IIASA

International Institute for Applied Systems Analysis • A-2361 Laxenburg • Austria
Tel: +43 2236 807 • Fax: +43 2236 71313 • E-mail: info@iiasa.ac.at • Web: www.iiasa.ac.at

INTERIM REPORT IR-97-021 / April

Stochastic generalized gradient method with application to insurance risk management[a]

Yuri M. Ermoliev (ermoliev@iiasa.ac.at)
Vladimir I. Norkin (norkin@umc.kiev.ua)

[a]We would like to thank Gordon MacDonald and Joanne Linnerooth-Bayer for their helpful comments.

Approved by
GordonMacDonald (macdon@iiasa.ac.at)
Director, *IIASA*

Abstract

Recently [9] we analyzed important classes of nonsmooth and nonconvex risk control problems which can not be solved by standard optimization techniques. The aim of this article is to develop computational procedures enabling us to bypass some of the obstacles identified in this paper. We illustrate this by using insurance risk processes with insolvency (stopping time).

Key words: Discrete event system, stochastic gradient method, generalized differentiable function, risk processes, insurance.

Contents

Stochastic generalized gradient method with application to insurance risk management[1]

Yuri M. Ermoliev (ermoliev@iiasa.ac.at)
Vladimir I. Norkin (norkin@umc.kiev.ua)

1 Introduction

In a rather general form the problems analyzed in [9] can be formulated in the following way:

$$\text{minimize}[F(x) = \mathbf{E}f(x,\theta)] \tag{1}$$

subject to

$$x \in X \subset R^n, \tag{2}$$

where x is a vector of decision (variable), θ is a random parameter, defined on a probability space $(\Theta, \Sigma, \mathbf{P})$, $f(x,\theta)$ is a random performance function, $F(x)$ is the expected performance function, X is a feasible set. The essential feature of the problems is the lack of the analytical structure of $f(\cdot,\theta)$, in particular its highly discontinuous character, which makes the deterministic approximation meaningless:

$$\text{minimize}[F_N(x) = \frac{1}{N}\sum_{i=1}^{N} f(x,\theta_i)] \tag{3}$$

subject to

$$x \in X \subset R^n, \tag{4}$$

where $\theta_i,\ i = 1,\dots,N$, are i.i.d. observations of θ, since $F_N(x)$ lacks analytical structure. Nonconvex and nonsmooth character of random function f, leads to a highly multiextremal nonsmooth and even discontinuous function $F_N(x)$ with local minimums having nothing in common with local minimums of $F(x)$, which can be a continuously differentiable and even convex function. In such a case random search procedures based on direct estimation of $F(x)$ and its derivatives are required. The case of continuously differentiable expectation functions $F(x)$ was considered by Glynn [14], Ho and Cao [17], Suri [26], Gaivoronski [13], Rubinstein and Shapiro [24].

In the case of nonsmooth stochastic systems an important factor is the concept of Lipschitz expectation functions (see Gupal [15], Ermoliev and Gaivoronski [8], Gaivoronski [13]). Moreover, as it was shown in [9] we often 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. These functions belong to the class of so-called generalized differentiable functions. In Section 2 we briefly discuss important insurance risk control problems with such functions. Section 3 introduces formally the class of generalized differentiable functions. In Sections 4, 5 we prove convergence of the deterministic and stochastic generalized gradient methods with orthogonal projection on nonconvex feasible sets. Section 6 concludes.

[1]We would like to thank Gordon MacDonald and Joanne Linnerooth-Bayer for their helpful comments.

2 Insurance risk control processes

Even a simple situation illustrates the complexity of insurance risk control problems. Assume that an insurer has the initial capital x_1. Claims arrive at random time moments $\tau_1,\ \tau_2,\ \ldots$ with random sizes $L_1, L_2, \ldots$. The risk reserve $R(t)$ at time t is the difference between accumulated premium $P(t)$, initial capital x_1 and aggregated claim $C(t)$:

$$R(x,t) = x_1 + P(x_2,t) - C(x_3,t), \quad 0 \le t \le T\ ,$$

where the premium income is $P(x_2,t) = x_2 t$. The aggregated claim

$$C(x_3,t) = \sum_{k=1}^{N(t)} \min\{L_k, x_3\}\ ,$$

where $N(t)$ is the random number of claims in $[0,t)$, x_3 is the variable defined by excess-of-loss reinsurance. The ruin occurs at the random stopping time $\tau(x) = \min\{0 < t \le T :\ \ R(x,t) < 0\}$; if $R(x,t) \ge 0$ for all $t \in [0,T]$ then by convention $\tau(x) = T+1$. The ruin can be mitigated by the choice of policy variables $x = (x_1,\ x_2,\ x_3)$ from a feasible set. Assume that $\tau_1,\ \tau_2, \ldots$ and $L_1, L_2, \ldots$ are defined on some probability space $(\Theta, \Sigma, \mathbf{P})$. An important performance indicator of this process is the following risk function $F(x) = \mathbf{E} f(x,\theta)$, where θ denotes all random variables involved in the problem and

$$f(x,\theta) = R(x,\tau).$$

The function $f(x,\theta)$ is defined by means of *min* and *-min* operations. It becomes more evident from further simplification of the problem. Consider the case of two time epochs: current time moment and the future. For a fixed current policy variable $x = (x_1,\ x_2,\ x_3)$ the future risk reserve

$$R(x) = x_1 + x_2 - \min\{L, x_3\},$$

where L is a random claim. The risk function

$$F(x) = \mathbf{E} f(x,\theta),\ f(x,\theta) = min\{0, x_1 + x_2 - min[L, x_3]\}$$

is nonconvex and nonsmooth. The random function $f(x,\theta)$ is generated by *min* and *-min* operations from linear functions.

Assume now that $\mathbf{Prob}\{R(x,t) = 0\} = 0$ for all x and t (we can always achieve this by adding some independent small random noise with density to $R(x,t)$). Then with probability 1 function $f(x,\theta)$ is generalized differentiable (see next section) with generalized gradients:

$$g(x,\theta) = \begin{cases} \begin{pmatrix} 1 \\ \tau(x) \\ -n(x_3) \end{pmatrix}, & \tau(x) \le T, \\ 0 \in R^3, & \tau(x) > T, \end{cases}$$

where $n(x_3)$ is the number of cases when $L_t > x_3,\ 0 < t \le \tau(x)$.

Stochastic jumping process $R(x,t)$ has a rather complicated structure and purely analytical analysis of its characteristics and appropriate policy variables x is only possible in special cases. In realistic situations parameters of these processes may be time dependent and there may be a variety of policy variables interconnecting different lines of insurance industry. Extreme and catastrophic events such as fires, floods, windstorms, human-made

accidents and disasters produce highly correlated claims, which should be properly diversified in time and space. All these require the analysis of multidimensional interdependent insurance risk processes that is formally often equivalent to the analysis of large number integro-differential equations with "trajectories" depending on policy variables. These equations are analytically tractable only in very special cases. Of course, it is possible to use Monte-Carlo simulation techniques in a straightforward manner for any given collection of policy variables, but unfortunately the number of possible combinations exponentially approaches infinity. For example for 10 policy alternatives (say, levels of contracts with reinsurance) and 10 scenarios the number of combinations is 10^{10}. Procedure (35)-(37) confronts this complexity. It allows us to simulate stochastic processes directly without solving differential equations and generate feedbacks to policy variables after each random simulation forcing these variables to converge towards better values, for example such that decrease insolvencies of companies, increase their profits and satisfactions of individuals. We analyze these aspects in [12].

Let us discuss a simple example. Consider process $R(x,t)$ and assume for the sake of simplicity that variables x_1, x_2 are fixed say $x_1 = R_0, x_2 = a$. Hence the policy variable is the level of contract with reinsurance $x_3 = x$ and let $c(x)$ be related cost. A decrease in x reduces the chance of insolvency but at the same time it increases the cost $c(x)$. Consider the following risk function

$$F(x) = c(x) + r\mathbf{E}R(x, \tau(x)),$$

where the expectation is taken with respect to the randomness involved in τ and r is a risk parameter. The function $F(x)$ reflects in a sense a trade-off between the risk of insolvency and costs on the risk reduction measure x. It is possible to show that for a given $r > 0$ the minimization of $F(x)$ can be viewed as the minimization of $c(x)$ subject to constraint: the probability of insolvency does not exceed a given level. The minimization of $F(x)$ is not in general possible by using standard techniques. Thus deterministic approximation (3) is impossible because $\tau(x)$ is an implicit random function of x. Procedure (35)-(37) starts with a given initial values of reinsurance contract x^0 and sequentially updates this value after each simulation run. Assume x^k is the value of x^0 after k simulations. New value x^{k+1} is calculated as the following. For given x^k the random process $R(x^k, t), 0 \le t \le T$, is simulated and $\tau(x^k)$ is observed. The value x^k is adjusted according to the feedback:

$$x^{k+1} = \left\{ \begin{array}{ll} \min\{0, x^k - \frac{c}{k+1}[c'(x^k) - n(x^k]\}, & \tau(x^x) \le T, \\ \min\{0, x^k - \frac{c}{k+1}c'(x^k)\}, & \tau(x^k) > T, \end{array} \right\}$$

where c is a positive constant. Since the situation $\tau(x^k) < T$ may be rather rare for some levels x^k, special measures are required to increase the frequency of cases $\tau(x^k) \le T$. We discuss it with more details in [12]. After a finite number of adjustments k the value x^k is stabilized around the desirable value. It is important that the number of simulations required for such type adaptive adjustments usually has the same order of magnitude as the estimation of $F(x)$ at a given initial value x^0.

3 Generalized differentiable functions

Let us introduce a class of functions that is closed under operations *min* and *max (-min)* and smooth transformations. Continuous differentiable functions belong to this class. As we can see in section 4, 5 there is simple gradient type procedure for the optimization of such functions.

Definition 3.1 *(Norkin [21]) Function $f : R^n \longrightarrow R$ is called generalized differentiable (GD) at $x \in R^n$ if in a vicinity of x there exists upper semicontinuous multivalued mapping $\overline{\partial} f$ with closed convex compact values $\overline{\partial} f(x)$ such that*

$$f(y) = f(x) + < g, y - x > + o(x, y, g), \tag{5}$$

where $< \cdot, \cdot >$ denotes the inner product of two vectors in R^n, $g \in \overline{\partial} f(y)$ and

$$\lim_k \frac{|o(x, y^k, g^k)|}{\|y^k - x\|} = 0 \tag{6}$$

for any sequences $y^k \longrightarrow x$, $g^k \longrightarrow g$, $g^k \in \overline{\partial} f(y^k)$. The function f is called generalized differentiable if it is generalized differentiable at each point $x \in R^n$; $\overline{\partial} f(x)$ is called a subdifferential of f at x.

Example 3.1 *Function $|x|$, $x \in R$, is generalized differentiable with*

$$\overline{\partial}|x| = \begin{cases} +1, & x > 0, \\ [-1, +1] & x = 0, \\ -1, & x < 0 \end{cases}$$

Its expansion (5) at $x = 0$ has the form

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

Generalized differentiable (GD) functions possess the following properties (see Norkin [21], Mikhalevich, Gupal and Norkin [19]):

They are locally Lipschitzian, but generally not directionally differentiable; continuously differentiable, convex and concave functions are generalized differentiable, gradients and subgradients of these functions can be taken as generalized gradients; class GD-functions is closed with respect to finite *max*, *min* operations and superpositions;

$$\overline{\partial} \max(f_1(x), f_2(x)) = co\{\overline{\partial} f_i(x) |\ \ f_i(x) = \max(f_1(x), f_2(x))\}, \tag{7}$$

and subdifferential $\overline{\partial} f_0(f_1, \ldots, f_m)$ of a composite function $f_0(f_1, \ldots, f_m)$ is calculated by the chain rule; class of GD-functions is closed with respect to taking expectation: $\overline{\partial} F(x) = \mathbf{E}\overline{\partial} f(x, \omega)$ for $F(x) = \mathbf{E} f(x, \omega)$, where $f(\cdot, \omega)$ is a generalized differentiable function. Thus the expectation functions discussed in Section 2 are indeed generalized differentiable; the subdifferential $\overline{\partial} f(x)$ is defined not uniquely, but Clarke subdifferential $\partial f(x)$ always satisfy Definition 3.1, and $\partial f(x) \subseteq \overline{\partial} f(x)$ for any $\overline{\partial} f(x)$ and $\overline{\partial} f(x)$ is a singleton almost everywhere in R^n; some elements of $\partial f(x)$ for a composite function $f(x)$ such as $f(x) = \max(f_1(x), f_2(x))$, $f(x) = \min(f_1(x), f_2(x))$, and $f(x) = f_0(f_1(x), \ldots, f_m(x))$ can be calculated by the lexicographic method (Nesterov [20]); there is the following analog of Newton-Leibnitz formula

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

where $g((1-t)x + ty) \in \overline{\partial} f((1-t)x + ty)$.

These properties of generalized differentiable functions make them suitable for modeling various nonsmooth stochastic systems.

4 Deterministic generalized gradient method with projection on a nonconvex feasible set

Let us first analyze the deterministic procedure to demonstrate the convergence analysis technique. Consider a problem:

$$f(x) \longrightarrow \min_{x \in X}, \tag{8}$$

where

$$X = \{x \in R^n |\ \psi(x) \leq 0\}, \tag{9}$$

$f(x)$ and $\psi(x)$ are generalized differentiable functions. Let $\overline{\partial} f(x)$ and $\overline{\partial}\psi(x)$ be subdifferentials of $f(x)$ and $\psi(x)$, in particular they may coincide with Clarke's subdifferentials $\partial f(x)$ and $\partial\psi(x)$. Assume that

$$\rho(0, \overline{\partial}\psi(x)) = \inf_{g \in \overline{\partial}\psi(x))} \|g\| > 0 \tag{10}$$

for all x such that $\psi(x) = 0$.

The necessary optimality condition for this problem has the form [19]:

$$0 \in \overline{\partial} f(x) + N_X(x),$$

where

$$N_X(x) = \begin{cases} \{\lambda \overline{\partial}\psi(x) |\ \lambda \geq 0\}, & \psi(x) = 0, \\ 0, & \psi(x) < 0. \end{cases}$$

Let $X^* = \{x \in X |\ 0 \in \overline{\partial} f(x) + N_X(x)\}$ and $f^* = \{f(x) |\ x \in X^*\}$. Consider the following conceptual iterative search procedure:

$$x^0 \in X, \tag{11}$$

$$x^{k+1} \in \Pi_X(x^k - \rho_k g^k), \tag{12}$$

$$g^k \in \overline{\partial} f(x^k) \quad k = 0, 1, \ldots, \tag{13}$$

where Π_X is a (multivalued) projection operator on the set X, i.e. $z \in \Pi_X(y)$ iff $y - z \in N_X(z)$; nonnegative numbers ρ_k satisfy conditions

$$\lim_{k \to \infty} \rho_k = 0, \quad \sum_{k=0}^{\infty} \rho_k = \infty. \tag{14}$$

Remark 4.1 *Method (11)-(13) is an extension of the projection subgradient method by Shor, Ermoliev, Polyak (see further references in [1], pp.143-144) for nonconvex problems. Dorofeev [4], [5] studied a similar method for the class of subdifferentially regular (quasidifferentiable) functions, which do not cover important applications (for instance, this class includes convex, weakly convex [22] and max- functions, but does not include concave and min- functions).*

Theorem 4.1 *Sequence $\{x^k\}$ generated by method (11)-(13) converges to the solution of problem (8): minimal in function f cluster points of $\{x^k\}$ belong to X^* and all cluster points of $\{f(x^k\}$ constitute an interval in f^*. If the set f^* does not contain intervals (for instance, f^* is finite or countable), then all cluster points of $\{x^k\}$ belong to a connected subset of X^* and $\{f(x^k)\}$ has a limit in f^*.*

The proof of convergence is based on using nonsmooth nonconvex Lyapunov functions and techniques developed by Nurminski [22], Ermoliev [7], Dorofeev [5], Mikhalevich, Gupal and Norkin [19].

Lemma 4.1 *Assume that* $\lim_{s\to\infty} x^{k_s} = y \overline{\in} X^*$. *Then for any* $\epsilon > 0$ *there must exist indices* $l_s > k_s$ *such that* $\|x^k - y\| \le \epsilon$ *for all* $k \in [k_s, l_s)$ *and*

$$\limsup_s f(x^{l_s}) > f(y) = \lim_s f(x^{k_s}). \tag{15}$$

Proof. Denote $\overline{x}^{k+1} = x^k - \rho_k g^k$ and represent

$$x^{k+1} = \Pi_X(x^k - \rho_k g^k) = x^k - \rho_k(g^k + h^k) = x^k - \rho_k Q^k,$$

where

$$Q^k = g^k + h^k,$$

$$h^k = h^k(\overline{x}^{k+1}) = \frac{1}{\rho_k}(\overline{x}^{k+1} - \Pi_X(\overline{x}^{k+1})) \in N_X(x^{k+1}), \tag{16}$$

Then:

$$\|h^k\| = \frac{1}{\rho_k}\|\overline{x}^{k+1} - \Pi_X(\overline{x}^{k+1})\| \le \frac{1}{\rho_k}\|\overline{x}^{k+1} - x^k\| = \|g^k\|,$$

$$\|Q^k\| = \frac{1}{\rho_k}\|x^{k+1} - x^k\| \le \frac{1}{\rho_k}\|\overline{x}^{k+1} - x^k\| = \|g^k\|.$$

We have to consider two cases: $\psi(y) < 0$ and $\psi(y) = 0$. In the first case for $k \ge k_s$ method (12) operates in a sufficiently small vicinity of y as an unconstrained subgradient method and the statement of the lemma is known (see [21],[19]). In what follows we consider a new case $\psi(y) = 0$ (the case $\psi(x) < 0$ may be considered as a simplification of the case $\psi(y) = 0$). For $y = \lim_s x^{k_s}$ define

$$\mu = \rho(0, \overline{\partial}\psi(y)) = \inf_g\{\|g\| |\ g \in \overline{\partial}\psi(y)\}, \tag{17}$$

$$\nu = \rho(0, \overline{\partial}f(y) + N_X(y)) = \inf_g\{\|g\| |\ g \in (\overline{\partial}f(y) + N_X(y))\}; \tag{18}$$

$$\gamma = \sup_g\{\|g\| |\ g \in \overline{\partial}f(y)\}. \tag{19}$$

Due to upper semicontinuity of $\overline{\partial}f$, $\overline{\partial}\psi(x)$ there exists ϵ_1-vicinity of y such that

$$\sup_{g,z}\{\|g\| |\ g \in \overline{\partial}f(z),\ \ \|z - y\| \le \epsilon_1\} \le 2\gamma = \Gamma; \tag{20}$$

$$\sup_{g,z}\{\|g\| |\ g \in \overline{\partial}\psi(z),\ \ \|z - y\| \le \epsilon_1\} \le 2\gamma = \Gamma; \tag{21}$$

Define

$$\overline{N}(z) = \{g \in N_X(z) |\ \|g\| \le \Gamma\},$$
$$\overline{G}(z) = \overline{\partial}f(z) + \overline{N}(z).$$

Obviously,

$$\rho(0, \overline{G}(y)) = \inf_g\{\|g\| |\ g \in \overline{G}(y)\} \ge \nu.$$

Due to upper semicontinuity of $\overline{\partial}\psi$ and $\overline{G}(y)$ there exists ϵ_2-vicinity ($\epsilon_2 \le \epsilon_1$) of y such that for all z, $\|z - y\| \le \epsilon_2$,

$$\rho(\overline{\partial}\psi(z), \overline{\partial}\psi(y)) \le \mu/2, \tag{22}$$

where $\rho(\cdot,\cdot)$ is the Hausdorff distance between sets.

Due to generalized differentiability of f and ψ for $c = \frac{\nu^2}{64\Gamma(1+2\Gamma/\mu)}$ there exists $\epsilon_3 \le \epsilon_2$ such that for $\|z-y\| \le \epsilon_3$:

$$f(z) \le f(y) + < g, z-y > + c\|z-y\|, \tag{23}$$

$$\begin{aligned} \psi(z) &\le \psi(y) + < d, z-y > + c\|z-y\| \\ &= < d, z-y > + c\|z-y\|, \end{aligned} \tag{24}$$

for all $g \in \overline{\partial} f(z)$, $d \in \overline{\partial}\psi(z)$. Now set $\overline{\epsilon} = \epsilon_3$ and fix some $\epsilon \le \overline{\epsilon}$. Set $\overline{\rho}_1 = \epsilon/(3\Gamma)$. Let $\|y^s - y\| \le \epsilon/3$, and $\rho_s \le \overline{\rho}_1$ for $s \ge S$. Denote

$$m_s = \sup\{m |\ \|x^k - y\| \le 2\epsilon/3 \ \ \forall\ k \in [k_s, m)\}.$$

We now show that $m_s < \infty$ for $s \ge S$. Indeed, if for all k $\|x^k - y\| \le 2\epsilon/3$ then we obtain the contradiction:

$$2\epsilon/3 \ge \|x^k - y\| \ge \|x^k - x^{k_s}\| - \|x^{k_s} - y\| \ge \nu/2 \sum_{r=k_s}^{k-1} \rho_r - \epsilon/3 \longrightarrow \infty$$

as $k \longrightarrow \infty$. Furthermore

$$\|x^{m_s} - y\| \le \|x^{m_s-1} - y\| + \rho_{m_s-1}\|Q_s^{m_s-1}\| \le \epsilon.$$

Since

$$\epsilon/3 \le \|\sum_{k=k_s}^{m_s-1} \rho_k Q^k\| \le \Gamma \sum_{k=k_s}^{m_s-1} \rho_k,$$

then

$$\sum_{k=k_s}^{m_s-1} \rho_k \ge \frac{\epsilon}{3\Gamma}.$$

For $k \in [k_s, m_s]$, $s \in S$, x^k and $g^k \in \overline{\partial} f(x^k)$ from (23) follows that:

$$\begin{aligned} f(x^k) &\le f(y) + < g^k, x^k - y > + c\|x^k - y\| \\ &\le f(y) + < g^k, x^k - x^{k_s} > + c\|x^k - x^{k_s}\| + (\Gamma + c)\|x^{k_s} - y\| \\ &= f(y) + < g^k + h^k, x^k - x^{k_s} > - < h^k, x^k - x^{k_s} > + \\ &\quad c\|x^k - x^{k_s}\| + (\Gamma + c)\|x^{k_s} - y\|, \end{aligned} \tag{25}$$

where h^k is defined by (16), and let us estimate the term $u_k = - < h^k, x^k - x^{k_s} >$. If $\psi(\overline{x}^k) \le 0$ then $h^k = 0$ and $u_k = 0$. Consider the case $\psi(\overline{x}^k) > 0$, i.e. $h^k \ne 0$. Since

$$h^k \in N_X(x^k) = \{\lambda g |\ g \in \overline{\partial}\psi(x^k),\ \lambda \ge 0\},$$

then

$$h^k = \lambda_k d^k, \quad d^k \in \overline{\partial}\psi(x^k), \quad \lambda_k > 0,$$

and

$$0 < \lambda_k = \|h^k\|/\|d^k\| \le \Gamma/(\mu/2) = 2\Gamma/\mu.$$

Substitute $x^k = \Pi_X(\overline{x}^k)$ and d^k into (24):

$$0 = \psi(x^k) \le < d^k, x^k - y > + c\|x^k - y\|. \tag{26}$$

Now multiplying (26) by λ_k, we obtain:

$$\begin{aligned} - < h^k, x^k - y > &\leq \lambda_k c\|x^k - y\| \leq (2c\Gamma/\mu)\|x^k - y\| \\ &\leq (2c\Gamma/\mu)\|x^k - x^{k_s}\| + (2c\Gamma/\mu)\|x^{k_s} - y\|. \end{aligned} \tag{27}$$

Using inequality (27) we can rewrite (25) in the following form:

$$\begin{aligned} f(x^k) &\leq f(y) + < g^k + h^k, x^k - x^{k_s} > + \\ &\quad (1 + 2\Gamma/\mu)c\|x^k - x^{k_s}\| + (\Gamma + c + 2c\Gamma/\mu)\|x^{k_s} - y\|. \end{aligned} \tag{28}$$

Now we have to estimate scalar products

$$< g^k + h^k, x^k - x^{k_s} > = < g^k + h^k, \sum_{i=k_s}^{k-1} (g^i + h^i) > .$$

Lemma 4.2 *(see Mikhalevich, Gupal and Norkin [19]). Let P be a convex set in R^n such that $0 < \gamma_0 \leq \|p\| \leq \Gamma_0 < +\infty$ for all $p \in P$. Then for an arbitrary collection of vectors $\{p^r \in P |\ r = k, \ldots, m\}$ and any collection of non-negative numbers $\{\rho_r \in R^1 |\ r = k, \ldots, m-1\}$ such that*

$$\sum_{r=k}^{m-1} \rho_r \geq \sigma_0 > 0, \quad \sup_{k \leq r \leq m} \rho_r \leq \frac{\sigma_0 \gamma_0^2}{6\Gamma_0^2},$$

there exists index $l \in (k, m]$ such that

$$< p^l, \sum_{r=k}^{l-1} \rho_r p^r / \sum_{r=k}^{l-1} \rho_r > \geq \frac{\gamma_0^2}{4}, \quad \sum_{r=k}^{l-1} \rho_r \geq \frac{\sigma_0 \gamma_0}{3\Gamma_0}.$$

Proof. For completeness we give the proof of the lemma. Let

$$\sum_{r=k}^{t-1} \rho_r < \frac{\gamma\sigma}{3\Gamma} \leq \sum_{r=k}^{t} \rho_r, \quad \sum_{r=k}^{m'-1} \rho_r < \sigma \leq \sum_{r=k}^{m'} \rho_r. \tag{29}$$

Suppose the opposite to the statement of the lemma is true, i.e. for all $l \in (t, m']$

$$\left(p^l, \sum_{r=k}^{l-1} \rho_r p^r \right) \Big/ \sum_{r=k}^{l-1} \rho_r \Bigg) < \frac{\gamma^2}{4}. \tag{30}$$

We have

$$\left\| \sum_{r=k}^{l} \rho_r p^r \right\|^2 = \left\| \sum_{r=k}^{l-1} \rho_r p^r \right\|^2 + 2\rho_l \left(p^l, \sum_{r=k}^{l-1} \rho_r p^r \right) + \rho_l^2 \|p^l\|^2$$

and

$$\left\| \sum_{r=k}^{m'} \rho_r p^r \right\|^2 = \left\| \sum_{r=k}^{t} \rho_r p^r \right\|^2 + 2 \sum_{l=t+1}^{m'} \rho_l \left(p^l, \sum_{r=k}^{l-1} \rho_r p^r \right) + \sum_{l=t+1}^{m'} \rho_l^2 \|p^l\|^2. \tag{31}$$

Substituting (29), (30) into (31), we obtain:

$$\begin{aligned} \gamma^2\sigma^2 &\leq \Gamma^2 \left(\textstyle\sum_{r=k}^{t} \rho_r p^r \right)^2 + \frac{\gamma^2}{2} 2 \sum_{l=t+1}^{m'} \rho_l \sum_{r=k}^{l-1} \rho_r + \Gamma^2 \sup_{k \leq r \leq m'} \rho_r \sum_{l=t+1}^{m'} \rho_l \\ &\leq \Gamma^2 \left(\frac{\gamma}{3\Gamma} + \frac{\gamma^2}{6\Gamma^2} \right) \sigma^2 + \frac{\gamma^2}{2}\sigma^2 + \Gamma^2 \frac{\gamma^2\sigma}{6\Gamma^2}\sigma \leq \frac{11}{12}\gamma^2\sigma^2. \end{aligned}$$

This contradiction proves the lemma. □

Now let us come back to the proof of Lemma 4.1. Set

$$P = co\{\overline{G}(z) |\ \|z - y\| \le \epsilon\},$$
$$p^r = g^r + h^r, \quad k = k_s \le r \le m = m_s,$$
$$\gamma_0 = \nu/2, \quad \Gamma_0 = \Gamma.$$

We have

$$\sum_{k=k_s}^{m_s} \rho_k \ge \sum_{k=k_s}^{m_s-1} \rho_k \ge \frac{\|x^{m_s} - x^{k_s}\|}{\Gamma} \ge \frac{\epsilon}{3\Gamma} = \sigma_0 > 0,$$

$$\lim_{s\to\infty} \sup_{k\ge k_s} \rho_k = 0.$$

By Lemma 4.2 for all sufficiently large s there exist indices $l_s,\ k_s < l_s \le m_s$, such that

$$\left\langle g^{l_s} + h^{l_s}, \sum_{k=k_s}^{l_s-1} \rho_k (g^k + h^k) / \sum_{k=k_s}^{l_s-1} \rho_k \right\rangle \ge \frac{\nu^2}{16},$$

$$\sum_{k=k_s}^{l_s-1} \rho_k \ge \frac{\epsilon\nu}{18\Gamma^2}.$$

Substituting these estimates for $k = l_s$ into inequality (28), we obtain the final estimate with $c = \frac{\nu^2}{64\Gamma(1+2\Gamma/\mu)}$:

$$\begin{aligned} f(x^{l_s}) &\le f(y) - \frac{\nu^2}{16} \sum_{k=k_s}^{l_s-1} \rho_k + \Gamma(1 + 2\Gamma/\mu) c \sum_{k=k_s}^{l_s-1} \rho_k \\ &\quad + (\Gamma + c + 2c\Gamma/\mu)\|x^{k_s} - y\| \\ &\le f(y) - \frac{\nu^2}{600\Gamma^2}\epsilon\nu + (\Gamma + c + 2c\Gamma/\mu)\|x^{k_s} - y\|. \end{aligned} \tag{32}$$

Thus we have proved that for all sufficiently small $\epsilon \le \overline{\epsilon}$ and sufficiently large s there exist indices l_s such that $\|x^k - y\| \le \epsilon$ for $k \in [k_s, l_s)$ and $f(x^{l_s})$ satisfies (32). From here the statement of the lemma follows. □

Proof of Theorem 4.1. The proof is based on Lemma 4.1.

1^0. Obviously, the sequence $\{x^k\}$ belongs to a compact set X.

2^0. By boundedness of subgradients $\overline{\partial} f(x)$ on a compact set X we obtain

$$\lim_{k\to\infty} \|x^{k+1}(\omega) - x^k(\omega)\| \le \sup_{g\in\overline{\partial} f(x),\ x\in X} \|g\| \lim_{k\to\infty} \rho_k = 0.$$

From here it follows that cluster points of $\{x^k\}$ constitute a connected set in X.

3^0. Sequence $\{x^k\}$ from compact set X has a closed set of limit points X'. The continuous function $f(x)$ achieves its minimum on X', say, at some point x'. The point $x' = \lim_{s\to\infty} x^{k_s}$ belongs to X^* because otherwise due to Lemma 4.1 it is not minimal in the above sense. Thus $\liminf_{k\to\infty} f(x^k) \in f^*$.

4^0. Now prove that limit points of the sequence $\{f(x^k)\}$ constitute an interval in f^*. If $\limsup_{k\to\infty} f(x^k) = \liminf_{k\to\infty} f(x^k)$ then the statement follows from 3^0. Suppose

$$\limsup_{k\to\infty} f(x^k) > \liminf_{k\to\infty} f(x^k) = f_0^* \in f^*.$$

Assume the opposite to the statement of the theorem. Then there exists some number $f_1 \overline{\in} f^*$ such that $f_1 < \limsup_{k\to\infty} f(x^k(\omega))$. Let us choose number f_2 such that

$$\liminf_{k\to\infty} f(x^k) = f^* < f_1 < f_2 < \limsup_{k\to\infty} f(x^k).$$

Sequence $\{f(x^k)\}$ intersects interval (f_1, f_2) from below infinitely many times, so there exist subsequences $\{x^{k_s}\}$ and $\{x^{n_s}\}$ such that

$$f(x^{k_s}) \le f_1 < f(x^k) < f_2 \le f(x^{n_s}), \quad k_s < k < n_s. \tag{33}$$

Without loss of generality we can consider that $x^{k_s} \longrightarrow x'$. Due to 2^0 and continuity of f we have

$$\lim_{s\to\infty} f(x^{k_s}) = f(x') = f_1 \overline{\in} f^*.$$

Hence $\lim_{s\to\infty} x^{k_s} = x' \overline{\in} X^*$. Now we can apply Lemma 4.1 to subsequences $\{x^k\}_{k=k_s}^{\infty}$. Choose ϵ such that

$$\sup_{\{y:\|y-x'\|\le\epsilon\}} f(y) < f_2.$$

Then (15) contradicts to inequalities (33). Hence

$$\left[\liminf_{k\to\infty} f(x^k), \limsup_{k\to\infty} f(x^k)\right] \subseteq f^*.$$

Since X^* and f^* are closed sets then

$$\left[\liminf_{k\to\infty} f(x^k), \limsup_{k\to\infty} f(x^k)\right] \subseteq f^*.$$

5^0. Suppose now that f^* does not contain intervals, for instance, f^* is finite or countable. From statement 4^0 we have

$$\lim_{k\to\infty} f(x^k) = f_0^* \in f^*. \tag{34}$$

If a cluster point $x' = \lim_{s\to\infty} x^{k_s}$ does not belong to X^*, then due to Lemma 4.1 we would have a contradiction $\{f(x^k)\}$ stated in (34). □

Remark 4.2 *The convergence result of Theorem 4.1 remains true for generalized gradient method (11), (12), where*

$$g^k \in \overline{\partial} f(\tilde{x}^k), \quad \|\tilde{x}^k - x^k\| \le \delta_k, \quad \lim_k \delta_k = 0.$$

In this case the basic Lemma 4.1 follows from the stability result of Lemma 5.4. If points $\tilde{x}^k$ are taken at random then with probability one $\overline{\partial} f(\tilde{x}^k) = \partial f(\tilde{x}^k)$ and the method converges to $X^* = \{x|\ 0 \in \partial f(x) + N_X(x)\}$. In the last case we can use formula (7) and the chain rule to calculate $g^k \in \overline{\partial} f(\tilde{x}^k)$. The use of $\overline{\partial} f(\tilde{x}^k)$ resembles the concept of mollifier gradient [9].

5 Stochastic generalized gradient method

Consider now stochastic optimization problem (1), (2), where the objective function $F(x)$ is generalized differentiable, the set $X = \{x|\ \psi(x) \leq 0\}$ is given by a generalized differentiable function $\psi(x)$, satisfying regularity condition (10). Define $X^* = \{x|\ 0 \in \overline{\partial} F(x) + N_X(x)\}$ and $F^* = \{F(x)|\ x \in X^*\}$.

Consider the following procedure

$$x^0 \in X, \tag{35}$$

$$x^{k+1}(\omega) \in \Pi_X(x^k - \rho_k s^k(\omega)), \quad k = 0, 1, \ldots, \tag{36}$$

$$s^k(\omega) = \frac{1}{n_k} \sum_{i=r_k}^{k} \xi^i(\omega), \quad n_k = k - r_k + 1 \geq 0, \tag{37}$$

where all random quantities $x^k(\omega)$, $\xi^k(\omega)$, $s^k(\omega)$, $k = 0, 1, \ldots$, are defined on some probability space $(\Omega, \Sigma, \mathbf{P})$, $\xi^i(\omega)$, $i = 0, 1, \ldots$, are random vectors (stochastic generalized gradients) such that

$$\mathbf{E}\{\xi^i(\omega)|\ x^0(\omega), \ldots, x^i(\omega)\} = g^i(\omega) \in \overline{\partial} f(x^i(\omega)),$$

$$\|\xi^i(\omega)\| \leq C < +\infty; \tag{38}$$

Π_X is a (multivalued) projection operator on the set X, i.e. $z \in \Pi_X(y)$ iff $y - z \in N_X(z)$; non-negative numbers r_k, n_k and ρ_k satisfy conditions

$$n_k = k + 1 - r_k \leq m < +\infty; \tag{39}$$

$$\sum_{k=0}^{\infty} \rho_k = +\infty, \quad \sum_{k=0}^{\infty} \rho_k^2 < +\infty. \tag{40}$$

Remark 5.1 *Method (35)-(37) combines ideas of projection stochastic quasigradient method by Ermoliev (see details and further references in [11], pp. 142-185) and stochastic gradient averaging method [1], [5], [7], [15], [19], It is easy to generalize the convergence analysis to biased estimates of generalized gradients – stochastic quasigradients.*

Theorem 5.1 *Let $f(x)$ and $\psi(x)$ be generalized differentiable functions, sequence $x^k(\omega)$ is generated by method (35)-(37), where r_k, n_k, ρ_k satisfy (39), (40). Then minimal (in function F) cluster points of $\{x^k(\omega)\}$ a.s. belong to X^* and all cluster points of $\{F(x^k(\omega)\}$ a.s. constitute an interval in F^*. If the set F^* does not contain intervals (for instance, F^* is finite or countable) then all cluster points of $\{x^k(\omega)\}$ a.s. belong to a connected subset of X^* and $\{F(x^k(\omega))\}$ has a limit in F^*.*

Proof. Denote $\overline{x}^{k+1} = x^k - \rho_k s^k$ and represent

$$x^{k+1} = \Pi_X(x^k - \rho_k s^k) = x^k - \rho_k(s^k + h^k) = x^k - \rho_k Q^k,$$

where

$$Q^k = s^k + h^k,$$

$$h^k = h^k(\overline{x}^{k+1}) = \frac{1}{\rho_k}(\overline{x}^{k+1} - \Pi_X(\overline{x}^{k+1})) \in N_X(x^{k+1}),$$

$$\|h^k\| = \frac{1}{\rho_k}\|\overline{x}^{k+1} - \Pi_X(\overline{x}^{k+1})\| \leq \frac{1}{\rho_k}\|\overline{x}^{k+1} - x^k\| = \|s^k\|.$$

$$\|Q^k\| = \frac{1}{\rho_k}\|x^{k+1} - x^k\| \le \frac{1}{\rho_k}\|\overline{x}^{k+1} - x^k\| = \|s^k\|.$$

Now fix a subsequence $\{x^{k_s}(\omega)\}$. For $k > k_s$

$$\begin{aligned} x^{k+1}(\omega) &= x^{k_s}(\omega) - \sum_{t=k_s}^{k} \rho_t Q^t(\omega) \\ &= x^{k_s}(\omega) - \sum_{t=k_s}^{k} \rho_t \overline{Q}^t(\omega) - \zeta_{k_s}^{k+1}(\omega) \\ &= y_{k_s}^{k+1}(\omega) - \zeta_{k_s}^{k+1}(\omega), \end{aligned} \tag{41}$$

where

$$y_{k_s}^{k_s}(\omega) = x^{k_s}(\omega), \tag{42}$$

$$y_{k_s}^{k+1}(\omega) = \sum_{k=k_s}^{k} \rho_t \overline{Q}^t(\omega) = y_{k_s}^{k}(\omega) - \rho_k \overline{Q}^k(\omega), \;\; k \ge k_s; \tag{43}$$

$$\overline{Q}^k(\omega) = \frac{1}{n_k}\sum_{r=r_k}^{k} (\overline{g}^r(\omega) + h^r(\omega)), \tag{44}$$

$$\overline{g}^r(\omega) = \mathbf{E}\{\xi^r(\omega) |\, x^0(\omega), \ldots, x^r(\omega)\} \in \overline{\partial} f(x^r(\omega)), \tag{45}$$

$$h^r(\omega) = \frac{1}{\rho_r}(\overline{x}^r(\omega) - \Pi_X(\overline{x}^r(\omega)) \in N_X(x^{r+1}(\omega), \tag{46}$$

$$\zeta_n^m(\omega) = \sum_{t=n}^{m-1} \rho_t \frac{1}{n_t}\sum_{r=r_t}^{t} (\xi^r(\omega) - \overline{g}^r(\omega)). \tag{47}$$

Instead of $\{x^k(\omega)\}$ we shall study the behavior of the close sequence $\{y_{k_s}^k(\omega)\}_{k \ge k_s}$, $s = 0, 1, \ldots$, generated by deterministic (under fixed ω) procedure (43)-(47). This procedure uses subgradients $\overline{g}^r(\omega)$ of function F taken not at points $y_{k_s}^r(\omega)$ but at close points $x^r(\omega)$. Besides, the vector $h^r(\omega)$ is normal to X not at the point $y_{k_s}^{r+1}(\omega)$, but at a close point $x^{r+1}(\omega)$. We have an estimate:

$$\|y_{k_s}^k(\omega) - x^k(\omega)\| = \|\zeta_{k_s}^k(\omega)\| \le \sup_{k \ge k_s} \|\zeta_{k_s}^k(\omega)\| = \delta_{k_s}(\omega).$$

Let us show (Lemma 5.1) that $\lim_{s\to\infty} \delta_{k_s}(\omega) = 0$ a.s. Notice that

$$|f(x^k(\omega)) - f(y_{k_s}^k(\omega))| \le L_f \|x^k(\omega) - y_{k_s}^k(\omega)\| = L_f \delta_{k_s}(\omega), \tag{48}$$

where L_f is a Lipschitz constraint of function f over set X. Then the difference $|f(x^k(\omega)) - f(y_{k_s}^k(\omega))|$, $k \ge k_s$, is arbitrary small for s sufficiently large. The remaining part of the proof we subdivide into several separate lemmas.

Lemma 5.1 *Random sequence* $\{\zeta_0^k(\omega)\}_{k=0}^{\infty}$,

$$\zeta_0^k(\omega) = \sum_{t=0}^{k-1} \rho_t \frac{1}{n_t}\sum_{r=r_t}^{t} (\xi^r(\omega) - \overline{g}^r(\omega)), \quad n_t \le m, \tag{49}$$

a.s. has a limit.

Proof. Denote

$$\lambda_{tr} = \begin{cases} \frac{1}{n_t}, & r_t \le r \le t, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\begin{array}{rcl} \zeta_0^k & = & \sum_{t=0}^{k-1} \rho_t \sum_{r=r_t}^{t} \lambda_{tr}(\xi^r - \overline{g}^r) = \sum_{t=0}^{k-1}(\sum_{t=r}^{k-1} \lambda_{tr}\rho_t)(\xi^r - \overline{g}^r) \\ & = & \sum_{t=0}^{k-1}(\sum_{t=r}^{\infty} \lambda_{tr}\rho_t)(\xi^r - \overline{g}^r) - \sum_{t=0}^{k-1}(\sum_{t=k}^{\infty} \lambda_{tr}\rho_t)(\xi^r - \overline{g}^r). \end{array}$$

Sequence

$$\overline{\zeta}_0^k = \sum_{t=0}^{k-1}(\sum_{t=r}^{\infty} \lambda_{tr}\rho_t)(\xi^r - \overline{g}^r) \tag{50}$$

is a martingale with respect to σ-field generated by $\{x^k(\omega)\}_{k=0}^{\infty}$. Denote

$$\Gamma = \sup\{\|g\| \mid g \in \overline{\partial} f(x),\ x \in X\} < +\infty.$$

Then

$$\begin{array}{rcl} \mathbf{E}\|\overline{\zeta}_0^k(\omega)\|^2 & \le & (\Gamma + C)^2 \sum_{r=0}^{\infty}(\sum_{t=r}^{\infty} \lambda_{tr}\rho_t)^2 \le (\Gamma + C)^2 \sum_{r=0}^{\infty}(\sum_{t=r}^{r+m} \rho_t)^2 \\ & \le & (\Gamma + C)^2 m^2 \sum_{r=0}^{\infty} \rho_t^2 < +\infty \end{array}$$

and

$$\mathbf{E}\|\overline{\zeta}_0^k(\omega)\| \le 1 + \mathbf{E}\|\overline{\zeta}_0^k(\omega)\|^2 < +\infty.$$

Hence the martingale (50) a.s. has a finite limit. For the remainder term

$$\alpha^k(\omega) = \sum_{t=0}^{k-1}(\sum_{t=k}^{\infty} \lambda_{tr}\rho_t)(\xi^r - \overline{g}^r)$$

the following estimates hold true:

$$\begin{array}{rcl} \alpha^k(\omega) & \le & \sum_{r=0}^{k-1}(\sum_{t=k}^{\infty} \lambda_{tr}\rho_t)(\|\xi^r\| + \|\overline{g}^r\|) \\ & \le & (\Gamma + C) \sum_{r=0}^{k-1}(\sum_{t=k}^{\infty} \lambda_{tr}\rho_t) = (\Gamma + C) \sum_{t=k}^{\infty} \rho_t(\sum_{r=0}^{k} \lambda_{tr}) \\ & = & (\Gamma + C) \sum_{t=k}^{\infty} \rho_t(\sum_{r=r_t}^{k} \lambda_{tr}) \\ & \le & (\Gamma + C) \sum_{t=k}^{k+m} \rho_t \longrightarrow 0 \ \text{ as } \ k \longrightarrow \infty. \end{array}$$

Hence the sequence $\{\zeta_0^k(\omega) = \overline{\zeta}_0^k(\omega) + \alpha^k(\omega)\}$ a.s. has a limit.□

Corollary 5.1 *For any subsequence of indices* $\{k_s\} \longrightarrow \infty$

$$\delta_{k_s}(\omega) = \sup_{k \ge k_s} \|\zeta_{k_s}^k(\omega)\| \longrightarrow 0 \ \ \textit{a.s. as} \ \ s \longrightarrow \infty.$$

Remark 5.2 *Lemma 5.1 and Corollary 5.1 remain true if* $r_k = k$ *in (37) and (38) is replaced by*

$$\mathbf{E}\|\xi^i(\omega)\|^2 < +\infty.$$

Lemma 5.2 *Let* ω *be such that* $\{\zeta_0^k(\omega)\}_{k=0}^{\infty}$ *has a limit. Assume that* $\lim_{s\to\infty} x^{k_s}(\omega) = x(\omega) \overline{\in} X^*$. *Denote*

$$m_s(\epsilon, \omega) = \sup\{m \mid \|x^k(\omega) - x(\omega)\| \le \epsilon \ \textit{ for } \ k \in \{k_s, m\}.$$

Then a.s. there exists $\overline{\epsilon}(\omega)$ *such that for any* $\epsilon \in (0, \overline{\epsilon}]$ *there exist indices* $l_s(\omega) \in [k_s(\omega), m_s(\epsilon, \omega)]$, *and*

$$f(x(\omega)) = \lim_{s\to\infty} f(x^{k_s}(\omega)) > \limsup_{s\to\infty} f(x^{l_s}(\omega)). \tag{51}$$

Lemma 5.2 due to (41), (48) and Corollary 5.1 follow from the similar property of the sequences $\{y_{k_s}^k(\omega)\}_{k\geq k_s}$, generated by (43)-(45). We formulate this property as a separate lemma.

Lemma 5.3 *Let ω be such that $\{\zeta_0^k(\omega)\}_{k=0}^{\infty}$ has a limit. Assume that $\lim_{s\to\infty} x^{k_s}(\omega) = x(\omega)\overline{\in}X^*$. Denote*

$$m_s(\epsilon,\omega) = \sup\{m|\ \|y_{k_s}^k(\omega) - x(\omega)\| \leq \epsilon \ \text{ for } \ k \in [k_s, m)\}.$$

Then a.s. there exists $\overline{\epsilon}(\omega)$ such that for any $\epsilon \in (0,\overline{\epsilon}]$ there exist indices $l_s(\omega) \in [k_s(\omega), m_s(\epsilon,\omega)]$, and

$$f(x(\omega)) = \lim_{s\to\infty} f(x^{k_s}(\omega)) > \limsup_{s\to\infty} f(y_{k_s}^{l_s}(\omega)). \tag{52}$$

Lemma 5.3 follows from the following stability property of the deterministic subgradient method.

Lemma 5.4 *Let some sequence of starting points $\{y^s\}$ converge to $y = \lim_{s\to\infty} y^s$. For each s consider a sequence $\{y_{k_s}^k\}_{k=k_s}^{n_s}$ such that*

$$\begin{array}{l} y_s^{k_s} = y^s, \\ y_s^{k+1} = y_s^k - \rho_k(g_s^k + h_s^k), \quad s \leq k < n_s; \\ g_s^k \in G_{\delta_s^k}(y_s^k) = co\{g \in \overline{\partial} f(y)|\ \|y - y_s^k\| \leq \delta_s^k\}, \\ h_s^k \in \{\frac{y - \Pi_X(y)}{\rho_k}|\ \|y - \overline{y}_s^k\| \leq \delta_s^k\}, \\ \overline{y}_s^k = y_s^k - \rho_s^k g_s^k. \end{array}$$

Denote

$$\rho_s = \sup_{k_s\leq k\leq n_s} \rho_s^k, \quad \delta_s = \sup_{k_s\leq k\leq n_s} \delta_s^k, \quad \sigma_s = \sum_{k=k_s}^{n_s-1} \rho_s^k.$$

If $0\overline{\in}\overline{\partial} f(y) + N_X(y)$ and $\sigma_s \geq \sigma > 0$ then for any sufficiently small ϵ there exist $\overline{\rho} = \overline{\rho}(y,\epsilon)$ and $\overline{\delta} = \overline{\delta}(y,\epsilon)$ such that for $\{y_s^k\}_{k=k_s}^{n_s}$ with $\delta_s^k \leq \overline{\delta}$ and $\rho_s^k \leq \overline{\rho}$ there exist indices l_s such that $\|y_s^k - y\| \leq \epsilon$ for $k \in [k_s, l_s)$ and

$$f(y) = \lim_{s\to\infty} f(y^s) > \limsup_{s\to\infty} f(y_s^{l_s}).$$

Proof. The proof is similar to the proof of Lemma 4.1. We have to consider again two cases: $\psi(y) < 0$ and $\psi(y) = 0$. In the first case the subgradient method operates in a sufficiently small vicinity of y as an unconstrained method and the statement of the lemma is known (see [19]). In what follows we consider a new case $\psi(y) = 0$ (the case $\psi(x) < 0$ may also be considered as a simple repetition of the case $\psi(y) = 0$). As in proof of Lemma 4.1 for $y = \lim_s y^s$ define μ, ν, γ by (17)- (19) and ϵ_1, ϵ_2, ϵ_3, c such that (20)-(24) hold.

Now set

$$\overline{\epsilon} = \min\{\epsilon_3, \sigma\nu/2\}$$

and fix some $\epsilon \leq \overline{\epsilon}$. Set $\overline{\delta}_1 = \epsilon/4$, $\overline{\rho}_1 = \epsilon/(4\Gamma)$. Let for $\|y^s - y\| \leq \epsilon/4$, $\delta_s \leq \overline{\delta}_1$, $\rho_s \leq \overline{\rho}_1$ for $s \geq S$.

Define the index

$$m_s = \sup\{m|\ \|y_s^r - y\| \leq \epsilon/2 \ \ \forall\, r \in [k_s, m)\}.$$

We now show that $\epsilon/2 \le \|y_s^{m_s} - y\| \le 3\epsilon/4$. Firstly we shall prove the left inequality. If $\|y_s^{m_s} - y\| \le \epsilon/2$ then $m_s = n_s$ and we obtain a contradiction:

$$\epsilon_2 > 3\epsilon/4 \ge \|y_s^{n_s} - y_s\| \ge \sigma\nu/2.$$

Furthermore

$$\|y_s^{m_s} - y\| \le \|y_s^{m_s-1} - y\| + \rho_s^{m_s-1}\|g_s^{m_s-1} + h_s^{m_s-1}\| \le 3\epsilon/4.$$

Since

$$\epsilon/4 \le \|\sum_{k=k_s}^{m_s-1} \rho_s^k(g_s^k + h_s^k)\| \le \Gamma \sum_{k=k_s}^{m_s-1} \rho_s^k,$$

then

$$\sum_{k=k_s}^{m_s-1} \rho_s^k \ge \frac{\epsilon}{4\Gamma}.$$

Let $g_s^k \in G_{\delta_s^k}(y_s^k)$, then

$$g_s^k = \sum_{i=1}^{n+1} \lambda_s^{ki} g_s^{ki}, \quad \sum_{i=1}^{n+1} \lambda_s^{ki} = 1;$$

$$g_s^{ki} \in \overline{\partial} f(y_s^{ki}), \quad \|y_s^{ki} - y_s^k\| \le \delta_s^k.$$

If $\|y^s - y\| \le \epsilon/4$, $\delta_s \le \epsilon/4$, $k_s \le k \le m_s$, $1 \le i \le n+1$, then

$$\|y_s^{ki} - y\| \le \|y_s^{ki} - y_s^k\| + \|y_s^k - y\| \le \delta_s^k + 3\epsilon/4 \le \epsilon \le \epsilon_3.$$

For y_s^{ki} we can use (23):

$$f(y_s^{ki}) \le f(y) + < g_s^{ki}, y_s^{ki} - y^s > + c\|y_s^{ki} - y^s\| + (\Gamma + c)\|y^s - y\|.$$

If we replace y_s^{ki} $(1 \le i \le n+1)$ by a close point y_s^k, then

$$\begin{aligned} f(y_s^k) \quad &\le \quad f(y) + < g_s^{ki}, y_s^k - y^s > \\ &\quad + \quad + c\|y_s^k - y\| + (2\Gamma + c)\delta_s + (\Gamma + c)\|y^s - y\|. \end{aligned}$$

Multiplying these inequalities by λ_s^{ki} and summing in i, we obtain

$$\begin{aligned} f(y_s^k) \quad &\le \quad f(y) + < g_s^k, y_s^k - y^s > \\ &\qquad + c\|y_s^k - y^s\| + (2\Gamma + c)\delta_s + (\Gamma + c)\|y^s - y\| \\ &= \quad f(y) + < g_s^k + h_s^k, y_s^k - y^s > - < h_s^k, y_s^k - y^s > + \\ &\qquad + c\|y_s^k - y^s\| + (2\Gamma + c)\delta_s + (\Gamma + c)\|y^s - y\|, \end{aligned} \tag{53}$$

where

$$h_s^k = (\tilde{y}_s^k - z_s^k)/\rho_s^k, \quad \|\tilde{y}_s^k - \overline{y}_s^k\| \le \delta_s^k, \quad z_s^k = \Pi_X(\tilde{y}_s^k).$$

Let us evaluete the term $u_s^k = - < h_s^k, y_s^k - y^s >$. If $\psi(y_s^k) \le 0$ then $h_s^k = 0$ and $u_s^k = 0$. Consider the case $\psi(y_s^k) > 0$, i.e. $u_s^k \ne 0$. Since

$$h_s^k \in N_X(z_s^k) = \{\lambda g |\ g \in \overline{\partial}\psi(z_s^k),\ \lambda \ge 0\},$$

then

$$h_s^k = \lambda_s^k d_s^k, \quad d_s^k \in \overline{\partial}\psi(z_s^k), \quad \lambda_s^k > 0.$$

We have

$$0 < \lambda_s^k = \|h_s^k\|/\|d_s^k\| \le \Gamma/(\mu/2) = 2\Gamma/\mu.$$

Substitute z_s^k and d_s^k into expansion (24):

$$\psi(z_s^k) \le < d_s^k, z_s^k - y > + c\|z_s^k - y\|. \tag{54}$$

Since

$$y_s^k = \overline{y}_s^k - \rho_s^k h_s^k = \overline{y}_s^k - \tilde{y}_s^k + z_s^k,$$

then

$$\|y_s^k - z_s^k\| = \|\overline{y}_s^k - \tilde{y}_s^k\| \le \delta_s^k.$$

Replacing z_s^k in (54) by a close point y_s^k, we obtain:

$$0 = \psi(y_s^k) \le < d_s^k, y_s^k - y > + c\|y_s^k - y\| + (2\Gamma + c)\delta_s^k. \tag{55}$$

Now multiply (55) by $\lambda_s^k \le 2\Gamma/\mu$:

$$0 \le < h_s^k, y_s^k - y^s > + (2\Gamma c/\mu)\|y_s^k - y^s\| + (\Gamma + 2c\Gamma/\mu)\|y^s - y\| + 2\Gamma(2\Gamma + c)\delta_s^k/\mu.$$

i.e.

$$- < h_s^k, y_s^k - y^s > \le (2\Gamma c/\mu)\|y_s^k - y^s\| + \Gamma(1 + 2c/\mu)\|y^s - y\| + 2\Gamma(2\Gamma + c)\delta_s^k/\mu. \tag{56}$$

Using inequality (56) we can rewrite (53) in the following form:

$$\begin{aligned} f(y_s^k) \quad \le \quad & f(y) + < g_s^k + h_s^k, y_s^k - y^s > + (c + 2\Gamma c/\mu)\|y_s^k - y^s\| \\ & + (2\Gamma + c + 2\Gamma c/\mu)\|y^s - y\| + (2\Gamma + c)(1 + 2\Gamma c/\mu)\delta_s^k. \end{aligned} \tag{57}$$

Now we evaluate scalar products

$$< g_s^k + h_s^k, y_s^k - y^s > = < g_s^k + h_s^k, \sum_{i=k_s}^{k-1} (g_s^i + h_s^i) > .$$

by means of Lemma 4.2. Set

$$\begin{aligned} & P = co\{\overline{G}(z) |\ \|z - y\| \le \epsilon\}, \\ & p^r = g_s^r + h_s^r, \quad k = k_s \le r \le m = m_s, \\ & \gamma_0 = \nu/2, \quad \Gamma_0 = 2\Gamma, \end{aligned}$$

then

$$\sum_{k=k_s}^{m_s} \rho_s^k \ge \sum_{k=k_s}^{m_s - 1} \rho_s^k \ge \frac{\|y_s^{m_s} - y^s\|}{2\Gamma} \ge \frac{\epsilon}{4\Gamma} = \sigma_0 > 0 \ \text{ for } s \ge S,$$

$$\lim_{s\to\infty} \sup_{k \ge k_s} \rho_s^k = \lim_{s\to\infty} \sigma_s = 0.$$

By Lemma 4.2 for all sufficiently large s there exist indices $l_s,\ k_s < l_s \le m_s$, such that

$$\left\langle g_s^{l_s} + h_s^{l_s}, \sum_{k=k_s}^{l_s - 1} \rho_s^k Q_s^k / \sum_{k=k_s}^{l_s - 1} \rho_s^k \right\rangle \ge \frac{\nu^2}{16},$$

$$\sum_{k=k_s}^{l_s - 1} \rho_s^k \ge \frac{\epsilon\nu}{48\Gamma^2}.$$

Substituting these values for $k = l_s$ into inequality (57), we obtain finally for $c = \frac{\nu^2}{64\Gamma(1+2\Gamma/\mu)}$:

$$\begin{aligned} f(y_s^{l_s}) &\leq f(y) - \frac{\nu^2}{16}\sum_{k=k_s}^{l_s-1}\rho_s^k + 2\Gamma(1+2\Gamma/\mu)c\sum_{k=k_s}^{l_s-1}\rho_s^k \\ &\quad +(2\Gamma+c)(1+2\Gamma c/\mu)\delta_s + (2\Gamma+c+2\Gamma c/\mu)\|y^s-y\| \\ &\leq f(y) - \frac{\nu^2}{1600\Gamma^2}\epsilon\nu + +(2\Gamma+c)(1+2\Gamma c/\mu)\delta_s \\ &\quad +(2\Gamma+c+2\Gamma c/\mu)\|y^s-y\|. \end{aligned} \tag{58}$$

Hence we have proved that for all sufficiently small $\epsilon \leq \overline{\epsilon}$ and sufficiently large s there exist indices l_s such that $\|y_s^k - y\| \leq \epsilon$ for $k \in [k_s, l_s)$ and $f(y_s^{l_s})$ satisfies (58). From here the statement of the Lemma follows.□

Lemma 5.5 *In method (35)-(37)*

$$\lim_{k\to\infty}\|x^{k+1}(\omega) - x^k(\omega)\| = 0 \quad a.s.$$

Proof.

$$\begin{aligned} \|x^{k+1}(\omega) - x^k(\omega)\| &\leq \|s^k(\omega)\| \\ &\leq \rho_k\|\tfrac{1}{n_k}\textstyle\sum_{r=r_k}^k g^r(\omega)\| + \rho_k\|\tfrac{1}{n_k}\textstyle\sum_{r=r_k}^k(\xi^r(\omega) - g^r(\omega))\| \\ &\leq \rho_k\|\tfrac{1}{n_k}\textstyle\sum_{r=r_k}^k g^r(\omega)\| + \rho_k\|\zeta_k^k(\omega)\|, \end{aligned}$$

where

$$\begin{aligned} g^r(\omega) &= \mathbf{E}\{\xi^r(\omega)|\, x^0(\omega), \ldots, x^k(\omega)\}, \\ \zeta_k^k(\omega) &= \tfrac{1}{n_k}\textstyle\sum_{r=r_k}^k(\xi^r(\omega) - g^r(\omega)), \quad n_k = k - r_k + 1. \end{aligned}$$

Here $g^r \in \overline{\partial} f(x^r(\omega))$ are uniformly bounded and hence

$$\lim_{k\to\infty}\rho_k\|\frac{1}{n_k}\sum_{r=r_k}^k g^r(\omega)\| = 0.$$

By Lemma 5.1 sequence $\{\zeta_0^k(\omega)\}$ (see (49)) a.s. has a limit, hence $\lim_{k\to\infty}\rho_k\zeta_k^k(\omega) = 0$ a.s. □

Now we can complete the proof of Theorem 5.1. Consider the set $\Omega' \subseteq \Omega$ such that the sequence $\{\zeta_0^k(\omega)\}_{k=0}^\infty$ defined by (49) converges. By Lemma 5.1 $\mathbf{P}(\Omega') = 1$. Let us fix some $\omega \in \Omega'$. The remaining part of the proof repeats steps $1^0 - 5^0$ of Theorem 4.1 and uses Lemma 5.2 instead of Lemma 4.1.

6 Concluding Remarks

In this paper we developed easy implementable computational procedures which may naturally incorporate fast Monte Carlo simulation and thus deal with nonsmooth and nonconvex problems analyzed in our recent paper [9]. These techniques aim at specific classes of problems with so-called generalized differentiable performance functions that do not require the concept of stochastic mollifier gradients introduced in [9] primarily in the connection with discontinuous performance functions. As we illustrated in Section 2 models with such performance functions cover important area of risk control involving stopping time (ruin) problems. Interesting enough that proposed search procedures may incorporate (see Remark 4, 2) random mechanisms of stochastic mollifiers with their additional

power to bypass local solutions without joint constraints on step-sizes of the procedure and the "stochasticity" of mollifiers. The estimates of generalized gradients in $(35)-(37)$ may also be biased and called stochastic quasigradient similar to how this notion is used in such situations for other (i.e. convex) functions. The convergence requirements are the same as for stochastic quasigradient procedures with convex functions and constraints, although the technique of proofs is essentially different and it is based on stopping time arguments (Lemma $4,1$). The stochastic optimization procedure (35)-(37) is based on (see Remark 5,1) averaging operation. This type of operations were introduced by Ermoliev (see [6] and discussions in [11] p.150) for so-called compound stochastic optimization problems. Averaging stochastic gradients as a particular case was firstly studied by Bagenov and Gupal [1], Chepurnoi [2], Ermoliev [7] (pp.214-215), Ruszczyński and Syski [25].

This paper deals only with the analysis of conceptual (basic) procedure that can be modified correspondingly for a particular case of problem. For example, in cases of insurance risk processes (as we showed in Section 2) there exist simple algorithms for the calculation of stochastic generalized gradients that can be easily embedded into Monte Carlo scenario simulation techniques. This algorithm provides feedbacks to policy variables in such a way that drives them towards improvements with respect to given collections of performance indicators. Let us also notice that proposed procedures have also been motivated by the study of the stochastic tâtonnement process [10] where the question of convergence plays an essential role. The use of $(35)-(37)$ type procedures combined with fast Monte Carlo simulations show a rather fast improvement of initial policy variables (see, for example [23]).

References

[1] Bajenov L.G. and Gupal A.M. (1972), Stochastic analog of conjugate gradients method, *Kibernetika*, **1**.

[2] Chepurnoi N.D. (1982), Methods of nondifferentiable optimization with averaged subgradients. Abstract of disserttion, Press of the Institute of Cybernetics, Kiev.

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

[4] Dorofeev P.A. (1985), On some properties of the generalized gradient method, Zh. Vychisl. Mat. i Mat. Fiz., 25, 2, 181-189 (In Russian, English translation: U.S.S.R. Comp. Maths. Math. Phys., Vol. 25, No. 1, pp. 117-122).

[5] Dorofeev P.A. (1986), A scheme of iterative minimization methods, Zh. Vychisl. Mat. i Mat. Fiz., 26, 4, 536-544 (In Russian, English translation: U.S.S.R. Comput. Math. Math. Phys., Vol. 26, No. 2, pp.131-136.

[6] Ermoliev Yu.M. (1971), General problem of stochastic programming, *Kibernetika*, **3**.

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

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

[9] Ermoliev Yu.M. and Norkin V.I. (1995), On Nonsmooth Problems of Stochastic Systems Optimization, Working paper WP-95-096, Int. Inst. for Appl. Syst. Anal., Laxenburg, Austria.

[10] Ermoliev Yu.M., Keyzer M.A., and Norkin V.I. (1996), Global Convergence of the Stochastic Tâtonnement Process, IIASA (draft of W.P.).

[11] Ermoliev Yu.M. and Wets R.J-B. (Eds.) (1988), Numerical Techniques for Stochastic Optimization, *Computational Mathematics*, **10**, Springer, Berlin.

[12] Ermolieva T.Y., Ermoliev Y.M., and Norkin V.I. (1997), Computational approaches to catastrophic risk management problems, IIASA (draft of Interim Report).

[13] 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.

[14] Glynn P. (1989), Optimization of Stochastic Systems via Simulation, Technical report No. 43, Stanford University, California.

[15] Gupal A.M. (1979), Stochastic methods for solving nonsmooth extremal problems, Naukova Dumka, Kiev.

[16] 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.

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

[18] Krivulin N.K. (1990), *Optimization of dynamic discrete event systems through simulations*, Candidate Dissertation, Leningrad, Leningrad University.

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

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

[21] Norkin V.I. (1978), On nonlocal algorithms for optimization of nonsmooth functions, *Kibernetika*, No. 5, pp. 75-79.

[22] Nurminski E.A. (1979), *Numerical methods for solving deterministic and stochastic nonlinear minimax problems*, Kiev, Naukova Dumka.

[23] Oortmarssen G. and Ermoliev Yu.M. (1994), Stochastic Optimization of Cancer Screening Strategies, Working Paper WP-94-124, Int. Inst. for Appl. System Anal., Laxenburg, Austria.

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

[25] Ruszczynski A. and Syski W. (1983), Stochastic approximation algorithm with gradient averaging for unconstrained problems, *IEEE Transactions on Automatic Control* AC-28, pp. 1097-1105.

[26] 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.