Cybernetics and Systems Analysis, Vol. 34, No. 2, 1998

SYSTEMS ANALYSIS

STOCHASTIC GENERALIZED GRADIENT METHOD FOR NONCONVEX NONSMOOTH STOCHASTIC OPTIMIZATION

Yu. M. Ermol'ev and V. I. Norkin

UDC 519.95

1. NONCONVEX NONSMOOTH STOCHASTIC PROGRAMMING PROBLEMS

Nonconvex nonsmooth stochastic optimization problems for discrete-event systems considered in [1] can be stated in general form as

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

subject to

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

where x is the solution vector (a variable), θ is a random parameter defined on the probability space $(\Theta, \Sigma, \mathbf{P})$, $f(x, \theta)$ is a random function evaluating the quality of the solution x given the random parameter θ, $\mathbf{E}$ is the expectation symbol, X is the feasible set.

An essential feature of the problem is that the function $f(\cdot, \theta)$ is not endowed with good analytical properties. In particular, it may be nonconvex, nonsmooth, and even discontinuous.

A standard approach to the solution of the problem involves approximation of $F(x)$ by its empirical mean:

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

subject to

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

where θ_i, $i = 1, 2, \ldots, N$, are independent identically distributed observations of θ. As shown in [1], this standard approach is often meaningless, because the functions $f(x, \theta)$ and thus $F_N(x)$ may have a poor analytical structure even if $F(x)$ is a smooth function. Moreover, the function $F(x)$ may be nonconvex and nonsmooth, which leads to a highly nonsmooth or discontinuous $F_N(x)$ whose multiple local minima have nothing in common with the true minima of $F(x)$. In this case, our only choice is to use stochastic search procedures based on direct evaluation of the function $F(x)$ and its derivatives. Application of the smoothing method for this purpose is considered in [2].

If the (generalized) differentiation and expectation operators are interchangeable

$$\partial F(x) = \partial \mathrm{E}_\theta f(x, \theta) = \mathrm{E}_\theta \partial f(x, \theta),$$

Translated from Kibernetika i Sistemnyi Analiz, No. 2, pp. 50-71, March-April, 1998. Original article submitted June 26, 1997.

1060-0396/98/3402-0196$20.00

then for the stochastic generalized gradients of $F(x)$ we may use the subgradients $\xi^k = g(x^k, \theta) \in \partial f(x^k, \theta)$. The case when the function $F(x)$ is continuously differentiable although $f(x, \theta)$ are nonsmooth is considered in [3-7].

The case when the expectation function $F(x)$ is locally Lipschitzian is considered in [6, 8, 9]. Moreover, as shown in [1], we are often dealing not with general Lipschitzian functions, but with functions formed from some basic (continuously differentiable) functions by the operations of taking maximum or minimum and smooth transformations. They are included in the class of so-called generalized differentiable functions [10].

Section 2 briefly discusses the main practical problems with such functions. Section 3 formally introduces the class of generalized differentiable functions and examines its properties. Sections 4 and 5 prove convergence of deterministic and stochastic generalized gradient methods (with projection of the current approximation on a nonconvex feasible set). These procedures generalize the results of Dorofeev [11, 12] obtained for the case of quasi-differentiable functions $F(x)$, which are not applicable to the problems of Section 2.

2. EXAMPLES OF NONSMOOTH STOCHASTIC SYSTEMS

Let us consider some examples of stochastic systems with nonconvex and nonsmooth performance functions. In these examples, the system is exposed to the action of discrete events (a discrete-event system, DES).

2.1. Controlled Risk Processes

Consider the simplest model describing the evolution of capital of an insurance company. Assume that the initial capital of the insurance company is x_1, the insurance claims are received at random time moments $\tau_1, \tau_2, \ldots$ and for random amounts $L_1, L_2, \ldots$. The reserve $R(x, t)$ of the insurance company at time t is the difference between the initial capital x_1 plus accumulated premiums $P(x_2, t)$ and the aggregated claims $C(x_3, t)$ plus reinsurance payments $c(x_3)t$:

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

The parameters x_2, x_3 are specifies in the insurance and reinsurance contracts.

The premium $P(x_2, t)$ collected during the time interval $[0, t)$ is $x_2 t$. The sum of claims is

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

where $N(t)$ is the random number of claims during the time interval $[0, t)$; x_3 is the reinsurance threshold specified in the reinsurance contract. Bankruptcy occurs at the moment $\tau(x) = \min\{0 < t \leq T: R(x, t) < 0\}$; if $R(x, t) \geq 0$ for all $t \in [0, T]$, then we take $\tau(x) = T + 1$. This event can be averted by selecting appropriate values of the variables $x = (x_1, x_2, x_3)$ from the 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 indicator of this process is the risk function $F(x) = \mathbf{E} \min\{0, R(x, \tau)\}$.

The function $f(x, \theta) = \min\{0, R(x, \tau)\}$ is obviously constructed using the operations of taking minimum and maximum.

Now assume that $\text{Prob}\{R(x, t) = 0\} = 0$ for all x and t (this always can be achieved by adding a random noise to the parameters of the process $R(x, t)$). Then with probability 1 the function $f(x, \theta)$ is generalized differentiable (see Section 3) with the subgradients

$$g(x, \theta) = \begin{cases} \begin{pmatrix} 1 \\ \tau(x) \\ -n(x_3) - \tau \dfrac{d}{dx_3} c(x_3) \end{pmatrix}, & \tau(x) \leq T, \\ 0 \in R^3, & \tau(x) > T, \end{cases}$$

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

2.2. Communication Networks with Failure

Consider a network of interconnected elements, which may be in a "normally operating" state or in a "faulty" state. The network has an input and an output and is regarded as normally operating if there exists a path from input to output consisting of normally operating elements. Denote by $\tau_i(x, \theta)$ the time of fault-free operation of element i, where $x \in R^n$ is the vector of controlled parameters and θ the vector of uncontrolled random parameters. Then the network lifetime $f(x, \theta)$ is expressed as the maxmin of $\tau_i(x, \theta)$:

$$f(x, \theta) = \max_{P_l \in \mathcal{P}} \min_{e \in P_l} \tau_e(x, \theta),$$

where $\mathcal{P}$ is the set of paths from network input to output; the subscript e identifies a path element.

For a sufficiently large general network, the function $f(x, \theta)$ cannot be computed analytically (it is difficult to enumerate all the paths P_l from $\mathcal{P}$). This rules out the deterministic approximation (3), (4). However, a simple (wave-front) algorithm exists that evaluates the function $f(x, \theta)$ and its stochastic (quasi)gradients $g(x, \theta)$ for each observation of the vector θ.

Simple conditions (see, e.g., [7, 13, 14]) guarantee differentiability of $F(x) = \mathbf{E}f(x, \theta)$ and the equality $\nabla F(x) = \mathbf{E}\partial f(x, \theta)$ even for nondifferentiable $f(x, \theta)$. These conditions, however, do not ensure continuous differentiability of $F(x)$, and this function may be nondifferentiable even in very simple practical cases (see [1, 13, 14]).

2.3. Simple Conveyer Line

A conveyer line [15] consists of n serially connected machines. A part moving down the line is sequentially served by each machine, if it is powered. Denote by x_i the time when machine i is powered; by y_i the time when the part leaves machine i'; by $y_0(\theta)$ the time of arrival of the part in the conveyer line; by $\tau_i(\theta)$ the (random) processing time of the part by machine i. Let a_i be the unit cost incurred when a part is waiting for machine i to be powered, b_i the cost incurred when a powered machine i is waiting for a part to arrive. Then the random cost associated with waiting for a machine to be powered or for a part to arrive for processing is calculated by the following recurrences:

$$f^0(\theta) = 0, \quad y_0 = 0,$$

$$f^i(x, y, \theta) = f^{i-1}(x, y, \theta) + \max\{b_i(y_{i-1} - x_i), a_i(x_i - y_{i-1})\},$$

$$y_i = \max\{y_{i-1}, x_i\} + \tau_i(\theta), \quad i = 1, 2, \ldots$$

The functions $f^i(x, y, \theta)$ are again constructed using the operations of taking maximum and minimum, and are nonconvex and nonsmooth.

2.4. Systems with Queues

Consider a network consisting of m servers handling messages or streams of messages. At each instant, a server may process only one message, which is then passed to another server in accordance with a known routing rule. If the next server is busy, the message is enqueued to be served according to the first-in, first-out rule.

For each server $i = 1, 2, \ldots, m$, we introduce the following notation: n_i is the initial queue length; $\tau_{ij}(x, \theta)$ is the (random) processing time of message j dependent on the controlled parameter x and the uncontrolled (random) parameter θ; $\alpha_{ij}(x, \theta)$ is the time of arrival of message j in sever i; $\beta_{ij}(x, \theta)$ is the time when server i starts processing message j; $\gamma_{ij}(x, \theta)$ is the time when server i finishes processing message j.

The algorithm for server i is described by the following recurrences:

$$\alpha_{i1} = \ldots = \alpha_{in_i} = 0, \quad \beta_{i1} = 0,$$

$$\beta_{ik} = \sum_{j=1}^{k-1} \tau_{ij}, \quad \gamma_{ik} = \sum_{j=1}^{k} \tau_{ij}, \quad k = 1, \ldots, n_i,$$

$$\beta_{ij} = \max \{\gamma_{i(j-1)}, \alpha_{ij}\},$$

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

Message streams are modeled in this system by introducing special servers that do not receive new messages, have an infinite message queue, and emit (processed) messages at appropriate intervals.

Note that each instant α when a message arrives at a server or each instant β when a server starts processing a new message coincides with the instant γ when some message completes processing on some server. It is therefore sufficient to consider only the processing completion moments γ.

The message routing procedure is defined by integer-valued functions $\mu_{ij}(x, \theta)$ that define the destination of message j processed by server i.

Note that many important efficiency measures of this network are nondifferentiable functions of x, despite the continuous differentiability of the functions $\tau_{ij}(x, \theta)$.

THEOREM 2.1 [13, 14]. Assume that $\mu_{ij}(x, \theta) = \mu_{ij}$. Then the function $\gamma_{ij}(x, \theta)$ is expressible in terms of $\tau_{ij}(x, \theta)$ by the operations of taking maximum and minimum and forming positive linear combinations.

Consider a particular case of the theorem, when the message paths μ_{ij} are fixed, $\mu_{ij} = \mu_i$. Denote by $I_i = \{\text{servers } r | \mu_r = i\}$ the set of correspondents of server i. Then

$$\gamma_{ij} = \max \left(\gamma_{i(j-1)}, \min_{i' \in I_i, j' \le j} \max (\alpha_{i(j-1)}, \gamma_{i'j'}) \right) + \tau_{ij},$$

which explains the assertion of Theorem 2.1.

The main operating measures (criteria) of this network are expressible in terms of the times $\gamma_{ik}(x, \theta)$. For instance, the mean queue length of server i is

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

the utilization rate of server i is

$$f(x, \theta) = \sum_{j=1}^{k} \tau_{ij}(x, \theta) / \gamma_{ik}(x, \theta)$$

and so on.

For a sufficiently general network configuration it is difficult to express the efficiency measures $f(x, \theta)$ explicitly in terms of the functions $\tau_{ij}(x, \theta)$. In general, these efficiency measures are obviously complex nonconvex nonsmooth functions of the network parameters.

3. GENERALIZED DIFFERENTIABLE FUNCTIONS

So-called generalized differentiable functions provide an appropriate model for the performance functions of discrete-event systems considered in the previous section.

Definition 3.1 [10]. The function f: $R^n \to R$ is called generalized differentiable (GD) at the point $x \in R^n$ if in some neighborhood of x there exists an upper semicontinuous multivalued mapping $\bar{\partial} f$ with closed convex values $\bar{\partial} f(x)$ such that

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

where $\langle \cdot, \cdot \rangle$ is the scalar product of two vectors in R^n, $g \in \bar{\partial} f(y)$, and the residual term satisfies the condition

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

for every sequence $y^k \to x$, $g^k \in \bar{\partial} f(y^k)$. The function f is called generalized differentiable if it is generalized differentiable at every point $x \in R^n$; $\bar{\partial} f(x)$ is the subdifferential of f at the point x.

Example 3.1. The function $|x|$, $x \in R$, is generalized differentiable with the subdifferential

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

Its decomposition (5) at the point $x = 0$ is given by

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

Generalized differentiable functions have the following properties (see [10, 16]):

— generalized differentiable functions are locally Lipschitzian, but in general are directionally nondifferentiable;

— continuously differentiable, convex and concave functions are generalized differentiable; the gradients and subgradients of these functions can be used as generalized gradients;

— the class of generalized differentiable functions is closed under the finite operations of taking maximum and minimum and under superposition;

— we have the calculus of generalized gradients

$$\bar{\partial} \max(f_1(x), f_2(x)) = \operatorname{conv}\{\bar{\partial} f_i(x) \mid f_i(x) = \max(f_1(x), f_2(x))\}, \tag{7}$$

where $\operatorname{conv}\{\cdot\}$ is the convex hull of the set $\{\cdot\}$, and the subdifferential $\bar{\partial} f_0(f_1, \ldots, f_m)$ of the compound function $f_0(f_1, \ldots, f_m)$ is evaluated by the standard chain rule;

— the class of generalized differentiable functions is closed under the expectation operation, and $\bar{\partial} F(x) = \mathbf{E}\bar{\partial} f(x, \omega)$ for $F(x) = \mathbf{E} f(x, \omega)$, where $f(\cdot, \omega)$ is a generalized differentiable function;

— the subdifferential $\bar{\partial} f(x)$ is nonuniquely defined by Definition 3.1, but the Clarke subdifferential [17] $\partial f(x)$ always satisfies Definition 3.1; for every $\bar{\partial} f(x)$ we have $\partial f(x) \subseteq \bar{\partial} f(x)$; $\bar{\partial} f(x)$ is a point almost everywhere in R^n;

— some elements $\partial f(x)$ for compound functions of the form $f(x) = \max(f_1(x), f_2(x))$, $f(x) = \min(f_1(x), f_2(x))$, and $f(x) = f_0(f_1(x), \ldots, f_m(x))$ can be computed by Nesterov's lexicographic method [18];

— we have the following analogue of the Newton–Leibniz formula:

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

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

These properties of generalized differentiable functions suggest that they are an appropriate model for performance functions of various nonsmooth stochastic systems (see Section 2).

4. DETERMINISTIC GENERALIZED GRADIENT METHOD WITH PROJECTION ONTO A NONCONVEX FEASIBLE SET

Let us consider a deterministic analogue of the stochastic problem (1), (2) to demonstrate the technique that we use to prove convergence of the generalized gradient method.

Consider the problem

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

where

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

$f(x)$ and $\psi(x)$ are generalized differentiable functions. Let $\bar{\partial} f(x)$ and $\bar{\partial}\psi(x)$ be subdifferentials of $f(x)$ and $\psi(x)$ at the point x, respectively. In particular, they may be identical with the Clarke subdifferentials $\partial f(x)$ and $\partial\psi(x)$.

Assume that constraint (9) is regular, i.e.,

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

for all x such that $\psi(x) = 0$. A necessary condition of optimality for this problem has the form (see [16])

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

where

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

Denote the solution set by $X^* = \{x \in X \mid 0 \in \bar{\partial} f(x) + N_X(x)\}$ and the set of optimal values by $f^* = \{f(x) \mid x \in X^*\}$.

Consider the following conceptual iterative method:

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

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

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

where Π_X is the (multivalued) projector on the set X, i.e., $z \in \Pi_X(y)$ if and only if $y - z \in N_X(z)$; the nonnegative numbers ρ_k are defined by the 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 a generalization to the nonconvex case of the subgradient methods of Shor [19], Ermol'ev [20], and Polyak [21] (originally developed for convex functions $f(x)$ and a convex set X). A similar method is considered in [11, 12] for the class of subdifferentially regular (quasidifferentiable) functions, which are not applicable for the important applications of Section 2 (for instance, they include convex and weakly convex [22] functions and the maximum function, but do not include concave functions and the minimum function).

THEOREM 4.1. The sequence $\{x^k\}$ generated by method (11)-(14) converges to the solution of problem (8) in the function, i.e., the minimum (by the function f) limit points of $\{x^k\}$ are contained in X^*, and all the limit points of the numerical sequence $\{f(x^k)\}$ constitute an interval in the set $f^* = \{f(x) \mid x \in X^*\}$. If the set f^* does not contain intervals (for instance, f^* is finite or countable), then $\{x^k\}$ converges to the solution of the problem, i.e., all the limit points of $\{x^k\}$ constitute a connected subset of the set X^* and $\{f(x^k)\}$ has a limit in f^*.

The convergence theorem is proved by contradiction using a nonconvex nonsmooth Lyapunov function $f(x)$ for the sequence $\{x^k\}$ (as in [11, 12, 16, 20, 22]). The proof repeatedly uses the following assertion concerning the sequence $\{x^k\}$ generated by the algorithm (11)-(14).

LEMMA 4.1. If $\lim_{s\to\infty} x^{k_s} = y \in X^*$, then for every $\varepsilon > 0$ there exist indices $l_s > k_s$ such that $\|x^k - y\| \leq \varepsilon$ for all $k \in [k_s, l_s)$ and

$$\limsup_{s\to\infty} f(x^{l_s}) < f(y) = \lim_s f(x^{k_s}). \tag{15}$$

Proof. Let $\bar{x}^{k+1} = x^k - \rho_k g^k$ and write

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

where

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

$$h^k = h^k(\bar{x}^{k+1}) = \frac{1}{\rho_k}\left(\bar{x}^{k+1} - \Pi_X(\bar{x}^{k+1})\right) \in N_X(x^{k+1}). \tag{16}$$

We have the bounds

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

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

Two cases have to be considered: $\psi(y) < 0$ and $\psi(y) = 0$. In the first case, for $k \geq k_s$ the method (12)-(14) functions in a sufficiently small neighborhood of the point y as a subgradient method in the unconstrained problem, and the assertion of the lemma is well known (see [10, 16]).

We will consider the case $\psi(y) = 0$ (the case $\psi(y) < 0$ can be treated as a simpler repetition of the case $\psi(y) = 0$). For $y = \lim_{s\to\infty} x^{k_s}$ let

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

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

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

By upper semicontinuity of $\bar{\partial}f$, $\bar{\partial}\psi$, there exists an ε_1-neighborhood of the point y such that

$$\sup_{g,z} \{\|g\| \mid g \in \bar{\partial}f(z), \quad \|z - y\| \leq \varepsilon_1\} \leq 2\gamma = \Gamma, \tag{20}$$

$$\sup_{g,z} \{\|g\| \mid g \in \bar{\partial}\psi(z), \quad \|z - y\| \leq \varepsilon_1\} \leq 2\gamma = \Gamma. \tag{21}$$

Let

$$\bar{N}(z) = \{g \in N_X(z) \mid \|g\| \leq \Gamma\},$$

$$\bar{G}(z) = \bar{\partial}f(z) + \bar{N}(z).$$

Clearly,

$$\rho\,(0, \bar{G}(y)) = \inf_{g} \{\|g\| \mid g \in \bar{G}\,(y)\} \geq \nu.$$

By upper semicontinuity of $\bar{\partial}\psi$ and $\bar{G}$ there exists an ε_2-neighborhood ($\varepsilon_2 \leq \varepsilon_1$) of the point y such that for all z, $\|z - y\| \leq \varepsilon_2$,

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

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

By generalized differentiability of f and ψ, for the constant $c = \nu^2/(64\,\Gamma(1 + 2\,\Gamma/\mu))$ there exists $\varepsilon_3 \leq \varepsilon_2$ such that given $\|z - y\| \leq \varepsilon_3$ we have

$$f\,(z) \leq f\,(y) + \langle g, z - y\rangle + c\,\|z - y\|, \tag{23}$$

$$\psi\,(z) \leq \psi\,(y) + \langle d, z - y\rangle + c\,\|z - y\| = \langle d, z - y\rangle + c\,\|z - y\| \tag{24}$$

for all $g \in \bar{\partial}f(z)$, $d \in \bar{\partial}\psi(z)$.

Now let $\bar{\varepsilon} = \varepsilon_3$, $\bar{\rho}_1 = \varepsilon/(4\Gamma)$ and take some $\varepsilon \leq \bar{\varepsilon}$. Let $\|x^{k_s} - y\| \leq \varepsilon/4$ and $\rho_s \leq \bar{\rho}_1$ for all $s \geq S$.

Let

$$m_s = \sup\{m \mid \|x^k - y\| \leq \varepsilon / 2 \quad \forall k \in [k_s, m)\}.$$

We will show that $m_s < \infty$ for all $s \geq S$. Indeed, if for all k we have $\|x^k - y\| \leq \varepsilon/2$, we get a contradiction:

$$\varepsilon / 2 \geq \|x^k - y\| \geq \|x^k - x^{k_s}\| - \|x^{k_s} - y\| \geq \nu / 2 \sum_{r=k_s}^{k-1} \rho_r - \varepsilon / 4 \to \infty,$$

when $k \to \infty$. Now,

$$\|x^{m_s} - y\| \leq \|x^{m_s - 1} - y\| + \rho_{m_s - 1}\,\|Q_s^{m_s - 1}\| \leq 3\varepsilon / 4.$$

Since

we have

$$\varepsilon / 4 \leq \Big\|\sum_{k=k_s}^{m_s - 1} \rho_k\, Q^k\Big\| \leq \Gamma \sum_{k=k_s}^{m_s - 1} \rho_k,$$

$$\sum_{k=k_s}^{m_s - 1} \rho_k \geq \frac{\varepsilon}{4\Gamma}.$$

For $k \in [k_s, m_s]$, $s \geq S$, substitute the approximations x^k and the subgradients $g^k \in \bar{\partial}f(x^k)$, generated by algorithm (11)-(13) in decomposition (23):

$$\begin{aligned} f\,(x^k) &\leq f\,(y) + \langle g^k, x^k - y\rangle + c\,\|x^k - y\| \leq \\ &\leq f\,(y) + \langle g^k, x^k - x^{k_s}\rangle + c\,\|x^k - x^{k_s}\| + (\Gamma + c)\,\|x^{k_s} - y\| = \\ &= f\,(y) + \langle g^k + h^k, x^k - x^{k_s}\rangle - \langle h^k, x^k - x^{k_s}\rangle + c\,\|x^k - x^{k_s}\| + (\Gamma + c)\,\|x^{k_s} - y\|, \end{aligned} \tag{25}$$

where h^k is defined by Eq. (16). Let us bound the term $u_k = -\langle h^k, x^k - x^{k_s}\rangle$.

If $\psi(\bar{x}^k) \leq 0$, then $h^k = 0$ and $u_k = 0$. Consider the case $\psi(\bar{x}^k) > 0$, i.e., $h^k \neq 0$. Since

$$h^{k+1} \in N_X(x^{k+1}) = \{\lambda g \mid g \in \bar{\partial}\psi(x^{k+1}), \lambda \geq 0\},$$

we obtain

$$h^k = \lambda_k d^k, \quad d^k \in \bar{\partial}\psi(x^{k+1}), \quad \lambda_k > 0,$$

and

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

Substitute in Eq. (24) $x^{k+1} = \Pi_X(\bar{x}^{k+1})$ for z and d^k for d:

$$0 = \psi(x^{k+1}) \leq \langle d^k, x^{k+1} - y\rangle + c\,\|x^{k+1} - y\|. \tag{26}$$

Now, multiplying Eq. (26) by λ_k we obtain

$$\begin{aligned} -\langle h^k, x^k - y\rangle &\leq \lambda_k c\,\|x^{k+1} - y\| + \Gamma\|x^{k+1} - x^k\| \leq (2c\,\Gamma / \mu)\|x^k - y\| + \\ &+ \Gamma(1 + 2c/\mu)\|x^{k+1} - x^k\| \leq \\ &\leq (2c\,\Gamma/\mu)\|x^k - x^{k_s}\| + (2c\,\Gamma/\mu)\|x^{k_s} - y\| + \Gamma(1 + 2c/\mu)\,\|x^{k+1} - x^k\|. \end{aligned} \tag{27}$$

Using inequality (27), we rewrite Eq. (25) in the form

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

We now have to bound the scalar product

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

To this end, we use the following lemma.

LEMMA 4.2 [16]. 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 set of vectors $\{p^r \in P \mid r = k, \ldots, m\}$ and every set of nonnegative numbers $\{\rho^r \in R^1 \mid r = k, \ldots, m-1\}$ 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 an index $l \in (k, m]$ for which

$$\left\langle p^l, \sum_{r=k}^{l-1} \rho^r p^r \Big/ \sum_{r=k}^{l-1} \rho^r \right\rangle \geq \frac{\gamma_0^2}{4}, \quad \sum_{r=k}^{l-1} \rho^r \geq \frac{\sigma_0 \gamma_0}{3\Gamma_0}.$$

We now continue the proof of Lemma 4.1. Let

$$P = \operatorname{conv}\{\bar{G}(z) \mid \|z - y\| \leq \varepsilon\},$$

$$p^r = g^r + h^r, \quad k = k_s \leq r \leq m = m_s,$$

$$\gamma_0 = \nu / 2, \quad \Gamma_0 = \Gamma.$$

We have the inequalities

$$\sum_{k=k_s}^{m_s} \rho_s^k \geq \sum_{k=k_s}^{m_s - 1} \rho_s^k \geq \frac{\|x^{m_s} - x^{k_s}\|}{\Gamma} \geq \frac{\varepsilon}{4\Gamma} = \sigma_0 > 0$$

and

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

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

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

$$\sum_{k=k_s}^{l_s - 1} \rho_s^k \geq \frac{\varepsilon \nu}{24 \Gamma^2}.$$

Substituting these bounds in inequality (28) for $k = l_s$, we obtain the final bound

$$f(x^{l_s}) \leq f(y) - \frac{\nu^2}{16} \sum_{k=k_s}^{l_s - 1} \rho_k + \Gamma(1 + 2\Gamma/\mu)\, c \sum_{k=k_s}^{l_s - 1} \rho_k + (\Gamma + c + 2c\Gamma/\mu)\|x^{k_s} - y\| +$$

$$+ \Gamma^2(1 + 2c/\mu)\rho_{l_s} \leq \tag{29}$$

$$\leq f(y) - \frac{\nu^2}{600\,\Gamma^2} \varepsilon \nu + (\Gamma + c + 2c\Gamma/\mu)\|x^{k_s} - y\| + \Gamma^2(1 + 2c/\mu)\rho_{l_s},$$

where $c = \nu^2/(64\Gamma(1 + 2\Gamma/\mu))$.

We have thus proved that for all sufficiently small $\varepsilon \leq \bar{\varepsilon}$ and sufficiently large s there exist indices l_s such that $\|x^k - y\| \leq \varepsilon$ for $k \in [k_s, l_s)$ and $f(x^{l_s})$ satisfies Eq. (29). This completes the proof of Lemma 4.1. Q.E.D.

Proof of Theorem 4.1. The proof involves multiple applications of Lemma 4.1. in the framework of the convergence proof procedure described in [22] and generalized in [11, 16, 20]. The proof consists of the following steps.

1^0. The sequence $\{x^k\}$ is obviously contained in the compact set X.

2^0. By boundedness of the generalized gradients $\partial f(x)$ on the compact set X, we obtain

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

Hence it follows that the limit points of the sequence $\{x^k\}$ form a connected set in X.

3^0. The sequence $\{x^k\}$, for example, from the compact set X has a compact set of limit points X'. The continuous function $f(x)$ attains its minimum on X' at the point x'. Note that the point $x' = \lim_{s\to\infty} x^{k_s}$ is contained in X^*, because otherwise by Lemma 4.1 it is not a minimum point of f. Thus, $\lim_{k\to\infty} \inf f(x^k) \in f^*$.

4^0. Let us show that the limit points of the sequence $\{f(x^k)\}$ form an interval in f^*. If $\lim_{k\to\infty} \sup f(x^k) = \lim_{k\to\infty} \inf f(x^k)$, then the assertion follows from 3^0. Let

$$\lim_{k\to\infty} \sup f(x^k) > \lim_{k\to\infty} \inf f(x^k) = f_0^* \in f^*.$$

Assume that the assertion of the theorem is false. Then there exists a number $f_1 \in \bar{f}^*$ such that $f_1 < \lim_{k\to\infty} \sup f(x^k(\omega))$. Choose the number f_2 so that

$$\lim_{k\to\infty} \inf f(x^k) = f_0^* < f_1 < f_2 < \lim_{k\to\infty} \sup f(x^k).$$

The sequence $\{f(x^k)\}$ intersects the interval (f_1, f_2) from bottom up infinitely many times. Thus there exists subsequences $\{x^{k_s}\}$ and $\{x^{n_s}\}$ such that

$$f(x^{k_s}) \leq f_1 < f(x^k) < f_2 \leq f(x^{n_s}), \quad k_s < k < n_s, \quad s = 1, 2, \ldots \tag{30}$$

Without loss of generality we may assume that $x^{k_s} \to x'$. By 2^0 and continuity of f we have

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

Thus, $\lim_{s\to\infty} x^{k_s} = x' \in \bar{X}^*$. We can now apply Lemma 4.1 to the subsequences $\{x^k\}_{k\geq k_s}$, $s = 1, 2, \ldots$. Take ε such that

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

Then property (15) contradicts inequalities (30). The contradiction proves that

$$\left[\lim_{k\to\infty} \inf f(x^k), \lim_{k\to\infty} \sup f(x^k)\right) \subseteq f^*.$$

Since X^* and f^* are closed sets, we have

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

5^0. Now assume that f^* does not contain intervals, for instance, f^* is finite or countable. From 4^0 we obtain

$$\lim_{k\to\infty} f(x^k) = f_0^* \in f^*. \tag{31}$$

If some limit point $x' = \lim_{s\to\infty} x^{k_s}$ is not contained in X^*, then by Lemma 4.1 we obtain a contradiction with the convergence of the sequence $\{f(x^k)\}$ stated in (31). Q.E.D.

Remark 4.2. Theorem 4.1 remains valid also for the randomized generalized gradient method of the type (11)-(14), where the gradient g^k is evaluated not at the current point x^k, but instead at a close point $\tilde{x}^k$, i.e.,

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

In this case, Lemma 4.1 follows from the stability property of the method stated in Lemma 5.4 below. If the points $\bar{x}^k$ are chosen at random, then $\bar{\partial} f(\bar{x}^k) = \partial f(\bar{x}^k)$, with probability 1 and the method converges to the stationary (in Clarke's sense [17]) points $X^* = \{x \mid 0 \in \partial f(x) + N_X(x)\}$. In the last case we may apply formula (7) and the chain rule to evaluate $g^k \in \bar{\partial} f(\bar{x}^k)$.

5. STOCHASTIC GENERALIZED GRADIENT METHOD WITH PROJECTION ONTO A NONCONVEX FEASIBLE SET

Consider the stochastic programming problem (1), (2), where the objective function $F(x)$ is generalized differentiable and the set $X = \{x \mid \psi(x) \leq 0\}$ is defined by a generalized differentiable function $\psi(x)$ that satisfies the regularity condition (10). Let $X^* = \{x \mid 0 \in \bar{\partial} F(x) + N_X(x)\}$ and $F^* = \{F(x) \mid x \in X^*\}$.

Let us consider a generalization of the stochastic quasigradient method of [20] to the case when the objective function $F(x) = \mathbf{E} f(x, \theta)$ and the constraint set X are nonconvex:

$$x^0 \in X, \tag{32}$$

$$x^{k+1}(\omega) \in \Pi_X(x^k - \rho_k s^k(\omega)), \quad k = 0, 1, \ldots, \tag{33}$$

$$s^k(\omega) = \frac{1}{n_k} \sum_{i=r_k}^{k} \xi^i(\omega), \quad n_k = k - r_k + 1 \geq 0, \tag{34}$$

Here the random variables $x^k(\omega)$, $\xi^k(\omega)$, $s^k(\omega)$, $k = 0, 1, \ldots$, 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 the conditional means are

$$\mathbf{E}\{\xi^i(\omega) \mid x^0(\omega), \ldots, x^i(\omega)\} = g^i(\omega) \in \bar{\partial} f(x^i(\omega)), \tag{35}$$

$$\|\xi^i(\omega)\| \leq C < +\infty;$$

Π_X is the (multivalued) projector on X, i.e., $z \in \Pi_X(y)$ if and only if $y - z \in N_X(z)$; the nonnegative numbers r_k, n_k and the monotone decreasing sequence of nonnegative numbers ρ_k satisfy the conditions

$$n_k = k + 1 - r_k \leq m < +\infty; \tag{36}$$

$$\sum_{k=0}^{\infty} \rho_k = +\infty, \quad \sum_{k=0}^{\infty} \rho_k^2 < +\infty. \tag{37}$$

Remark 5.1. The method (32)-(34) combines the ideas of the stochastic quasigradient projection method [20] (for a convex function $F(x)$ and a convex set X) and averaged stochastic gradient methods [8, 12, 16, 23-25] (for nonconvex functions $F(x)$). In [11, 12] a similar method is studied for the case of subdifferentially regular (quasidifferentiable) functions $F(x)$ and $\psi(x)$. This method, however, is not applicable to the examples from Section 2.

THEOREM 5.1. Assume that $F(x)$ and $\psi(x)$ are generalized differentiable functions, and the sequence $x^k(\omega)$ is generated by method (32)-(34), where r_k, n_k, ρ_k satisfy (36), (37). Then the minimum (by the function F) limit points of the sequence $\{x^k(\omega)\}$ are almost surely contained in X^* and all limit points of the numerical sequence $\{F(x^k(\omega))\}$ almost surely form an interval in the set F^*. If the set F^* contains no intervals (for instance, it is finite or countable), then all the limit points of $\{x^k(\omega)\}$ almost surely form a connected subset of the set X^* and the sequence $\{F(x^k(\omega))\}$ has a limit in the set F^*.

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

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

where

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

$$h^k = \frac{1}{\rho_k}\left(\bar{x}^{k+1} - \Pi_X(\bar{x}^{k+1})\right) \in N_X(x^{k+1}).$$

The norms are bounded as follows:

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

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

Now take a subsequence $\{x^{k_s}(\omega)\}$. For $k \geq k_s$,

$$x^{k+1}(\omega) = x^{k_s}(\omega) - \sum_{t=k_s}^{k} \rho_t Q^t(\omega) =$$

$$= x^{k_s}(\omega) - \sum_{t=k_s}^{k} \rho_t \bar{Q}^t(\omega) - \zeta_{k_s}^{k+1}(\omega) = y_{k_s}^{k+1}(\omega) - \zeta_{k_s}^{k+1}(\omega), \tag{38}$$

where

$$y_{k_s}^{k+1}(\omega) = \sum_{t=k_s}^{k} \rho_t \bar{Q}^t(\omega) = y_{k_s}^{k}(\omega) - \rho_k \bar{Q}^k(\omega), \quad k \geq k_s, \tag{39}$$

$$y_{k_s}^{k_s}(\omega) = x^{k_s}(\omega), \tag{40}$$

$$\bar{Q}^k(\omega) = \frac{1}{n_k} \sum_{r=r_k}^{k} (\bar{g}^r(\omega) + h^r(\omega)), \tag{41}$$

$$\bar{g}(\omega) = \mathrm{E}\{\xi^r(\omega) \mid x^0(\omega), \ldots, x^r(\omega)\} \in \bar{\partial} F(x^r(\omega)), \tag{42}$$

$$h^r(\omega) = \frac{1}{\rho_r}\left(\bar{x}^{r+1}(\omega) - \Pi_X(\bar{x}^{r+1}(\omega))\right) \in N_X(x^{r+1}(\omega)), \tag{43}$$

$$\zeta_{k_s}^{k+1}(\omega) = \sum_{t=k_s}^{k} \rho_t \frac{1}{n_t} \sum_{r=r_t}^{t} (\xi^r(\omega) - \bar{g}^r(\omega)). \tag{44}$$

Instead of the sequence $\{x^k(\omega)\}_{k \geq k_s}$, consider the close sequence $\{y_{k_s}^k(\omega)\}_{k \geq k_s}$, $s = 0, 1, \ldots$, generated by the deterministic procedure (39)-(44) (with a fixed ω). This procedure uses the generalized gradients $\bar{g}^r(\omega)$ of the function F evaluated not at the points $y_{k_s}^k(\omega)$ but at the close points $x^r(\omega)$. Moreover, the vector $h^r(\omega)$ is the normal to X not at the point

$y_{k_s}^k(\omega)$, but at the close point $x^{r+1}(\omega)$. As a result, we obtain the bound

$$\| y_{k_s}^k(\omega) - x^k(\omega) \| = \| \xi_{k_s}^k(\omega) \| \leq \sup_{k \geq k_s} \| \xi_{k_s}^k(\omega) \| = \delta_{k_s}(\omega).$$

In Lemma 5.1 we will show that $\lim_{s \to \infty} \delta_{k_s}(\omega) = 0$ almost surely. Note that

$$| F(x^k(\omega)) - F(y_{k_s}^k(\omega)) | \leq L_F \| x^k(\omega) - y_{k_s}^k(\omega) \| = L_F \delta_{k_s}(\omega), \tag{45}$$

where L_F is the Lipschitz constant of the function F on the set X. Hence it follows that the differences $|F(x^k(\omega)) - F(y_{k_s}^k(\omega))|$, $k \geq k_s$, may be arbitrarily small when s is sufficiently large. The rest of the proof is divided into several lemmas.

LEMMA 5.1. The random sequence $\{\zeta_0^k(\omega)\}_{k=0}^{\infty}$, where

$$\zeta_0^k(\omega) = \sum_{t=0}^{k-1} \rho_t \frac{1}{n_t} \sum_{r=r_t}^{t} (\xi^r(\omega) - \bar{g}^r(\omega)), \quad n_t \leq m, \tag{46}$$

almost surely has a limit.

Proof. Let

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

Then

$$\zeta_0^k = \sum_{t=0}^{k-1} \rho_t \sum_{r=r_t}^{t} \lambda_{tr} (\xi^r - \bar{g}^r) = \sum_{r=0}^{k-1} \left(\sum_{t=r}^{k-1} \lambda_{tr} \rho_t \right) (\xi^r - \bar{g}^r) =$$

$$= \sum_{r=0}^{k-1} \left(\sum_{t=r}^{\infty} \lambda_{tr} \rho_t \right) (\xi^r - \bar{g}^r) - \sum_{r=0}^{k-1} \left(\sum_{t=k}^{\infty} \lambda_{tr} \rho_t \right) (\xi^r - \bar{g}^r).$$

The sequence

$$\bar{\zeta}_0^k = \sum_{r=0}^{k-1} \left(\sum_{t=r}^{\infty} \lambda_{tr} \rho_t \right) (\xi^r - \bar{g}^r) \tag{47}$$

is a martingale with respect to the stream of σ-algebras generated by the sequence $\{x^k(\omega)\}_{k=0}^{\infty}$. Let

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

Then

$$\mathrm{E} \| \bar{\zeta}_0^k(\omega) \|^2 \leq (\Gamma + C)^2 \sum_{r=0}^{\infty} \left(\sum_{t=r}^{\infty} \lambda_{tr} \rho_t \right)^2 \leq (\Gamma + C)^2 \sum_{r=0}^{\infty} \left(\sum_{t=r}^{r+m} \lambda_{tr} \rho_t \right)^2 \leq$$

$$\leq (\Gamma + C)^2 m^2 \sum_{r=0}^{\infty} \rho_t^2 < + \infty,$$

$$\mathrm{E} \| \bar{\zeta}_0^k(\omega) \| \leq 1 + \mathrm{E} \| \bar{\zeta}_0^k(\omega) \|^2 \leq \text{const} < + \infty .$$

The martingale (47) thus almost surely has a limit. For the residual term

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

we have the following bounds

$$\alpha^k(\omega) \leqslant \sum_{r=0}^{k-1}\left(\sum_{t=k}^{\infty}\lambda_{tr}\rho_t\right)(\|\xi^r\| + \|\bar{g}^r\|) \leqslant$$

$$\leqslant (\Gamma + C)\sum_{r=0}^{k-1}\left(\sum_{t=k}^{\infty}\lambda_{tr}\rho_t\right) = (\Gamma + C)\sum_{t=k}^{\infty}\rho_t\left(\sum_{r=0}^{k}\lambda_{tr}\right) =$$

$$= (\Gamma + C)\sum_{t=k}^{\infty}\rho_t\left(\sum_{r=r_k}^{k}\lambda_{tr}\right) \leqslant (\Gamma + C)\sum_{t=k}^{k+m}\rho_t \to 0 \quad \text{as} \quad k \to \infty.$$

Thus, the sequence $\{\zeta_0^k(\omega) = \bar{\zeta}_0^k(\omega) + \alpha^k(\omega)\}$ almost surely has a limit. Q.E.D.

COROLLARY 5.1. For every sequence of indices $\{k_s\} \to \infty$ we have

$$\delta_{k_s}(\omega) = \sup_{k \geq k_s}\|\zeta_{k_s}^k(\omega)\| \to 0 \quad \text{almost surely as } s \to \infty$$

Remark 5.2. Lemma 5.1 and Corollary 5.1 remain valid if $r_k = k$ in Eqs. (34) and (35) is replaced with $\mathbf{E}\|\xi^i(\omega)\|^2 < C < +\infty$.

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

$$m_s(\varepsilon, \omega) = \sup\{m \mid \|x^k(\omega) - x(\omega)\| \leqslant \varepsilon \quad \text{for all} \quad k \in [k_s, m)\}$$

Then $\bar{\varepsilon}(\omega)$ exists almost surely such that for every $\varepsilon \in (0, \bar{\varepsilon}]$ there are indices $l_s(\omega) \in [k_s(\omega), m_s(\varepsilon, \omega)]$, for which

$$F(x(\omega)) = \lim_{s\to\infty} F(x^{k_s}(\omega)) > \limsup_{s\to\infty} F(x^{l_s}(\omega)). \tag{48}$$

Lemma 5.2. follows by Eqs. (38), (45) and Corollary 5.1 from a similar property of the sequences $\{y_{k_s}^k(\omega)\}_{k\geq k_s}$, generated by Eqs. (39)-(42). Let us state this property in the form of a separate lemma.

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

$$m_s(\varepsilon, \omega) = \sup\{m \mid \|y_{k_s}^k(\omega) - x(\omega)\| \leqslant \varepsilon \quad \text{for all} \quad k \in [k_s, m)\}.$$

Then $\bar{\varepsilon}(\omega)$ exists almost surely such that for every $\varepsilon \in (0, \bar{\varepsilon}]$ there are indices $l_s(\omega) \in [k_s(\omega), m_s(\varepsilon, \omega)]$, for which

$$F(x(\omega)) = \lim_{s\to\infty} F(x^{k_s}(\omega)) > \limsup_{s\to\infty} F(y_{k_s}^{l_s}(\omega)). \tag{49}$$

Lemma 5.3 in turn follows from a stability property of the deterministic subgradient method (11)-(13).

LEMMA 5.4. Assume that the sequence of initial points $\{y^s\}$ converges to $y = \lim_{s\to\infty} y^s$. For each s consider the sequence $\{y_s^k\}_{k=k_s}^{n_s}$ such that

$$y_s^{k_s} = y^s,$$

$$y_s^{k+1} = y_s^k - \rho_k(g_s^k + h_s^k), \quad s \le k < n_s,$$

$$g_s^k \in G_{\delta_s^k}(y_s^k) = \operatorname{co}\{g \in \bar{\partial} f(y) \mid \|y - y_s^k\| \le \delta_s^k\},$$

$$h_s^k \in \left\{\frac{y - \Pi_X(y)}{\rho_k} \mid \|y - \bar{y}_s^k\| \le \delta_s^k\right\},$$

$$\bar{y}_s^k = y_s^k - \rho_s^k g_s^k,$$

where $\operatorname{co}\{\cdot\}$ is the convex hull of the set $\{\cdot\}$.

Let

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

If $0 \in \bar{\partial} f(y) + N_X(y)$ and $\sigma_s \ge \sigma \ge 0$, then for every sufficiently small ε there exist $\bar{\rho} = \bar{\rho}(y, \varepsilon)$ and $\bar{\delta} = \bar{\delta}(y, \varepsilon)$ such that for $\{y_s^k\}_{k=k_s}^{n_s}$ with the parameters $\delta_s^k \le \bar{\delta}$ and $\rho_s^k \le \bar{\rho}$ there are indices l_s for which $\|y_s^k - y\| \le \varepsilon$ when $k \in [k_s, l_s)$ and $f(y) = \lim_{s \to \infty} f(y^s) > \limsup_{s \to \infty} f(y_s^{l_s})$.

Proof is similar to that of Lemma 4.1. We have to consider two cases: $\psi(y) < 0$ and $\psi(y) = 0$. In the first case the generalized gradient method operates in a sufficiently small neighborhood of the point y as for an unconstrained problem, and the assertion of the lemma is well known (see [16]). In what follows we consider the case when $\psi(y) = 0$ (the case $\psi(y) < 0$ may be treated as a simpler repetition of the case $\psi(y) = 0$). As in the proof of Lemma 4.1, for $y = \lim_s y^s$ we define μ, ν, γ by relationships (17)-(19) and choose ε_1, ε_2, ε_3, c so that relationships (20)-(24) are satisfied.

Now, set $\bar{\varepsilon} = \min\{\varepsilon_3, \sigma\nu/2\}$ and take some $\varepsilon \le \bar{\varepsilon}$. Let $\bar{\delta}_1 = \varepsilon/4$, $\bar{\rho}_1 = \varepsilon/(4\Gamma)$ and assume that $\|y^s - y\| \le \varepsilon/4$, $\delta_s \le \bar{\delta}_1$, $\rho_s \le \bar{\rho}_1$ for $s \ge S$.

Define the index

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

We will show that $\varepsilon/2 \le \|y_s^{m_s} - y\| \le 3\varepsilon/4$.

First let us prove the left-hand inequality. If $\|y_s^{m_s} - y\| \le \varepsilon/2$, then $m_s = n_s$, and we obtain a contradiction:

$$\varepsilon_2 > 3\varepsilon/4 \ge \|y_s^{n_s} - y_s\| \ge \sigma\nu/2.$$

Now,

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

Since

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

we have

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

Let $g_s^k \in G_{\delta_s k}(y_s^k)$. Then

$$g_s^k = \sum_{i=1}^{n+1} \lambda_s^{ki} g_s^{ki}, \quad \sum_{i=1}^{n+1} \lambda_s^{ki} = 1, \quad g_s^{ki} \in \bar{\partial} f(y_s^{ki}), \quad \|y_s^{ki} - y_s^k\| \leq \delta_s^k.$$

If $\|y^s - y\| \leq \varepsilon/4$, $\delta_s \leq \varepsilon/4$, $k_s \leq k \leq m_s$, $1 \leq i \leq n+1$, then $\|y_s^{ki} - y\| \leq \|y_s^{ki} - y_s^k\| + \|y_s^k - y\| \leq \delta_s^k + 3\varepsilon/4 \leq \varepsilon \leq \varepsilon_3$.

Using decomposition (23) for $z = y_s^{ki}$ we write

$$f(y_s^{ki}) \leq f(y) + \langle g_s^{ki}, y_s^{ki} - y^s \rangle + c \|y_s^{ki} - y^s\| + (\Gamma + c)\|y^s - y\|.$$

Here y_s^{ki} $(1 \leq i \leq n+1)$ can be approximately replaced with y_s^k. Thus,

$$f(y_s^k) \leq f(y) + \langle g_s^{ki}, y_s^k - y^s \rangle + c \|y_s^k - y\| + (2\Gamma + c)\delta_s + (\Gamma + c)\|y^s - y\|.$$

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

$$\begin{aligned} f(y_s^k) &\leq f(y) + \langle g_s^k, y_s^k - y^s \rangle + c \|y_s^k - y^s\| + (2\Gamma + c)\delta_s + (\Gamma + c)\|y^s - y\| = \\ &= f(y) + \langle g_s^k + h_s^k, y_s^k - y^s \rangle - \langle h_s^k, y_s^k - y^s \rangle + \\ &+ c \|y_s^k - y^s\| + (2\Gamma + c)\delta_s + (\Gamma + c)\|y^s - y\|, \end{aligned} \tag{50}$$

where

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

Let us bound the term $u_s^k = -\langle h_s^k, y_s^k - y^s \rangle$.

If $\psi(y_s^k) \leq 0$, then $h_s^k = 0$ and $u_s^k = 0$. Consider the case when $\psi(y_s^k) > 0$, i.e., $u_s^k \neq 0$. Since

$$h_s^k \in N_X(z_s^k) = \{\lambda g \mid g \in \bar{\partial}\psi(z_s^k), \lambda \geq 0\},$$

we have

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

We have the bounds

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

Substitute z_s^k and d_s^k in decomposition (24):

$$\psi(z_s^k) \leq \langle d_s^k, z_s^k - y \rangle + c \|z_s^k - y\|. \tag{51}$$

Note that $y_s^{k+1} = \bar{y}_s^k - \rho_s^k h_s^k = \bar{y}_s^k - \bar{y}_s^k + z_s^k$ and thus $\|y_s^{k+1} - z_s^k\| = \|\bar{y}_s^k - \bar{y}_s^k\| \leq \delta_s^k$.

Replacing z_s^k in (51) with the close point y_s^{k+1}, we obtain

$$0 = \psi(y_s^{k+1}) \leq \langle d_s^k, y_s^{k+1} - y \rangle + c \|y_s^{k+1} - y\| + (2\Gamma + c)\delta_s^k. \tag{52}$$

Now multiply Eq. (52) by $\lambda_s^k \leq 2\Gamma/\mu$:

$$\begin{aligned} -\langle h_s^k, y_s^k - y^s \rangle \leq (2\Gamma c/\mu)\|y_s^k - y^s\| + \Gamma(1 + 2c/\mu)\|y_s^{k+1} - y_s^k\| + \\ + \Gamma(1 + 2c/\mu)\|y^s - y\| + 2\Gamma(2\Gamma + c)\delta_s^k/\mu. \end{aligned} \tag{53}$$

Using inequality (53), we rewrite Eq. (50) in the form

$$\begin{aligned} f(y_s^k) \leq f(y) + \langle g_s^k + h_s^k, y_s^k - y^s \rangle + (c + 2\Gamma c/\mu)\|y_s^k - y^s\| + (1 + 2c/\mu)\Gamma^2 \rho_s^k + \\ + (2\Gamma + c + 2\Gamma c/\mu)\|y^s - y\| + (2\Gamma + c)(1 + 2\Gamma c/\mu)\delta_s^k. \end{aligned} \tag{54}$$

Now use Lemma 4.2 to bound the scalar products

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

Let

$$P = \operatorname{conv}\{\bar{G}(z) \mid \|z - y\| \leq \varepsilon\},$$

$$p^r = g_s^r + h_s^r, \quad k = k_s \leq r \leq m = m_s,$$

$$\gamma_0 = \nu/2, \quad \Gamma_0 = 2\Gamma,$$

Then

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

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

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

$$\langle g_s^{l_s} + h_s^{l_s}, \sum_{k=k_s}^{l_s - 1} \rho_s^k Q_s^k / \sum_{k=k_s}^{l_s - 1} \rho_s^k \rangle \geq \frac{\nu^2}{16}, \quad \sum_{k=k_s}^{l_s - 1} \rho_s^k \geq \frac{\varepsilon\nu}{48\Gamma^2}.$$

Substituting these bounds in inequality (54) for $k = l_s$, we obtain a final bound for $c = \nu^2/(64\Gamma(1 + 2\Gamma/\mu))$.

$$\begin{aligned} f(y_s^{l_s}) \leq f(y) - \frac{\nu^2}{16} \sum_{k=k_s}^{l_s - 1} \rho_s^k + 2\Gamma(1 + 2\Gamma/\mu)c \sum_{k=k_s}^{l_s - 1} \rho_s^k + \\ + \Gamma^2(1 + 2c/\mu)\rho_s + (2\Gamma + c)(1 + 2\Gamma c/\mu)\delta_s + (2\Gamma + c + 2\Gamma c/\mu)\|y^s - y\| \leq \\ \leq f(y) - \frac{\nu^2}{1600\Gamma^2}\varepsilon\nu + \Gamma^2(1 + 2c/\mu)\rho_s + (2\Gamma + c)(1 + 2\Gamma c/\mu)\delta_s + \\ + (2\Gamma + c + 2\Gamma c/\mu)\|y^s - y\|. \end{aligned} \tag{55}$$

We have thus proved that for all sufficiently small $\varepsilon \leq \bar{\varepsilon}$ and sufficiently large s there are indices l_s such that $\|y_s^k - y\| \leq \varepsilon$ for $k \in [k_s, l_s)$ and $f(y_s^{l_s})$ satisfies condition (55). This proves the lemma. Q.E.D.

We need another lemma concerning the length of the steps $\|x^{k+1} - x^k\|$ in method (32)-(34).

LEMMA 5.5. For method (32)-(34) almost surely

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

Proof. Using (32)-(34), we obtain

$$\|x^{k+1}(\omega) - x^k(\omega)\| \leq \|s^k(\omega)\| \leq \rho_k \|\frac{1}{n_k} \sum_{r=r_k}^{k} g^r(\omega)\| +$$

$$+ \rho_k \|\frac{1}{n_k} \sum_{r=r_k}^{k} (\xi^r(\omega) - g^r(\omega))\| \leq \rho_k \|\frac{1}{n_k} \sum_{r=r_k}^{k} g^r(\omega)\| + \rho_k \|\Delta^k(\omega)\|,$$

where

$$g^r(\omega) = \mathrm{E}\{\xi^r(\omega) \mid x^0(\omega), \ldots, x^k(\omega)\},$$

$$\Delta^k(\omega) = \frac{1}{n_k} \sum_{r=r_k}^{k} (\xi^r(\omega) - g^r(\omega)), \quad n_k = k - r_k + 1.$$

Here $g^r \in \bar{\partial} F(x^r(\omega))$ are uniformly bounded and thus

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

By Lemma 5.1, the sequence $\{\zeta_0^k(\omega)\}$ (see Eq. (46)) almost surely has a limit. Thus for $\{\rho_k \Delta^k(\omega)\}$ we have $\lim_{k \to \infty} \Delta^k(\omega) = 0$ almost surely, which completes the proof of the lemma. Q.E.D.

Now we can complete the proof of Theorem 5.1. Consider the set $\Omega' \subseteq \Omega$ such that the series $\{\zeta_0^k(\omega)\}_{k=0}^{\infty}$ (46) converges. By Lemma 5.1, $\mathbf{P}(\Omega') = 1$. Take some $\omega \in \Omega'$. The rest of the proof repeats steps 1^0-5^0 from the proof of Theorem 4.1 and uses Lemma 5.2 instead of Lemma 4.1.

REFERENCES

1. Yu. M. Ermoliev and V. I. Norkin, "On nonsmooth and discontinuous problems of stochastic systems optimization," Eur. J. Oper. Res., **101**, 230-244 (1997).
2. Yu. M. Ermoliev, V. I. Norkin, and R. J.-B. Wets, "The minimization of semicontinuous functions: mollifier subgradients," SIAM J. Contr. Optim., **33**, No. 1, 149-167 (1995).
3. P. Glynn, Optimization of Stochastic Systems via Simulation, Technical Report No. 43, Stanford University, Palo Alto, CA (1989).
4. Y. G. Ho and X. R. Cao, Discrete Event Dynamic Systems and Perturbation Analysis, Kluwer, Boston (1991).
5. R. Suri, "Perturbation analysis: the state of the art and research issues explained via the GI/G/1 queue," Proc. IEEE, **77**, No. 1,114-137 (1989).

6. A. A. Gaivoronski, "Optimization of stochastic discrete event dynamic systems: a survey of some recent results," in: Simulation and Optimization, Lect. Notes Econ. Math. Sys., Vol. 374, G. Pflug and U. Dieter (eds.), Springer, Berlin, (1992), pp. 24-44.
7. R. Y. Rubinstein and A. Shapiro, The Optimization of Discrete Event Dynamic Systems by the Score Function Method, Wiley, New York (1993).
8. A. M. Gupal, Stochastic Methods for Solving Nonsmooth Extremal Problems [in Russian], Naukova Dumka, Kiev (1979).
9. Yu. Ermoliev and A. Gaivoronski, "Stochastic programming techniques for optimization of discrete event systems," Ann. Oper. Res., **39**, 120-135 (1992).
10. V. I. Norkin, "Nonlocal optimization algorithms for nonsmooth functions," Kibernetika, No. 5, 75-79 (1978).
11. P. A. Dorofeev, "Some properties of the generalized gradient method," Zh. Vychisl. Matem. Mat. Fiz., **25**, No. 2, 181-189 (1985).
12. P. A. Dorofeev, "General scheme of iterative minimization methods," Zh. Vychisl. Matem. Mat. Fiz., **26**, No. 4, 536-544 (1986).
13. N. K. Krivulin, Optimization of Discrete Event Dynamic Systems by Simulation [in Russian], Abstract of thesis, Leningrad Univ. (1990).
14. N. K. Krivulin, "Optimization of complex systems by simulation," Vestnik Leningrad. Univ., No. 8, 100-102 (1990).
15. F. Mirzoakhmedov, "Optimization of a queueing system and a numerical solution method," Kibernetika, No. 3, 73-75 (1990).
16. V. S. Mikhalevich, A. M. Gupal, and V. I. Norkin, Nonconvex Optimization Methods [in Russian], Nauka, Moscow (1987).
17. F. Clarke, Optimization and Nonsmooth Analysis [Russian translation], Nauka, Moscow (1988).
18. Yu. E. Nesterov, Effective Methods in Nonlinear Programming [in Russian], Radio i Svyaz', Moscow (1989).
19. N. Z. Shor, Methods for Minimization of Nondifferentiable Functions and Their Applications [in Russian], Naukova Dumka, Kiev (1979).
20. Yu. M. Ermol'ev, Stochastic Programming Methods [in Russian], Nauka, Moscow (1976).
21. B. T. Polyak, An Introduction to Optimization [in Russian], Nauka, Moscow (1983).
22. E. A. Nurminskii, Numerical Methods for Solving Deterministic and Stochastic Minmax Problems [in Russian], Naukova Dumka, Kiev (1979).
23. A. M. Gupal and L. G. Bazhenov, "Stochastic analogue of the conjugate gradient method," Kibernetika, No. 1, 125-126 (1972).
24. A. M. Gupal and L. G. Bazhenov, "Stochastic linearization method," Kibernetika, No. 3, 116-117 (1972).
25. A. Ruszczynski, "A method of feasible directions for solving nonsmooth stochastic programming problems," in: Lect. Notes Contr. Inform. Sci., F. Archetti, G. Di Pillo, and M. Lucertini (eds.), Springer, Berlin (1986), pp. 258-271.