Working Paper

ON NONSMOOTH PROBLEMS OF STOCHASTIC SYSTEMS OPTIMIZATION

Yuri M. ERMOLIEV
Vladimir I. NORKIN

WP-95-96
December 1995

IIASA International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

ON NONSMOOTH PROBLEMS OF STOCHASTIC SYSTEMS OPTIMIZATION

Yuri M. ERMOLIEV
Vladimir I. NORKIN

WP-95-96
December 1995

Working Papers are interim reports on work of the International Institute for Applied Systems Analysis and have received only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.

IIASA

International Institute for Applied Systems Analysis □ A-2361 Laxenburg □ Austria
Telephone: +43 2236 807 □ Fax: +43 2236 71313 □ E-Mail: info@iiasa.ac.at

Abstract

A class of stochastic optimization problems is analyzed that cannot be solved by deterministic and standard stochastic approximation methods. We consider risk control problems, optimization of stochastic networks and discrete event systems, screening irreversible changes, pollution control. The results of Ermoliev, Norkin, Wets [11] are extended to the case of problems involving random variables and general constraints. It is shown that the concept of mollifier subgradient leads to easily implementable computational procedures for stochastic systems with Lipschitz and discontinuous expectation functions. New optimality conditions are formulated enabling to design stochastic search procedures for constrained optimization of discontinuous systems.

Contents

ON NONSMOOTH PROBLEMS OF STOCHASTIC SYSTEMS OPTIMIZATION

Yuri M. ERMOLIEV
Vladimir I. NORKIN

1 Introduction

A tendency towards a stability, the minimization of unbalances and the search for an equilibrium and efficiency are natural features of man-made systems. Optimization is needed on various stages of system analysis: in the collection and reconciliation of initial data, parameters identification, sensitivity analysis and policy assessment. A realistic search of policies enforcing changes for better, for example, reducing vital violations, deviations from limiting resources, social and environmental standards, requires rigorous or heuristic optimization tools. Smooth (classical) optimization techniques have been motivated by applications in mechanics, physics and statistics. The analysis of man-made systems with complex interactions between man, nature and technology calls for new approaches that do not rely on the smooth behavior of the system and exact information on its performance.

In this paper we analyse problems arising in optimization of complex stochastic systems exhibiting nonsmooth behavior, abrupt and possibly catastrophic changes. Nonsmooth and discontinuous behavior is typical for systems undergoing through structural changes and new developments. The discontinuity is an inherent feature of systems with discrete variables (indivisibilities), such as manufacturing systems, communication networks, neural nets. In the impulse control, the discontinuity (the size of a jump) itself is a control variable. The lack of scientific information on gradual changes of a system forces analysts to deal with the so-called data-based models where actual changes are represented as transformations between a discrete set of observable states. In risk control, a possibility of an abrupt change is the nature of the problem. A failure may trigger jumps of the system from one space to another and the main dilemma for a control policy is to "hit or miss" an appropriate point in the evolution of the system in order to prevent irreversibility.

The concept of nonsmooth and abrupt change is emphasized in the study of environmental systems by such notions as critical loads, surprise, chemical time bomb phenomenon. There are excellent reviews of discontinuous, imperfectly reversible change in ecological systems (Holling [24]) and sociotechnical systems (Brooks [5]). The significance of "extreme events" arguments in climate impact studies was developed by Parry [38] and has been summarized by Wigley [55] as follows: "Impacts accrue... not so much from slow fluctuations in the mean, but from the tails of the distributions, from extreme events. In many cases, an extreme can be defined as an event where a... variable exceeds some "threshold". Clark [7] argued that such a nonlinearity requires risk based approaches to asses and control possible impacts and the deviation of extremes from threshold levels may be significantly important.

There are a number of methodological challenges involved in the control of abruptly changing (nonsmooth) stochastic systems. One obvious obstacle is the lack of scientific information on involving uncertainties and thresholds. Often less evident challenge is the lack of analytical tools to assess the propagation of abrupt changes, uncertainties and related risks through the system. The main problem is to analyze the interactive roles played by uncertainties, changes and policy responses across the spectrum of spatial and temporal scales.

In this article we analyse practical problems and key shortcomings of existing methods calling for new evolvements of optimization tools. Behavior of a nonsmooth system at local points may not allow to predict its behavior (in contrast to classical smooth systems) even outside an arbitrary small neighborhood. In such a case, the main idea is to develop approaches relying on a "global view" of systems behavior, or, as Ho [23] argued, bird's-eye viewpoint of system responses. The concept of mollifier subgradients (see Ermoliev, Norkin, Wets [11]) provides such an opportunity.

The rest of the article is organized as follows: In Section 2 some important classes of stochastic systems with nonsmooth performance indicators are analysed. Two types of discontinuities are distinguished: discontinuities of sample performance indicators (sample functions) and discontinuities of expected performance indicators (expectation functions). Section 3 introduces the complexity of nonsmooth problems even in the cases where the interchange of integration and differentiation operations is possible. This case already imposes essential obstacles to deterministic and standard stochastic approximation methods. As we see, the concept of mollifier subgradient enables to bypass local stabilities, optima or equilibria. In particular it allows to use finite-difference approximation type procedures for locally Lipschitz and discontinuous functions. This section and the next discuss also the infinitesimal perturbation analysis for discrete event systems (Ho and Cao [22], Suri [52]). In Section 4 notions of mollifier and cosmic convergence (Rockafellar and Wets [45]) are used to formulate the optimality conditions for discontinuous problems in a form that opens up the way for stochastic search procedures under rather general constraints. Proofs are given in Section 5. Section 6 includes concluding remarks indicating some numerical experiments and further research directions.

2 Nonsmooth Stochastic Systems

Let us consider some important cases of stochastic systems, where performance indicators have nonsmooth character. By analyzing them we identify necessary evolvements for the existing optimization techniques.

2.1 Hit or miss control policy: basic optimization procedures

The main difficulties in the optimization of discontinuous systems are easily illustrated by the following simplest example of "hit or miss" decision problems arising in risk control. Assume that at some point in the evolution of a system (ecosystem, nuclear power plant, economic system), if the policymaker does not intervene and control ongoing processes, it could lead to a "failure" with considerable and possibly irreversible damages. Suppose that the system can be used during time interval $[0, T]$, but the actual life time τ may be shorter: we can shut down the system at time $x \leq T$, or there may be a failure at $\omega \leq x$, hence $\tau = \min(x, \omega)$. Therefore the performance of the system critically depends on two events: "failure" and "no failure". The profit of the system without failure is proportional to τ, but the failure at $\omega \leq x$ leads to high losses with the cost c_f. Suppose ω is distributed on the interval $[0, T]$ with a continuous density function $\mu(\omega)$ and the loss function (performance function) is defined as

$$f(x, \omega) = \begin{cases} -c_p x & \text{if } \ 0 \leq x < \omega, \\ c_f - c_p \omega & \text{if } \ \omega \leq x \leq T. \end{cases}$$

The sample performance function $f(x, \omega)$ is discontinuous with respect to both variables. The expected cost (performance) function which can serve a role of a risk indicator for a feasible solution $x \in [0, T]$ has the form of expectation

$$F(x) = \mathbf{E} f(x, \omega) = \mathbf{E}[f(x, \omega) I_{x<\omega}] + \mathbf{E}[f(x, \omega) I_{x \geq \omega}], \tag{1}$$

where I_A is the indicator function of the event A:

$$I_A = \begin{cases} 1 & \text{if } \omega \in A, \\ 0, & \text{otherwise.} \end{cases}$$

The minimization of $F(x)$ is an example of stochastic optimization problems (see, for example, Ermoliev and Wets [13]). The main complexity is the lack of exact information on $F(x)$. In general problems, function $F(x)$ has the form of a multiple integral with an implicitly given probability distribution.

Let us now use (1) to outline possible approaches to the optimization of stochastic systems. One common approach is to approximate the expectation $F(x)$ by sample mean

$$F^N(x) = \frac{1}{N} \sum_{k=1}^{N} f(x, \omega_k), \tag{2}$$

where ω_k are independent samples of ω. Thus the original problem with the expectation function $F(x)$ is approximated by a deterministic problem with objective function $F^N(x)$ that could be solved, if possible, by a broad variety of deterministic methods. There is a number of shortcomings of this approach.

(i) It cannot be used when the underlying probability distribution depends on decision variable x or functions $f(x,\omega)$ are given implicitly;

(ii) As in problem (1) the sample performance function $f(\cdot,\omega_k)$ is discontinuous, although the expectation $F(x)$ is a continuously differentiable function. Since functions $f(\cdot,\omega_k)$, $k = 1,\ldots,N$, are discontinuous, the function $F^N(x)$ is also discontinuous at each point of a local extremum $x = \omega_k$. The number of extrema tends to infinity if $N \longrightarrow \infty$ and the use of the global optimization techniques is practically impossible.

(iii) The deterministic approximation $F^N(x)$ may destroy the convexity of $F(x)$ as the following example shows:

$$F(x) = \sum_{i=1}^{N} \mathbf{E}a_i(x_i - b_i)^2,$$

where $\omega = \{a_i, b_i\}$ are normally distributed random variables and $\mathbf{E}a_i > 0$. The sample mean approximation $F^N(x)$ may be nonconvex.

(iv) The convergence of $\min F^N(x)$ to $\min F(x)$ as $N \longrightarrow \infty$ is established in all practically important cases. Despite this, the use of gradient type procedures for minimizing $F^N(x)$ may result in local solutions which have nothing in common with local solutions of the original problem. It may occur in cases when the interchange of differentiation and expectation operators is impossible (see Section 3), therefore the use of $F^N(x)$ leads to solution sets described in terms of subdifferentials $\partial f(x,\omega_k)$ which may have no connections with the set $\{x|\ 0 \in \partial F(x)\}$.

Nevertheless, there is a remarkable feature of the performance function (1) which can be successfully utilized in the design of the solution strategy: despite the discontinuity of $f(x,\omega)$ the function $F(x)$ is continuous and smooth. The function $F(x)$ may also be convex, although $f(x,\omega)$ is not convex for some ω. Therefore it is advantageous to use stochastic search procedures dealing directly with the optimization of the original function $F(x)$:

$$x^{k+1} = x^k - \rho_k \xi^k, \quad k = 0, 1, \ldots, \tag{3}$$

where ξ^k is in general a biased statistical estimate (stochastic quasigradient) of the $\nabla F(x^k)$ at current point x^k and ρ_k is a step-size multiplier. Unbiased estimates ξ^k are also called stochastic gradients or subgradients, generalized gradients depending on whether $F(x)$ is a continuously differentiable or nonsmooth function.

Let us note that the gradient of the sample performance function $\nabla f(\cdot,\omega)$ exists everywhere except $x = \omega$. Unfortunately the interchange formula for the gradient and the mathematical expectation is not valid:

$$\nabla F(x) \neq \mathbf{E}[\nabla f(x,\omega)].$$

Indeed the direct differentiation of both sides in (1) yields

$$\begin{aligned}\nabla F(x) &= f(x,x_{-0})\mu(x) + \textstyle\int_0^x \nabla f(x,\omega)\mu(\omega)d\omega \\ &\quad -f(x,x)\mu(x) + \textstyle\int_x^T \nabla f(x,\omega)\mu(\omega)d\omega \\ &= (f(x,x_{-0}) - f(x,x))\mu(x) + \mathbf{E}\nabla f(x,\omega),\end{aligned}$$

where $f(x,x_{-o}) = \lim_{y \to x-o} f(x,y)$. Therefore the discontinuity of $f(x,\omega)$ results in a new additional term in $\nabla f(x,\omega)$ and we have the following unbiased estimate of the gradient $\nabla F(x)$:

$$\xi(x,\omega) = (f(x,x_{-0}) - f(x,x))\mu(x) + \nabla f(x,\omega). \tag{4}$$

The estimate ξ (stochastic gradient) can be used in stochastic methods (3). It is impossible to use straightforward one run stochastic finite-difference approximation of $\nabla f(x,\omega)$:

$$\xi^k = \sum_{j=1}^{n} \frac{1}{\Delta_k}[f(x^k + \Delta_k \mathbf{e}^j, \omega^k) - f(x^k,\omega^k)]\mathbf{e}^j, \tag{5}$$

where $\mathbf{e}^j$ is the ort of j-th coordinate axe. This is due to the additional term $(f(x,x_{-0}) - f(x,x))\mu(x)$ in (4). Of course, it is possible to use the standard stochastic approximation procedure when $F(x)$ is a twice differentiable function:

$$\xi^k = \sum_{j=1}^{n} \frac{1}{\Delta_k}[f(x^k + \Delta_k \mathbf{e}^j, \omega^{kj}) - f(x^k,\omega^{k0})]\mathbf{e}^j, \tag{6}$$

where $\omega^{k0}, \omega^{k1}, \ldots, \omega^{kn}$ are independent samples of ω. Unfortunately, the variance of such stochastic quasigradient ξ^k tends to infinity as $k \longrightarrow \infty$. In contrast the variance of the single-run estimate (5) tends to 0 as $k \longrightarrow \infty$. But this type of procedures is applicable only after introduction of smoothing effects to the problem by adding an auxiliary random variable (see Ermoliev and Gaivoronski [10]). From the general idea of mollifier subgradients (Ermoliev, Norkin and Wets [11]) it roughly follows (see next section) that the single run estimate $\xi^k = \nabla f(x^k,\omega^k)$ and the finite difference approximation (5) can be used in (3) with slight modifications: these vectors must be calculated not at the current point x^k but at a point $\overline{x}^k$ randomly chosen in a neighborhood of x^k. In other words $\overline{x}^k = x^k + \epsilon^k$, where ϵ^k is a stochastic vector such that $||\epsilon^k|| \longrightarrow 0$ as $k \longrightarrow \infty$. We clarify this statement in the next sections.

2.2 Stochastic networks with failures

Consider a network of connected elements which can be in "working" or "not working" states. The network has the entry and the exit and is considered operating if there exists a path from the entry to the exit. Denote by $\tau_i(x,\omega)$ the random time for element i to work without failure, where $x \in R^n$ denotes a vector of control parameters and ω is a vector of uncontrolled random factors. Then lifetime $\tau(x,\omega)$ of the system is expressed through times $\tau_i(x,\omega)$ by means of *max* and *min* operations:

$$\tau(x,\omega) = \max_{P_i \in \mathcal{P}} \min_{e \in P_i} \tau_e(x,\omega),$$

where $\mathcal{P}$ is the set of paths from the entry to the exit of the network; index e denotes an element within a path.

It is obvious that for rather general networks the function $\tau(x,\omega)$ cannot be calculated analytically (it is difficult to enumerate paths) in order to implement deterministic approach (2). But there exists a simple algorithm which allows to calculate $\tau(x,\omega)$ and its (generalized) gradients for each observation (simulation) ω. This information is sufficient for the implementation of stochastic approximation type estimates (6) in the case of continuously differentiable function $\mathbf{E}f(\cdot,\omega)$ when the interchange formula for the gradient and the mathematical expectation is valid. Unfortunately, performance function $\tau(x,\omega)$ is nonsmooth and nonconvex even for smooth functions $\tau_i(x,\omega)$. The following simple examples show that the expected performance function $\mathbf{E}f(\cdot,\omega)$ may also be nondifferentiable even if the distribution of ω has a density.

Example 2.1 *Let* $\tau_1(x,\omega) = \omega_1 f_1(x) + \omega_2$, $\tau_2(x,\omega) = \omega_1 f_2(x) + \omega_2$, *where* $\omega = (\omega_1,\omega_2)$ *is uniformly distributed on* $[0,1]^2$ *and* $f_1(x)$, $f_2(x)$ *are some smooth functions, then*

$$f(x,\omega) = \max\{\tau_1(x,\omega), \tau_2(x,\omega)\} = \omega_1 \max(f_1(x), f_2(x)) + \omega_2,$$

and

$$\mathbf{E}f(x,\omega) = \frac{1}{2}\max(f_1(x), f_2(x)) + \frac{1}{2}.$$

are nondifferentiable functions.

In the above example random variables $\tau_1(x,\omega)$ and $\tau_2(x,\omega)$ were dependent, but the same takes place even when random variables are independent and have the probability density.

Example 2.2 *Let (see Krivulin [28])*

$$\tau_1(x,\omega_1) = \begin{cases} x/2, & \omega_1 \le 1/2, \\ 3/2, & \omega_1 > 1/2; \end{cases}$$

$$\tau_2(x,\omega_2) = \begin{cases} x^2, & \omega_2 \le 1/2, \\ 1, & \omega_2 > 1/2; \end{cases}$$

where $x \in [0,1]$, $\omega = (\omega_1, \omega_2)$ and random variables ω_1, ω_2 are independent and uniformly distributed on the interval $[0,1]$. Then random function

$$\begin{aligned} f(x,\omega) &= \max(\tau_1(x,\omega_1), \tau_2(x,\omega_2)) \\ &= \begin{cases} \max(x/2, x^2), & \omega_1 \leq 1/2,\ \omega_2 \leq 1/2; \\ 1, & \omega_1 \leq 1/2, \omega_2 > 1/2; \\ 3/2, & \omega_1 > 1/2; \end{cases} \end{aligned}$$

and its mathematical expectation

$$F(x) = \mathbf{E}_\omega f(x,\omega) = \frac{1}{4}\max(x/2, x^2) + 1$$

is a nondifferentiable function of x. Observe that in this example functions $\tau_i(x,\omega)$ are continuously differentiable in x but are discontinuous in ω.

The composite function $f(x,\omega)$ defined by means of *max* and *min* operations have a rather complicated nondifferentiable character. The calculation of a subgradient of $f(\cdot,\omega)$ is impossible in the case when the chain rule is not valid. For example, for Lipschitz continuous functions $f_1(\cdot,\omega)$, $f_2(\cdot,\omega)$ if $f(\cdot,\omega) = \max\{f_1(\cdot,\omega), f_2(\cdot,\omega)\}$ then for Clarke's subdifferential $\partial f(\cdot,\omega)$ (see Clarke [6]) we have only the inclusion $\partial f(\cdot,\omega) \subseteq conv\{\partial f_i(\cdot,\omega)|\ f_i(\cdot,\omega) = f(\cdot,\omega)\}$. Therefore the interchange formula $\partial \mathbf{E} f(\cdot,\omega) = \mathbf{E}\partial f(\cdot,\omega)$ is not valid. Despite this we can use the single-run estimate (5) and many other similar estimates (based on the general concept of mollifier subgradients) with x^k substituted by $\overline{x}^k$ (see fomulas (12), (13), (27)).

2.3 A simple transfer line

A transfer line (see Mirzoahmedov [33]) consists of n consequentially connected devices. A customer who enters the line is sequentially served by each device if the last has been switched on beforehand. Denote by x_i the moment of switching on device i, y_i – the moment when the customer leaves device i, $y_0(\omega)$ – the moment the customer comes to the line, $\tau_i(\omega)$ – (random) service time by i-th device. Let a_i and b_i denote costs associated with waiting a unit of time by the customer for switching on i-th device and by operating i-th device for a customer correspondingly. Then random costs of waiting by the customer and device i are calculated as:

$$\begin{aligned} & f^0(\omega) = 0, \\ & f^i(x,y,\omega) = f^{i-1}(x,y,\omega) + max\{a_i(y_i - x_i), b_i(x_i - y_i)\}, \\ & y_i = max\{y_{i-1}, x_i\} + \tau_i(\omega), \qquad i = 1, 2, \ldots, \end{aligned}$$

Therefore functions $f^i(x,y,\omega)$ are again constructed by means of *max* and *min* operations and are nonconvex and nonsmooth. The discontinuous problems are encountered in the case of periodically operating devices or devices which may fail and have to be restored (Ermoliev, Uryas'ev, and Wessels [14]).

2.4 Pollution control under extreme events

A feature common to most of models applied for the design of pollution control polices is the use of transfer coefficients a_{ij} linking the amount of pollution x_i emitted by source i to the resulting pollution concentrations y_j at receptor location j as

$$y_j = \sum_{i=1}^{n} a_{ij} x_i, \quad j = 1, \ldots, m.$$

The coefficients a_{ij} are often computed with Gaussian type diffusion equations. These equations are run over all the possible meteorological conditions, and the outputs are then weighted by the frequencies of occurrence of the meteorological input over a given time interval, yielding average transfer coefficients. The deterministic models determine cost-effective emission strategies subject to achieving exogenously specified environmental goals, such as ambient average standards at receptors. The natural improvement of deterministic models is the inclusion of chance constraints that account for the random nature of meteorological conditions in order to reduce extreme events:

$$\mathbf{Prob}(\sum_{i=1}^{n} a_{ij}(\omega) x_i \leq q_j) \geq p_j, \quad j = 1, \ldots, m,$$

i.e. the probability that the deposition level at each receptor (country) j should not exceed the critical load (threshold) q_j with a given probability (acceptable risk level) p_j. Let us denote

$$A_j(x, \omega) = \{\omega | \sum_{i=1}^{n} a_{ij}(x, \omega) x_i \leq q_j\}.$$

Then these constraints can be rewritten by using the mathematical expectation of a discontinuous function

$$I_{A_j(x,\omega)} = \begin{cases} 1, & x \in A_j(x, \omega), \\ 0, & \text{otherwise.} \end{cases}$$

If there is a finite number of possible values (scenarios) of ω, reflecting prevailing weather conditions, then the function

$$F_j(x) = \mathbf{E} I_{A_j(x,\omega)}$$

is piecewise constant. The gradients of such functions are almost everywhere 0 and hence the conventional optimization techniques cannot be used.

2.5 Screening irreversible changes

There may be different interpretations of the following problem in terms of environmental monitoring or the inspection of engineering systems with deteriorating components.

In a simplified model of cervical cancer screening (Oortmarssen and Ermoliev [37]) a life history sample ω is represented by the following. Suppose there are time moments: T_P – the time of entry into the progressive screen-detectable stage, and T_D – the time of clinical diagnosis of a cancer, $T_D = T_P + Z_{PD}$, where T_P and Z_{PD} are independent non-negative random variables with probability distribution functions F_P and F_{PD}. The disease can be prevented by screening examination at time x such that $T_P \leq x \leq T_D$. In this case, the life expectancy is defined by a random variable T_L with distribution $F_L(t)$. Otherwise the survival time following clinical diagnosis and treatment is described by a nonnegative random variable Z_{DC} with distribution $F_{DC}(t)$. A sample of the life history is $\omega = (T_P, T_D, T_C, T_L)$. The life time gained is defined as

$$f(x,\omega) = \begin{cases} T_L - T_C, & \text{if } T_L > T_C \text{ and } T_P \leq x < T_D \\ 0, & \text{otherwise.} \end{cases}$$

Therefore the expected performance

$$F(x) = \mathbf{E}[q(T_D)|T_P \leq x < T_D]\,\mathbf{Pr}[T_P \leq x < T_D] = \mathbf{E}q(T_D)I_{\{T_P \leq x < T_D\}},$$

where $q(T_D)$ denotes the expected number of lifeyears gained for given T_D. The sample performance is again nondifferentiable and implicitly given function. The additional complexity is that the positive values of these function occur with low probability.

2.6 Queuing networks

A network consisting of L devices which "serve" messages or flows. At any moment the device serves only one message, which is then transferred to another node in accordance with a certain routing procedure. If the device is busy, then the message is waiting in the queue and is served according to the rule: first come - first served.

For each node $i = 1, 2, \ldots, L$ we introduce the following notations: n_i - length of the queue at the initial moment; $\tau_{ij}(x,\omega)$ - (random) service time of message j depending on some control parameter x and uncontrolled (random) parameter ω; $\alpha_{ij}(x,\omega)$ - the time when message j comes to node i; $\beta_{ij}(x,\omega)$ - the time when device i starts to serve message j; $\gamma_{ij}(x,\omega)$ - the time when device i finishes servicing message j; the message routing procedure is given by integer functions $\mu_{ij}(x,\omega)$ defining a destination node for j-th message served at i-th node.

The logic of a node operation is described by the following recurrent relations:

$$\begin{aligned} &\alpha_{i1} = \ldots = \alpha_{in_i} = 0, \quad \beta_{i1} = 0, \\ &\beta_{ik} = \textstyle\sum_{j=1}^{k-1} \tau_{ij}, \gamma_{ik} = \sum_{j=1}^{k} \tau_{ij}, \quad k = 1, \ldots, n_i. \end{aligned}$$

$$\begin{aligned} &\beta_{ij} = \max\{\gamma_{i(j-1)}, \alpha_{ij}\}, \\ &\gamma_{ij} = \beta_{ij} + \tau_{ij} = \max\{\gamma_{i(j-1)}, \alpha_{ij}\} + \tau_{ij}, \quad j = 1, 2, \ldots. \end{aligned}$$

Inflows of messages are modeled by introducing special nodes which have infinite queue and do not receive messages from other nodes.

Notice that each moment α when a message arrives to a given node or a moment β when a message is started to be served coincides with some moment of finishing servicing a message at some node. That's why it is sufficient to consider only moments γ of finishing servicing messages at nodes.

It is interesting enough that important indicators of this network are again nondifferentiable functions despite the continuously differentiable $\tau_{ij}(x,\omega)$.

Theorem 2.1 *Assume $\mu_{ij}(x,\omega) = \mu_{ij}$. Then functions $\gamma_{ij}(x,\omega)$ can be expressed through functions $\tau_{ij}(x,\omega)$ by means of max, min operations and by positive linear combinations.*

For proof see the Appendix.

In order to illustrate the statement of this theorem, assume $\mu_{ij} = \mu_i$. Denote $I_i = \{\text{nodes } r |\ \mu_r = i\}$. Then

$$\gamma_{ik} = \max\left(\gamma_{i(k-1)}, \min_{i' \in I_i, k' \leq k} \max(\alpha_{i(k-1)}, \gamma_{i'k'})\right) + \tau_{ik}. \tag{7}$$

The result of Theorem 2.1 can be extended to more general networks with the following nodes: nodes with parallel servicing by identical devices; nodes like transfer line with limited buffer capacity and blocking; assembling nodes; message splitting nodes; nodes with synchronized servicing.

For the described network main performance indicators (criteria) are expressed through times $\gamma_{ik}(x,\omega)$:

1. total waiting time at node i

$$f_1(x,\omega) = \gamma_{ik}(x,\omega) - \sum_{j=1}^{k} \tau_{ij}(x,\omega);$$

2. mean message presence at node i

$$f_2(x,\omega) = \frac{1}{k}\sum_{j=1}^{k}(\gamma_{ij}(x,\omega) - \alpha_{ij}(x,\omega));$$

3. mean time of waiting for k messages at node i

$$f_3(x,\omega) = \frac{1}{k}\sum_{j=1}^{k}(\beta_{ij}(x,\omega) - \alpha_{ij}(x,\omega));$$

4. maximal presence time of k messages at node i

$$f_4(x,\omega) = \max_{1 \leq j \leq k}(\gamma_{ij}(x,\omega) - \alpha_{ij}(x,\omega));$$

5. maximal waiting time of k messages at node i

$$f_5(x,\omega) = \max_{1\leq j\leq k}(\beta_{ij}(x,\omega) - \alpha_{ij}(x,\omega));$$

6. node utilization coefficient

$$f_6(x,\omega) = \sum_{j=1}^{k} \tau_{ij}(x,\omega)/\gamma_{ik}(x,\omega);$$

7. node idleness coefficient

$$f_7(x,\omega) = (\gamma_{ik}(x,\omega) - \sum_{j=1}^{k} \tau_{ij}(x,\omega))/\gamma_{ik}(x,\omega);$$

8. mean number of messages at node i

$$f_8(x,\omega) = \sum_{j=1}^{k}(\gamma_{ij}(x,\omega) - \alpha_{ij}(x,\omega))/\gamma_{ik}(x,\omega));$$

9. mean queue length at node i

$$f_9(x,\omega) = \sum_{j=1}^{k}(\beta_{ij}(x,\omega) - \alpha_{ij}(x,\omega))/\gamma_{ik}(x,\omega));$$

Thus this theorem shows that the use of standard stochastic approximation methods for the optimization of discrete event systems is generally restricted, since the expected value of indicators 1-9 may easily be a nonsmooth function (see examples 2.1, 2.2). The possibility to use them is extensively discussed by Ho and Cao [22], Suri [52], Gaivoronski [15], Rubinstein and Shapiro [50].

For a general network configuration it is very difficult to express explicitly performance functions $f_m(x,\omega), m = \overline{1,9}$ through functions $\tau_{ij}(x,\omega)$ and apply the deterministic approximation (2). This approach is restricted by the lack of analytical structure of $F^N(x)$. If it were not the case then for optimization of $F^N(x)$ conventional deterministic procedures could be used (see Plambeck *et al.* [41], [42], Robinson [43], Gürkan, Özge and Robinson [20]).

Next sections are devoted to the development of stochastic methods enabling to deal with admitted complexities.

3 Nonsmooth Sample Functions

In this section we further analyse the shortcomings of existing optimization techniques for problems with nonsmooth performance functions and introduce the concept of stochastic mollifier gradients to deal with outlined deficiencies.

3.1 Differentiability of expectation function

Subsection 2.1 shows that nonsmooth sample performance functions do not necessarily lead to nonsmooth expectation functions. Unfortunately, even the case when the interchange of differentiation and integration operations is possible may still be infeasible for the standard optimization procedures. Consider the situation when a variable transformation smoothes discontinuity of the sample performance function and in principle allows to speak about standard approaches.

Consider the following general stochastic optimization problem:

$$\text{minimize} \quad F(x) = \mathbf{E}_\omega f(x,\omega) \tag{8}$$

$$\text{subject to} \quad x \in X \subseteq R^n, \tag{9}$$

where $\omega \in \Omega$, $(\Omega, \Sigma, \mathbf{P})$ is some probability space, $\mathbf{E}_\omega$ denotes symbol of mathematical expectation, $f : X \times \Omega \longrightarrow R^1$ is a random (i.e. measurable in ω under fixed x) integrable function which can be nonconvex, nonsmooth and even discontinuous. Nevertheless, the expectation function $F(x)$ may still be continuously differentiable. The smoothing effects of sample performance are achieved by variable transformations, as the following example illustrates.

Example 3.1 *Suppose that function $f(x,\omega)$ has the following structure*

$$f(x,\omega) = f(\phi_1(x)\omega_1 + \psi_1(x), \ldots, \phi_m(x)\omega_m + \psi_m(x)),$$

where locally integrable function $f(\tau_1, \ldots, \tau_m)$ can be nonsmooth or discontinuous and grows at the infinity not faster than some polynomial of $\tau = (\tau_1, \ldots, \tau_m)$; functions $\phi_i(x) > 0$ and $\psi_i(x)$ are continuously differentiable; random variable $\omega = (\omega_1, \ldots, \omega_m)$ has continuously differentiable density $p(\omega_1, \ldots, \omega_m)$ on a rectangular $\Omega = \{\omega \in R^m | \ a_i \leq \omega_i \leq b_i, \ i = 1, \ldots, m\}$, $p(\omega_1, \ldots, \omega_m)$ tends (together with its derivatives) to zero (in the infinity) faster than any polynomial of ω. Introducing new variables $\tau_i = \phi_i(x)\omega_i + \psi_i(x), \ i = 1, \ldots, m$ we can represent

$$F(x) = \int_\Omega g(x,\omega)p(\omega)d\omega = \int_{T(x)} f(\tau)\rho(x,\tau)d\tau,$$

where

$$\begin{aligned}
\rho(x,\tau) &= p((\tau_1 - \psi_1(x))/\phi_1(x), \ldots, (\tau_m - \psi_m(x))/\phi_m(x)),\\
T(x) &= \{\tau \in R^m | \ \alpha_i(x) \leq \tau_i \leq \beta_i(x), \ i = 1, \ldots, m\},\\
\alpha_i(x) &= \begin{cases} \phi_i(x)a_i + \psi_i(x), & a_i > -\infty,\\ -\infty, & a_i = -\infty; \end{cases}\\
\beta_i(x) &= \begin{cases} \phi_i(x)b_i + \psi_i(x), & b_i < +\infty,\\ +\infty, & b_i = +\infty. \end{cases}
\end{aligned}$$

If $-\infty < \alpha_i(x)$, $\beta_i(x) < +\infty$ and function $f(\tau)$ is continuous then $F(x)$ is continuously differentiable and

$$\nabla F(x) = \int_{T(x)} f(\tau)\nabla\rho(x,\tau)d\tau + \int_{S(x)} f(\tau)\rho(x,\tau)n(x,\tau)ds,$$

where $S(x)$ is a surface of the set $T(x)$, $n(x,\tau)$ is a unit normal vector to the surface $S(x)$ at $\tau \in S(x)$, ds is a symbol of integration over surface $S(x)$.

If $\alpha_i(x) = -\infty$, $\beta_i(x) = +\infty$, $i = 1,\ldots,m$, then mathematical expectation function $F(x)$ is continuously differentiable and (see Schwartz [51], Archetti and Betrò [2], Rubinstein and Shapiro [50])

$$\nabla F(x) = \int_{R^m} f(\tau)\nabla\rho(x,\tau)d\tau.$$

There are many similar results (see Uryas'ev [53], Marti [30], Pflug [39], [40]) for replacing the differentiation from nonsmooth or discontinuous function $f(x,\omega)$ to continuously differentiable density $p(\omega)$ leading to differentiability of $F(x)$.

The admitted formula for $\nabla F(x)$ suggests to use the deterministic approximation (2) to the function

$$F(x) = \int_{R^m} f(\tau)\rho(x,\tau)d\tau,$$

that may have a serious obstacle since the replacement of decision variable x from function f to the density ρ in general destroys convexity of the sample function. Hence the sample mean approximation (2) and deterministic methods may produce local solutions that have nothing in common with the solution of the original optimization problem.

The interchange formula is sometimes also derived from the following simple fact (see, for instance, Krivulin [28], Glasserman [16], Rubinstein and Shapiro [50]).

If $f(x,\omega)$ is

(a) differentiable at given x for almost all ω and

(b) satisfies local Lipschitz condition in x with integrable in ω Lipschitz constant,

then

$$\nabla F(x) = \mathbf{E}_\omega \nabla f(x,\omega). \tag{10}$$

The proof follows from Lebesgue's dominant convergence theorem. The following important result Krivulin [28], [29] provides a calculus of nonsmooth functions $f(x,\omega)$ satisfying **(a)**, **(b)**, including *min, max* operations which is essential for stochastic networks and queuing systems (Sections 2.2, 2.6).

Proposition 3.1 *(Krivulin [28]). Let random functions $f(x,\omega)$, $g(x,\omega)$ be a.s. differentiable and satisfy local Lipschitz condition in x with integrable (in ω) Lipschitz constants. Assume that for fixed x random variables $f(x,\omega)$, $g(x,\omega)$ are independent (for instance, depend on different components of ω) and at least one of them has a continuous distribution function. Then functions* $\max(f(x,\omega),g(x,\omega))$ *and* $\min(f(x,\omega),g(x,\omega))$ *satisfy conditions* **(a)**, **(b)**.

Obviously, the sum, difference, product, fraction and superposition of functions satisfying conditions **(a)**, **(b)** also satisfy **(a)**, **(b)** under not strong additional assumptions on Lipschitz constants.

As stated in Theorem 2.1, for queuing networks performance functions satisfy conditions **(a)**, **(b)** and it is possible to calculate a stochastic gradient $\nabla f(x,\omega)$ satisfying (10). This approach corresponds to the infinitesimal perturbation analysis for discrete event systems (see Ho and Cao [22], Suri [52]), invoking the following questions.

Firstly, if (10) is valid but $f(x,\omega)$ is not continuously differentiable, then the convergence of stochastic gradient method (3) with $\xi^k = \nabla f(x^k,\omega)$ can only be studied within a general context of nonsmooth optimization techniques considered further.

Secondly, Examples 2.1, 2.2, 3.1, and subsections 2.2, 2.3, 2.6 show practical limitation of conditions ensuring the smoothness (continuous differentiability) of $F(x)$.

3.2 Lipschitzian expectations

Consider now the problem (8) - (9), where $f : R^n \times \Omega \longrightarrow R^1$ is a locally Lipschitz in x function with integrable in square Lipschitz constant $L_X(\omega)$, X is some compact set in R^n, $(\Omega, \Sigma, \mathbf{P})$ is a probability space and $\mathbf{E}_\omega$ is a symbol of a mathematical expectation operator. Then expectation function $F(x)$ is also Lipschitzian with the constant

$$L_X = \int_\omega L_X(\omega)\mathbf{P}(d\omega).$$

Denote $\partial F(x)$, $\partial f(x,\omega)$ Clarke's subdifferentials [6] of Lipschitz in x functions $F(x)$ and $f(x,\omega)$. The main difficulty concerns the estimation of a subgradient from $\partial F(x)$. There is in fact no calculus of such a vector, for example, by using a chain rule. The interchange formula for differentiation and integration operator

$$\partial F(\cdot) = \int_\Omega \partial f(\cdot,\omega)\mathbf{P}(d\omega)$$

is generally not valid and therefore it is impossible to estimate an element from $\partial F(\cdot)$ assuming we can calculate elements of $\partial f(\cdot,\omega)$. Usually only a set $G_f(\cdot,\omega)$ is known containing $\partial f(\cdot,\omega)$.

Let $\phi : R^n \longrightarrow R^1$ be some probability density function on R^n such that $\phi(x) = 0$ outside some bounded set in R^n. Consider parametric family of mollifiers (see Ermoliev, Norkin and

Wets [11] and the next section for exact definition)

$$\phi_\alpha(x) = \frac{1}{\alpha^n}\phi(\frac{x}{\alpha}), \qquad \alpha > 0,$$

and a family of smoothed (averaged) functions

$$F_\alpha(x) = \int_{R^n} F(y)\phi_\alpha(y - x)dy, \qquad \alpha > 0.$$

Let us note that $F_\alpha(x)$ incorporates a global information on slopes of function $F(x)$ in a vicinity defined by "weights" $\phi_\alpha(\cdot)$. Functions $F_\alpha(x)$ are considered in optimization theory (see Yudin [54], Hasminski [21], Antonov and Katkovnik [1], Zaharov [56], Katkovnik and Kulchitski [26], Archetti and Betrò [2], Katkovnik [25], Gupal [17], [18], Gupal and Norkin [19], Rubinstein [49], Batuhtin and Maiboroda [4], Mayne and Polak [31], Kreimer and Rubinstein [27], Batuhtin [3], Ermoliev, Norkin and Wets [11]). The convolution with appropriate mollifier improves differentiability, but on the other hand increases computational complexity of resulting problems since it transfers a deterministic function $F(x)$ into an expectation function defined as multiple integral. Therefore, this operation is meaningful only in combination with appropriate stochastic optimization techniques.

If function $\phi(x)$ is continuously differentiable (or a constant inside some convex set and equals to zero outside it) then smoothed functions $F_\alpha(x)$, $\alpha > 0$, are continuously differentiable and uniformly in X converge to $F(x)$ as $\alpha \longrightarrow 0$. Suppose random functions $f(x,\omega)$ are measurable in both variables (x,ω) then

$$F_\alpha(x) = \mathbf{E}_\omega f_\alpha(x,\omega) = \mathbf{E}_\omega \int_{R^n} f(y,\omega)\phi_\alpha(y - x)dy, \qquad \alpha > 0,$$

where

$$f_\alpha(x,\omega) = \int_{R^n} f(y,\omega)\phi_\alpha(y - x)dy, \qquad \alpha > 0.$$

Functions $f_\alpha(x,\omega)$ are Lipschitzian in x (with the same Lipschitz constant $L_X(\omega)$) and even continuously differentiable in x. Therefore, functions $F_\alpha(x)$, $\alpha > 0$, are also continuously differentiable and the following differentiation formula is true

$$\nabla F_\alpha(x) = \mathbf{E}_\omega \nabla f_\alpha(x,\omega). \tag{11}$$

From here one can obtain different representations for $\nabla F_\alpha(x)$ depending on the form of the mollifier. If it is uniform in a cube probability density as in Gupal [17], [18]

$$\phi_\alpha(x) = \begin{cases} \frac{1}{\alpha^n}, & \max_{1\le i \le n}|x_i| \le \alpha/2, \\ 0, & \max_{1\le i \le n}|x_i| > \alpha/2, \end{cases}$$

then

$$\nabla F_\alpha(x) = \mathbf{E}_\omega \int_{-1/2}^{+1/2} d\eta_1 \ldots \int_{-1/2}^{+1/2} d\eta_n \cdot \xi_\alpha(x,\eta,\omega) = \mathbf{E}_\omega \mathbf{E}_\eta \xi_\alpha(x,\eta,\omega), \tag{12}$$

where

$$\xi_\alpha(x, \eta, \omega) =$$
$$\frac{1}{\alpha}\sum_{i=1}^{n}(f(x_1 + \alpha\eta_1, \ldots, x_{i-1} + \alpha\eta_{i-1}, x_i + \tfrac{1}{2}\alpha, x_{i+1} + \alpha\eta_{i+1}, \ldots, x_n + \alpha\eta_n, \omega) -$$
$$f(x_1 + \alpha\eta_1, \ldots, x_{i-1} + \alpha\eta_{i-1}, x_i - \tfrac{1}{2}\alpha, x_{i+1} + \alpha\eta_{i+1}, \ldots, x_n + \alpha\eta_n, \omega))\mathbf{e}_i,$$

$\mathbf{e}_i$ are unite coordinate vectors. It means that $\nabla F_\alpha(x)$ is a mathematical expectation of the finite difference approximation $\xi_\alpha(x, \eta, \omega)$, where ω has distribution $\mathbf{P}$ and $\eta = (\eta_1, \ldots, \eta_n)$ is a random vector with components uniformly distributed on the interval $(-\frac{1}{2}, +\frac{1}{2})$. In other words $\xi_\alpha(x, \eta, \omega)$ is an unbiased estimate of the gradient ∇F_α at point x. Each such vector we can call a stochastic mollifier gradient of $F(x)$. The vector ξ_α requires calculation of function $f(x, \omega)$ at $2n$ points. Of course, there may be various other finite difference estimators for $\nabla F_\alpha(x)$ (see Gupal [18], Katkovnik [25], Rubinstein [47], [48], Ermoliev and Gaivoronski [10] and the next section).

If we know the analytical structure of Lipschitz function $f(\cdot, \omega)$ and its Clarke's subgradient $g(x, \omega)$, then (11) can be rewritten as

$$\nabla F_\alpha(x) = \mathbf{E}_\omega \mathbf{E}_\eta g(x + \alpha\eta, \omega). \tag{13}$$

The stochastic quasi-gradient method of unconstrained ($X = R^n$) optimization of Lipschitzian function $F(x)$ has the form (3) with $\xi^k = g(x^k + \alpha_k\eta^k, \omega^k)$, or $\xi^k = \xi_{\alpha_k}\left(x^k, \eta^k, \omega^k\right)$ and non-negative step multipliers ρ_k and smoothing parameters α_k satisfy conditions:

$$\sum_{k=0}^{\infty} \rho_k = +\infty, \quad \sum_{k=0}^{\infty} \rho_k^2 < +\infty, \tag{14}$$

$$\lim_{k\to\infty} \rho_k = \lim_{k\to\infty} \alpha_k = \lim_{k\to\infty} \rho_k/\alpha_k = \lim_{k\to\infty} |\alpha_k - \alpha_{k+1}|/\rho_k = 0. \tag{15}$$

The procedure uses optimization steps concurrently with approximation steps as proposed in Ermoliev and Nurminski [12] and Katkovnik and Kulchitsky [26].

Theorem 3.1 *(Gupal [18]). Assume that random trajectories $\{x^k\}$ generated by (13) are bounded. Suppose also that the set of function values $F(x)$ on the stationary set $X^* = \{x \in R^n | \, 0 \in \partial F(x)\}$ is finite or countable. Then under the above mentioned conditions cluster points of almost all trajectories $\{x^k\}$ belong to X^* and the sequence $\{F(x^k)\}$ has a limit as $k \longrightarrow \infty$.*

Conditions (14) are typical for standard stochastic approximation type algorithms. Additional requirements (15) are not very restrictive (for instance, $\rho_k = C/k^p$, $\alpha_k = C/k^q$ with $1/2 < p < 1$ and $0 < q < p, C > 0$ satisfy them). Thus procedure (3) with (14), (15) generalizes standard stochastic approximation methods for nonsmooth functions. The case $\xi^k = \xi_{\alpha_k}(x^k, h^k, \omega^k)$ provides a general purpose approach. In the case $\xi^k = g(x^k + d_k h^k, \omega_k)$ there remains a question to answer: how to calculate Clarke's subgradients $g(x, \omega)$ of Lipschitz functions $f(x, \omega)$. Consider an important case.

3.3 Generalized differentiability

The calculus of subgradients (see Clarke [6]) only states that

$$\partial f(x,\omega) \subseteq G_f(x,\omega), \tag{16}$$

where $G_f(x,\omega)$ is some extended subgradient set determined by the structure of f. The equality holds true for a special case of subdifferentially regular functions which does not cover important applications. In many cases, as it was shown in Section 2 we deal not with a general class of Lipschitz functions but with a subclass generated from some basic (continuously differentiable) functions by means of maximum, minimum or smooth transformation operations. Then appropriate Clarke's subgradients can in principle be calculated by the lexicographic method (see Nesterov [34]). Another possibility is to prove that $G_f(x,\omega)$ in (16) is a singleton for almost all x what is true for so-called generalized differentiable functions.

Definition 3.1 *(Norkin [35]) Function $f : R^n \longrightarrow R$ is called generalized differentiable (GD) at $x \in R^n$ if in some vicinity of x there exists upper semicontinuous multivalued mapping G_f with closed convex compact values $G_f(x)$ such that*

$$f(y) = f(x) + < g, y - x > + o(x, y, g), \tag{17}$$

where $< \cdot, \cdot >$ denotes an inner product of two vectors, $g \in G_f(y)$ and the remainder term satisfies the following condition:

$$\lim_k \frac{|o(x, y^k, g^k|}{\|y^k - x\|} = 0 \tag{18}$$

for any sequences $y^k \longrightarrow x$, $g^k \longrightarrow g$, $g^k \in G_f(y^k)$. Function f is called generalized differentiable if it is generalized differentiable at each point $x \in R^n$.

Example 3.2 *For instance, function $|x|$ is generalized differentiable with*

$$G_{|\cdot|}(y) = \begin{cases} +1, & y > 0, \\ [-1, +1] & y = 0, \\ -1, & y < 0 \end{cases}$$

Its expansion at $x = 0$ has the form

$$|y| = |0| + sign(y) \cdot (y - 0) + 0.$$

Generalized differentiable (GD) functions possess the following properties (see Norkin [35], Mikhalevich, Gupal and Norkin [32]): They are locally Lipschitzian; continuously differentiable, convex and concave functions are generalized differentiable; class of GD-functions is closed with respect to *max* and *min* and operators superpositions; there is a calculus of subgradients:

$$G_{\max(f_1,f_2)}(x) = convexhull\{G_{f_i} | \; f_i(x) = \max(f_1(x), f_2(x))\}, \tag{19}$$

and subdifferential $G_{f_0(f_1,\dots,f_m)}$ of a composite function $f_0(f_1,\dots,f_m)$ is calculated by the chain rule; class of GD-functions is closed with respect to taking expectation with $G_F(x) = \mathbf{E}G_f(x,\omega)$ for $F(x) = \mathbf{E}f(x,\omega)$, where $f(\cdot,\omega)$ is a GD-function; $\partial_{Clarke}f(x) \subseteq G_f(x)$ and $G_f(x)$ is a singleton almost everywhere in R^n; the following analog of Newton-Leibnitz formula holds true:

$$F(y) - F(x) = \int_0^1 < g((1-t)x + ty), y - x > dt,$$

where $g((1-t)x + ty) \in G_F((1-t)x + ty)$.

Finally, for minimization of an GD-expectation function $F(x) = \mathbf{E}f(x,\omega)$ over convex set K the following stochastic generalized gradient method is applicable:

$$x^0 \in K, \quad x^{k+1} = \Pi_K(x^k - \rho_k g(\tilde{x}^k,\omega^k)), \;\; k = 0,1,\dots, \tag{20}$$

$$g(\tilde{x}^k,\omega^k) \in G_f(\tilde{x}^k,\omega^k), \tag{21}$$

$$\|\tilde{x}^k - x^k\| \le \delta_k, \;\; \lim_k \delta_k = 0, \tag{22}$$

$$\rho_k \ge 0, \; \lim_k \rho_k = 0, \; \sum_{k=0}^{\infty} \rho_k^2 < +\infty, \; \sum_{k=0}^{\infty} \rho_k = +\infty, \tag{23}$$

where $\Pi(y)$ is orthogonal projection of y on K.

From Section 2 follows that generalized differentiable functions may be important for queuing and other discrete event systems. Therefore we can view calculus (19) together with procedure (20) - (23) as an extension of smooth perturbation analysis (Ho and Cao [22], Suri [52]) to nonsmooth cases.

4 Stochastic Discontinuous Optimization

In this section we extend the results of Ermoliev, Norkin and Wets [11] to discontinuous stochastic optimization problems. These results are essentially based on the notion of discontinuity preventing the system from instantenuous jumps and returns to normal states (strongly lower semicontinuous functions). In the case of stochastic systems this notion requires further elaboration.

4.1 Classes of discontinuous functions

Assume that objective function $F(x)$ of problem (8) - (9) is lower semicontinuous (lsc) that guarantees the existence of the solution.

Definition 4.1 *A function $F: R^n \longrightarrow R^1$ is called strongly lower semicontinuous at x, if it is lower semicontinuous at x and there exists a sequence $x^k \longrightarrow x$ with F continuous at x^k (for all k) such that $F(x^k) \longrightarrow F(x)$. The function F is called strongly lower semicontinuous on $X \subseteq R^n$ if this holds for all $x \in X$.*

To give a sufficient condition for mathematical expectation $F(x) = \mathbf{E}f(x, \omega)$ to be strongly lower semicontinuous we introduce subclasses of directionally continuous and piecewise continuous functions.

Definition 4.2 *Lower semicontinuous function $F : R^n \longrightarrow R^1$ is called directionally continuous at x if there exists an open (direction) set $D(x)$ such that it contains sequences $x^k \in D(x)$, $x^k \longrightarrow x$ and for each such sequence $F(x^k) \longrightarrow F(x)$. Function $F(x)$ is called directionally continuous if this holds for any $x \in R^n$.*

Definition 4.3 *Function $F(x)$ is called piecewise continuous if for any open set $A \subset R^n$ there is another open set $B \subset A$ on which $F(x)$ is continuous.*

Proposition 4.1 *If function $F(x)$ is piecewise continuous and directionally continuous then it is strongly lower semicontinuous.*

Proof. By definition of piecewise continuity for any open vicinity $V(x)$ of x we can find an open set $B \subset D(x) \cap V(x)$ on which function F is continuous. Hence there exists sequence $x^k \in D(x)$, $x^k \longrightarrow x$ with F continuous at x^k. By definition of directional continuity $F(x^k) \longrightarrow F(x)$.□

Proposition 4.2 *If function $F(x)$ has the form*

$$F(x) = F_0(F_1(x_1), \ldots, F_m(x_m)),$$

where $x = (x_1, \ldots, x_m)$, $x_i \in R^{n_i}$, function $F_0(\cdot)$ is continuous and functions $F_i(x_i)$, $i = 1, \ldots, n$ are strongly lsc (directionally continuous), then the composite function $F(x)$ is also strongly lsc (directionally continuous).

Function $F(x) = F_0(F_1(x), \ldots, F_m(x))$, $x \in R^n$, where $F_0(\cdot)$ is continuous and $F_i(x)$, $i = 1, \ldots, m$, are piecewise continuous, is also piecewise continuous.

The properties of strong lsc, directional continuity and piecewise continuity are preserved under continuous transformations.

Proof is evident.

Proposition 4.3 *Assume function $f(\cdot, \omega)$ is locally bounded around x by an integrable (in ω) function, piecewise continuous around x and a.s. directionally continuous at x with direction set $D(x, \omega) = D(x)$ (not depending on ω). Suppose ω takes only a finite or countable number of values. Then mathematical expectation function $F(x) = \mathbf{E}f(x, \omega)$ is strongly lsc at x.*

For proof see the Appendix.

4.2 Mollifier subgradients

Averaged functions are defined specifically to a family of mollifiers. It is important that, roughly speaking, any family of mollifier can be used for the purpose of optimization. Let us introduce necessary notions and facts which are generalized in the next section to the case of constrained problems.

Definition 4.4 *Given a locally integrable (discontinuous) function $F: R^n \longrightarrow R^1$ and a family of mollifiers $\{\psi_\theta : R^n \longrightarrow R_+,\ \theta \in R_+\}$ that by definition satisfy*

$$\int_{R^n} \psi_\theta(z)dz = 1, \quad supp\psi_\theta := \{z \in R^n |\ \psi_\theta(z) > 0\} \subseteq \rho_\theta \mathbf{B} \text{ with a unit ball } \mathbf{B},\ \rho_\theta \downarrow 0 \text{ as } \theta \downarrow 0,$$

the associated family $\{F_\theta,\ \theta \in R_+\}$ of averaged functions is defined by

$$F_\theta(x) := \int_{R^n} F(x-z)\psi_\theta(z)dz = \int_{R^n} F(z)\psi_\theta(x-z)dz. \tag{24}$$

Mollifiers may also have unbounded support. If function $F(x)$ grows in the infinity not faster than some polynomial of x then one can take mollifiers with infinite support like

$$\psi_\theta(z) = \frac{1}{(\sqrt{2\pi}\theta)^n} \exp\left(-\frac{|y|^2}{2\theta^2}\right),$$

which tends to zero in the infinity faster than any polynomial of y. In this case an estimate of gradient $\nabla F_\theta(x)$ has a particular simple form (see Archetti and Betrò [2], Rubinstein [47], Schwartz [51] for justification of the under integral differentiation).

Assume now $F(x) = Ef(x,\omega)$. If $f(x,\omega)$ is such that $\mathbf{E}_\omega|f(x,\omega)|$ exists and grows in the infinity not faster than some polynom of x and random vector η has standard normal distribution, then for

$$\xi_\theta(x,\eta,\omega) = \frac{1}{\theta}[f(x+\theta\eta,\omega) - f(x,\omega)]\eta \tag{25}$$

or

$$\xi_\theta(x,\eta,\omega) = \frac{1}{2\theta}[f(x+\theta\eta,\omega) - f(x-\theta\eta,\omega)]\eta, \tag{26}$$

we have

$$\nabla F_\theta(x) = \mathbf{E}_{\eta\omega}\xi_\theta(x,\eta,\omega), \tag{27}$$

$\mathbf{E}_{\eta\omega}$ denotes mathematical expectation over joint random variable (η,ω). The finite difference approximations $\xi_\theta(x,\eta,\omega)$ are unbiased estimates of $\nabla F_\theta(x)$. As before, we can call them stochastic mollifier gradient of $F(x)$.

Definition 4.5 *(See, for example, Rockafellar and Wets [44], [46]). A sequence of functions* $\{F^k : R^n \longrightarrow \overline{R}\}$ *epi-converges to* $F : R^n \longrightarrow \overline{R}$ *relative to* $X \subseteq R^n$ *if for any* $x \in X$

(i) $\liminf_{k\to\infty} F^k(x^k) \geq F(x)$ *for all* $x^k \longrightarrow x,\ x^k \in X$;

(ii) $\lim_{k\to\infty} F^k(x^k) = F(x)$ *for some sequence* $x^k \longrightarrow x,\ x^k \in X$.

The sequence $\{F^k\}$ *epi-converges to* F *if this holds relative to* $X = R^n$.

For example, if $g : R^n \times R^m \longrightarrow \overline{R}$ is (jointly) lsc at $(\overline{x}, \overline{y})$ and is continuous in y at $\overline{y}$, then for any sequence $y^k \longrightarrow \overline{y}$, the corresponding sequence of functions $F^k(\cdot) = g(\cdot, y^k)$ epi-converges to $F(\cdot) = g(\cdot, y)$.

We use further the following important property of epi-convergent functions.

Theorem 4.1 *If sequence of functions* $\{F^k : R^n \longrightarrow \overline{R}\}$ *epi-converges to* $F : R^n \longrightarrow \overline{R}$ *then for any compact* $K \subset R^n$

$$\lim_{\epsilon\downarrow 0}(\liminf_{k}(\inf_{K_\epsilon} F^k)) = \lim_{\epsilon\downarrow 0}(\limsup_{k}(\inf_{K_\epsilon} F^k)) = \inf_{K} F, \tag{28}$$

where

$$K_\epsilon = K + \epsilon \mathbf{B}, \quad \mathbf{B} = \{x \in R^n | \|x\| \leq 1\}. \tag{29}$$

If

$$F^k(x^k_\epsilon) \leq \inf_{K_\epsilon} F^k + \delta_k, \quad x^k_\epsilon \in K_\epsilon, \quad \delta_k \downarrow 0 \; as \; k \longrightarrow \infty,$$

then

$$\limsup_{\epsilon\downarrow 0}(\limsup_{k} x^k_\epsilon) \subseteq argmin_K F, \tag{30}$$

where $(\limsup_k x^k_\epsilon)$ *denotes the set* X_ϵ *of cluster points of the sequence* $\{x^k_\epsilon\}$ *and* $(\limsup_{\epsilon\downarrow 0} X_\epsilon)$ *denotes the set of cluster points of the family* $\{X_\epsilon,\ \epsilon \in R_+\}$ *as* $\epsilon \downarrow 0$.

For proof see the Appendix.

Jointly with Propositions 4.1, 4.3 the following statement gives sufficient conditions for averaged functions to epi-converge to original discontinuous expectation function.

Proposition 4.4 *(Ermoliev et.al [11]). For any strongly lower semicontinuous, locally integrable function* $F : R^n \longrightarrow R$, *any associated sequence of averaged functions* $\{F_{\theta_k},\ \theta_k \downarrow 0\}$ *epi-converges to* F.

Thus in principle we can solve constrained discontinuous optimization problems via epi-convergent approximations. Unfortunately it involves repeated limits, as we can see from (28).

Propositions 4.1, 4.3, 4.4, next theorem 4.2 justify the use of stochastic mollifiers of section 3.2 and such as (25)-(26) in procedure (3).

Definition 4.6 *Let function $F: R^n \longrightarrow R$ be locally integrable and $\{F^k := F_{\theta_k}\}$ be a sequence of averaged functions generated from F by means of the sequence of mollifiers $\{\psi^k := \psi_{\theta_k} : R^n \longrightarrow R\}$ where $\theta_k \downarrow 0$ as $k \longrightarrow \infty$. Assume that the mollifiers are such that the averaged functions F^k are smooth (of class C^1). The set of ψ-mollifier subgradients of F at x is by definition*

$$\partial_\psi F(x) := \limsup_k \{(\nabla F(x^k) \| \; x^k \longrightarrow x\},$$

i.e. $\partial_\psi F(x)$ consists of the cluster points of all possible sequences $\{\nabla F^k(x^k)\}$ such that $x^k \longrightarrow x$.

Theorem 4.2 *(Ermoliev et.al [11]). Suppose that $F: R^n \longrightarrow R$ is strongly lower semicontinuous and locally integrable. Then for any sequence $\{\psi_{\theta_k}\}$ of smooth mollifiers, we have*

$$0 \in \partial_\psi F(x)$$

whenever x is a local minimizer of F.

4.3 Constrained discontinuous optimization

Theorem 4.2 can be used for constrained optimization problems if exact penalties are applicable. Unfortunately, in stochastic system optimization exact values of constraints are often not available. Besides, we also encounter the following difficulties. Consider the problem

$$\min\{\sqrt[3]{x} | \; x \geq 0\}.$$

In any reasonable definition of gradients the gradient of the function $\sqrt[3]{x}$ at point $x = 0$ equals to $+\infty$. Hence to formulate necessary optimality conditions for such kind of constrained optimization problems and possibly involving discontinuities we need a special language which incorporates infinite quantities. An appropriate notion is a cosmic vector space $\overline{R^n}$ introduced by Rockafellar and Wets [45], [46]. Denote $R_+ = \{x \in R | \; x \geq 0\}$ and $\overline{R_+} = R_+ \cup \{+\infty\}$.

Definition 4.7 *Define a (cosmic) space $\overline{R^n}$ as a set of pairs $\overline{x} = (x, a)$, where $x \in R^n$, $\|x\| = 1$ and $a \in \overline{R_+}$. All pairs of the form $(x, 0)$ are considered identical and are denoted as $\overline{0}$.*

A topology in the space $\overline{R^n}$ is defined by means of cosmically convergent sequences.

Definition 4.8 *Sequence $(x_k, a_k) \in \overline{R^n}$ is called (cosmically) convergent to an element $(x, a) \in \overline{R^n}$ (denoted c-lim$_{k \to \infty}(x_k, a_k)$) if either* $\lim_k a_k = a = 0$ *or there exist both limits* $\lim_k x_k \in R^n$, $\lim_k a_k \in \overline{R^n}$ *and* $x = \lim_k x_k$, $a = \lim_k a_k \neq 0$, *i.e.*

$$\text{c-lim}_k(x_k, a_k) \quad = \quad \begin{cases} (\lim_k x_k, \lim_k a_k) & \textit{if} \;\; (\lim_k a_k) < +\infty, \\ (\lim_k x_k, +\infty) & \textit{if} \;\; a_k \longrightarrow +\infty, \\ (\lim_k x_k, +\infty) & \textit{if} \;\; a_k = +\infty. \end{cases}$$

Denote

$$\text{c-Limsup}_k(x_k, a_k) = \{(x,a) \in \overline{R^n} |\ \exists\{k_m\} : (x,a) = \text{c-lim}_{k\to\infty}(x_{k_m}, a_{k_m})\}.$$

For convex set $K \in R^n$ let $N_K(x)$ be a normal cone and

$$\overline{N}_K(x) = \{(y,b) \in \overline{R^n} |\ y \in N_K(x),\ \|y\| = 1,\ b \in \overline{R_+}\}$$

be an extended normal cone to K at point $x \in K$.

Let us now extend the notion of mollifier subgradients (Definition 4.6).

Definition 4.9 *Let function $F : R^n \longrightarrow R$ be locally integrable and $\{F^k := F_{\theta_k}\}$ be a sequence of averaged functions generated from F by convolution with mollifiers $\{\psi^k := \psi_{\theta_k} : R^n \longrightarrow R\}$ where $\theta_k \downarrow 0$ as $k \longrightarrow \infty$. Assume that the mollifiers are such that the averaged functions F^k are smooth (of class C^1). The set of the extended ψ-mollifier subgradients of F at x is by definition*

$$\overline{\partial}_\psi F(x) := \textit{c-Limsup}_k\{(N^k_\nabla(x^k), \|\nabla F^k(x^k)\|)|\ x^k \longrightarrow x\},$$

where

$$N^k_\nabla(x^k) = \begin{cases} \textit{arbitrary unit vector}, & \textit{if } \|\nabla F^k(x^k)\| = 0, \\ \frac{\nabla F^k(x^k)}{\|\nabla F^k(x^k)\|}, & \textit{otherwise}; \end{cases}$$

i.e. $\overline{\partial}_\psi F(x)$ consists of the cluster points of all possible sequences $\{(N^k_\nabla(x^k), \|\nabla F^k(x^k)\|)\}$ such that $x^k \longrightarrow x$. The full (extended) Ψ-mollifier subgradient set is

$$\partial_\psi F(x) := \cup_\psi \partial_\psi F(x)$$

where ψ ranges over all possible sequences of mollifiers that generate smooth averaged functions.

The extended mollifier subdifferential $\overline{\partial}_\psi F(x)$ is always a non-empty closed set.

Now we can formulate necessary optimality conditions for constrained discontinuous optimization problem:

$$\min\{F(x)|\ x \in K\}, \tag{31}$$

where $F(x)$ may have the form of the expectation.

Theorem 4.3 *Let K be a convex closed set in R^n. Assume that a locally integrable function F has a local minimum relative to K at some point $x \in K$ and there is a sequence $x^k \in K$, $x^k \longrightarrow x$ with F continuous at x^k and $F(x^k) \longrightarrow F(x)$. Then, for any sequence $\{\psi^k\}$ of smooth mollifiers, one has*

$$-\overline{\partial}_\psi F(x) \cap \overline{N}_K(x) \neq \emptyset, \tag{32}$$

where $-\overline{\partial}_\psi F(x) = \{(-g,a) \in \overline{R^n} |\ (g,a) \in \overline{\partial}_\psi F(x)\}$.

For proof see the Appendix.

Example 4.1 *Consider an optimization problem*

$$\min\{\sqrt[3]{x} \mid x \geq 0\}.$$

Then we have

$$\overline{\partial}_\psi \sqrt[3]{x}|_{x=0} = (+1, +\infty), \quad \overline{N}_{x\geq 0}(0) = \cup_{a\in\overline{R_+}}(-1, a)$$

and thus

$$-\overline{\partial}_\psi \sqrt[3]{x}|_{x=0} \cap \overline{N}_{x\geq 0}(0) = (-1, +\infty) \neq \emptyset.$$

Next proposition shows that optimality conditions (32) are also satisfied for limits X' of some local minimizers x_ϵ of relaxed problems

$$\min\{F(x) \mid x \in K_\epsilon = K + \epsilon\mathbf{B}\}. \tag{33}$$

Proposition 4.5 *Let x_ϵ be a local minimizer of (33) such that there exists sequence $x_\epsilon^k \longrightarrow x_\epsilon$, $x_\epsilon^k \in K_\epsilon$ with F continuous at x_ϵ^k and $F(x_{\epsilon^k}) \longrightarrow F(x_\epsilon)$ as $k \longrightarrow \infty$. Assume $x_{\epsilon_m} \longrightarrow x$ for some $\epsilon_m \downarrow 0$ as $m \longrightarrow \infty$. Then (32) is satisfied at x.*

Proof follows from Theorem 4.3 and closeness of (extended) mollifier subdifferential mapping $x \longrightarrow \overline{\partial}_\psi F(x)$ and (extended) normal cone mapping $(x, \epsilon) \longrightarrow \overline{N}_{K_\epsilon}(x)$.

Proposition 4.6 *If in (31) the objective function F is strongly lsc and the constraint set K is convex compact then the set X^* of points, satisfying necessary optimality condition (32), is nonempty and contains at least one global minimizer of (31).*

Proof follows from Theorem 4.1 and Proposition 4.4.

Theorem 4.3 and Propositions 4.5, 4.6 immediately give at least the following idea for approximate solving the problem. Let us fix a small smoothing parameter θ and a small constraint relaxation parameter ϵ and instead of original discontinuous optimization problem consider a relaxed smoothed optimization problem:

$$\min[F_\theta(x) \mid x \in K_\epsilon], \tag{34}$$

where F_θ and K_ϵ are defined by (24), (29).

Then stochastic gradient method for solving (34) has the form:

x^0 is an arbitrary starting point;

$$x^{k+1} = \Pi(x^k - \rho_k \xi_\theta(x^k)), \quad k = 0, 1, \ldots; \tag{35}$$

where $\mathbf{E}\{\xi_\theta(x^k)|x^k\} = \nabla F_\theta(x^k)$, $\prod$ denotes the orthogonal projection operator on the convex set K_ϵ, step multipliers ρ_k satisfy (14).

The convergence of stochastic gradient method with projection on a convex compact set for smooth nonconvex objective function F_θ was studied in [8]. Thus Theorem 4.3, Propositions 4.3, 4.5, 4.6, together with procedure (35) justify the use of stochastic gradient (quasi-gradient) methods for general constrained nonsmooth stochastic problems. Vectors $\xi_\theta(x^k)$ can be called stochastic mollifiers gradients similar to subsections 3.2, 4.2.

5 Conclusions

The analysis of practical nonsmooth stochastic problems in Section 2 shows the importance of random search methods to directly confront their inherent complexity. In particular, we mentioned the following reasons.

If expectations are approximated as usual by the sample mean, a deterministic procedure cannot provide an asymptotic convergence rate faster than $\frac{1}{\sqrt{k}}$, where k is the number of samples. This follows from the central limit theorem. Stochastic methods have the same asymptotic rate of convergence.

The deterministic approximation may destroy important features of the original problem such as smoothness and even convexity. The number of local solutions and discontinuities may rapidly increase with the attempt to improve the accuracy of the approximation.

Although the convergence of optimal values of the approximate "geterministic" problem is proved under rather general assumptions, the use of subgradient deterministic methods may produce local solutions that have nothing in common with solutions of the original problem.

Stochastic procedures have the same asymptotic rate. But since they directly confront the stochasticity and complexity of the original problem they avoid the above mentioned features.

We outlined nonsmooth perturbation analysis (Section 3.3) as a possible approach to special cases of practical problems in Section 2. A promising direction seems to be the use of mollifiers (Sections 3.2, 4.2). This concept incorporates two fundamental approaches to the differentiation of "non-classical" functions: the theory of distributions (theory of generalized functions) and nonsmooth analysis. Resulting random search procedures use a global view on "landscape" of performance functions enabling to bypass local solutions and discontinuities. Numerical experiments with realistic discontinuous problems, (Oortmarssen and Ermoliev, [37]), indicate fast convergence to a practically important neighborhood of optimal solutions. The use of mollifiers seems to be important in discrete stochastic optimization (Norkin, Ermoliev and Ruszczyńsky [36]) to calculate upper and lower bounds in stochastic branch and bound method.

Proposed new optimality conditions (Section 4.3) point out on general approaches to the optimization of rather complex stochastic discontinuous systems.

Of course, there are still more questions than answers. For example, further elaboration of optimality conditions, the development of appropriate calculus and classes of computational procedures. An important task is the study of specific classes of problems and the choice of most suitable classes of mollifiers. We hope to answer some of these questions in the near future.

6 Appendix

6.1 Proof of Theorem 2.1

If $\mu_{ij} = \mu_i$ then the proof can be performed by induction in k with sequential use of representation (7). The following proof for a general case is a modification of the proof from Krivulin [29]. Denote

$$T = \{\tau_{ij}, \ \ i = 1, \ldots, L; \ \ j = 1, 2, \ldots\}.$$

Let M be a class of function obtained from functions $\tau_{ij} \in T$ by means of *min*, *max* and *sum* operations. The theorem states that for all k and $i = 1, \ldots, L$ $\gamma_{ik} \in M$.

Let us introduce some notations:

$\Gamma = \{\gamma_{ij}, \ \ i = 1, \ldots, L; \ \ j = 1, 2, \ldots\}$;

$\Gamma^0 = \{\gamma_{ij} | \ n_i > 0, \ \ i = 1, \ldots, L; \ \ j = 1, 2, \ldots, n_i\}$.

$\Gamma_i = \{\gamma_{rs} | \ \mu_{rs} = i; \ \ r = 1, \ldots, L; \ \ s = 1, 2, \ldots\}$;

$j_{ri} = \sup\{j | \ \mu_{rj} = i\}$;

$\Gamma_{ik} = \{\gamma_{rs} \in \Gamma_i | \ s \leq \min(k, j_{ri})\}$;

$\Gamma'_{ik} = \{\gamma'_t = \gamma_{r_t s_t} \in \Gamma_{ik} | \ \gamma'_t \leq \gamma'_{t+1}, \ t = 1, 2, \ldots, t' - 1\}$;

$B_{ik} = \Gamma_{ik} \cup \gamma_{i(k-1)}$;

$\Delta_{ik} = \Gamma \setminus \{\gamma_{ij} | \ j \geq k\}$.

For elements of Γ^0 the statement of the theorem is obviously true.

Choose $\gamma_{ik} \in \Gamma \setminus \Gamma^0$, $k > n_i$. Since $\gamma_{ik} = \max(\gamma_{i(k-1)}, \alpha_{ik}) + \tau_{ik}$, then to prove the theorem one has to show that $\alpha_{ik} \in M$ and $\gamma_{i(k-1)} \in M$.

If $\gamma_{i(k-1)} > \alpha_{ik}$ then γ_{ik} is defined through $\gamma_{i(k-1)}$. Otherwise γ_{ik} is defined through α_{ik}. The moment α_{ik} coincides with $(k - n_i)$-th element in the ordered set Γ'_{ik}. In this case

$$\alpha_{ik} = \min_{1 \leq t_1 < \ldots < t_{k'} \leq t'} \max(\gamma'_{t_1}, \ldots, \gamma'_{t_{k'}}), \quad k' = k - n_i,$$

i.e. $\alpha_{ik} \in M$. Indeed, $\max(\gamma'_{t_1}, \ldots, \gamma'_{t_{k'}}) \geq \max(\gamma'_1, \ldots, \gamma'_{k'})$ and minimum is achieved on the row of the first k' elements of Γ'_{ik}.

Denote $[B_{ik}]$ algebra of objects obtained from $B_{ik} = \Gamma_{ik} \cup \gamma_{i(k-1)}$ by means of *min*, *max* and *sum* operations. It follows from above that γ_{ik} belongs to $[B_{ik} \cap \Delta_{ik}]$. We shall show that any element of $[B_{ik} \cap \Delta_{ik}]$ belongs to M.

Choose any $\gamma_{i_1 k_1} \in B_{ik} \cap \Delta_{ik}$. Then either $\gamma_{i_1 k_1} \in \Gamma^0$ or $\gamma_{i_1 k_1} = \max(\gamma_{i_1 (k_1 - 1)}, \alpha_{i_1 k_1}) + \tau_{i_1 k_1}$. Analogously it follows that $\gamma_{i_1 k_1}$ belongs to $[B_{i_1 k_1} \cap \Delta_{ik} \cap \Delta_{i_1 k_1}]$. For any $\gamma_{i_2 k_2} \in B_{i_1 k_1} \cap \Delta_{ik} \cap \Delta_{i_1 k_1}$, $\gamma_{i_2 k_2}$ belongs to $[B_{i_2 k_2} \setminus \Delta_{ik} \setminus \Delta_{i_1 k_1} \cap \Delta_{i_2 k_2}]$ and so on.

In a similar way we can construct a chain of events γ_{ik}, $\gamma_{i_1 k_1}, \ldots, \gamma_{i_n k_n}$ with $\gamma_{i_n k_n}$ from $[B_{i_n k_n} \cap \Delta_{ik} \cap \ldots \cap \Delta_{i_n k_n}]$. This chain cannot be infinite because in the sequence $\Delta_{ik}, \ldots, \Delta_{i_n k_n}$

each node i' can occur only a finite number of times (once appeared in $\Delta_{i'k'}$ node i' is repeated at most k' times). Therefore for some n we obtain $\gamma_{i_n k_n} \in \Gamma_0$.

The above analysis shows that γ_{ik} is expressed through elements of T by means of *min*, *max* and *sum* operations, i.e. $\gamma_{ik} \in M$.

6.2 Proof of Proposition 4.3

Lower semicontinuity of F follows from Fatu lemma. The convergence of $F(x^k)$ to $F(x)$ for $x^k \longrightarrow x$, $x^k \in D(x)$ follows from Lebesgue's dominant convergence theorem. Hence F is directionally continuous at x in the direction $D(x)$. It remains to show that in any open set $A \subset R^n$ which is close to x there are points of continuity of F. For the case when ω takes finite number of values $\omega_1, \ldots, \omega_m$ with probabilities $p_1, \ldots, p_m$ the function $F(\cdot) = \sum_{i=1}^m p_i f(\cdot, \omega_i)$ is clearly piece-wise continuous. For the case when ω takes a countable number of values there is a sequence of closed balls $B_i \subseteq B_{i-1} \subset A$ convergent to some point $y \in A$ with $f(\cdot, \omega_i)$ continuous on B_i. We shall show that $F(\cdot) = \sum_{i=1}^\infty p_i f(\cdot, \omega_i)$ is continuous at y. By assumption $|f(x, \omega_i)| \leq C_i$ for $x \in A$ and $\sum_{i=1}^\infty p_i C_i < +\infty$. We have

$$\begin{aligned} F(x) - F(y) &= \textstyle\sum_{i=1}^\infty p_i(f(x,\omega_i) - f(y,\omega_i)) \\ &= \textstyle\sum_{i=1}^m p_i(f(x,\omega_i) - f(y,\omega_i)) + \delta_m(x,y); \\ |\delta_m(x,y)| &\leq \textstyle\sum_{i=m+1}^\infty 2p_iC_i \quad x, y \in A. \end{aligned}$$

Then for any $x^k \longrightarrow y$

$$\limsup_k F(x^k) \leq F(y) + \sum_{i=m+1}^\infty 2p_iC_i$$

$$\liminf_k F(x^k) \geq F(y) - \sum_{i=m+1}^\infty 2p_iC_i.$$

Since $\sum_{i=m+1}^\infty 2p_iC_i \longrightarrow 0$ as $m \longrightarrow \infty$ then $\lim_k F(x^k) = F(y)$.□

6.3 Proof of Theorem 4.1

Note that $(\inf_{K_\epsilon} F^k)$ monotonously increases (non decreases) as $\epsilon \downarrow 0$, hence the same holds for $\liminf_{k\to\infty} \inf_{K_\epsilon} F^k$ and $\limsup_{k\to\infty} \inf_{K_\epsilon} F^k$. Thus limits over $\epsilon \downarrow 0$ in (28) exist.

Let us take arbitrary sequence $\epsilon_m \downarrow 0$, indices k_m^s and points x_m^s such that under fixed m

$$\liminf_k (\inf_{K_{\epsilon_m}} F^k) = \lim_{s\to\infty} (\inf_{K_{\epsilon_m}} F^{k_m^s}) = \lim_{s\to\infty} F^{k_m^s}(x_m^s).$$

Thus

$$\lim_{\epsilon\downarrow 0}(\limsup_k (\inf_{K_\epsilon} F^k)) \geq \lim_{\epsilon\downarrow 0}(\liminf_k (\inf_{K_\epsilon} F^k)) = \lim_{m\to\infty} \lim_{s\to\infty} F^{k_m^s}(x_m^s) = \lim_{m\to\infty} F^{k_m^{s_m}}(x_m^{s_m})$$

for some indices s_m. By property (i) of epi-convergence $\lim_{m\rightarrow\infty} F^{k^{s_m}_m}(x^{s_m}_m) \geq \inf_K F$. Hence

$$\lim_{\epsilon\downarrow 0}(\limsup_k(\inf_{K_\epsilon} F^k)) \geq \lim_{\epsilon\downarrow 0}(\liminf_k(\inf_{K_\epsilon} F^k)) \geq \inf_K F.$$

Let us proof the opposite inequality. Since F is lower semicontinuous, then $F(x) = \inf_K F$ for some $x \in K$. By condition (ii) of epi-convergence there exists sequence $x^k \longrightarrow x$ such that $F^k(x^k) \longrightarrow F(x)$. For k sufficiently large $x^k \in K_\epsilon$, hence $\inf_{K_\epsilon} F^k \leq F^k(x^k)$ and

$$\lim_{\epsilon\downarrow 0}(\liminf_k(\inf_{K_\epsilon} F^k)) \leq \lim_{\epsilon\downarrow 0}(\limsup_k(\inf_{K_\epsilon} F^k)) \leq F(x) = \inf_K F.$$

The proof of (28) completed.

Now prove (30). Let $x^k_\epsilon \in K_\epsilon$ and $F^k(x^k_\epsilon) \leq \inf_{K_\epsilon} F^k + \delta_k, \quad \delta_k \downarrow 0$. Denote $X_\epsilon = \limsup_k x^k_\epsilon \subseteq K_\epsilon$. Let $\epsilon_m \downarrow 0$, $x_{\epsilon_m} \in X_{\epsilon_m}$ and $x_{\epsilon_m} \longrightarrow x \in K$ as $m \longrightarrow \infty$. By construction of X_ϵ for each fixed m there exist sequences $x^{k^s_m}_m \longrightarrow x_{\epsilon_m}$ satisfying

$$F^{k^s_m}(x^{k^s_m}_m) \leq \inf_{K_{\epsilon_m}} F^{k^s_m} + \delta_{k^s_m}, \quad \delta_{k^s_m} \downarrow 0 \text{ as } s \longrightarrow \infty.$$

By property (i)

$$F(x_{\epsilon_m}) \leq \liminf_s F^{k^s_m}(x^{k^s_m}_m) \leq \liminf_s(\inf_{K_{\epsilon_m}} F^{k^s_m}) \leq \limsup_k(\inf_{K_{\epsilon_m}} F^k).$$

Due to lower semicontinuity of F and (28) we obtain

$$F(x) \leq \liminf_{m\longrightarrow\infty} F(x_{\epsilon_m}) \leq \liminf_{\epsilon_m\downarrow 0}(\limsup_k(\inf_{K_{\epsilon_m}} F^k)) = \inf_K F,$$

hence $x \in \mathrm{argmin}_K F$, that proves (30).□

6.4 Proof of Proposition 4.5

Let x be a local minimizer of F on K. For a sufficiently small compact neighborhood V of x, define $\phi := F(z) + ||z - x||^2$. The function ϕ achieves its global minimum on $(K \cap V)$ at x. Consider also the averaged functions

$$\phi^k(z) = \int_{R^n} \phi(y - z)\psi^k(y)dy = F^k(z) + \beta^k(x, z),$$

where

$$F^k(z) = \int_{R^n} F(y - z)\psi^k(y)dy,$$

$$\beta^k(x, z) = \int |y - z - x|^2\psi^k(y)dy.$$

In [11] it is shown that (i) functions ϕ^k are continuously differentiable, (ii) they epi-converge to ϕ relative to $K \cap V$ and (iii) their global minimums z^k on $K \cap V$ converge to x as $k \longrightarrow \infty$. For sufficiently large k the following necessary optimality condition is satisfied:

$$-\nabla F^k(z^k) = n(z^k) \in N_K(z^k), \quad z^k \in K.$$

If $\nabla F^{k_m}(z^{k_m}) = 0$ for some $\{z^{k_m} \longrightarrow x\}$ then also $\overline{0} \in \overline{\partial}_\psi F(x)$ and $\overline{0} \in \overline{N}_K(x)$. If $\nabla F^{k_m}(z^{k_m}) \longrightarrow g \neq 0$ for some $\{z^{k_m} \longrightarrow x\}$ then

$$-\frac{\nabla F^{k_m}(z^{k_m})}{||\nabla F^{k_m}(z^{k_m})||} \longrightarrow -\frac{g}{||g||} \in N_K(x),$$

and $(\frac{g}{||g||}, ||g||) \in \overline{\partial}_\psi F(x)$, $(-\frac{g}{||g||}, ||g||) \in \overline{N}_K(x)$. If $\limsup_k ||\nabla F^k(z^k)|| = +\infty$ then for some $\{z^{k_m} \longrightarrow x\}$

$$-\frac{\nabla F^{k_m}(z^{k_m})}{||\nabla F^{k_m}(z^{k_m})||} \longrightarrow -g \in N_K(x),$$

and $(g, +\infty) \in \overline{\partial}_\psi F(x)$, $(-g, +\infty) \in \overline{N}_K(x)$.□

References

[1] Antonov G.E. and Katkovnik V.Ya. (1970), Filtration and smoothing in extremum search problems for multivariable functions, Avtomatika i vychislitelnaya tehnika, N.4, Riga. (In Russian).

[2] Archetti F. and Betrò B. (1975), Convex programming via stochastic regularization, *Quaderni del Dipartimento di Ricerca Operativa e Scienze Statistiche*, N 17, Università di Pisa.

[3] Batuhtin,B.D. (1994), On one approach to solving discontinuous extremal problems, Izvestia AN Rossii. Tehnicheskaia kibernetika (Communications of Russian Academy of Sciences. Technical Cybernetics), No. 3, pp.37-46. (In Russian).

[4] Batuhtin,B.D. and Maiboroda L.A. (1984), *Optimization of discontinuous functions*, Moscow, Nauka. (In Russian).

[5] Brooks, H. (1985), The topology of surprises in technology, institutions and development, In *Sustainable development of the biosphere*, eds. W.C.Clark and R.E.Munn, Cambridge, Cambridge Univ. Press.

[6] Clarke F.H. (1983), *Optimization and Nonsmooth Analysis*, Wiley, NY.

[7] Clark, W.C. (1985), On the Practical Implication of the Carbon Dioxide Question, Working Paper, International Institute fir Applied System Analysis, Laxenburg, Austria.

[8] Dorofeev P.A. (1986), A scheme of iterative minimization methods, U.S.S.R. Comput. Math. Math. Phys., Vol. 26, No. 2, pp.131-136. (In Russian).

[9] Ermoliev Yu.M. (1976), *Methods of Stochastic Programming*, Nauka, Moscow. (In Russian).

[10] Ermoliev Yu. and Gaivoronski A. (1992), Stochastic programming techniques for optimization of discrete event systems, *Annals of Operations Research*, Vol. 39, pp.120-135.

[11] Ermoliev Yu.M, Norkin V.I. and Wets R.J-B. (1995), The minimization of semi-continuous functions: Mollifier subgradients, SIAM J. Contr. and Opt., No.1, pp.149-167.

[12] Ermoliev Yu.M. and Nurminski E.A. (1973), Limit extremal problems, Kibernetika, No. 4, pp. 130-132. (In Russian).

[13] Ermoliev Yu.M. and Wets R.J-B. (Eds.) (1988), *Numerical Techniques for Stochastic Optimization*, Springer, Berlin.

[14] Ermoliev Yu.M., Uryas'ev S. and Wessels J. (1992), On Optimization of Dynamical Material Flow Systems Using Simulation, Working Paper WP-92-76, October 1992, IIASA, Laxenburg, Austria.

[15] Gaivoronski A.A. (1992), Optimization of stochastic discrete event dynamic systems: a survey of some recent results, In: *Simulation and Optimization*, Eds. G.Pflug and U.Dieter, Lecture Notes in Economics and Mathematical Systems 374, Springer-Verlag, pp.24-44.

[16] Glasserman, P. (1991), *Gradient Estimation via Perturbation Analysis*, Kluwer, Norwell, Mass.

[17] Gupal A.M. (1977), On a method for the minimization of almost differentiable functions, Kibernetika, No. 1, pp.114-116. (In Russian, English translation in: Cybernetics, Vol. 13, N. 1).

[18] Gupal A.M. (1979), *Stochastic methods for solving nonsmooth extremal problems*, Naukova dumka, Kiev. (In Russian).

[19] Gupal A.M. and Norkin V.I. (1977), An algorithm for minimization of discontinuous functions, Kibernetika, No. 2, 73-75. (In Russian, English translation in: Cybernetics, Vol. 13, N. 2).

[20] Gürkan G., Özge A.Yo. and Robinson S.M. (1994), Sample-Path Optimization in Simulation, Working paper WP-94-70, Int. Inst. for Appl. System Anal., Laxenburg, Austria.

[21] Hasminski R.Z. (1965), Application of random noise in optimization and recognition problems, Problemy peredachi informatzii, Vol. 1, N. 3. (In Russian).

[22] Ho Y.C. and Cao X.R. (1991), *Discrete Event Dynamic Systems and Perturbation Analysis*, Kluwer, Norwell, Mass.

[23] Ho Y.C. (1994), Heuristics, Rules of Thumb, and the 80/20 Proposition, *IEEE Transactions on Automatic Control*, Vol. 39, N 5, p. 1025-1027.

[24] Holling, C.S. (1985), Resistance of ecosystems: local surprise and global change, In *Sustainable development of the biosphere*, eds. W.C.Clark and R.E.Munn, Cambridge, Cambridge Univ. Press.

[25] Katkovnik V.Ya. (1976), *Linear Estimates and Stochastic Optimization Problems*, Nauka, Moscow. (In Russian).

[26] Katkovnik V.Ya. and Kulchitsky Yu. (1972), Convergence of a class of random search algorithms, *Automat. Remote Control*, No. 8, pp. 1321-1326. (In Russian).

[27] Kreimer J. and Rubinstein R.Y. (1992), Nondifferentiable optimization via smooth approximation: general analytical approach, Annals of Oper. Res., Vol. 39, pp.97-119.

[28] Krivulin N.K. (1990), *Optimization of dynamic discrete event systems through simulations*, Candidate Dissertation, Leningrad, Leningrad university. (In Russian).

[29] Krivulin N.K. (1990), On optimization of complex systems by means of simulations, *Vestnik Leningradskogo Universiteta*, Leningrad, pp.100-102. (In Russian).

[30] Marti K. (1995), Differentiation of Probability Functions: The Transformation Method. *Computers Math. Applications*, Vol. 30, No 3-6, pp. 361-382.

[31] Mayne D.Q. and Polak E. (1984), Nondifferentiable optimization via adaptive smoothing, J. of Opt. Theory and Appl., Vol. 43, pp.601-613.

[32] Mikhalevich V.S., Gupal A.M. and Norkin V.I. (1987), *Methods of nonconvex optimization*, Nauka, Moscow. (In Russian).

[33] Mirzoahmedov F. (1990), The queuing system optimization problem and a numerical method for its solution, *Kibernetika*, No. 3, pp.73-75. (In Russian, English translation in *Cybernetics*, Vol. 26, N. 3).

[34] Nesterov Yu.E. (1989), *Effective Methods in Nonlinear Programming*, Radio & Svyaz, Moscow. (In Russian).

[35] Norkin V.I. (1978), On nonlocal algorithms for optimization of nonsmooth functions, Kibernetika, No. 5, pp. 75-79. (In Russian, English translation in *Cybernetics*, Vol. 14, N. 5).

[36] Norkin V.I., Ermoliev Yu.M. and Ruszczyński (1995). On Optimal Allocation of Indivisibles Under Uncertainty. Working Paper WP-94-021, April 1994, revised October 1995, IIASA, Laxenburg, Austria.

[37] Oortmarssen G. and Ermoliev Yu.M. (1994), Stochastic Optimization of Screening Strategies for Preventing Irreversible Changes, Working paper WP-94-124, Int. Inst. for Appl. System Anal., Laxenburg, Austria.

[38] Parry, M.L. (1978), *Climate change, agriculture and settlement*, Folkestone, U.K. Dawson.

[39] Pflug G.Gh. (1988), Derivatives of probability measures - concepts and applications to the optimization of stochastic systems, in *Discrete event systems: Models and Applications* (P.Varaiya and A.B.Kurzhanski, eds.), *Lecture Notes in Control and Information Sciences*, Springer Verlag, pp.162-178.

[40] Pflug G.Gh. (1990), On-line optimization of of simulated Markovian processes, *Mathematics of Operations Research*, Vol. 15, No. 3, pp.381-3

[41] Plambeck E.L., B.-R. Fu, S.M. Robinson, and R. Suri (1993), Throughput optimization in tandem production lines via nonsmooth programming, In *Proceedings of 1993 Summer Computer Simulation Conference*, ed. J.Schoen, pp. 70-75. Society for Computer Simulation, San Diego, California.

[42] Plambeck E.L., B.-R. Fu, S.M. Robinson, and R. Suri (1994), *Sample-path optimization of convex stochastic performance functions*, Preprint, Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin.

[43] Robinson S.M. (1994), *Analysis of sample-path optimization*, Preprint, Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin.

[44] Rockafellar R.T. and Wets R.J-B. (1984), Variational systems, an introduction, in: *Multifunctions and Integrands*, G.Salinetti, ed., Lecture Notes in Mathematics 1091, Springer-Verlag, Berlin, pp.1-54.

[45] Rockafellar R.T. and Wets R.J-B. (1991), Cosmic convergence, in: *Optimization and Nonlinear Analysis*, eds. A.Ioffe, M.Marcus and S.Reich, Pitman Research Notes in Mathematics Series 244, Longman Scientific & Technical, Essex, U.K., pp. 249-272.

[46] Rockafellar R.T. and Wets R.J-B. (1995), *Variational Analysis*, a monograph to be published in Springer-Verlag.

[47] Rubinstein R.Y. (1981), *Simulation and the Monte-Carlo Method*, John Wiley & Sons, NY.

[48] Rubinstein R.Y. (1986), *Monte-Carlo Optimization, Simulation and Sensitivity of Queuing Networks*, Wiley, NY.

[49] Rubinstein R.Y. (1983), Smoothed functionals in stochastic optimization, Math. Oper. Res, Vol. 8, pp.26-33.

[50] Rubinstein R.Y. and A. Shapiro (1993), *The optimization of discrete event dynamic systems by the score function method*, Wiley, NY.

[51] Schwartz L. (1966), *Théorie des Distributions*, Heerman, Paris.

[52] Suri R. (1989), Perturbation Analysis: The State of the Art and Research Issues Explained via the GI/G/1 Queue, *Proc. of the IEEE*, Vol. 77, No. 1, pp. 114-137.

[53] Uryas'ev S. (1995), Derivatives of Probability Functions and Some Applications. *Annals of Operation Research*, 56, 287-311.

[54] Yudin D.B. (1965), Qualitative methods for analysis of complex systems I, Izvestia AN SSSR, Tehnich. Kibernetika, No. 1. (In Russian).

[55] Wigley, T.M.L. (1985), Impact of extreme events, *Nature*, No. 386, pp.106-107.

[56] Zaharov V.V. (1970), Integral smoothing method in multi-extremal and stochastic problems, Izvestia AN SSSR, Tehnich. Kibernetika, No. 4. (In Russian).