UDK 517.2+519.977.58+519.8

V.I. NORKIN[1]

GENERALIZED GRADIENTS IN PROBLEMS OF DYNAMIC OPTIMIZATION, OPTIMAL CONTROL, AND MACHINE LEARNING[2]

Abstract. In this work, nonconvex nonsmooth problems of dynamic optimization, optimal control in discrete time (including feedback control), and machine learning are considered from a common point of view. An analogy is observed between tasks of controlling discrete dynamic systems and training multilayer neural networks with nonsmooth target function and connections. Methods for calculating generalized gradients for such systems based on Hamilton-Pontryagin functions are substantiated. Stochastic generalized gradient algorithms are extended for optimal controlling and learning nonconvex nonsmooth dynamic systems.

Keywords: dynamic optimization, optimal control, machine learning, multilayer neural networks, deep learning, nonconvex nonsmooth optimization, stochastic optimization, stochastic generalized gradient.

INTRODUCTION

Nonlinear optimal control problems with discrete time can be considered as large-scale optimization problems, for the solution of which gradient-type methods can be applied. For this, it is necessary to have formulas and rules for calculating gradients of the target functional over controls. Such formulas for the problem with a free right end were obtained, for example, in [1–5] using the Lagrange multiplier method under the assumption of continuous differentiability of the functions involved in the problem (references to other and earlier papers are given in [6, Section 5.5]. In [2, 3] these results are extended to stochastic discrete optimal control problems. For nonsmooth convex optimal control problems with linear equations of motion, similar formulas for calculating subgradients of the objective function were obtained in [2–4]. However, in the presence of nonlinear motion equations, the dependence of the target function on controls can be nonconvex. In this regard, in [3] these formulas are also substantiated for the case of a weakly convex [7] objective functional and smooth equations of motion. All these formulas use the procedures of direct calculation of the trajectory of motion and the backward calculation of auxiliary conjugate variables, essentially adopted from the Pontryagin maximum principle [8, 9]. Similar procedures for calculating gradients in the space of weights are widely used in training multilayer neural networks [6, 10 - 12]. However, the problem is that the learning quality functions are not only nonconvex but often turn out to be nonsmooth of the model parameters. Nonsmoothness, in particular, can be caused by the use of nonsmooth activation functions of neurons. In this paper, we generalize these results (regarding the calculation of generalized gradients) to nonconvex nonsmooth problems of discrete optimal control and learning with the so-called generalized

[1] V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine, Kiev & Faculty of Applied Mathematics of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute” Email: vladimir.norkin@gmail.com

[2] The work is partially supported by grant CPEA-LT-2016/10003 funded by the Norwegian Agency for International Cooperation and Quality Enhancement in Higher Education (Diku).

differentiable functions [13]. Moreover, nonsmooth functions can enter both in the objective function and the equations of motion.

Nonsmooth optimal control problems with a nondifferentiable objective functional arise, for example, when the norm of deviation from the target is used as the objective function, when nonsmooth penalty functions are used to eliminate differential and phase constraints [14], and also in the presence of maximum and minimum operations in the formulation of the problem. Nonsmooth motion equations arise, for example, if the area of motion is bounded, if threshold restrictions on the path of motion are applied, such as conditions of non-negativity of the stock and boundedness of the warehouse in the inventory theory [15], such as non-ruin conditions in the theory of risk processes [16]. Nonsmooth optimal control problems, in particular, general necessary conditions of extremum for such problems, were studied in [4, 14, 17–21] and in others. Reviews of iterative gradient and related methods for finding optimal controls are given in [1–5, 22].

Note that a multilayer neural network can be formally interpreted as a dynamic system that gradually (layer by layer) converts the input signal into the output one. The task of machine learning consists in the identification of parameters of a model, for example, the weights of a multilayer neural network, by using a set of input-output examples. Nonsmooth tasks of machine learning arise when using nonsmooth indicators of learning quality (such as a module), when applying non-smooth regularization functions, and when using non-smooth (for example, piecewise linear) activation functions in multilayer neural networks. To solve smooth problems of training neural networks, the BackProp method [6, 10 - 12, 23, 24] is widely used, i.e. a special method for calculating gradients of the quality of learning functional over various and numerous parameters. The history of the discovery, development, and applications of the BackProp method was traced in [6]. However, in applied deep neural networks, along with smooth sigmoidal neuron activation functions, nonsmooth linear rectification functions are also widely used (for example, $f(x) = \max\{0, x\}$) [12, section 6.3.1; 24, section 3.3]. Such functions generate essentially non-convex nonsmooth functionals of the quality of training.

There is another area of computational mathematics that develops and studies methods for calculating derivatives and gradients of complex composite objective functions: this is the theory of automatic differentiation, presented, for example, in [5, 25]. In this approach, the algorithmic process of calculating the objective function is presented in the form of a network diagram, at the initial vertices of which there are elementary functions of optimization variables with known gradients, and at intermediate vertices, there are composition operators with known rules for calculating gradients of the composition. With a direct pass of the graph, the value of the complex objective function is calculated, and with the opposite pass, it is possible to calculate the gradients

of the objective function over optimization variables. Concerning neural networks, the automatic differentiation method reproduces the BackProp method and is implemented in numerous software libraries [24, section 2.6; 25]. However, here there is a problem of justifying the method for nonconvex nonsmooth functions, which is discussed only at the informal level [25].

Training methods for smooth neural networks are discussed in [10–12, 23, 24, 26–29]. Basically, these are the method of stochastic gradients and its modifications, adopted from the theory of stochastic approximation [30] and the theory of convex stochastic programming [2, 31, 32], since only these methods are applicable for training deep neural networks. This method can also be substantiated for optimizing nonconvex nonsmooth functions [33, 34], thereby filling the gap in the theory of training nonsmooth neural networks. Thus, the development and justification of effective methods for calculating generalized gradients for nonsmooth optimal control and learning problems, firstly, expands the range of numerically solvable problems, and secondly, it opens up a wide field of application of numerical methods of nonconvex nonsmooth optimization developed in [2, 35 - 48]. In particular, for the so-called generalized differentiable functions, for which the rules for calculating generalized gradients are justified in this article, various methods of local optimization of the gradient type are considered in [33, 42, 48–50]. Note that generalized differentiable functions include convex, concave, weakly convex and weakly concave [7], semismooth [36], [39], and piecewise smooth [51] functions and are closed with respect to the finite operations of maximum, minimum, superposition, and mathematical expectation [13, 48, 50, 52].

Sections 1, 2 summarize the theory of generalized differentiable functions and methods for their optimization. In Section 3, we obtained the main result (Theorem 6), a method for calculating generalized gradients of objective functionals in nonconvex nonsmooth dynamic optimization problems, and in Sections 4 - 6 it is applied to computing generalized gradients in problems of optimal control, stochastic optimal control, and machine learning. Section 7 concludes the article.

1. GENERALIZED DIFFERENTIABLE FUNCTIONS

Definition 1 [13, 50]. A function $f:\mathrm{R}^n \to \mathrm{R}^1$ is called generalized differentiable at point $x \in \mathrm{R}^n$, if in some ε-neighborhood $\left\{y \in \mathrm{R}^n : \|y-x\| < \varepsilon\right\}$ of the point x it is defined an upper semicontinuous at x multivalued mapping $\partial f(\cdot)$ with convex compact values $\partial f(y)$ and such that the following expansion holds true:

$$f(y) = f(x) + \langle g, y-x\rangle + o(x,y,g), \qquad (1)$$

where $d \in \partial f(y)$, $\langle \cdot, \cdot \rangle$ denotes a scalar product of two vectors, and the remainder term $o(x, y, g)$ satisfies the condition: $\lim_{k \to \infty} o(x, y^k, g^k) / \|y^k - x\| = 0$ for all sequences $g^k \in \partial f(y^k)$, $y^k \to x$ as $k \to \infty$. A function f is called generalized differentiable if it is generalized differentiable at each point $x \in \mathrm{R}^n$; the mapping $\partial f(\cdot)$ is called the generalized gradient mapping of the function f; the set $\partial f(x)$ is called a generalized gradient set of the function $f(\cdot)$ at point x; vectors $g \in \partial f(x)$ are called generalized gradients of the function $f(\cdot)$ at point x.

Any generalized differentiable function is Lipschitzian and its Clark subdifferential $\partial_C f(x)$ [19] is minimal (with respect to inclusion) a generalized gradient mapping for $f(x)$. For almost everyone $x \in \mathrm{R}^n$ it holds $\partial f(x) = \partial_C f(x)$ [42]. The class of generalized differentiable functions contains continuously differentiable, convex, concave, semi-smooth [36] and some other piecewise smooth functions [51] and is closed with respect to the operations of maximum, minimum, superposition and mathematical expectation (see [13, 42, 48, 50, 52]). In [53], the concept of generalized-differentiable functions was extended to vector-valued functions. A related approach to the differentiation of functions was proposed in the book [54, p. 175, 444], where a function is called differentiable at a point x if the following representation $f(y) = f(x) + \langle \varphi(y), y - x \rangle$ with some continuous at x vector-function $\varphi(y)$ holds true. In contrast to the latter definition, expansion (1) contains a uniformly small additional term $o(x, y, g)$, and the gradient is also multi-valued (as in related works [51, 55, 56]). These modifications significantly expand the class of functions under consideration.

Theorem 1 (generalized differentiability of composite functions [50]). Let $f_0(z)$, $z \in \mathrm{R}^m$, and $f_i(x)$, $x \in \mathrm{R}^n$, $i = 1, ..., m$, are generalized differentiable functions with generalized gradient mappings $\partial_z f_0(\cdot)$ и $\partial f_i(\cdot)$. Them the maximum function $f(x) = \max\{f_1(x), ..., f_m(x)\}$ and the minimum function $f(x) = \min\{f_1(x), ..., f_m(x)\}$ are generalized differentiable functions with the generalized gradient mapping $\partial f(x) = \mathrm{conv.hull}\{\partial f_i(x) : f_i(x) = f(x)\}$, and a composite function $f(x) = f_0(z(x)) = f_0(f_1(x), ..., f_m(x))$ is also generalized differentiable, and its generalized gradient mapping is calculated by the chain rule:

$$\partial f(x) = \mathrm{conv.hull}\{g = [g_1 ... g_m] g_0 : g_0 \in \partial_z f_0(z(x)),\ g_i \in \partial f_i(x),\ i = 1, ..., m\}$$

where $[g_1 ... g_m]$ is a matrix composed of column-vectors $g_1, ..., g_m$, conv.hull denotes the convex hull, $\partial_z f_0(z(x))$ designates generalized gradient set of function $f_0(z)$ at point $z(x) = (f_1(x), ..., f_m(x))$.

Thus, to calculate the generalized gradients of the sum, product, quotient, and other complex functions, the usual rules for differentiating complex smooth functions are applicable.

Theorem 2 (generalized differentiability of the mathematical expectation, [52]). Suppose that

function $f(x, \theta)$ is generalized differentiable in $x \in V$ and integrable in $\theta \in \Theta$, where V is an open subset in R^n, and θ is an elementary event of a probability space $(\Theta, \Sigma, \mathrm{P})$;

the generalized gradient mapping $\partial_x f(x, \theta)$ of the function $f(\cdot, \theta)$ is measurable in θ for any $x \in \mathrm{R}^n$;

for the open set $V \subset \mathrm{R}^n$ there exists an integrable function $L_V(\theta)$ such that

$$\sup\{\|g\| : g \in \partial_x f(x, \theta),\ x \in V\} \le L_V(\theta).$$

Then the mathematical expectation function $F(x) = \mathrm{E} f(x, \theta) = \int_S f(x, \theta) \mathrm{P}(d\theta)$ is generalized differentiable in V with the generalized gradient mapping $\partial F(x) = \mathrm{E} \partial_x f(x, \theta)$.

Thus, when differentiating the mathematical expectation (Lebesgue integral) under the conditions of Theorem 2, we can introduce the differentiation operator under the sign of the integral. Here, the integral $\mathrm{E} \partial_x f(x, \theta)$ of the multi-valued mapping $\theta \to \partial_x f(x, \theta)$ is understood as a collection of integrals of integrable selectors of this mapping (under fixed x).

Example 1 (stock evolution). Stock evolution in a warehouse is described by the following recurrent dynamic relationship:

$$x_{t+1} = \max\{0, x_t + u_t - \theta_t\}, \quad t = 0, 1, ..., m, \quad x_0 = a.$$

Here x_t is the stock level at the beginning of the time period $[t, t+1]$, u_t is the replenishment request at the beginning of the time period $[t, t+1]$, θ_t is random demand for the goods in the time period $[t, t+1)$, M is the maximum warehouse volume, a is the initial stock in the warehouse. The quality of the functioning of the warehouse under the control program $u = \{u_0, ..., u_m\}$ can be described by the criterion

$$F(a, u) = \mathrm{E}\left(\sum_{t=0}^{m} \max\{\alpha(x_t - \theta_t), \beta(\theta_t - x_t)\} + \max\{\alpha(x_{m+1} - b), \beta(b - x_{m+1})\} \right) \to \min_u,$$

where α, β are the penalty factors for excess and shortage of goods in the period of time $[t, t+1)$; b is the desired level of goods in the stock at the end of the planning period; E is the sign of mathematical expectation.

Here is a complex random function

$$\varphi(a,u,\theta_0,...,\theta_k) = \sum_{t=0}^{m} \max\{\alpha(x_t - \theta_t), \beta(\theta_t - x_t)\} + \max\{\alpha(x_{m+1} - b), \beta(b - x_{m+1})\}$$

is generalized differentiable with respect to its arguments (a,u), and under the conditions of Theorem 2, the expectation function $F(a,u)$ is also generalized differentiable with respect to its variables.

2. GENERALIZED GRADIENT METHODS FOR OPTIMIZATION OF NON-CONVEX NONSMOOTH FUNCTIONS

It is not enough to be able to calculate the generalized gradients of nonsmooth functions; it is still necessary to justify the possibility of their application for the analysis and optimization of these functions.

Generalized gradient and stochastic generalized gradient methods for minimizing generalized differentiable functions were substantiated in [13, 16, 33, 42, 48–50, 52]. They are generalizations to the nonconvex nonsmooth case of the generalized gradient method [2, 35, 40, 57] and the stochastic quasigradient method [2], designed to solve convex optimization problems.

Let consider a problem

$$F(x) \to \min_{x \in X} \tag{2}$$

Of the minimization of a generalized differentiable function $F(x)$ over a compact set $X \subset \mathrm{R}^n$, given by some generalized differentiable function $G(x)$, i.e.

$$X = \{x \in \mathrm{R}^n : G(x) \le 0\}. \tag{3}$$

The necessary optimality condition for problem (2) has the form: $0 \in \partial F(x) + N_X(x)$, where the normal cone $N_X(x)$ to the set X at point $x \in X$ is given by the relations: $N_X(x) = \{\lambda g : g \in \partial G(x), \lambda \ge 0\}$ if $G(x) = 0$ and $N_X(x) = 0$ if $G(x) < 0$. In case of a convex set X the cone $N_X(x)$ coincides with the cone of normals to X at point x.

The basic iterative generalized gradient method for solving problem (2) has the form:

$$x^{k+1} \in \Pi_X\left(x^k - \rho_k d^k\right), \qquad d^k \in \partial F(x^k), \qquad x^0 \in X, \qquad k = 0,1,..., \tag{4}$$

where $\Pi_X(\cdot)$ is the (multivalued) projection operator onto a nonconvex feasible set X; non-negative quantities ρ_k, δ_k satisfy the conditions:

$$\lim_{k\to\infty}\delta_k = \lim_{k\to\infty}\rho_k = 0, \qquad \sum_{k=0}^{\infty}\rho_k = +\infty. \tag{5}$$

In another form, method (4) has the form:

$$\begin{aligned} x^{k+1} \in \Pi_X\left(x^k - \rho_k d^k\right) &= \arg\min_{x\in X}\left\|x-(x^k-\rho_k d^k)\right\|^2 = \\ &= \arg\min_{x\in X}\left(\left\|x-x^k\right\|^2 + 2\rho_k\left\langle d^k, x-x^k\right\rangle\right) = \\ &= \arg\min_{x\in X}\left(\left\langle d^k, x-x^k\right\rangle + \frac{1}{2\rho_k}\left\|x-x^k\right\|^2\right). \end{aligned} \tag{6}$$

In the absence of constraints, method (4)-(6) turns into an ordinary generalized gradient descent:

$$x^{k+1} = x^k - \rho_k d^k, \qquad d^k \in \partial F(x^k), \qquad x^0 \in X, \qquad k = 0,1,\ldots.$$

Theorem 3 (convergence of the nonconvex generalized gradient method [33]). Under conditions (5), the minimum (in the function F) limit points of the sequence $\{x^k\}$ belong to the set $X^* = \{x \in X : 0 \in \partial F(x) + N_X(x)\}$ of points satisfying the necessary optimality conditions and all limit points of the numerical sequence $\{F(x^k)\}$ constitute an interval in the set $F^* = F(X^*)$. If the set F^* does not contain intervals (for example, if it is finite or countable), then all the limit points of the sequence $\{x^k\}$ belong to the connected subset of X^*, and there is a limit $\lim_{k\to\infty} F(x^k) \in F^*$.

Note that the set $F^* = \{F(x) : 0 \in \partial F(x)\}$ does not contain intervals for sufficiently smooth functions by virtue of Sard's theorem [58, Section 2.3].

Similar results on the convergence of the generalized gradient method for nonconvex nonsmooth functions with the use of Clark subgradients were obtained in [34].

The randomized generalized gradient method is defined by the following relationships:

$$x^{k+1} \in \Pi_X\left(x^k - \rho_k \tilde{d}^k\right), \quad \tilde{d}^k \in \partial F(\tilde{x}^k), \qquad \left\|\tilde{x}^k - x^k\right\| \le \delta_k, \qquad x^0 \in X, \qquad k = 0,1,\ldots. \tag{7}$$

where $\tilde{x}^k$ is a randomly (e.g., uniformly) taken point in the δ_k-neighborhood of the point x_k.

Denote $X_C^* = \{x \in X : 0 \in \partial_C F(x) + N_X(x)\} \subseteq X^*$, where $\partial_C F(x)$ is the Clarke subdifferential of function F at point x. Remark that $\partial F(x) = \partial_C F(x)$ for almost all $x \in \mathrm{R}^n$ [42], so in the randomized method (7) almost sure the Clarke subgradients of the generalized differentiable

function $F(\cdot)$ are used. A similar idea of the randomization of the generalized gradient decent method is exploited in [59].

Theorem 4 (convergence of the randomized generalized gradient method [16, 33, 48, 49]). Let conditions (5) and $\lim_{k\to\infty}\delta_k = 0$ are satisfied. Then for almost all trajectories $\{x^k\}$ of process (7) the minimal (in function *F*) limit points of the sequence $\{x^k\}$ belong to the set X_C^* of points that satisfy Clarke's necessary optimality conditions, and almost sure all limit point of the number sequence $\{F(x^k)\}$ constitute an interval in the set $F_C^* = F(X_C^*)$. If the set F_C^* does not contain intervals (for example, it is finite or countable), then all limit point of the sequence $\{x^k\}$ belong to a connected subset of the set X_C^*, and there exist a limit $\lim_{k\to\infty} F(x^k) \in F_C^*$.

Thus, the randomized method converges, generally speaking, to a narrower set of critical points X_C^* that X^*, since $\partial_C F(\cdot) \subseteq \partial F(\cdot)$.

The randomized subgradient method also admits the following interpretation. We introduce the so-called smoothed functions

$$F_k(x) = \frac{1}{V_{\delta_k}} \int\limits_{\{\tilde{x}:\|\tilde{x}-x\|\le\delta_k\}} F(\tilde{x})d\tilde{x}, \tag{8}$$

where V_{δ_k} is the volume of the δ_k-neighborhood of the zero point. If we introduce a random vector $\tilde{x}^k$ uniformly distributed in a δ_k-neighborhood of a point x, then the smoothed function $F_k(x)$ and its gradient $\nabla F_k(x)$ can be represented respectively in the form $F_k(x) = \mathrm{E}F(\tilde{x}^k)$ and $\nabla F_k(x) = \mathrm{E}\partial F(\tilde{x}^k)$, where E denotes the mathematical expectation with respect to $\tilde{x}^k$, $\mathrm{E}\partial F(\tilde{x}^k)$ denotes the mathematical expectation of a random multi-valued mapping $\tilde{x}^k \to \partial F(\tilde{x}^k)$.

Thus, the randomization in method (7) plays a threefold role: on the one hand, it allows us to narrow the convergence set of the generalized gradient method to the set $X_C^* \subseteq X^*$, and on the other hand, it gives to method (7) some global properties due to the fact that it minimizes the sequence of smoothed functions $F_k(x)$. Besides, the randomization ensures that the method does not stick at critical points that are not local minima. In case $\delta_k = \delta$ the randomized generalized gradient method (7) becomes a stochastic gradient method for minimizing the same non-changing smoothed function $F_k(x)$. To strengthen global properties of method (7), it is possible to use the estimate of the gradient of the smoothed function (8) by means of several independent realizations of a random point $\tilde{x}$ from the δ_k-vicinity of the current point x^k.

We now consider an analog of methods (4), (7) to minimize the generalized differentiable expectation function:

$$F(x) = \mathrm{E}_\theta f(x,\theta) \to \min_{x \in X} \tag{9}$$

over a nonconvex set X (defined by (3)):

$$x^{k+1} \in \Pi_X \left(x^k - \rho_k d(\tilde{x}^k, \theta^k) \right), \quad d(\tilde{x}^k, \theta^k) \in \partial_x f(\tilde{x}^k, \theta^k), \tag{10}$$

$$\left\| \tilde{x}^k - x^k \right\| \le \delta_k, \qquad k = 0,1,...,$$

where $\Pi_X(\cdot)$ is the (multivalued) projection operator onto a nonconvex feasible set X; $g(x,\theta)$ is a (x,θ)-measurable selector of the mapping $\partial_x f(x,\theta)$ of a generalized differentiable random function $f(\cdot,\theta)$; $\{\theta^k\}$ are independent identically distributed observations of a random variable θ; points $\tilde{x}^k$ are randomly uniformly selected from the sets $\{x : \left\| x - x^k \right\| \le \delta_k\}$; non-negative quantities ρ_k, δ_k are measurable with respect to σ-algebra $\sigma\left\{x^0,...,x^k\right\}$, and with probability one satisfy the conditions:

$$\lim_{k\to\infty} \delta_k = \lim_{k\to\infty} \rho_k = 0, \qquad \sum_{k=0}^{\infty} \rho_k = +\infty, \qquad \sum_{k=0}^{\infty} \rho_k^2 < +\infty. \tag{11}$$

Theorem 5 (convergence with probability 1 of the non-convex method of stochastic generalized gradients [49, 33, 49]). For almost all trajectories $\{x^k\}$ of the method (10), (11) the assertions of Theorem 4 hold. If in algorithm (10), (11) all $\delta_k \equiv 0$, then the statement of Theorem 5 holds for the $X^* = \left\{ x \in X : 0 \in \partial F(x) + N_X(x) \right\}$.

Book [42] considers several other stochastic nonconvex nonsmooth optimization methods (methods with averaging the trajectory, averaging the generalized gradients, methods of the ravine step, reduced gradient, heavy ball, and others).

3. CALCULATION OF GENERALIZED GRADIENTS IN DYNAMIC OPTIMIZATION PROBLEMS

Consider the following optimization problem for a dynamic system with a free final state in discrete time:

$$J(u) = \sum\nolimits_{i=0}^{m} F_i(x_i, u) + \Phi(x_{m+1}) \to \min_{u \in U} \tag{12}$$

Subject to equations

$$x_{i+1} = G_i(x_i, u), \qquad i = 0,1,...,m, \quad x_0 \in \mathrm{R}^{n_0}, \tag{13}$$

where index $i = 0,1,...,m+1$ designates discrete time; $x_i = \left(x_i^1,...,x_i^{n_i}\right)^T \in \mathrm{R}^{n_i}$ is the state of the controlled system at moment i; $u = \left(u^1,...,u^l\right)^T \in \mathrm{R}^l$ is the vector of the optimized parameters of the dynamic system; functions $G_i = \left(G_i^1,...,G_i^{n_{i+1}}\right)^T$, F_i, Φ are given; U is a given set in R^l; $m \geq 1$ is a natural number; x_0 is an initial point in R^{n_0}. Note that problems with restrictions on the final state or the trajectory can be reduced to form (12) using nonsmooth penalties. Note that in (12), (13) a change of the dimension n_i of the phase space with the passage of discrete time $i = 0,1,...,m$ is allowed.

Tasks of type (12), (13) arise, for example, in optimization of feedback control systems. In this case a control of a certain functional form $u_t = g(x_t, y)$ is substituted in the standard functional equation $x_{i+1} = G_i(x_i, u_t)$. The feedback control $u_t = g(x_t, y)$ depends on the current state x_t of the system and on the finite-dimensional vector y of the desired parameters to be optimized. Similar settings also arise in the problems of training recurrent neural networks (with the same weights for all layers of the network) [11; 12, ch. 10; 24, ch. 6].

Let us introduce notation [4]:

$$G_{ix} = \begin{pmatrix} G_{ix^1}^1 & ... & G_{ix^n}^1 \\ ... & ... & ... \\ G_{ix^1}^n & ... & G_{ix^n}^n \end{pmatrix} = \begin{pmatrix} G_{ix}^1 \\ ... \\ G_{ix}^n \end{pmatrix}, \qquad G_{iu} = \begin{pmatrix} G_{iu^1}^1 & ... & G_{iu^l}^1 \\ ... & ... & ... \\ G_{iu^1}^n & ... & G_{iu^l}^n \end{pmatrix} = \begin{pmatrix} G_{iu}^1 \\ ... \\ G_{iu}^n \end{pmatrix},$$

$$\left(F_{ix}\right)^T = \left(F_{ix^1},...,F_{ix^n}\right), \qquad \left(F_{iu}\right)^T = \left(F_{iu^1},...,F_{iu^l}\right), \qquad \left(\Phi_x\right)^T = \left(\Phi_{x^1},...,\Phi_{x^n}\right),$$

where $\left(F_{ix}, F_{iu}\right)^T$, $\left(G_{ix}^j, G_{iu}^j\right)^T$ are some generalized gradients over (x,u) of functions F_i, G_i^j; Φ_x is some generalized gradient of function Φ; $(\cdot)^T$ designates the transposition of the matrix $(\cdot)$.

Theorem 6. Let the functions F_i, G_i, Φ in problem (12), (13) be generalized differentiable over totality of their arguments $x_i \in \mathrm{R}^{n_i}$, $u \in V$, where V is an open neighborhood of the set U, $i = 0,1,...,m$. Then the function $J(x_0,u)$ is generalized differentiable over $\left(x_0,u\right) \in \mathrm{R}^{n_0} \times V$, and vectors $\left(J_{x_0}, J_u\right)^T = \left(J_{x_0^1},...,J_{x_0^{n_0}}, J_{u^1},...,J_{u^l}\right)^T$ with components

$$J_{x_0^j} = H_{0x_0^j}(x_0, \psi_0, u), \qquad J_{u^j} = \sum\nolimits_{i=0}^{m} H_{iu^j}(x_i, \psi_i, u), \tag{14}$$

are generalized gradients of the function $J(x_0,u)$ at point (x_0,u), where $H_i(x_i,\psi_i,u_i)=F_i(x_i,u_i)+G_i^T(x_i,u_i)\cdot\psi_i$, $i=0,1,...,m$, is a discrete (over i) Hamilton-Pontryagin function; $x=(x_0,...,x_{m+1})$ is a discrete trajectory of process (13), which corresponds to chosen parameter $u\in U$; a sequence of auxiliary (conjugate) vector-functions $(\psi_m,...,\psi_0)=\psi$ is defined by the back propagation equations:

$$\begin{aligned}&\psi_{i-1}=H_{ix_i}(x_i,\psi_i,u)=F_{ix_i}(x_i,u)+G_{ix_i}^T(x_i,u)\cdot\psi_i=F_{ix_i}(x_i,u)+\sum_{j=1}^{n_i}G_{ix_i}^j(x_i,u)\psi_i^j,\\&\psi_m=\Phi_{x_{m+1}}(x_{m+1})=\left(\Phi_{x_{m+1}^1}(x_{m+1}),...,\Phi_{x_{m+1}^{n_{m+1}}}(x_{m+1})\right)^T,\qquad i=m,m-1,...,1,0.\end{aligned}\tag{15}$$

Proof. We have

$$x_1=x_1(x_0,u)=G_0(x_0,u),$$

$$x_2=x_2(x_0,u)=G_1(x_1,u)=G_1\left(G_0(x_0,u),u\right),$$

…

$$x_{m+1}=G_m(x_m,u)=G_m(G_{m-1}(...(G_0(x_0,u),u),...),u);$$

$$\begin{aligned}J(x_0,u)=J(x_0,u)&=F_0(x_0,u)+F_1(x_1,u)+F_2(x_2,u)+...+F_m(x_m,u)+\Phi(x_{m+1})=\\&=F_0(x_0,u)+F_1(G_0(x_0,u),u)+F_2(G_1(G_0(x_0,u),u),u)+...\\&\quad+F_m(G_{m-1}(G_{m-2}(...(G_0(x_0,u),u),...),u)+\\&\quad+\Phi(G_m(G_{m-1}(...(G_0(x_0,u),u),...),u).\end{aligned}$$

Let $\left(F_{ix_i},F_{iu}\right)^T$ and $\left(G_{ix_i}^j,G_{iu}^j\right)^T$, $j=1,...,n_i$, be some generalized gradients of the functions F_i and G_i^j, $j=1,...,n_i$, at point (x_i,u), $i=1,...,m$; $\Phi_{x_{m+1}}(x_{m+1})$ be some generalized gradient of the function Φ at point x_{m+1}; $\{x_0,x_1,...,x_{m+1}\}$ be the trajectory of process (13) under given (x_0,u).

By virtue of Theorem 1, complex (vector) functions $x_1,x_2,...,x_m$ and the objective function J are (component-wise) generalized differentiable with respect to their (vector) arguments (x_0,u), and by virtue of the chain rule (from Theorem 1) of generalized differentiation of composite functions, the vector $\left(J_{x_0},J_u\right)^T=\left(J_{x_0^1},...,J_{x_0^{n_0}},J_{u^1},...,J_{u^l}\right)^T$ is some generalized gradient of the function J at the point (x_0,u), where the components $J_{x_0^i}$, $i=1,..,n_0$, are calculated as follows:

$$J_{x_0^i} = F_{0x_0^i}(x_0,u) + \sum_{j_1=1}^{n_1} F_{1x_1^{j_1}}(x_1,u)\cdot G_{0x_0^i}^{j_1}(x_0,u) + \sum_{j_2=1}^{n_2} F_{x_2^{j_2}}(x_2,u)\sum_{j_1=1}^{n_1} G_{1x_1^{j_1}}^{j_2}(x_1,u)\cdot G_{0x_0^i}^{j_1}(x_0,u) + ...$$
$$... + \sum_{j_m=1}^{n_m} F_{mx_m^{j_m}}(x_m,u)\sum_{j_{m-1}=1}^{n_{m-1}} G_{(m-1)x_{m-1}^{j_{m-1}}}^{j_m}(x_{m-1},u)\sum_{j_{m-2}=1}^{n_{m-2}} ... \sum_{j_1=1}^{n_1} G_{1x_1^{j_1}}^{j_2}(x_1,u)\cdot G_{0x_0^i}^{j_1}(x_0,u) +$$
$$+ \sum_{j_{m+1}=1}^{n_{m+1}} \Phi_{x_{m+1}^{j_{m+1}}}(x_{m+1})\sum_{j_m=1}^{n_m} G_{mx_{m-1}^{j_m}}^{j_{m+1}}(x_m,u)\sum_{j_{m-1}=1}^{n_{m-1}} ... \sum_{j_1=1}^{n_1} G_{1x_1^{j_1}}^{j_2}(x_1,u)\cdot G_{0x_0^i}^{j_1}(x_0,u).$$

Changing the order of summation in the products and taking out the common factors out of brackets, we obtain:

$$J_{x_0^i} = F_{0x_0^i}(x_0,u) + \sum_{j_1=1}^{n_1} G_{0x_0^i}^{j_1}(x_0,u)\left(F_{1x_1^{j_1}}(x_1,u) + \sum_{j_2=1}^{n_2} G_{1x_1^{j_1}}^{j_2}(x_1,u)\left(F_{2x_2^{j_2}}(x_2,u) + ... \right.\right.$$
$$\left.\left. ... + \sum_{j_m=1}^{n_m} G_{(m-1)x_{m-1}^{j_{m-1}}}^{j_m}(x_{m-1},u)\left(F_{mx_m^{j_m}}(x_m,u) + \sum_{j_{m+1}=1}^{n_{m+1}} G_{mx_m^{j_m}}^{j_{m+1}}(x_m,u)\Phi_{x_{m+1}^{j_{m+1}}}(x_{m+1}) \right)\right) ... \right).$$

Sequentially using the vectors $\{\psi_m, \psi_{m-1}, ..., \psi_0\}$ from (15), we obtain the final result for $J_{x_0^i}$:

$$J_{x_0^i} = F_{0x_0^i}(x_0,u) + \sum_{j_1=1}^{n_1} G_{0x_0^i}^{j_1}(x_0,u)\left(F_{1x_1^{j_1}}(x_1,u) + \sum_{j_2=1}^{n_2} G_{1x_1^{j_1}}^{j_2}(x_1,u)\left(F_{2x_2^{j_2}}(x_2,u) + ... \right.\right.$$
$$\left.\left. ... + \sum_{j_m=1}^{n_m} G_{(m-1)x_{m-1}^{j_{m-1}}}^{j_m}(x_{m-1},u)\left(F_{mx_m^{j_m}}(x_m,u) + G_{mx_m^{j_m}}^{T}(x_m,u)\cdot\psi_m \right) ... \right) =$$
$$= F_{0x_0^i}(x_0,u) + \sum_{j_1=1}^{n_1} G_{0x_0^i}^{j_1}(x_0,u)\left(F_{1x_1^{j_1}}(x_1,u) + \sum_{j_2=1}^{n_2} G_{1x_1^{j_1}}^{j_2}(x_1,u)\left(F_{2x_2^{j_2}}(x_2,u) + ... \right.\right.$$
$$\left.\left. ... + \sum_{j_{m-1}=1}^{n_{m-1}} G_{(m-2)x_{m-2}^{j_{m-2}}}^{j_{m-1}}(x_{m-2},u)\left(F_{(m-1)x_{m-1}^{j_{m-1}}}(x_{m-1},u) + G_{(m-1)x_{m-1}^{j_{m-1}}}^{T}(x_{m-1},u)\cdot\psi_{m-1} \right)\right) ... \right) =$$
$$= ... = F_{0x_0^i}(x_0,u) + \sum_{j_1=1}^{n_1} G_{0x_0^i}^{j_1}(x_0,u)\left(F_{1x_1^{j_1}}(x_1,u) + G_{1x_1^{j_1}}^{T}(x_1,u)\cdot\psi_1 \right) =$$
$$= F_{0x_0^i}(x_0,u) + G_{0x_0^i}^{T}(x_0,u)\cdot\psi_0 = H_{0x_0^i}(x_0,\psi_0,u).$$

Components J_{u^i}, $i = 1,...,m$, are calculated by the chain rules (Theorem 1) of generalized differentiation of composite functions:

$$J_{x_0^i} = F_{0u^i}(x_0,u) + F_{1u^i}(x_1,u) + \sum_{j_1=1}^{n_1} F_{1x_1^{j_1}}(x_1,u)G_{0u^i}^{j_1}(x_0,u) +$$
$$+ F_{2u^i}(x_2,u) + \sum_{j_2=1}^{n_2} F_{2x_2^{j_2}}(x_2,u)G_{1u^i}^{j_2}(x_1,u) +$$
$$+ \sum_{j_2=1}^{n_2} F_{2x_2^{j_2}}(x_2,u)\sum_{j_1=1}^{n_1} G_{1x_1^{j_1}}^{j_2}(x_1,u)G_{0u_i}^{j_1}(x_0,u) + ...$$
$$... + F_{mu^i}(x_m,u) + \sum_{j_m=1}^{n_m} F_{mx_m^{j_m}}(x_m,u)G_{(m-1)u^i}^{j_m}(x_{m-1},u) +$$
$$+ \sum_{j_m=1}^{n_m} F_{mx_m^{j_m}}(x_m,u)\sum_{j_{m-1}=1}^{n_{m-1}} G_{(m-1)x_{m-1}^{j_{m-1}}}^{j_m}(x_{m-1},u)G_{(m-2)u^i}^{j_{m-1}}(x_{m-2},u) + ...$$
$$... + \sum_{j_m=1}^{n_m} F_{mx_m^{j_m}}(x_m,u)\sum_{j_{m-1}=1}^{n_{m-1}} G_{(m-1)x_{m-1}^{j_{m-1}}}^{j_m}(x_{m-1},u)\sum_{j_{m-2}=1}^{n_{m-2}} G_{(m-2)x_{m-2}^{j_{m-2}}}^{j_{m-1}}(x_{m-2},u)\times ...$$
$$... \times \sum_{j_1=1}^{n_1} G_{1x_1^{j_1}}^{j_2}(x_1,u)G_{0u^i}^{j_1}(x_0,u) +$$

$$+\sum_{j_{m+1}=1}^{n_{m+1}}\Phi_{x_{m+1}^{j_{m+1}}}(x_{m+1})G_{mu^i}^{j_{m+1}}(x_m,u)+$$
$$+\sum_{j_{m+1}=1}^{n_{m+1}}\Phi_{mx_{m+1}^{j_{m+1}}}(x_{m+1},u)\sum_{j_m=1}^{n_m}G_{mx_m^{j_m}}^{j_{m+1}}(x_m,u)G_{(m-1)u^i}^{j_m}(x_{m-1},u)+...$$
$$...+\sum_{j_{m+1}=1}^{n_{m+1}}\Phi_{mx_{m+1}^{j_{m+1}}}(x_{m+1},u)\sum_{j_m=1}^{n_m}G_{mx_m^{j_m}}^{j_m}(x_m,u)\sum_{j_{m-1}=1}^{n_{m-1}}G_{(m-1)x_{m-1}^{j_{m-1}}}^{j_m}(x_{m-1},u)\times...$$
$$...\times\sum_{j_1=1}^{n_1}G_{1x_1^{j_1}}^{j_2}(x_1,u)G_{0u^i}^{j_1}(x_0,u).$$

Rearranging the summation signs in all products of the form $\sum_{j_s=1}^{n_s}...\times\sum_{j_{s-1}=1}^{n_{s-1}}...\times...\times\sum_{j_1=1}^{n_1}...$ in the reverse order, collecting the sums with the first summation over j_1, then the sums with the first summation over j_2, etc., and taking out common factors out of brackets, we get:

$$J_{u^i}=F_{0u^i}(x_0,u)+\sum_{j_1=1}^{n_1}G_{0u^i}^{j_1}(x_0,u)\Big(F_{1x_1^{j_1}}(x_1,u)+\sum_{j_2=1}^{n_2}G_{1x_1^{j_1}}^{j_2}(x_1,u)\Big(F_{2x_2^{j_2}}(x_2,u)+...$$
$$...+\sum_{j_k=1}^{n_k}G_{(k-1)x_{k-1}^{j_k}}^{j_k}(x_{k-1},u)\Big(F_{kx_k^{j_k}}(x_k,u)+\sum_{j_{k+1}=1}^{n_{k+1}}G_{kx_k^{j_k}}^{j_{k+1}}(x_k,u)\Phi_{x_{k+1}^{j_{k+1}}}(x_{k+1})\Big)\Big)...\Big)+$$
$$+F_{1u^i}(x_1,u)+\sum_{j_2=1}^{n_2}G_{1u^i}^{j_2}(x_1,u)\Big(F_{2x_2^{j_2}}(x_2,u)+\sum_{j_3=1}^{n_3}G_{3x_3^{j_2}}^{j_3}(x_3,u)\Big(F_{3x_3^{j_3}}(x_3,u)+...\Big)+$$
$$...+F_{(m-1)u^i}(x_m,u)+\sum_{j_m=1}^{n_m}G_{(m-1)u^i}^{j_m}(x_{m-1},u)\Big(F_{mx_m^{j_m}}(x_m,u)+\sum_{j_{m+1}=1}^{n_{m+1}}G_{mx_m^{j_m}}^{j_{m+1}}(x_m,u)\Phi_{x_{m+1}^{j_{m+1}}}(x_{m+1})\Big)+$$
$$+F_{mu^i}(x_m,u)+\sum_{j_{m+1}=1}^{n_{m+1}}G_{mu^i}^{j_{m+1}}(x_m,u)\Phi_{x_{m+1}^{j_{m+1}}}(x_{m+1}).$$

Now sequentially using the vectors $\{\psi_m,\psi_{m-1},...,\psi_0\}$ from (15), we obtain the final result for J_{u^i} :

$$J_{u^i}=...+F_{mu^i}(x_m,u)+\sum_{j_{m+1}=1}^{n_{m+1}}G_{mu^i}^{j_{m+1}}(x_m,u)\Phi_{x_{m+1}^{j_{m+1}}}(x_{m+1})=...+F_{mu^i}(x_m,u)+G_{mu^i}^T(x_m,u)\cdot\psi_m=$$
$$=...+F_{(m-1)u^i}(x_m,u)+\sum_{j_m=1}^{n_m}G_{(m-1)u^i}^{j_m}(x_{m-1},u)\Big(F_{mx_m^{j_m}}(x_m,u)+G_{mx_m^{j_m}}^T(x_m,u)\cdot\psi_m\Big)+H_{mu^i}(x_m,\psi_m,u)=$$
$$=...+F_{(m-1)u^i}(x_m,u)+\sum_{j_m=1}^{n_m}G_{(m-1)u^i}^{j_m}(x_m,u)\Big(F_{mx_m^{j_m}}(x_m,u)+G_{mx_m^{j_m}}^T(x_m,u)\cdot\psi_m\Big)+H_{mu^i}(x_m,\psi_m,u)=$$
$$=...+F_{(m-1)u^i}(x_m,u)+G_{(m-1)u^i}^T(x_m,u)\cdot\psi_{m-1}+H_{mu^i}(x_m,\psi_m,u)=$$
$$=...+H_{(m-1)u^i}(x_{m-1},\psi_{m-1},u)+H_{mu^i}(x_m,\psi_m,u)=$$
$$=...=H_{0u^i}(x_0,\psi_0,u)+...+H_{mu^i}(x_m,\psi_m,u).$$

The proof is complete.

4. CALCULATION OF GENERALIZED GRADIENTS IN CONTROL PROBLEMS

Consider the following optimal control problem (Lagrange problem) in discrete time [1 - 5]:

$$J(x_0,u)=\sum_{i=0}^{m}F_i(x_i,u_i)+\Phi(x_{m+1})\to\min_u \tag{16}$$

subject to motion equations

$$x_{i+1}=G_i(x_i,u_i),\qquad i=0,1,...,m,\quad x_0\in\mathrm{R}^{n_0}, \tag{17}$$

and constraints

$$u=(u_0,...,u_m),\qquad u_i\in U_i\subset \mathrm{R}^{l_i},\qquad i=0,1,...,m, \tag{18}$$

where index $i=0,1,...,m$ designates a discrete time; $x_i=\left(x_i^1,...,x_i^{n_i}\right)^T\in \mathrm{R}^{n_i}$ is the state of the control system at moment i; $u_i=\left(u_i^1,...,u_i^{l_i}\right)^T\in \mathrm{R}^{l_i}$ is the control vector at time i; functions $G_i=\left(G_i^1,...,G_i^{n_{i+1}}\right)^T$, F_i, Φ are given; U_i is given set; $m\geq 0$ is a natural number; x_0 is a given point in R^{n_0}.

Theorem 7. Let functions F_i, G_i, Φ in problem (16)-(18) be generalized differentiable over the totality of their arguments $x_i\in \mathrm{R}^{n_i}$, $u_i\in V_i$, where V_i is an open neighborhood of the set U_i, $i=0,1,...,m$. Then the function $J(x_0,u)$ is generalized differentiable over $\left(x_0,u_0,u_1,...,u_m\right)\in \mathrm{R}^{n_0}\times V_0\times...\times V_m$, and the vector

$$\begin{aligned}&\left(H_{0x_0}(x_0,\psi_0,u_0),H_{0u_0}(x_0,\psi_0,u_0),...,H_{mu_m}(x_m,\psi_m,u_m)\right)^T=\\&=\Big(H_{0x_0^1}(x_0,\psi_0,u_0),...,H_{0x_0^{n0}}(x_0,\psi_0,u_0),\\&\quad H_{0u_0^1}(x_0,\psi_0,u_0),...,H_{0u_0^{l0}}(x_0,\psi_0,u_0),...,H_{mu_m^1}(x_m,\psi_m,u_m),...,H_{mu_m^{lm}}(x_m,\psi_m,u_m)\Big)^T,\end{aligned} \tag{19}$$

is a generalized gradient of the function $J(x_0,u)$ at a given point (x_0,u), where $H_i(x_i,\psi_i,u_i)=F_i(x_i,u_i)+G_i^T(x_i,u_i)\cdot\psi_i$, $i=0,1,...,m$, is a discrete (over i) Hamilton-Pontryagin function; $x=\left(x_0,...,x_{m+1}\right)$ is the discrete trajectory of process (17) that corresponds to the chosen control $u\in U=U_0\times U_1\times...\times U_m$; the sequence of auxiliary (conjugate) vector functions $\left(\psi_0,...,\psi_m\right)=\psi$ is defined by the back propagation equations:

$$\psi_{i-1}=H_{ix_i}(x_i,\psi_i,u_i)=F_{ix_i}(x_i,u_i)+G_{ix_i}^T(x_i,u_i)\cdot\psi_i,\quad i=k,k-1,...,0,\qquad \psi_m=\Phi_{x_{m+1}}(x_{m+1}).$$

Proof. This theorem is a consequence of Theorem 6 due to the fact that problem (16) - (18) can be interpreted as problem (12)-(13) with a vector variable $u=(u_0,u_1,...,u_m)\in U=U_0\times...\times U_m$. Since the Hamilton-Pontryagin function $H_i(x_i,\psi_i,u)$ in this case actually depends only on u_i, the components $J_{x_0^j}=H_{0x_0^j}(x_0,\psi_0,u_0)$ of the generalized gradient set of the function $J(x_0,u)$ are calculated as in Theorem 6, and when calculating the components $J_{u_i^j}$ by formula (14), only one term $H_{iu_i^j}(x_i,\psi_i,u_i)=J_{u_i^j}$ remains. The theorem is proved.

Comment. In [1 - 5], formula (19) was obtained as a consequence of the differentiation rules for complex smooth functions. For nonsmooth convex optimal control problems with linear equations of motion, similar formulas for calculating the subgradients of the objective function were obtained in [2–4]. In [3], these formulas were also substantiated for the case of a weakly convex [7] objective functional and smooth equations of motion. In Theorem 7, we generalize these results to a much wider class of nonsmooth nonconvex optimal control problems, including those with nonsmooth equations of motion.

5. CALCULATION OF STOCHASTIC GENERALIZED GRADIENTS IN THE PROBLEM OF STOCHASTIC OPTIMAL CONTROL

Let us consider the following stochastic optimal control problem in discrete time [2, 3]:

$$J(a_0,u)=\mathbf{E}\left(\sum_{i=0}^{m} f_i(x_i,u_i,\theta)+\varphi(x_{m+1},\theta)\right)\to\min_u \tag{20}$$

subject to conditions

$$x_{i+1}=g_i(x_i,u_i,\theta),\qquad i=0,1,...,m,\quad x_0=(a_0,b_0(\theta)), \tag{21}$$

$$u=(u_0,...,u_m)^T,\qquad u_i\in U_i,\qquad i=0,1,...,m, \tag{22}$$

where $x_i=\left(x_i^1,...,x_i^{n_i}\right)\in \mathrm{R}^{n_1}\times...\times\mathrm{R}^{n_i}$; $u_i=\left(u_i^1,...,u_i^{l_i}\right)\in \mathrm{R}^{m_1}\times...\times\mathrm{R}^{l_i}$; functions $g_i=\left(g_i^1,...,g_i^{n_{i+1}}\right)$, f_i, φ are given; U_i is a given set in R^{l_i}, $i=0,1,...,m$; $m\ge 1$ is a natural number; $a_0\in\mathrm{R}^{n_0}$ is a given deterministic vector; $b_0(\theta)$ is a random vector; θ is an elementary event of some probability space $(\Theta,\Sigma,\mathrm{P})$; $\mathbf{E}$ is the sign of the mathematical expectation over the measure P. The deterministic sequence $u=(u_0,...,u_m)^T$ is called a program control of the random process (21).

Let us introduce notation: $F_i(x_i,u_i)=\mathbf{E}f_i(x_i,u_i,\theta)$, $G_i(x_i,u_i)=\mathbf{E}g_i(x_i,u_i,\theta)$;

$$g_{ix}=\begin{pmatrix} g_{ix^1}^1 & ... & g_{ix^n}^1\\ ... & ... & ...\\ g_{ix^1}^n & ... & g_{ix^n}^n\end{pmatrix}=\begin{pmatrix} g_{ix}^1\\ ...\\ g_{ix}^n\end{pmatrix},\qquad g_{iu}=\begin{pmatrix} g_{iu^1}^1 & ... & g_{iu^m}^1\\ ... & ... & ...\\ g_{iu^1}^n & ... & g_{iu^m}^n\end{pmatrix}=\begin{pmatrix} g_{iu}^1\\ ...\\ g_{iu}^n\end{pmatrix},$$

$$\left(f_{ix}\right)^T=\left(f_{ix^1},...,f_{ix^n}\right),\qquad \left(f_{iu}\right)^T=\left(f_{iu^1},...,f_{iu^m}\right),\left(\varphi_x\right)^T=\left(\varphi_{x^1},...,\varphi_{x^n}\right),$$

where $\left(f_{ix},f_{iu}\right)^T$, $\left(g_{ix}^j,g_{iu}^j\right)^T$ are some generalized gradients of functions $f_i(\cdot,\cdot,\theta)$, $g_i^j(\cdot,\cdot,\theta)$ over variables (x_i,u_i) under fixed parameter θ; $\varphi_x(\cdot,\theta)$ is some generalized gradient of function φ under fixed θ; expression $(\cdot)^T$ designates the transposition of the matrix $(\cdot)$.

Assumption A. (A1) Let in problem (20) - (22) functions $f_i(x_i,u_i,\theta)$, $g_i^j(x_i,u_i,\theta)$ and $\varphi(x_{m+1},\theta)$ be generalized differentiable by the totality of arguments $(x_i \in \mathrm{R}^{n_i}, u_i \in V_i)$ and $x_{m+1} \in \mathrm{R}^{n_{m+1}}$ for each fixed θ and these functions are measurable in θ under fixed (x_i,u_i) and x_{m+1}, $i = 0,1,...,m$. (A2) Let for each point (x_i,u_i) and x_{m+1} there exist neighborhoods $O_i \times V_i$ and O_{m+1}, where the functions $f_i(\cdot,\cdot,\theta)$, $g_i^j(\cdot,\cdot,\theta)$ and $\varphi(\cdot,\theta)$ are Lipschitzian with integrable Lipschits constants. (A3) The multivalued mappings $\partial_{(x_i,u_i)} f_i(\cdot,\cdot,\theta)$, $\partial_{(x_i,u_i)} g_i^j(\cdot,\cdot,\theta)$ and $\partial_{x_{m+1}} \varphi(\cdot,\theta)$ are measurable in θ, for example, these may be Clarke subdifferentials of the functions $f_i(\cdot,\cdot,\theta)$, $g_i^j(\cdot,\cdot,\theta)$ и $\varphi(\cdot,\theta)$ (concerning the measurability of the subdifferentials and the generalized gradient mapping, consult [52]).

From Theorems 1 and 2, similarly to Theorem 7, the following statement follows.

Theorem 8. Under Assumption A the objective function $J(a_0,u)$ of problem (20) - (22) is generalized differentiable over $(a_0, u=(u_0,...,u_m))$, the random objective function $f(a_0,u,\theta) = \sum_{i=0}^{m} f_i(x_i,u_i,\theta) + \varphi(x_{m+1},\theta)$ of problem (20) - (22) is measurable and integrable in θ and is generalized differentiable over $(a_0, u=(u_0,...,u_m))$, and the vector

$$
\begin{aligned}
h_{a_0,u}(a_0,u,s) &= \left(h_{0a_0}(x_0,\psi_0,u_0,\theta), h_{u_0}(x_0,\psi_0,u_0,\theta),...,h_{u_m}(x_m,\psi_m,u_m,\theta)\right)^T = \\
&= \Big(h_{0a_0^1}(x_0,\psi_0,u_0,\theta),...,h_{0a_0^{n_0}}(x_0,\psi_0,u_0,\theta), \\
&\quad h_{0u_0^1}(x_0,\psi_0,u_0,\theta),...,h_{0u_0^{l_0}}(x_0,\psi_0,u_0,\theta),..., \\
&\quad h_{mu_m^1}(x_m,\psi_m,u_m,\theta),...,h_{mu_m^{l_m}}(x_m,\psi_m,u_m,\theta)\Big)^T,
\end{aligned}
\tag{23}
$$

is a generalized gradient (under fixed θ) of the function $f(\cdot,\cdot,\theta)$ at point (x_0,u), where $h_i(x_i,\psi_i,u_i,\theta) = f_i(x_i,u_i,\theta) + g_i^T(x_i,u_i,\theta)\cdot\psi_i$, $i = 0,1,...,m$, is a stochastic Hamilton-Pontryagin function; $x = (x_0,...,x_{m+1})$ is the discrete trajectory of process (21) that corresponds to the chosen control $u \in U = U_0 \times U_1 \times ... \times U_m$; the sequence of auxiliary (conjugate) vector functions $(\psi_0,...,\psi_m) = \psi$ is defined by the following backpropagation equations:

$$\psi_m = \varphi_{x_{m+1}}(x_{m+1},\theta),$$

$$\psi_{i-1} = h_{ix_i}(x_i,\psi_i,u_i,\theta) = f_{ix_i}(x_i,u_i,\theta) + g_{ix_i}^T(x_i,u_i,\theta)\cdot\psi_i, \qquad i = m,m-1,...,1,$$

$$\psi_0 = h_{1a_0}(x_1,\psi_1,u_1,\theta) = f_{1a_0}(x_1,u_1,\theta) + g_{1a_0}^T(x_1,u_1,\theta)\cdot\psi_1.$$

Thus, the vector $h_{a_0,u}(a_0,u,\theta)$ is a stochastic generalized gradient of the function $J(a_0,u)$ such that $\mathbf{E}h_{a_0,u}(a_0,u,\theta \in \partial_{a_0,u}J(a_0,u)$, and it can be used in stochastic gradient minimization of $J(a_0,u)$.

Remark. For smooth problems, as well as convex stochastic optimal control problems, formulas for calculating stochastic gradients of the objective function, similar to (23), were obtained in [2, 3].

Example 2 (stock evolution, continuation of Example 1). Continuing Example 1 consider the following parametric replenishment strategy

$$u_t(x_t, y) = \min\left\{M, \max\{0, \gamma(y - x_t)\}\right\},$$

where the search parameters $\gamma \geq 0$ and $y \geq 0$ are such that $0 \leq \gamma \leq 1$, $0 \leq y \leq M$. The quality of the functioning of the warehouse with such (feedback) control can be described by the criterion

$$F(a,\gamma,y) = \mathrm{E}\left(\sum_{t=0}^{m} \max\left\{\alpha(x_t - \theta_t), \beta(\theta_t - x_t)\right\} + \max\{\alpha(x_{m+1} - b), \beta(b - x_{m+1})\}\right) \to \min_{\gamma,y},$$

where α, β are some penalty factors for excess and shortage of goods in a period of time $[t, t+1)$; E is the mathematical expectation sign.

Here the random function

$$\varphi(a,\gamma,y,\theta_0,...,\theta_m) = \sum_{t=0}^{m} \max\left\{\alpha(x_t - \theta_t), \beta(\theta_t - x_t)\right\} + \max\{\alpha(x_{m+1} - b), \beta(b - x_{m+1})\}$$

is generalized differentiable with respect to its arguments (a,γ,y), and under the conditions of Theorem 2, the expectation function $F(a,\gamma,y)$ is also generalized differentiable with respect to its variables (a,γ,y).

Denote

$$f(x_t,\theta_t) = \max\left\{\alpha(x_t - \theta_t), \beta(\theta_t - x_t)\right\}, \qquad \Phi(x_{m+1}) = \max\{\alpha(x_{m+1} - b), \beta(b - x_{m+1})\},$$

$$g_t(x_t,\gamma,y,\theta_t) = \max\left\{0, x_t + \min\left\{M, \max\{0, \gamma(y - x_t)\}\right\} - \theta_t\right\} =$$

$$= \max\left\{0, \min\left\{x_t - \theta_t + M, \max\{x_t - \theta_t, \gamma(y - x_t) + x_t - \theta_t\}\right\}\right\},$$

$$\psi_m = \begin{cases} \alpha, & x_{m+1} > b, \\ -\beta, & x_{m+1} \leq b; \end{cases} \qquad f_{x_t}(x_t,\omega_t) = \begin{cases} \alpha, & x_t > \theta_t, \\ -\beta, & x_t \leq \theta_t. \end{cases}$$

A generalized $(g_{tx_t}, g_{t\gamma}, g_{ty})$ of the random function $g_t(x_t, \gamma, y, \theta_t)$ is calculated according to Theorem 1 by means of the following algorithm:

if $\left(x_t + \min\left\{M, \max\{0, \gamma(y - x_t)\}\right\} - \theta_t < 0\right)$ *then*

$g_{tx_t} = g_{t\gamma} = g_{ty} = 0,$

else if $\left(x_t - \theta_t + M < \max\{x_t - \theta_t, \gamma(y - x_t) + x_t - \theta_t\}\right)$ *then*

$g_{tx_t} = 1,\ g_{t\gamma} = g_{ty} = 0,$

else if $\left(x_t - \theta_t > \gamma(y - x_t) + x_t - \theta_t\right)$ *then*

$g_{tx_t} = 1,\ g_{t\gamma} = g_{ty} = 0,$

else $g_{tx_t} = 1 - \gamma,\ g_{t\gamma} = y - x_t,\ g_{ty} = \gamma.$

We recursively calculate

$$\psi_{t-1} = f_{x_t}(x_t, \theta_t) + \psi_t g_{tx_t}\ , \qquad t = m, m-1, ..., 0\ .$$

Then, by virtue of Theorem 4, taking into account the fact that $f_\gamma(x_t, \theta_t) = f_y(x_t, \theta_t) = 0$, we obtain

$$\varphi_a(a, \gamma, y, \theta_0, ..., \theta_m) = \psi_0 g_{0x_0}\ , \qquad \varphi_\gamma(a, \gamma, y, \theta_0, ..., \theta_m) = \sum_{t=0}^{m} \psi_t g_{t\gamma}\ ,$$

$$\varphi_y(a, \gamma, y, \theta_0, ..., \theta_m) = \sum_{t=0}^{m} \psi_t g_{ty}\ .$$

CONCLUSIONS

In this article, the theory of generalized differentiation of nonsmooth functions is extended to non-differentiable functionals of nonsmooth discrete dynamical systems. Formulas are obtained for calculating the generalized gradients of these functionals based on discrete Hamilton-Pontryagin functions and procedures for the direct modeling of trajectories and the back propagation of conjugate variables. The connection between the tasks of controlling discrete dynamic systems and the problems of training multilayer neural networks is observed, while the dynamics in multilayer networks is interpreted as layer-by-layer pass and transformation of the input signal to the output one and backward propagation of gradient information from the output of the network to the input. The results obtained, on the one hand, allow us to consider more general nonsmooth control and training problems, and on the other hand, expand the scope of application of nonsmooth and stochastic optimization methods.

In the subsequent article, we extend the BackProp method to multilayer neural networks nonconvex nonsmooth learning problems and formulate it in terms of generalized gradients of nonsmooth Hamilton-Pontryagin functions. We will also consider an important version of the BackProp method for training the so-called recurrent neural networks, i.e. networks with feedbacks and with memory.

References

1. Bryson, A.E., Ho, Y-C. (1969). Applied Optimal Control. Optimization, Estimation, and Control. Waltham, Massachusetts: Blaisdell Publishing Company, A Division of Ginn and Company.
2. Ermoliev, Y.M. (1976). Methods of Stochastic Programming. Moscow: Nauka. (In Russian).
3. Ermoliev, Y.M., Gulenko, V.P., Tsarenko, T.I. (1978). Finite-Difference Method in the Problems of Optimal Control. Kyiv: Naukova Dumka. (In Russian).
4. Vasiliev, F.P. (1981). Solution Methods for Extremal Problems. Moscow: Nauka. (In Russian).
5. Evtushenko, Y.G. (2013). Optimization and Fast Automatic Differentiation. Moscow: A.A. Dorodnitzin computational center of the Russian Academy of Sciences. (In Russian).
6. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks.* Vol. 61, pp. 85-117.
7. Nurminski, E.A. (1979). Numerical methods for solving stochastic minimax problems. Kyiv: Naukova Dumka. (In Russian).
8. Pontryagin, L.S., Boltianski, V.G., Gamkelidze, R.V., Mischenko, E.F. (1962). The Mathematical Theory of Optimal Processes. John Wiley.
9. Boltianski, V.G. (1973). Optimal Control of Discrete Systems. Moscow: Nauka. (In Russian).
10. Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning Representations by Back-Propagating Errors. *Nature*. Vol. 323, pp. 533-536. DOI: https://doi.org/10.1038/323533a0
11. LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. 436. *Nature*. Vol. 521. DOI:10.1038/nature14539
12. Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning, MIT Press.
13. Norkin, V.I. Generalized differentiable functions. *Cybernetics.* Vol. 16, No.1, pp. 10-12. DOI: https://doi.org/10.1007/BF01099354
14. Demianov, V.F. (2005). Conditions of the Extremum and the Variational Calculous. Moscow: Vysshia Shkola. (In Russian).
15. Daduna, H., Knopov, P.S., Tur, L.P. (1999). Optimal strategies for an inventory system with cost functions of general form. *Cybern. Syst. Anal*. Vol. 35, Iss. 4, pp. 602–618. DOI https://doi.org/10.1007/BF02835856

16. Ermoliev, Yu.M., Norkin, V.I. (1997). Stochastic generalized gradient method with application to insurance risk management. *Interim Report* IR-97-021. Laxenburg, Austria: Int. Inst. for Appl. Syst. Anal.
17. Pschenichnyi, B.N. (1972). Necessary Conditions of Extremum, Marcel Decker.
18. Demyanov, V.F., Vinogradova, T.K., Nikulina, V.N. et al. (1982). Nonsmooth Problems of Optimization and Control. Leningrad: Leningrad University Press. (In Russian).
19. Clarke, F.H. (1990). Optimization and Nonsmooth Analysis. Volume 5 of Classics in Applied Mathematics. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), Second edition.
20. Morduhovich, B.S. (1988). Optimization Methods in Optimization and Control Problems. Moscow: Nauka. (In Russian).
21. Rockafellar, R.T., Wets, R.J-B. (1998). Variational Analysis. Berlin, Heidelberg: Springer.
22. Ahn, H-S., Moore, K.L., Chen, Y.Q. (2007). Iterative Learning Control. Robustness and Monotonic Convergence for Interval Systems. London: Springer-Verlag.
23. LeCun, Y.A, Bottou, L., Orr, G.B., Muller, K.-R. (2012). Efficient BackProp. G. Montavon et al. (Eds.): NN: Tricks of the Trade, 2nd edn., LNCS 7700, pp. 9 - 48. Berlin, Heidelberg: Springer-Verlag.
24. Nikolenko, S., Kadurin, A., Arhangelskaia, E. (2018). Deep Learning. Diving in the world of neural networks. St. Petersburg: Piter.
25. Griewank, A., Walther, A. (2008). Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation. Second Edition. Philadelphia: Society for Industrial and Applied Mathematics.
26. Hardt, M., Recht, B., Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. ICML'16 Proceedings of the 33rd International Conference on Machine Learning. Vol. 48. New York, NY, USA: JMLR.org. P. 1225-1234.
27. Zhang, C., Liao, Q., Rakhlin, A., Miranda, B., Golowich, N., Poggio, T. (2018). Theory of Deep Learning IIb: Optimization Properties of SGD. *CBMM Memo* No. 072. Cambridge, MA: Center for Brains, Minds, and Machines, McGovern Institute for Brain Research, Massachusetts Institute of Technology. 9 p. arXiv:1801.02254v1 [cs.LG] 7 Jan 2018
28. Soudry, D., Hoffer, E., Nacson, M.S., Gunasekar, S., Srebro, N. (2018). The Implicit Bias of Gradient Descent on Separable Data. arXiv:1710.10345v3 [stat.ML] 21 Mar 2018
29. Bottou, L., Curtis, F.E., Nocedal, J. (2018). Optimization Methods for Large-Scale Machine Learning. *SIAM Review*. Vol. 60(2), pp. 223–311. DOI:10.1137/16m1080173
30. Robbins, H., Monro, S. (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics*. Vol. 22(3), pp. 400-407.

31. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A. (2009). Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM J. Optim.* Vol. 19(4), pp. 1574-1609.
32. Shapiro, A., Dentcheva, D., Ruszczynski, A. (2009). Lectures on Stochastic Programming: Modeling and Theory. Philadelphia: SIAM.
33. Ermoliev, Y.M., Norkin, V.I. (1998). Stochastic generalized gradient method for solving nonconvex nonsmooth stochastic optimization problems. *Cybern. Syst. Anal.*, 34, No. 2, pp. 196-215. DOI: https://doi.org/10.1007/BF02742069
34. Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J.D. (2019). Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* P. 1-36. https://doi.org/10.1007/s10208-018-09409-5
35. Shor, N.Z. (1985). Minimization Methods for Non-Differentiable Functions. Springer.
36. Mifflin, R. (1977). An algorithm for constrained optimization with semi-smooth functions. *Math. Oper. Res.* Vol. 2, No. 2, pp.191-207.
37. Mifflin, R. (1977). Semismooth and semiconvex functions in constrained optimization. *SIAM J. Contr. Opt*. Vol. 15, No. 6, pp. 959-972.
38. Gupal, A. M. (1979). Stochastic Methods for Solution of Nonsmooth Extremal Problems, Kiev: Naukova Dumka. (In Russian).
39. Mifflin, R. A. (1982). Modification and an extension of Lemarechal's algorithm for nonsmooth minimization. In: Nondifferential and Variational Techniques in Optimization (D.C. Sorensen, J.B. Wets, eds.). *Math. Prog. Study*. Vol. 17, pp. 77-90.
40. Shor, N.Z. (1985). Minimization Methods for Nondifferentiable Functions. Berlin, Heidelberg: Springer.
41. Dorofeev, P.A. (1985). On some properties of the generalized gradient method, *Zh. Vych. Mat. Mat. Fiz.* 25, No. 2, pp. 181–189. (In Russian).
42. Mikhalevich, V.S., Gupal, A.M., Norkin, V.I. (1987). Methods of Nonconvex Optimization. Moscow: Nauka. (In Russian).
43. Zavriev, S.K., Perevozchikov, A.G. (1990). The stochastic method of the generalized gradient descent for solution of minimax problems with apparent variables. *Zh. Vych. Mat. Mat. Fiz.*, 30, No. 4, pp. 491–500. (In Russian).
44. Urjas'ev, S.P. (1990). Adaptive Algorithms of Stochastic Optimization and Game Theory. Moscow: Nauka. (In Russian).
45. Hiriart-Urruty, J.-B., Lemarechal, C. (1993). Convex analysis and minimization algorithms, Vol. II, Berlin, Heidelberg: Springer-Verlag.

46. Fukushima M., Qi L. (Eds.) (1999). Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods. Dordrecht, Boston, London: Kluwer Academic Publishers.
47. Stetsyuk, P.I. (2017). Theory and Software Implementations of Shor's *r*-Algorithms. *Cybern. Syst. Anal.* Vol. 53, Iss. 5, pp. 692–703. DOI: https://doi.org/10.1007/s10559-017-9971-1
48. Norkin, V.I. Stochastic methods for solving nonconvex stochastic optimization problems and their applications. *Extended abstract of the Doctor Thesis*. V.M.Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine. Kyiv, 1998. 32 p. Access (22.07.2019): http://library.nuft.edu.ua/ebook/file/01.05.01%20Norkin%20VI.pdf
49. Ermoliev, Y.M., Norkin, V.I. (2003). Solution of nonconvex nonsmooth stochastic optimization problems. *Cybern. Syst. Anal.*, 39, No. 5, pp. 701-715. DOI: https://doi.org/10.1023/B:CASA.0000012091.84864.65
50. Norkin, V.I. (1978). Nonlocal minimization algorithms of nondifferentiable functions. *Cybernetics*, 14, No. 5, pp. 704-707. DOI: https://link.springer.com/article/10.1007/BF01069307
51. Bolte, J., Daniilidis, A., Lewis, A. (2009). Tame functions are semismooth. *Math. Program., Ser. B*. Vol. 117, pp. 5-19. DOI 10.1007/s10107-007-0166-9
52. Norkin, V.I. (1986). Stochastic generalized-differentiable functions in the problem of nonconvex nonsmooth stochastic optimization. *Cybernetics*, 22, No.6, pp. 804-809. https://doi.org/10.1007/BF01068698
53. Mirică, S. (1980). A note on the generalized differentiability of mappings. *Nonl. Anal. Theory, Methods, Applications*. Vol. 4, No 3, pp. 567-575. DOI:10.1016/0362-546x(80)90092-9
54. Lyashko, I.I., Yemelianov, V.F., Boyarchuk, O.K. (1992). Mathematical Analysis. Part I. Kyiv: Vyscha Shkola. (In Ukrainian).
55. Qi, L., Sun, J. (1993). A nonsmooth version of Newton's method. *Math. Progr.* Vol. 58, pp. 353-368.
56. Qi, L. (1993). Convergence analysis of some algorithms for solving nonsmooth equations, *Math. Oper. Res.* Vol. 18, pp. 227-244.
57. Polyak, B.T. (1987). Introduction to Optimization. Optimization Software.
58. Sternberg, S. (1964). Lectures on Differential Geometry. Englewood Cliffs, N.J.: Prentice Hall.

59. Burke, J., Lewis, A. & Overton, M. (2005). A robust gradient sampling algorithm for nonsmooth nonconvex optimization, *SIAM J. on Opt.* Vol. 15(3), pp. 751-779. DOI:10.1137/030601296