ISSN 0005-1179, Automation and Remote Control, 2006, Vol. 67, No. 4, pp. 589–597. © Pleiades Publishing, Inc., 2006.
Original Russian Text © A.T. Vakhitov, O.N. Granichin, S.S. Sysoev, 2006, published in Avtomatika i Telemekhanika, 2006, No. 4, pp. 86–96.

A Randomized Stochastic Optimization Algorithm: Its Estimation Accuracy

A. T. Vakhitov, O. N. Granichin, and S. S. Sysoev

St. Petersburg State University, St. Petersburg, Russia

Received December 7, 2004

Abstract—For a randomized stochastic optimization algorithm, consistency conditions of estimates are slackened and the order of accuracy for a finite number of observations is studied. A new method of realization of this algorithm on quantum computers is developed.

PACS number: 02.50.Fz, 02.60.Pn, 03.67.Lx

DOI: 10.1134/S0005117906040060

1. INTRODUCTION

The problem of search for the minimum (maximum) of a function (functional) $f(\mathbf{x}) \to \min_{\mathbf{x}}$ is known since long. A large number of problems is reduced to this problem. Often the order of equations thus obtained and the number of unknowns are such that the solution cannot be determined analytically. In reality, analytical solution is not of great significance since it is distorted in application (for example, due to the limited digital capacity of the computer or inaccuracy of measuring devices).

For a continuously differentiable function, the problem is reduced to the determination of the roots of its derivative (or points at which the gradient vanishes). But if the function is not differentiable or the general form of the function is not known, the problem takes qualitatively a different nature. Nevertheless, there are algorithms capable of solving a wide range of problems with any given degree of accuracy. Such algorithms need no special application techniques and do not strongly depend on the type of the functional (if it belongs to a special class of applications). Such a kind of universality obviously results in the simplicity of their computer realization and iterative nature aids in refining the estimate at every new iteration. Here we mean recurrent stochastic optimization algorithms.

Most of pseudo-gradient optimization methods like the Kiefer–Wolfowitz procedure [1] require several measurements of the loss function at every iteration. As a result, the minimized function must be measured at several points in every iteration. But if the function changes with time or its measurement depends on the realization of some random variable and the function is to be minimized on the mean

$$\mathrm{E}_{\mathbf{w}}\{F(\mathbf{x},\mathbf{w})\} \to \min_{\mathbf{x}},$$

then multiple measurement of the function at a point is not possible. Such a situation arises, for instance, in optimization of systems in real time.

Algorithms satisfying such constraints were designed at the end of eighties and beginning of nineties [2–12]. They are called the randomized stochastic optimization algorithms since their input data are artificially randomized at every step. Their main advantages are the convergence under “almost arbitrary” perturbations [9–15] and a small number (one or two) of measurements of the loss function in iterations.

This paper is the continuation of [9–11]. Here we study new, but weaker, conditions for the convergence of the randomized stochastic optimization algorithm with one measurement of the loss function, estimate the result for a finite number of iterations, and design a scheme for realization of the main part of this algorithm on a quantum computer. Weaker convergence conditions widen the range of application of the algorithm and, consequently, the algorithm can be confidently applied even if the properties of the loss functions are known only in part.

2. FORMULATION OF THE PROBLEM AND MAIN ASSUMPTIONS

Let $F(\mathbf{x},\mathbf{w}):\ \mathbb{R}^q\times\mathbb{R}^p\to\mathbb{R}^1$ be a differentiable function with respect to x and let $\mathbf{x}_1,\mathbf{x}_2,\ldots$ be an experimentally chosen sequence of measurement points (observation plan) at which the value

$$y_n = F(\mathbf{x}_n,\mathbf{w}_n)+v_n,$$

of the function $F(\cdot,\mathbf{w}_n)$ is observed with additive noises v_n at instants $n=1,2,\ldots$, where $\{\mathbf{w}_n\}$ is an uncontrollable sequence of random variables belonging to $\mathbb{R}^p$ and having, in general, an unknown distribution $\mathrm{P}_w(\cdot)$ with a finite carrier.

Formulation of the problem. Using observations $y_1,y_2,\ldots$, we must construct a sequence of estimates $\{\widehat{\boldsymbol{\theta}}_n\}$ for the unknown vector $\boldsymbol{\theta}_*$ minimizing the function

$$f(\mathbf{x}) = \mathrm{E}_{\mathbf{w}}\{F(\mathbf{x},\mathbf{w})\} = \int\limits_{\mathbb{R}^p} F(\mathbf{x},\mathbf{w})\mathrm{P}_w(d\mathbf{w})$$

of the type of a mean-risk functional.

Usually, minimization of the function $f(\cdot)$ is studied with a simpler observation model

$$y_n = f(\mathbf{x}_n)+v_n$$

that fits within the general scheme. The generalization stipulated in the formulation of the problem is required to include the case of multiplicative noises in observations

$$y_n = \mathbf{w}_n f(\mathbf{x}_n)+v_n,$$

and this case is contained in the general scheme with the function $F(\mathbf{x},\mathbf{w})=\mathbf{w}f(\mathbf{x})$.

Let $\rho\in(1,2]$. In what follows, we denote expectation by $\mathrm{E}\{\cdot\}$, l_ρ-norm by $\|\cdot\|_\rho$ and scalar product in $\mathbb{R}^q$ by $\langle\cdot,\cdot\rangle$. Let us introduce a function

$$V(\mathbf{x}) = \|\mathbf{x}-\boldsymbol{\theta}_*\|_\rho^\rho = \sum_{i=1}^q |x^{(i)}-\theta_*^{(i)}|^\rho,$$

where $\boldsymbol{\theta}_*$ is an unknown vector.

Let us formulate the main assumptions.

(A.1) The function $f(\mathbf{x})$ has a unique minimum and

$$(\nabla V(\mathbf{x}),\nabla f(\mathbf{x}))\geq \mu V(\mathbf{x})\ \forall\mathbf{x}\in\mathbb{R}^q$$

with some constant $\mu>0$.

(A.2) For any $\mathbf{w}$, gradients of the functions $F(\cdot,\mathbf{w})$ satisfy the condition

$$\|\nabla_{\mathbf{x}}F(\mathbf{x},\mathbf{w})-\nabla_{\mathbf{x}}F(\mathbf{y},\mathbf{w})\|_{\frac{\rho}{\rho-1}}\leq M\|\mathbf{x}-\mathbf{y}\|_{\frac{\rho}{\rho-1}}\ \forall\,\mathbf{x},\mathbf{y}\in\mathbb{R}^q$$

with some constant $M>0$.

3. TEST PERTURBATION AND THE MAIN ALGORITHM

Let $\boldsymbol{\Delta}_n$, $n=1,2,\ldots$, be an observed sequence of independent random vectors in $\mathbb{R}^q$, called the *simultaneous test perturbation*. Vector components are also mutually independent and take values ±1 with identical probability $\frac{1}{2}$.

Taking an initial vector $\widehat{\boldsymbol{\theta}}_0 \in \mathbb{R}^q$, let us choose two sequences $\{\alpha_n\}$ and $\{\beta_n\}$ of positive numbers. In [10–12], the sequence of measurement points $\{\mathbf{x}_n\}$ and estimate sequence $\{\widehat{\boldsymbol{\theta}}_n\}$ are constructed with the following algorithm based on the use of one observation at every step (iteration):

$$\begin{cases} \mathbf{x}_n = \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_n, \quad y_n = F(\mathbf{x}_n, \mathbf{w}_n) + v_n \\ \widehat{\boldsymbol{\theta}}_n = \mathcal{P}_{\Theta_n}\left(\widehat{\boldsymbol{\theta}}_{n-1} - \dfrac{\alpha_n}{\beta_n}\boldsymbol{\Delta}_n y_n\right), \end{cases} \tag{1}$$

where $\mathcal{P}_{\Theta_n}(\cdot)$, $n = 1, 2, \ldots$, are projections on certain convex closed bounded subsets $\Theta_n \subset \mathbb{R}^q$ containing, beginning from some $n \geq 0$, the point $\boldsymbol{\theta}_*$. For a known convex closed bounded set Θ containing the point $\boldsymbol{\theta}_*$, we can take $\Theta_n = \Theta$. Otherwise, the sets $\{\Theta_n\}$ may be infinite.

4. CONVERGENCE

Let $\mathbb{W} = \mathrm{supp}(\mathrm{P}_{\mathbf{w}}(\cdot)) \subset \mathbb{R}^p$ be a finite carrier of the distribution $\mathrm{P}_{\mathbf{w}}(\cdot)$, let $\mathcal{F}_n$ be a σ-algebra of probability events generated by the random variables $\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_n$, formed by algorithm (1), let $d_n = \mathrm{diam}(\Theta_n)$ be the diameter of the set Θ_n in the metric $l_{\frac{\rho}{\rho-1}}$, and let

$$\gamma_n = \alpha_n \rho\mu - \alpha_n\beta_n(\rho-1)q^{\frac{\rho+1}{\rho}}M - 2^{2\rho-1}qc_n\psi_n,$$

$$\phi_n = \alpha_n\beta_n q^{\frac{\rho+1}{\rho}}M + 2^\rho q c_n \max_{\mathbf{w}\in\mathbb{W}} |F(\boldsymbol{\theta}_*, \mathbf{w})|^\rho + 2^{3\rho-2}q^2\beta_n^\rho\psi_n,$$

$$c_n = \alpha_n^\rho\beta_n^{-\rho}\rho, \quad \psi_n = M^\rho d_n^\rho + \max_{\mathbf{w}\in\mathbb{W}} \|\nabla_{\mathbf{x}}F(\boldsymbol{\theta}_*, \mathbf{w})\|^\rho_{\frac{\rho}{\rho-1}}.$$

Theorem 1. *Let* $\rho \in (1, 2]$ *and let the following conditions hold:*

the function $f(\mathbf{x}) = \mathrm{E}\{F(\mathbf{x}, \mathbf{w})\}$ *satisfies* (A.1),

the function $F(\cdot, \mathbf{w})$ $\forall \mathbf{w} \in \mathbb{W}$ *satisfies* (A.2),

the functions $F(\mathbf{x}, \mathbf{w})$ *and* $\nabla_{\mathbf{x}}F(\mathbf{x}, \mathbf{w})$ *are uniformly bounded on* $\mathbb{W}$,

$\forall n \geq 1$, *the random variables* $v_1, \ldots, v_n$ *and vectors* $\mathbf{w}_1, \ldots, \mathbf{w}_{n-1}$ *do not depend on* $\mathbf{w}_n$ *and* $\boldsymbol{\Delta}_n$, *and the random vector* $\mathbf{w}_n$ *does not depend on* $\boldsymbol{\Delta}_n$,

$$\mathrm{E}\{|v_n|^\rho\} \leq \sigma_n^\rho, \quad n = 1, 2, \ldots,$$

$\forall n$, $0 \leq \gamma_n \leq 1$, $\sum\limits_n \gamma_n = \infty$, $\mu_n \to 0$ *as* $n \to \infty$, *where*

$$\mu_n = \frac{\phi_n + 2qc_n\sigma_n^\rho}{\gamma_n}, \quad z_n = \left(1 - \frac{\mu_{n+1}}{\mu_n}\right)\frac{1}{\gamma_{n+1}}.$$

Then

(1) *the estimate sequence* $\{\widehat{\boldsymbol{\theta}}_n\}$ *generated by algorithm* (1) *tends to the point* $\boldsymbol{\theta}_*$ *in the sense*

$$\mathrm{E}\{V(\widehat{\boldsymbol{\theta}}_n)\} \to 0 \text{ as } n \to \infty,$$

(2) *if* $\overline{\lim\limits_{n\to\infty}}\, z_n \geq z > 1$, *then* $\mathrm{E}\{V(\widehat{\boldsymbol{\theta}}_n)\} = \mathcal{O}\left(\prod\limits_{i=0}^{n-1}(1-\gamma_i)\right)$,

(3) *if* $z_n \geq z > 1$ $\forall n$, *then* $\mathrm{E}\{V(\widehat{\boldsymbol{\theta}}_n)\} \leq \left(\mathrm{E}\{V(\widehat{\boldsymbol{\theta}}_0)\} + \dfrac{\mu_0}{z-1}\right)\prod\limits_{i=0}^{n-1}(1-\gamma_i)$,

(4) *if, additionally,*

$$\sum_n \phi_n + 2qc_n\mathrm{E}\{\sigma_n^\rho | \mathcal{F}_{n-1}\} < \infty,$$

then $\widehat{\theta}_n \to \theta_*$ *as* $n \to \infty$ *with probability 1, and*

$$\mathrm{P}\{V(\widehat{\boldsymbol{\theta}}_n) \le \varepsilon \ \forall n \ge n_0\} \ge 1 - \frac{\mathrm{E}\{V(\widehat{\boldsymbol{\theta}}_{n_0})\} + \sum\limits_{n=n_0}^{\infty} \phi_n + 2qc_n\sigma_n^{\rho}}{\varepsilon}.$$

The proof of Theorem 1 is given in the Appendix.

Remark (1) Conditions of Theorem 1 hold for the function $F(\mathbf{x}, \mathbf{w}) = \mathbf{w}f(\mathbf{x})$ if the function $f(\mathbf{x})$ satisfies assumptions (A.1) and (A.2).

(2) In Theorem 1, observation noises v_n can be said to be "almost arbitrary" since they may not be random, but unknown and bounded or realization of some stochastic process of arbitrary structure. Moreover, there is no need to assume that v_n and $\mathcal{F}_{n-1}$ are related to prove the assertions of Theorem 1.

(3) For Theorem 1 to hold, the components of the test perturbation $\boldsymbol{\Delta}_n$ need not necessarily take only the values ±1. It suffices to assume that their distribution carrier is symmetric and finite.

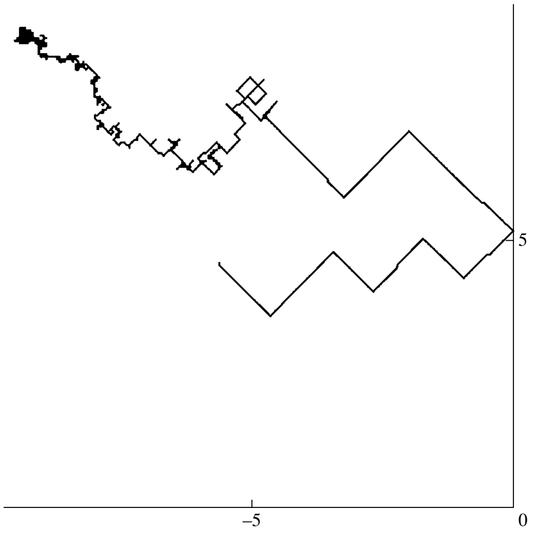
5. AN EXAMPLE

Let us demonstrate the performance of algorithm (1) with an example on its application to a two-dimensional optimization problem. Let us assume that a point $\boldsymbol{\theta}_*$ (target) lies on a plane and its location is not known. Let us assume that for any point $\mathbf{x}$, we can measure $|x^{(1)} - \theta_*^{(1)}|^{1.2} + |x^{(2)} - \theta_*^{(2)}|^{1.2}$ only with multiplicative and additive noises, i.e.,

$$y_n = \mathbf{w}_n \left(|x^{(1)} - \theta_*^{(1)}|^{1,2} + |x^{(2)} - \theta_*^{(2)}|^{1.2}\right) + v_n$$

are observable. The problem now is to find the location of the target.

In this case, $F(\mathbf{x}, \mathbf{w}) = \mathbf{w}\|\mathbf{x} - \boldsymbol{\theta}_*\|_{1.2}^{1.2}$. In the modeling experiment, noises were taken to be random independent sequences with a normal distribution $\mathcal{N}(1; 1)$ for multiplicative and bounded random variables, and deterministic sequences for additive variables $|v_n| \le \dfrac{1}{2}$. Number sequences $\{\alpha_n\}$ and $\{\beta_n\}$ were chosen such that $\alpha_n = \dfrac{0.15}{\sqrt{n}}$, $\beta_n = 1$. Projection was not used. The figure shows a typical result generated by the algorithm in one thousand iterations (vertices of the broken line are estimates). Coordinates of the target are $\boldsymbol{\theta}_* =(-6.58, 8.87)^{\mathrm{T}}$. Coordinates of the estimate are $\widehat{\boldsymbol{\theta}}_{1000} = (-6.76, 8.78)^{\mathrm{T}}$.



An estimate sequence.

Theorem 1 states only sufficient conditions. The main constraints on the minimized function are satisfied in our example. Though not all conditions are satisfied for the number sequences $\{\alpha_n\}$ and $\{\beta_n\}$ and the carrier $\mathbb{W}$ is not finite, the algorithm generated satisfactory estimate sequences.

6. QUANTUM COMPUTER AND ESTIMATION OF THE GRADIENT VECTOR OF A FUNCTION

Let us examine the choice of a best computer for implementing the randomized stochastic optimization algorithm with one measurement of the penalty function in iterations. Realization of algorithm (1) on a quantum computer is described in [10]. Recently, terminology and axiomatics of quantum computation models have been greatly refined. The realization method of [10] does not resemble a typical "quantum" algorithm and the earlier representation is not satisfactory. Below we describe a new method of representation of algorithms for implementation on a quantum computer, i.e., a method that is consistent with the general logic of quantum computation algorithms.

Till recently, quantum computer was regarded exclusively as a speculative mathematical model. It is no longer speculative owing to the NMR-based quantum computers developed by the IBM Corporation [16]. Of course, serious difficulties are still encountered in designing a quantum computer for everyday use. Nonetheless, intensive research and development projects are under way to surmount this problem.

We now briefly outline the mathematic model of a quantum computer and show how algorithm (1) is represented in this model. States in quantum mechanics are often denoted by vectors of unit length in a vector space over the field of complex numbers. Observed quantities are represented by self-conjugate operators in this complex space [17]. An observed quantity is a method of obtaining information on the state. Measurement of the quantum system changes this information. By measurement, we obtain one of the eigenvalues of the observed quantity and the system itself passes to the eigenvector corresponding to this eigenvalue. In measurement, an eigenvalue is chosen at random with a probability, equal to the square of the projection of the state on the corresponding eigenvector. Clearly, measurement of the quantum system yields complete information only if the state of the system coincides with one of the eigenvectors of the chosen observed quantity.

A quantum computer processes "qbits" ("quantum bits"), which form a quantum system of two states (a microscopic system corresponding to the description of an excited ion or a polarized photon, or spin of the atomic nucleus). The base of the qbit state space is usually denoted by $|0\rangle$ and $|1\rangle$ in analogy with the notation $\{0, 1\}$ used in the classical information theory. Such a system can exist not only in base states, but is also capable of storing more information than the corresponding classical system. Nevertheless, it passes to one of the base states during measurements (if the observed system is properly chosen) and the information it contains corresponds to some classical information. A quantum system of r qbits is represented as a tensor product. A set of base vectors of this state space can be parametrized by bit rows of length r. For example, for $r = 3$, the base vector $|0\rangle \otimes |0\rangle \otimes |0\rangle$ can be denoted by $|000\rangle$. Sometimes it is more convenient to use another form of expression $|0\rangle|0\rangle|0\rangle$ or $|00\rangle|0\rangle$. Therefore, the dimension of the space that a quantum computer uses grows exponentially with the number of qbits. This property underlies the "quantum parallelism." A rigorous quantum computation model is described in [18, 19].

Another important property of quantum states is their unique evolution. In other words, every transformation of qbits in a quantum computer is a unitary operator in the corresponding complex space. Hence every transformation of information (except for measurements) in a quantum computer must be invertible. Let $f : \mathbb{R}^q \to \mathbb{R}$ be a function satisfying the conditions of Theorem 1. Let us assume that the quantum computer is an r-bit machine. The unitary operation realizing the function $f(\mathbf{x})$ on a quantum computer can be defined on all classical binary chains $\mathbf{x}$ of length qr, defining the argument of the function $U_f : |\mathbf{x}\rangle|\mathbf{y}\rangle \to |\mathbf{x}\rangle|\mathbf{y} \oplus f(\mathbf{x})\rangle$, where $\mathbf{y}$ is an arbitrary binary chain of length r and $\oplus$ is a bit-by-bit operation of logical AND. This is a method of defining an operator on base vectors. On all other vectors, the operator is continued linearly. Clearly, the operator thus constructed is invertible and acts in a complex space of dimension 2^{qr+r}.

We estimate the minimum of a function, using algorithm (1). To feed the computer input, let us prepare a superposition of 2^q perturbed values of the current estimate vector

$$\mathbf{x}_n = \frac{1}{2^{\frac{q}{2}}} \sum_{\Delta_i \in \{-1,+1\}^q} |\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_i\rangle,$$

where ±1 are regarded as r-digit numbers. Applying the unitary operator U_f to $|\mathbf{x}_n\rangle|\mathbf{0}\rangle$, we obtain

$$U_f|\mathbf{x}_n\rangle|\mathbf{0}\rangle = \frac{1}{2^q} \sum_{\boldsymbol{\Delta}_i \in \{-1,+1\}^q} |\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_i\rangle |f(\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_i)\rangle.$$

By the general properties of the quantum computation model, after a state measurement, we obtain with probability $\dfrac{1}{2^q}$ a vector

$$|\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_i\rangle |f(\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_i)\rangle, \quad \boldsymbol{\Delta}_i \in \{-1,+1\}^q.$$

Using the first qr digits of this vector, we can easily determine a random perturbation vector $\boldsymbol{\Delta}_i$. According to algorithm (1), its coordinates of must be multiplied by the corresponding value of the loss function at a perturbed point, i.e., by the value at the last r digits of the measurement result.

7. CONCLUSIONS

There are several deterministic and stochastic iterative methods of optimization. Our method is superior in one respect—the loss function is measured only once in every iteration. Moreover, the only condition is that measurement noises must not depend on the simultaneous test perturbation.

For problems requiring several measurements of the loss function in iterations, algorithms with one measurement (for example, randomized algorithms with two or more measurements) are preferable since they rapidly converge to the point of minimum. Nevertheless, there are problems for which the one-measurement method is the only method. Therefore, the study of its applicability range is imperative.

APPENDIX

Proof of Theorem 1. Let us consider algorithm (1). Using the properties of the function $V(\mathbf{x})$, from the mean-value theorem we obtain for some $t \in (0,1)$

$$V(\widehat{\boldsymbol{\theta}}_n) = V\left(\mathcal{P}_{\Theta_n}\left(\widehat{\boldsymbol{\theta}}_{n-1} - \frac{\alpha_n}{\beta_n}\boldsymbol{\Delta}_n y_n\right)\right) \le V\left(\widehat{\boldsymbol{\theta}}_{n-1} - \frac{\alpha_n}{\beta_n}\boldsymbol{\Delta}_n y_n\right)$$
$$= V(\widehat{\boldsymbol{\theta}}_{n-1}) - \frac{\alpha_n}{\beta_n}\left\langle \nabla V(\widehat{\boldsymbol{\theta}}_{\mathrm{mid}}), \boldsymbol{\Delta}_n y_n \right\rangle = V(\widehat{\boldsymbol{\theta}}_{n-1}) - \frac{\alpha_n}{\beta_n}\left\langle \nabla V\left(\widehat{\boldsymbol{\theta}}_{n-1} - t\frac{\alpha_n}{\beta_n}\boldsymbol{\Delta}_n y_n\right), \boldsymbol{\Delta}_n y_n \right\rangle$$
$$= V(\widehat{\boldsymbol{\theta}}_{n-1}) - \rho\frac{\alpha_n}{\beta_n}\sum_{i=1}^{q}\left|\widehat{\theta}_{n-1}^{(i)} - \theta_*^{(i)} - t\frac{\alpha_n}{\beta_n}\Delta_n^{(i)} y_n\right|^{\rho-1} \operatorname{sgn}_n^{(i)}(t)\Delta_n^{(i)} y_n,$$

where $\operatorname{sgn}_n^{(i)}(t) = 0$ or ± 1, depending on the sign of the expression

$$\widehat{\theta}_{n-1}^{(i)} - \theta_*^{(i)} - t\frac{\alpha_n}{\beta_n}\Delta_n^{(i)} y_n.$$

Since the inequality

$$-\operatorname{sgn}(c-d)|c-d|^{\rho-1}b \le -\operatorname{sgn}(c)|c|^{\rho-1}b + 2^{2-\rho}|d|^{\rho-1}|b|$$

holds for b, c, and $d \in \mathbb{R}$, we obtain

$$\begin{aligned} V(\widehat{\boldsymbol{\theta}}_n) &\le V(\widehat{\boldsymbol{\theta}}_{n-1}) - \rho\frac{\alpha_n}{\beta_n}\sum_{i=1}^{q}\left|\widehat{\theta}^{(i)}_{n-1} - \theta^{(i)}_*\right|^{\rho-1}\operatorname{sgn}^{(i)}_n(t)\Delta^{(i)}_n y_n \\ &+ 2^{2-\rho}\rho\frac{\alpha_n}{\beta_n}\sum_{i=1}^{q}\left|t\frac{\alpha_n}{\beta_n}\Delta^{(i)}_n y_n\right|^{\rho-1}|\Delta^{(i)}_n y_n| \le V(\widehat{\boldsymbol{\theta}}_{n-1}) \\ &- \rho\frac{\alpha_n}{\beta_n}\sum_{i=1}^{q}\nabla V(\widehat{\theta}_{n-1})^{(i)}\Delta^{(i)}_n y_n + 2^{2-\rho}\rho\left(\frac{\alpha_n}{\beta_n}\right)^{\rho}\sum_{i=1}^{q}|\Delta^{(i)}_n y_n|^{\rho} \\ &= V(\widehat{\boldsymbol{\theta}}_{n-1}) - \rho\frac{\alpha_n}{\beta_n}\left\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \boldsymbol{\Delta}_n y_n\right\rangle + 2^{2-\rho}c_n\|\boldsymbol{\Delta}_n\|^{\rho}_{\rho}|y_n|^{\rho}. \qquad (+) \end{aligned}$$

Since the mean-value theorem also holds for the function $F(\cdot, \mathbf{w}_n)$, from the observation model we obtain for some $t' \in (0, 1)$

$$\begin{aligned} \boldsymbol{\Delta}_n y_n &= \boldsymbol{\Delta}_n\left(F(\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n\boldsymbol{\Delta}_n, \mathbf{w}_n) + v_n\right) \\ &= \boldsymbol{\Delta}_n F(\widehat{\boldsymbol{\theta}}_{n-1}, \mathbf{w}_n) + \boldsymbol{\Delta}_n v_n + \boldsymbol{\Delta}_n\left\langle\nabla_{\mathbf{x}}F(\widehat{\boldsymbol{\theta}}_{n-1} + t'\beta_n\boldsymbol{\Delta}_n, \mathbf{w}_n), \beta_n\boldsymbol{\Delta}_n\right\rangle. \end{aligned}$$

Applying the conditional expectation operation to the σ-algebra $\mathcal{F}_{n-1}$, since the test perturbation $\boldsymbol{\Delta}_n$ does not depend on noises v_n and vectors $\mathbf{w}_n$, we obtain

$$\begin{gathered} \mathrm{E}\{\boldsymbol{\Delta}_n v_n|\mathcal{F}_{n-1}\} = \mathrm{E}\{\boldsymbol{\Delta}_n|\mathcal{F}_{n-1}\}\mathrm{E}\{v_n|\mathcal{F}_{n-1}\} = 0, \\ \mathrm{E}\{\boldsymbol{\Delta}_n F(\widehat{\boldsymbol{\theta}}_{n-1}, \mathbf{w}_n)|\mathcal{F}_{n-1}\} = \mathrm{E}\{\boldsymbol{\Delta}_n|\mathcal{F}_{n-1}\}F(\widehat{\boldsymbol{\theta}}_{n-1}, \mathbf{w}_n) = 0. \end{gathered}$$

Consequently, the conditional expectation of the second term in formula (+) is

$$\begin{aligned} -\rho\frac{\alpha_n}{\beta_n}\mathrm{E}\left\{\left\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \boldsymbol{\Delta}_n y_n\right\rangle|\mathcal{F}_{n-1}\right\} &= -\rho\frac{\alpha_n}{\beta_n}\left\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \mathrm{E}\{\boldsymbol{\Delta}_n y_n|\mathcal{F}_{n-1}\}\right\rangle \\ = -\rho\frac{\alpha_n}{\beta_n}&\left\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \mathrm{E}\left\{\boldsymbol{\Delta}_n\left\langle\nabla_{\mathbf{x}}F(\widehat{\boldsymbol{\theta}}_{n-1} + t'\beta_n\boldsymbol{\Delta}_n, \mathbf{w}_n), \beta_n\boldsymbol{\Delta}_n\right\rangle|\mathcal{F}_{n-1}\right\}\right\rangle \\ &= -\rho\alpha_n\left\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \mathrm{E}\left\{\boldsymbol{\Delta}_n\left\langle\nabla_{\mathbf{x}}F(\widehat{\boldsymbol{\theta}}_{n-1}, \mathbf{w}_n), \boldsymbol{\Delta}_n\right\rangle|\mathcal{F}_{n-1}\right\}\right\rangle \\ +\rho\alpha_n\Big|\Big\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \mathrm{E}\Big\{&\boldsymbol{\Delta}_n\Big\langle\nabla_{\mathbf{x}}F(\widehat{\boldsymbol{\theta}}_{n-1} + t'\beta_n\boldsymbol{\Delta}_n, \mathbf{w}_n) - \nabla_{\mathbf{x}}F(\widehat{\boldsymbol{\theta}}_{n-1}, \mathbf{w}_n), \Delta_n\Big\rangle|\mathcal{F}_{n-1}\Big\}\Big\rangle\Big|. \end{aligned}$$

Since the function $\nabla_{\mathbf{x}}F(\cdot, \mathbf{w}_n)$ is uniformly bounded, using the Holder inequality [20], conditions (A.1) and (A.2), and the Young inequality [21] $a^{1/r}b^{1/s} \le \dfrac{1}{r}a + \dfrac{1}{s}b$, $r > 1$, $a, b > 0$, $\dfrac{1}{r} + \dfrac{1}{s} = 1$, we sequentially obtain

$$\begin{aligned} -\rho\frac{\alpha_n}{\beta_n}\mathrm{E}\left\{\left\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \boldsymbol{\Delta}_n y_n\right\rangle|\mathcal{F}_{n-1}\right\} &\le -\rho\alpha_n\left\langle\nabla V(\widehat{\boldsymbol{\theta}}_{n-1}), \nabla f(\widehat{\boldsymbol{\theta}}_{n-1})\right\rangle \\ +\rho\alpha_n V(\widehat{\boldsymbol{\theta}}_{n-1})^{\frac{\rho-1}{\rho}}q^{1/\rho}\Big|\mathrm{E}\Big\{\langle\nabla_{\mathbf{x}}F(\widehat{\boldsymbol{\theta}}_{n-1} &+ t'\beta_n\boldsymbol{\Delta}_n, \mathbf{w}_n) - \nabla_{\mathbf{x}}F(\widehat{\boldsymbol{\theta}}_{n-1}, \mathbf{w}_n), \boldsymbol{\Delta}_n\rangle|\mathcal{F}_{n-1}\Big\}\Big| \\ &\le -\rho\alpha_n\mu V(\widehat{\boldsymbol{\theta}}_{n-1}) + \alpha_n\rho V(\widehat{\boldsymbol{\theta}}_{n-1})^{\frac{\rho-1}{\rho}}q^{\frac{2}{\rho}}M\|t'\beta_n\boldsymbol{\Delta}_n\|_{\frac{\rho}{\rho-1}} \\ &\le -\alpha_n\rho\mu V(\widehat{\boldsymbol{\theta}}_{n-1}) + \alpha_n\rho\left(\frac{\rho-1}{\rho}V(\widehat{\boldsymbol{\theta}}_{n-1}) + \frac{1}{\rho}\right)q^{\frac{\rho+1}{\rho}}M\beta_n \\ &\le -\alpha_n\left(\rho\mu - \beta_n(\rho-1)q^{\frac{\rho+1}{\rho}}M\right)V(\widehat{\boldsymbol{\theta}}_{n-1}) + \alpha_n\beta_n q^{\frac{\rho+1}{\rho}}M. \end{aligned}$$

Let us evaluate the third term in the right side of inequality (+). For some point $\mathbf{x}_m$ on the interval joining $\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_n$ and $\boldsymbol{\theta}_*$, applying the mean-value theorem, Holder inequality, condition (A.2), and inequality $\left(\frac{a+b}{2}\right)^\rho \leq \frac{1}{2}(a^\rho + b^\rho)$, we obtain

$$\left|F(\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_n, \mathbf{w}_n)\right|^\rho = \left|F(\boldsymbol{\theta}_*, \mathbf{w}_n) + \langle \nabla_{\mathbf{x}} F(\mathbf{x}_m, \mathbf{w}_n), \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_n - \boldsymbol{\theta}_* \rangle\right|^\rho$$
$$\leq 2^{\rho-1} \max_{\mathbf{w}\in\mathbb{W}} |F(\boldsymbol{\theta}_*, \mathbf{w})^\rho| + 2^{\rho-1} \left(\|\nabla_{\mathbf{x}} F(\mathbf{x}_m, \mathbf{w}_n) - \nabla_{\mathbf{x}} F(\boldsymbol{\theta}_*, \mathbf{w}_n)\|_{\frac{\rho}{\rho-1}} + \max_{w\in\mathbb{W}} \|\nabla_{\mathbf{x}} F(\boldsymbol{\theta}_*, \mathbf{w})\|_{\frac{\rho}{\rho-1}} \right)^\rho$$
$$\times \left(\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}_*\|_\rho + \|\beta_n \boldsymbol{\Delta}_n\|_\rho \right)^\rho$$
$$\leq 2^{\rho-1} \max_{\mathbf{w}\in\mathbb{W}} |F(\boldsymbol{\theta}_*, \mathbf{w})|^\rho + 2^{3(\rho-1)} \left(M^\rho d_n^\rho + \max_{\mathbf{w}\in\mathbb{W}} \|\nabla_{\mathbf{x}} F(\boldsymbol{\theta}_*, \mathbf{w})\|^\rho_{\frac{\rho}{\rho-1}} \right) \left(V(\widehat{\boldsymbol{\theta}}_{n-1}) + q\beta_n^\rho \right).$$

Consequently, for the conditional expectation of the third term in (+), since Δ_n and v_n are independent, we obtain the estimate

$$2^{2-\rho} c_n \mathrm{E}\{\|\boldsymbol{\Delta}_n\|^\rho |y_n|^\rho | \mathcal{F}_{n-1}\} = 2^{2-\rho} c_n \mathrm{E}\{\|\boldsymbol{\Delta}_n\|^\rho |F(\mathbf{x}_n, \mathbf{w}_n) + v_n|^\rho | \mathcal{F}_{n-1}\}$$
$$\leq 2c_n \mathrm{E}\{\|\boldsymbol{\Delta}_n\|^\rho | \mathcal{F}_{n-1}\}(|F(\mathbf{x}_n, \mathbf{w}_n)|^\rho + \mathrm{E}\{|v_n|^\rho | \mathcal{F}_{n-1}\})$$
$$\leq q 2^{2\rho-1} c_n \psi_n V(\widehat{\boldsymbol{\theta}}_{n-1}) + 2^\rho q c_n \max_{\mathbf{w}\in\mathbb{W}} |F(\boldsymbol{\theta}_*, \mathbf{w})|^\rho + 2^{3\rho-2} q^2 \beta_n^\rho \psi_n + 2qc_n \mathrm{E}\{|v_n|^\rho | \mathcal{F}_{n-1}\}.$$

Using our notation and estimates obtained above, we can strengthen inequality (+) as

$$V(\widehat{\boldsymbol{\theta}}_n) \leq V(\widehat{\boldsymbol{\theta}}_{n-1})(1 - \gamma_n) + \phi_n + 2qc_n \mathrm{E}\{|v_n|^\rho | \mathcal{F}_{n-1}\}).$$

Taking the unconditional expectation of the left and right sides of the initial inequality, we obtain

$$\mathrm{E}\{V(\widehat{\boldsymbol{\theta}}_n)\} \leq \mathrm{E}\{V(\widehat{\boldsymbol{\theta}}_{n-1})\}(1 - \gamma_n) + \phi_n + 2qc_n \sigma_n^\rho.$$

If these inequalities and the conditions of Theorem 1 hold, then all assertions of this theorem follow directly from the corresponding assertions of [22]. This completes the proof of Theorem 1.

REFERENCES

1. Kiefer, J. and Wolfowitz, J., Statistical Estimation on the Maximum of a Regression Function, *Ann. Math. Stat.*, 1952, vol. 23, pp. 462–466.

2. Granichin, O.N., Stochastic Approximation with Input Perturbation under Dependent Observation Noises, *Vestn. Leningr. Gos. Univ.*, 1989, Ser. 1, no. 4, pp. 27–31.

3. Polyak, B.T. and Tsybakov, A.B., Optimal Accuracy Orders of Stochastic Approximation Algorithms, *Probl. Peredachi Inform.*, 1990, no. 2, pp. 45–53.

4. Polyak, B.T. and Tsybakov, A.V., On Stochastic Approximation with Arbitrary Noise (the KW Case), in *Topics in Nonparametric Estimation*, Khasminskii, R.Z., Ed., Providence: Am. Math. Soc., 1992, no. 12, pp. 107–113.

5. Spall, J.C., A One-Measurement Form of Simultaneous Perturbation Stochastic Approximation, *Automatica*, 1997, vol. 33, pp. 109–112.

6. Granichin, O.N., A Stochastic Approximation Procedure with Input Perturbation, *Avtom. Telemekh.*, 1992, no. 2, pp. 97–104.

7. Granichin, O.N., Estimation of the Maximum Point of an Unknown Function Observable on a Background of Dependent Noises, *Probl. Peredachi Inform.*, 1992, no. 2, pp. 16–20.

8. Chen, H.F., Duncan T.E., and Pasik-Duncan, B., A Kiefer–Wolfowitz Algorithm with Randomized Differences, *IEEE Trans. Automat. Control*, 1999, vol. 44, no. 3, pp. 442–453.

9. Granichin, O.N., Estimation of Linear Regression Parameters under Arbitrary Noises, *Avtom. Telemekh.*, 2002, no. 1, pp. 30–41.

10. Granichin, O.N., Randomized Stochastic Approximation Algorithms under Arbitrary Noises, *Avtom. Telemekh.*, 2002, no. 2, pp. 44–55.

11. Granichin, O.N., Optimal Convergence Rate of Randomized Stochastic Approximation Algorithms under Arbitrary Noises, *Avtom. Telemekh.*, 2003, no. 2, pp. 88–99.

12. Granichin, O.N. and Polyak, B.T., *Randomizirovannye algoritmy otsenivaniya i optimizatsii pri pochti proizvol'nykh pomekhakh* (Randomized Algorithms for Estimation and Optimization under Almost Arbitrary Noises), Moscow: Nauka, 2003.

13. Ljung, L. and Guo, L., The Role of Model Validation for Assessing the Size of the Unmodeled Dynamics, *IEEE Trans. Automat. Control*, 1997, vol. 42, no. 9, pp. 1230–1239.

14. Granichin, O.N., Nonminimax Filtration under Unknown Observation Noises, *Avtom. Telemekh.*, 2002, no. 9, pp. 125–133.

15. Granichin, O.N., Linear Regression and Filtering under Nonstandard Assumptions (Arbitrary noise), *IEEE Trans. Automat. Control*, 2004, vol. 49, no. 10, pp. 1830–1835.

16. Vandersypen, L., Steffen, M., Breyta, G., Yannoni, C.S., Sherwood, M.H., and Chuang, I.L., Experimental Realization of Shor's Quantum Factoring Algorithm using Nuclear Magnetic Resonance, *Nature*, 2001, vol. 414, pp. 883–887.

17. Faddeev, L.D. and Yakubovskii, O.A., *Lektsii po kvantovoi mekhanike dlya studentov matematikov* (Lectures on Quantum Mechanics for Students of Mathematics), Izhevsk: RKhD, 2001.

18. Shor, P.W., Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer, *SIAM J. Comput.*, 1997, vol. 26, pp. 1484–1509.

19. Kitaev, A., Shen', A., and Vyalyi, M., *Klassicheskie i kvantovye vychisleniya* (Classical and Quantum Calculus), Izhevsk: RKhD, 2004.

20. Korn, G.A. and Korn, T.M., *Mathematical Handbook for Scientists and Engineers*, New York: McGraw-Hill, 1968. Translated under the title *Spravochnik po matematike dlya nauchnykh rabotnikov i inzhenerov*, Moscow: Nauka, 1984.

21. Zorich, V.A., *Matematicheskii analiz* (Matematical Analysis), Moscow: MTsNMO, 2001.

22. Polyak, B.T., Convergence and Convergence Rates of Iterative Stochastic Algorithms. I, *Avtom. Telemekh.*, 1976, no. 12, pp. 83–94.

This paper was recommended for publication by B.T. Polyak, a member of the Editorial Board