). *In the case of piecewise linear tax the optimal income $y_{max}(\theta)$ is a piecewise constant function of type, i.e. the half-line $[0, +\infty)$ is divided into consecutive intervals $I_1, \ldots, I_{2N-1}$, where y_{max} is constant on intervals with even numbers and quadratically increases on those with odd numbers.*

Theorem 0.0.6 (On the form of maximal utility). *In the case of piecewise linear tax the maximum utility $U_{max}(\theta)$ is a continuous strictly increasing function of type. Moreover, $[0, +\infty)$ divides into consecutive intervals $I_1, \ldots, I_{2N-1}$, where the function U_{max} is convex on intervals with even numbers and concave on those with odd numbers.*

Note that the problem formulation echoes, but is not the same as the problem explored in [38].

3 (The duality in the monopoly problem)*.* When studying the Φ functional arising in the monopolist, multidimensional screening, and auction problems, the following results were obtained:

1. A new proof of the relation (5) is obtained based on a variant of the minimax principle, not assuming the compactness of one of the spaces
2. The following amplification of (5) has been proven:

$$\sup_{u \in \mathcal{U}(X)} \Phi(u) = \min_{m \preceq \pi} \mathrm{Beck}_{\rho,\varphi^*}(\pi). \tag{19}$$

   That is, the functional on the right side reaches a minimum.

More strictly, true is

Theorem 0.0.7. *Let φ be a convex lower semicontinuous function finite on $X = [0,1]^n$ and equal to $+\infty$ outside X, $m \in \mathcal{M}_0$, where $\mathcal{M}_0$ is the set of measures of finite variation with the property $m(X) = 0$.*

Then the following relation holds (part of the statement is that both the minimum and maximum are reached):

$$\max_{u \in \mathcal{U}(X)} \Phi(u) = \min_{\pi \in \mathcal{M}_0 : m \preceq \pi} \mathrm{Beck}_{\rho,\varphi^*}(\pi),$$

where

$$\Phi(u) = \Big(\int_X u dm - \int_X \varphi(\nabla u) \rho dx \Big), \ \mathrm{Beck}_{\rho,\varphi^*}(\pi) = \inf_{\pi + \mathrm{div}(c \cdot \rho) = 0} \int_X \varphi^*(c) \rho dx.$$

4. When investigating the resource allocation model in the form of a queuing system with resource sharing it is justified to consider as an element of the performance metric the queue size prediction functions. More specifically, proved is

Lemma 0.0.8. *It is true that*

$$\mathbb{E}\xi_i(t) = \mathbb{E}\mathcal{Q}_i(t) - \Delta_{i,1}(t) + \Delta_{i,2}(t), \tag{20}$$

where $\Delta_{i,1}$ *is the time density to calculate downtime bonus, and* $\Delta_{i,2}$ *is the data loss density over time. Concretely, for each* $u \in [0, T_{upd}]$ *it is true that*

$$\begin{aligned} \mathbb{E}B_i(u) &= \int_0^u \Delta_{i,1}(t)dt \\ \mathbb{E}L_i(u) &= \int_0^u \Delta_{i,2}(t)dt. \end{aligned} \tag{21}$$

5. When investigating forecasts in a queuing system model with the joint use of the resource, the following results were obtained.

Theorem 0.0.9 (On optimising of the uniform steadiness). *For a given steady state* (a, b) *the problem of minimising the sum of the mean local delays of all queues in the system among all uniform allocations of the given steady state is equivalent to the problem*

$$\min_{\substack{w_1+\cdots+w_N=1 \\ w_i^{*+}\le w_i \le w_i'}} \left(\frac{c_1}{w_1} + \cdots + \frac{c_N}{w_N} \right), \tag{22}$$

where

$$w_i^* = a_i - \frac{M - b_i}{T_{upd}}, \qquad w_i' = a_i + \frac{b_i}{T_{upd}}, \qquad c_i = \frac{a_i T_{upd}}{2} + b_i$$

and $x^+ = \max(0, x)$ *for real* x. *The algorithm below finds the exact solution to this optimisation problem in a finite number of iterations.*

Algorithm. 1. Take the point v with coordinates $v_i = \frac{\sqrt{c_i}}{\sum_i \sqrt{c_i}}$. If v satisfies the constraints of the problem, then v is the desired solution. Otherwise we move on to the next step.

2. If $N = 2$, the solution is chosen directly between the two ends of the segment, that corresponds to the constraints of the problem.

3. Let $v_{i_1}, \ldots, v_{i_l}$ be the components that violate the problem constraints, while all other components comply with the conditions of the problem. We will look for a solution on one of the $F_1, \ldots, F_l$ – faces of the parallelepiped $\mathcal{R}$, corresponding to the violated restrictions. Each of these faces corresponds to a hyperplane, obtained by fixing one of the variables w_i to w_i^{*+} or w_i'. We carry out the following steps for each face indicated.

   (a) Without loss of generality, assume that we have fixed the component w_N. If $w_N = 0,\ 1$, then the value of the target function on this face is infinite, so there is no point in looking for an optimum on it.

   (b) Otherwise, we have the following problem:

$$\min_{(w_1,\ldots,w_{N-1})\in\mathcal{P}^N_{w_N}} \left(\frac{c_1}{w_1} + \cdots + \frac{c_{N-1}}{w_{N-1}}\right), \tag{23}$$

   where $\mathcal{P}^N_{w_N} \subset \mathbb{R}^{N-1}$ is the set of points $x = (x_1, \ldots, x_{N-1})$ such that $(x_1, \ldots, x_{N-1}, w_N) \in \mathcal{P}$. After setting the new variables $y_i = \frac{x_i}{1-w_N}$ we have $y_1 + \cdots + y_{N-1} = 1$. The problem in variables $(y_1, \ldots, y_{N-1})$ is similar to the initial one. Let's run the algorithm from the beginning for a new problem in a smaller dimension.

4. After performing the previous step recursively on the faces $F_1, \ldots, F_l$ we have a minimum for each of them. Let's compare the values of the target function at these points and choose the minimum.

$\square$

Research methods

The methods used in the current work include probability theory and stochastic processes, real and functional analysis, measure theory, convex optimisation.

Scientific novelty and applicability of results

All of the above results of the study are new. The model explored in Chapters 1 and 2 is widely used in mathematical economics to study income taxation. The results obtained for the smooth version of this model allow in some cases obtain an explicit solution to the problem of maximising the total tax. The monopoly problem arises in questions of economic applications of various kinds. The resource sharing model is mainly applied to optimise various processes of queueing theory. For example we could mention the processing of jobs by the processor, the distribution and processing of incoming packets on the network router, network-wide connection management, etc.

Results approbation

The results of the thesis have been presented at the following conferences and seminars:

1. Conference "New Frontiers in High-Dimensional Probability and Applications to Machine Learning", 12-16 May 2021, Sirius University, Sochi, Russia. Report "Mathematical Problems of Optimal Scheduling".
2. Research Seminar "Stochastic Analysis and its Applications in Economics" led by Professor A.V. Kolesnikov and Professor V.D. Konakov, Higher School of Economics. Moscow, 2022. Report "Tax optimisation and probabilistic models".
3. Joint workshop on Network Theory between IITP RAS and Huawei Russian Research Institutes. A series of reports on the topic "Scheduling

Optimisation" in 2020 and 2021.

Publications

The results of the thesis have been presented in papers [11], [12], [39], published in journals indexed by citation systems Scopus and Web of Science.

1. Paper [11] is co-authored with S.N. Popova and published in the journal "Mathematical Notes" (Scopus Q2).
2. Paper [12] is published without co-authors in the journal "The Bulletin of Irkutsk State University. Series Mathematics". (Scopus Q2).
3. Paper [39] is co-authored with the supervisor A.V. Kolesnikov and is submitted to the journal "Mathematical Notes" (Scopus Q2).

Aside from journal publications the author prepared a preprint [36] published within the *arXiv* system.

Work structure and scope

The thesis is set out on 98 pages and consists of a table of contents, an introduction, four chapters, containing the results of the work and their proofs, the conclusion and the list of references, containing 40 items.

Work content

Chapter 1 investigates the one-dimensional version of the aggregate tax maximisation problem with smoothness conditions, which we will not duplicate here. In its original formulation, the problem consists of two stages of optimisation – each entity first optimises its utilitu, then the total tax is maximised. Under given conditions, however, the problem is reduced to a supremum over a specially constructed class of functions. This makes it easier to get an answer, because this method does not contain an intermediate maximum. In the special case where the cost function is

$$f(y, \theta) = \frac{y^2}{2\theta^2} \tag{24}$$

we get that

$$J(f, P) = \frac{1}{2} \int \frac{p^2(\theta)\theta^3}{p(\theta)\theta + 2F(\theta)} d\theta. \tag{25}$$

Chapter 2 shows for a start that the multidimensional problem is elementarily reduced to one-dimensional. Mostly, however, in chapter 2 we deal with the behaviour of agents in the model with piecewise linear taxation. An explicit view of the agent's optimal labour effort is obtained depending on its type of performance and the parameters of the tax function.

Chapter 3 deals with a dynamic process in a system with limited resources, which immediately distinguishes this formulation of the problem from the aggregate statistics in the first model. Consider a system of N queues, where each queue is formed by the jobs coming from corresponding flow. The flows are in the form of Gaussian processes of given intensities. The single processing centre handles all queues, allocating its resource between the queues at each point in time. After that, the choice of a specific system performance metric is supported by its probabilistic properties. Initial system parameters are classified and for each of the described classes the system optimisation problem is presented in the form of an explicit function minimisation. An algorithm is obtained for one class of system state (so-called uniform steadiness), which gives an exact solution to the optimisation problem. The correctness of the algorithm is proved analytically.

Chapter 4 considers the monopoly problem that arises in many models of mathematical economics. Previous results of this area are improved. In particular, the attainability of the maximum on convex functions in the monopoly problem is proved, as is the attainability of the minimum of the dual problem.

Bibliography

[1] Braess D. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung*, 1968, vol. 12, pp. 258–268. English translation: On a paradox of traffic planning. *Transportation Science*, 2005, vol. 39, no. 4, pp. 446–450. https://doi.org/10.1287/trsc.1050.0127

[2] Kameda H., Altman E., Pourtallier O., Li J., Hosokawa Y. Braess-like paradoxes in distributed computer systems. *IEEE Transactions on Automatic Control*, 2000, vol. 45, no. 9, pp. 1687–1691. https://doi.org/10.1109/9.880619

[3] Alekseev V.M., Tikhomirov V. M., Fomin S. V. Optimal control *(in Russian, Optimalnoe upravlenie). Nauka, M.*, 1979; 432 pp.

[4] Aseev S. M., Velov V. M. Another view of the maximum principle for infinite-horizon optimal control problems in economics *(in Russian, Drugoi vzglyad na printsip maksimuma dlya zadach optimalnogo upravleniya s beskonechnym gorizontom v ekonomike). Uspekhi Mat. Nauk*, 2019. vol. 74, no. 6. pp. 3–54.

[5] Harberger A. The Incidence of the Corporation Income Tax. *Journal of Political Economy*, 1962, vol. 70, pp. 215–240.

[6] Mirrlees J.A. An exploration in the theory of optimum income taxation. *Review of Economic Studies*, 1971, vol. 38, no. 2, pp. 175–208. https://doi.org/10.2307/2296779

[7] Mirrlees J.A. The theory of optimal taxation. *Handbook of Mathematical Economics*, 1986, vol. 3 (ed. by K.J. Arrow and M.D. Intriligator), Chapter 24, pp. 1197–1249. https://doi.org/10.1016/S1573-4382(86)03006-0

[8] Hellwig M.F. Incentive problems with unidimensional hidden characteristics: a unified approach. *Econometrica*, 2010. V. 78, No. 4, 1201–1237.

[9] Saez E. Using elasticities to derive optimal income tax rates. *Review of Economic Studies*, 2001. V. 68. P. 205–229.

[10] Atkinson A.B., Stiglitz J.E. The design of tax structure: Direct versus indirect taxation. *Journal of Public Economics*, 1976, vol. 6, no. 1-2, pp. 55–75. https://doi.org/10.1016/0047-2727(76)90041-4

[11] Bogachev T.V., Popova S.N. On optimization of tax functions. *Mathematical Notes*, 2021, vol. 109, no. 2, pp. 170–179. https://doi.org/10.1134/S000143462101020X

[12] Bogachev T.V. Optimal Behavior of Agents in a Piecewise Linear Taxation Environment. *The Bulletin of Irkutsk State University. Series Mathematics.*, 2022, vol. 42, pp. 17–26. https://doi.org/10.26516/1997-7670.2022.42.17

[13] Steinerberger S., Tsyvinski A. Tax mechanisms and gradient flows. *arXiv*:1904.13276v1.

[14] Sachs D., Tsyvinski A., Werquin N. Nonlinear tax incidence and optimal taxation in general equilibrium. *Econometrica*, 2020, vol. 88, no. 2, pp. 469–493. https://doi.org/10.3982/ECTA14681

[15] Rochet J.-P., Choné P.. Ironing, Sweeping, and Multidimensional Screening. *Econometrica*, 1998, 66(4), pp. 783–826.

[16] Hart S., Reny P.. Implementation of reduced form mechanisms: a simple approach and a new characterization. *Economic Theory Bulletin*, 2015, 3(1), pp. 1–8.

[17] Daskalakis C., Deckelbaum A., Tzamos, C., Strong Duality for a Multiple-Good Monopolist, *Econometrica*, 2017, 85(3), pp. 735–767.

[18] Kolesnikov A., Sandomirskiy F., Tsyvinski A., Zimin A. Beckmann's approach to multi-item multi-bidder auctions. *arXiv*:2203.06837

[19] Beckmann M. A continuous model of transportation. *Econometrica: Journal of the Econometric Society*, 1952, pp. 643–660.

[20] Santambrogio F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

[21] Villani C. Topics in optimal transportation. *Amer. Math. Soc. Providence*, Rhode Island, 2003; xvi+370 p.

[22] Brezic H. Analyse fonctionnelle, 1983

[23] Brasco, L., Petrache, M. A Continuous Model of Transportation Revisited. *J Math Sci*, 2014, 196, pp. 119–137.

[24] Erlang, A. K. A proof of Maxwell's law, the principal proposition in the kinetic theory of gases, 1925. In *Brockmeyer et al, The Life and Works of A. K. Erlang*, pp. 222–226, 1948.

[25] Thomson, W., Tait, P. G. Treatise on Natural Philosophy *Cambridge*, 1879.

[26] Kelly F. P. Network routing. In *Philosophical Transactions of the Royal Society*, Vol 337, Iss. 1647, pages 343–367. 1991.

[27] Raigorodskii A. M. Random graph models. *MCCME Moscow*, 2016.

[28] Avrachenkov K., Dreveton M. Statistical Analysis of Networks. *Boston-Delft: now publishers*, 2022. http://dx.doi.org/10.1561/9781638280514

[29] Kelly F. P., Elena Yudovina. Stochastic networks. In *Cambridge University Press*, 2014.

[30] Asmussen S. Applied Probability and Queues, 2nd edn. Springer New York, 2003. In *Stochastic Modelling and Applied Probability*, Vol 51.

[31] Erlang, A. K. Telephone waiting times, 1920. In *Brockmeyer et al, The Life and Works of A. K. Erlang*, pp. 156–171, 1948.

[32] Kelly F. P. Loss networks. *Annals of Applied Probability*, 1991, vol. 1, no. 3, pp. 319–378.

[33] Bonald T., Massoulié L., Proutiere A. and Virtamo J. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. In *Queueing Systems*, Vol 53, pages 65–84. 2006

[34] Bramson, M. Stability and Heavy Traffic Limits for Queueing Networks. *Springer*, 2006.

[35] Shah, D., Wischik, D. Switched networks with maximum weight policies: fluid approximation and multiplicative state space collapse. *Annals of Applied Probability*, 2012, vol. 22, iss. 1, pp. 70–127. http://dx.doi.org/10.1214/11-aap759

[36] Bogachev T. V. Optimization of the fluid model of scheduling: local predictions. *arXiv*:2209.04745v1, 2021.

[37] Avanzi B., Taylor G., Wong B., Xian A. Modelling and understanding count processes through a Markov-modulated non-homogeneous Poisson process framework. *arXiv*:2003.13888v2, 2020.

[38] Apps P., Long Ngo Van, Rees R. Optimal piecewise linear income taxation. *Journal of Public Economic Theory*, 2014, vol. 16, no. 4, pp. 523–545. https://doi.org/10.1111/jpet.12070

[39] Bogachev T.V., Kolesnikov A.V. On the monopolist problem and its dual. *Mathematical Notes*, 2023 (submitted to be published).

[40] Kolmogorov A. N., Fomin S.V. Elements of the theory of functions and functional analysis *(in Russian, Optimalnoe upravlenie)*. Fourth edition. *Nauka, Moscow*, 1976; 544 pp.