Automation and Remote Control, Vol. 63, No. 2, 2002, pp. 209–219. Translated from Avtomatika i Telemekhanika, No. 2, 2002, pp. 44–55.

STOCHASTIC SYSTEMS

Randomized Algorithms for Stochastic Approximation under Arbitrary Disturbances

O. N. Granichin

St. Petersburg State University, St. Petersburg, Russia

Received March 20, 2001

Abstract—New algorithms for stochastic approximation under input disturbance are designed. For the multidimensional case, they are simple in form, generate consistent estimates for unknown parameters under "almost arbitrary" disturbances, and are easily "incorporated" in the design of quantum devices for estimating the gradient vector of a function of several variables.

1. INTRODUCTION

In the design of optimization and estimation algorithms, noises and errors in measurements and properties of models are usually attributed with some useful statistical characteristics, which are used in demonstrating the validity of the algorithm. For example, noise is often assumed to be centered. Algorithms based on the ordinary least-squares method are used in engineering practice for simple averaging of observation data. If noise is assumed to centered without any valid justification, then such algorithms are unsatisfactory in practice and may even be harmful. Such is the state of affairs under "opponent's" counteraction. In particular, if noise is defined by a deterministic unknown function (opponent suppresses signals) or measurement noise is a dependent sequence, then averaging of observations does not yield any useful result. Analysis of typical results in such situations shows that the estimates generated by ordinary algorithms are unsatisfactory, observation sequences are believed to be degenerated and such problems are usually disregarded.

Another close problem in applying many recurrent estimation algorithms is that observation sequences do not have adequate variability. For example, the main aim in designing adaptive controls is to minimize the deviation of the state vector of a system from a given trajectory and this often results in a degenerate observation sequence. Consequently, identification is complicated and successful identification requires "diverse" observations.

Our new approach to estimation and optimization under poor conditions (for example, degenerate observations) is based on the use of *test disturbances*. The information in the observation channel can be "enriched" for solving many problems by introducing a test disturbance with known statistical properties into the input channels or algorithm. Sometimes the measured random process can be used as the test disturbance. In control systems, test actions can be introduced via the control channel, whereas in other cases, a randomized observation plan (experiment) can be used as the test action. In studying a renewed system with test disturbance, which is sometimes simply the old system in a different form, results on convergence and applicability fields of new algorithms can be obtained through traditional estimation methods. One remarkable characteristic of such algorithms is their convergence under "almost arbitrary" disturbances. An important constraint on their applicability is that the test disturbance and inherent noise are assumed to be independent. This constraint in many problems is satisfied. Such is the case if noise is taken to be an unknown bounded deterministic function or some external random perturbation generated by some statistical properties of the test disturbance unknown to the opponent counteracting our investigation.

0005-1179/02/6302-0209$27.00

Surprisingly, researchers did not notice for long that in case of noisy observations, search algorithms with sequential ($n = 1, 2, \dots$) changes in the estimate $\widehat{\boldsymbol{\theta}}_{n-1}$ along some random centered vector $\boldsymbol{\Delta}_n$

$$\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\Delta}_n \overline{y}_n,$$

might converge to the true vector of controlled parameters $\boldsymbol{\theta}^*$ under not only "good," but also "almost arbitrary" disturbances. This happens if observations $\overline{y}_n$ are taken at some point defined by the previous estimate $\widehat{\boldsymbol{\theta}}_{n-1}$ and vector $\boldsymbol{\Delta}_n$, called the *simultaneous test disturbance*. Such an algorithm is called the *randomized estimation algorithm*, because its convergence under "almost arbitrary" noises is demonstrated through the stochastic (probabilistic) properties of the test disturbance. In the near future, experimenters will radically change their present cautious attitude to stochastic algorithms and their results. Modern computing devices will be supplanted by quantum computers, which operator as stochastic systems due to the Heisenberg principle of uncertainty. By virtue of the possibility of quantum parallelism, randomized-type estimation algorithms will, most probably, form the underlying principle of future quantum computing devices.

A main feature of the stochastic approximation algorithms designed in paper is that the unknown maximized function is measured in each step not at the point defined by the previous estimate, but at its slightly perturbed position. Precisely this "perturbation" plays the part of simultaneous test disturbance, "enriching" the information arriving at the observation channel. The convergence of randomized stochastic approximation algorithms under "almost arbitrary" disturbances was first investigated in [1–4]. In [5], such algorithms are studied, but under "good" disturbances in observations, and the asymptotic optimality of recurrent algorithms in minimax sense is demonstrated. A similar algorithm is studied in [6, 7], in which it is shown that the number of iterations required for attaining the necessary estimation accuracy is not attained, though the number of measurements in the multidimensional case at each step is considerably less compared to the classical Kiefer–Wolfowitz procedure. In foreign literature, the new algorithm is called the simultaneous perturbation stochastic approximation. Similar algorithms are investigated in designing adjustment algorithms for neural networks [8, 9]. The consistency of recurrent algorithms under "almost arbitrary" perturbations are also investigated in [10, 11].

In this paper, we study a more general formulation of the problem of minimization of mean-risk functional with measurements of the cost function under "almost arbitrary" perturbations. We shall demonstrate that sequences of estimates generated by different algorithms converge to the true value of unknown parameters with probability 1 in the mean-square sense.

2. FORMULATION OF THE PROBLEM AND MAIN ASSUMPTIONS

Let $F(\mathbf{w}, \boldsymbol{\theta}) : \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^1$ be a θ-differentiable function and let $\mathbf{x}_1, \mathbf{x}_2 \dots$, be a sequence of measurement points chosen by the experimenter (observation plan), at which the value

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

of the function $F(w_n, \cdot)$ is accessible to observation at every instant $n = 1, 2, \dots$, with additive disturbances v_n, where $\{\mathbf{w}_n\}$ is a noncontrollable sequence of random variables ($\mathbf{w}_n \in \mathbb{R}^p$) having, in general, identical unknown distribution $\mathrm{P}_w(\cdot)$ with a finite carrier.

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

$$f(\boldsymbol{\theta}) = \int\limits_{\mathbb{R}^p} F(\mathbf{w}, \boldsymbol{\theta}) \mathrm{P}_w(dw)$$

of the type of mean-risk functional.

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

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

which easily matches within the general scheme. The generalization in the formulation is stipulated by the need to take account of multiplicative perturbations in observations:

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

which is contained in the general scheme with the function $F(w, \mathbf{x}) = wf(\mathbf{x})$, and the need to generalize the observation model [12], where

$$F(\mathbf{w}, \mathbf{x}) = \frac{1}{2}(\mathbf{x} - \boldsymbol{\theta}^* - \mathbf{w})^{\mathrm{T}}(\mathbf{x} - \boldsymbol{\theta}^* - \mathbf{w}).$$

Let us formulate the main assumptions, denoting the Euclidean norm and scalar multiplication in $\mathbb{R}^r$ by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$, respectively.

SA.1 The function $f(\cdot)$ is strongly convex, i.e., has a unique minimum in $\mathbb{R}^r$ at some point $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(f(\cdot))$ and

$$(\mathbf{x} - \boldsymbol{\theta}^*, \nabla f(\mathbf{x})) \geq \mu\|\mathbf{x} - \boldsymbol{\theta}^*\|^2 \ \forall \mathbf{x} \in \mathbb{R}^r$$

with some constant $\mu > 0$.

SA.2 The gradient of the function $f(\cdot)$ satisfies the Lipschitz condition

$$\|\nabla f(\mathbf{x}) - \nabla f(\boldsymbol{\theta})\| \leq A\|\mathbf{x} - \boldsymbol{\theta}\| \ \forall \mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^r$$

with some constant $A > \mu$.

3. TEST PERTURBATION AND MAIN ALGORITHMS

Let Δ_n, $n = 1, 2, \ldots$, be an observed sequence of independent random variables in $\mathbb{R}^r$, called the *simultaneous test perturbation*, with distribution function $\mathrm{P}_n(\cdot)$.

Let us take a fixed initial vector $\widehat{\theta}_0 \in \mathbb{R}^r$ and choose two sequences $\{\alpha_n\}$ and $\{\beta_n\}$ of positive numbers tending to zero. We design three algorithms for constructing sequences of measurement points $\{x_n\}$ and estimates $\{\theta_n\}$. The first algorithm uses at every step (iteration) one observation

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

whereas the second and third algorithms, which are randomized variants of the Kiefer–Wolfowitz procedure, use two observations

$$\begin{cases} \mathbf{x}_{2n} = \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_n,\ \mathbf{x}_{2n-1} = \widehat{\boldsymbol{\theta}}_{n-1} - \beta_n \boldsymbol{\Delta}_n \\ \widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_{n-1} - \dfrac{\alpha_n}{2\beta_n}\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)(y_{2n} - y_{2n-1}), \end{cases} \tag{2}$$

$$\begin{cases} \mathbf{x}_{2n} = \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_n, \quad \mathbf{x}_{2n-1} = \widehat{\boldsymbol{\theta}}_{n-1} \\ \widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_{n-1} - \dfrac{\alpha_n}{\beta_n}\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)(y_{2n} - y_{2n-1}). \end{cases} \tag{3}$$

All these three algorithms use some vector functions (kernels) $\boldsymbol{\mathcal{K}}_n(\cdot) : \mathbb{R}^r \to \mathbb{R}^r$, $n = 1, 2, \ldots$, with a compact carrier, which, along with distribution functions of the test perturbation $\mathrm{P}_n(\cdot)$, satisfy the conditions

$$\int \boldsymbol{\mathcal{K}}_n(x)\mathrm{P}_n(dx) = 0, \quad \int \boldsymbol{\mathcal{K}}_n(x)x^{\mathrm{T}}\mathrm{P}_n(dx) = \mathrm{I},$$
$$\sup_n \int \|\boldsymbol{\mathcal{K}}_n(x)\|^2 \mathrm{P}_n(dx) < \infty, \quad n = 1, 2, \ldots, \tag{4}$$

where I is an r-dimensional unit matrix.

Algorithm (2) was first developed by Kushner and Clark [13] for a uniform distribution and function $\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n) = \boldsymbol{\Delta}_n$. Algorithm (1) was first designed in [1] for constructing a sequence of consistent estimates under "almost arbitrary" perturbations in observations, using the same simple kernel function, but for a more general class of test perturbations. Polyak and Tsybakov [5] investigated algorithms (1) and (2) with a vector function $\boldsymbol{\mathcal{K}}_n(\cdot)$ of sufficiently general type for uniformly distributed test perturbation under the assumption that the observation perturbations are independent and centered. Spall [6, 7] studied algorithm (2) for a test perturbation distribution with finite inverse moments and vector function $\boldsymbol{\mathcal{K}}_n(\cdot)$ defined by the rule

$$\boldsymbol{\mathcal{K}}_n\left(\left(x^{(1)}, x^{(2)}, \ldots, x^{(r)}\right)^{\mathrm{T}}\right) = \left(\frac{1}{x^{(1)}}, \frac{1}{x^{(2)}}, \ldots, \frac{1}{x^{(r)}}\right)^{\mathrm{T}}.$$

Using this vector $\boldsymbol{\mathcal{K}}_n(\cdot)$ and constraints on the distribution of the simultaneous test perturbation, Chen et al. [10] studied algorithm (3).

4. CONVERGENCE WITH PROBABILITY 1 AND IN THE MEAN-SQUARE SENSE

Instead of algorithm (1), let us study a close algorithm with projection

$$\begin{cases} \mathbf{x}_n = \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n \boldsymbol{\Delta}_n,\ y_n = F(\mathbf{w}_n, \mathbf{x}_n) + v_n \\ \widehat{\boldsymbol{\theta}}_n = \mathcal{P}_{\Theta_n}\left(\widehat{\boldsymbol{\theta}}_{n-1} - \dfrac{\alpha_n}{\beta_n}\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)y_n\right). \end{cases} \tag{5}$$

For this algorithm, it is more convenient to prove the consistency of the estimate sequence. In this algorithm, $\mathcal{P}_{\Theta_n}$, $n = 1, 2, \ldots$, are operators of projection onto some convex closed bounded subsets $\Theta_n \subset \mathbb{R}^r$ containing the point $\boldsymbol{\theta}^*$, beginning from some $n \geq 1$. If the set Θ containing the point $\boldsymbol{\theta}^*$ is known in advance, then we can take $\Theta_n = \Theta$. Otherwise, the sets $\{\Theta_n\}$ may extend to infinity.

Let $\mathbb{W} = \mathrm{supp}(\mathrm{P}_w(\cdot)) \subset \mathbb{R}^p$ be the carrier of the distribution $\mathrm{P}_w(\cdot)$ and let $\mathcal{F}_{n-1}$ be the σ-algebra of probabilistic events generated by the random variables $\widehat{\boldsymbol{\theta}}_0, \widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_{n-1}$ formed by algorithm (5) (or (2), or (3)). In applying algorithm (2) or (3), we have

$$\overline{v}_n = v_{2n} - v_{2n-1}, \quad \overline{\mathbf{w}}_n = \begin{pmatrix} \mathbf{w}_{2n} \\ \mathbf{w}_{2n-1} \end{pmatrix}, \quad d_n = 1,$$

and in constructing estimates by algorithm (5), we have

$$\overline{v}_n = v_n, \quad \overline{\mathbf{w}}_n = \mathbf{w}_n, \quad d_n = \mathrm{diam}(\Theta_n),$$

where $\mathrm{diam}(\cdot)$ is the Euclidean diameter of the set.

Theorem 1. *Let the following conditions hold:*
(SA.1) *for the function* $f(\boldsymbol{\theta}) = \mathrm{E}\{F(\mathbf{w}, \boldsymbol{\theta})\}$,

(SA.2) *for the function* $F(\mathbf{w},\cdot)$ $\forall \mathbf{w}\in\mathbb{W}$,
(4) *for the functions* $\boldsymbol{\mathcal{K}}_n(\cdot)$ *and* $\mathrm{P}_n(\cdot)$, $n=1,2,\ldots,$
$\forall \boldsymbol{\theta}\in\mathbb{R}^r$ *the functions* $F(\cdot,\boldsymbol{\theta})$ *and* $\nabla_\theta F(\cdot,\boldsymbol{\theta})$ *are uniformly bounded on* $\mathbb{W}$,
$\forall n\geq 1$ *the random variables* $\overline{v}_1,\ldots,\overline{v}_n$ *and vectors* $\overline{\mathbf{w}}_1,\ldots,\overline{\mathbf{w}}_{n-1}$ *do not depend on* $\overline{\mathbf{w}}_n$ *and* $\boldsymbol{\Delta}_n$, *and the random vector* $\overline{w}_n$ *does not depend on* $\boldsymbol{\Delta}_n$, *and*
$\mathrm{E}\{\overline{v}_n^2\}\leq\sigma_n^2$, $n=1,2,\ldots$.

If $\sum\limits_n \alpha_n=\infty$, $\alpha_n\to 0$, $\beta_n\to 0$, *and* $\alpha_n^2\beta_n^{-2}(1+d_n^2+\sigma_n^2)\to 0$ *as* $n\to\infty$, *then the sequence of estimates* $\{\widehat{\boldsymbol{\theta}}_n\}$ *generated by algorithm* (5) (*or* (2), *or* (3)) *converges to the point* $\boldsymbol{\theta}^*$ *in the mean-square sense:* $\mathrm{E}\{\|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}^*\|^2\}\to 0$ *as* $n\to\infty$.

Moreover, if $\sum\limits_n \alpha_n\beta_n^2<\infty$ *and*

$$\sum_n \alpha_n^2\beta_n^{-2}\left(1+\mathrm{E}\{\overline{v}_n^2\,|\,\mathcal{F}_{n-1}\}\right)<\infty$$

with probability 1, then $\widehat{\boldsymbol{\theta}}_n\to\boldsymbol{\theta}^*$ *as* $n\to\infty$ *with probability 1.*

The proof of Theorem 1 is given the Appendix.

The conditions of Theorem 1 hold for the function $F(\mathbf{w},\mathbf{x})=\mathbf{w}f(\mathbf{x})$ if the function $f(\mathbf{x})$ satisfies conditions (SA.1) and (SA.2).

The observation perturbations v_n in Theorem 1 can be said to be "almost arbitrary" since they may also be nonrandom, but independent and bounded or the realization of some stochastic process with arbitrary dependencies. In particular, to prove the assertions of Theorem 1, there is no need to assume that $\overline{v}_n$ and $\mathcal{F}_{n-1}$ are dependent.

The assertions of Theorem 1 do not hold if the perturbations $\overline{v}_n$ depend on the observation point $\overline{v}_n=\overline{v}_n(\mathbf{x}_n)$ in a specific manner. If such a dependence holds, for example, due to rounding errors, then the observation perturbation must be subdivided into two parts $\overline{v}_n(\mathbf{x}_n)=\widetilde{v}_n+\xi(\mathbf{x}_n)$ in solving such a problem. The first part may satisfy the conditions of Theorem 1. If the second part results from rounding errors, then it, as a rule, has good statistical properties and may not prevent the convergence of algorithms.

The condition that the observation be independent of the test perturbation can be slackened. It suffices to assume that the conditional mutual correlation between $\overline{v}_n$ and $\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)$ tend to zero with probability 1 at a rate not less than $\alpha_n\beta_n^{-1}$ as $n\to\infty$.

Though algorithms (2) and (3) may look alike, algorithm (3) is more suitable for use in real-time systems if observations contain arbitrary perturbations. For algorithm (2), the condition that observation perturbation v_{2n} and the test perturbation $\boldsymbol{\Delta}_n$ be independent is rather restrictive, because the vector $\boldsymbol{\Delta}_n$ is used at the previous instant $(2n-1)$ in the system. In the operation of algorithm (3), perturbations v_{2n} and the test perturbation vector $\boldsymbol{\Delta}_n$ simultaneously appear in the system. Hence they can be regarded as independent.

For ℓ-times continuously differentiable functions $F(w,\cdot)$, the vector function $\boldsymbol{\mathcal{K}}_n(\cdot)$ is constructed in [5] with the help of Lagrange orthogonal polynomials $p_m(\cdot)$, $m=0,1,\ldots,\ell$, using a uniform test perturbation on an r-dimensional cube $[-1/2,1/2]^r$ as the probabilistic distribution. Moreover, it is shown that the mean-square convergence rate asymptotically behaves as $\mathcal{O}\left(n^{\frac{\ell}{\ell+1}}\right)$. A similar result can also be obtained for the general test perturbation distribution studied in this paper.

5. QUANTUM COMPUTER AND GRADIENT VECTOR ESTIMATION

Owing to the simplicity of representation, randomized stochastic approximation algorithms can be used not only in programmable computing devices, but also permit the incorporation of the

simultaneous perturbation principle into the design of classical-type electronic devices. A fundamental point in demonstrating the convergence of any method in practical application is that the observation perturbation must be independent of the generated test perturbation, and the components of the test perturbation vector must be mutually independent. In realizing the algorithm on a conventional computer implementing elementary operations in succession, the designed algorithms are not as effective as theoretical predictions. The very name "simultaneously perturbed" implies the imperative requirement for practical realization, viz., capable of being implemented in *parallel*. Nevertheless, if the vector arguments of a function are large dimensional (hundreds or thousands), it is not a simple matter to organize parallel computations on conventional computers.

We consider a model of a "hypothetical" quantum computer, which obviously generates not only a simultaneous test perturbation, but also realizes a fundamental nontrivial concept—*measurement of quantum computation result*—since the concept itself underlies the design of the randomized algorithm in the form of product of the computed value of a function and test perturbation. In other words, we give an example of the design of a quantum device for estimating the gradient vector of an unknown function of several variables in *one cycle*. The epithet "hypothetical" is purposefully put within quotes, because quantum computer is generally believed to be a machine of the near future after Shor has read his report at the Berlin Mathematical Congress in 1998 [14].

As in [14], let us describe the mathematical model of a quantum computer that computes by a specific circuit. We use, whenever possible, the generalizations of concepts describing the model of a conventional computer. A conventional computer handles bits, which take value from the set $\{0, 1\}$. It is based on a finite ensemble of circuits that can be applied to a set of bits. A quantum computer handles q-bits (quantum bits)—a quantum system of two states (a microscopic system describing, for example, an excited ion or polarized photon, or the spin of atomic kernel). The behavioral characteristics of this quantum system, such as interference, superposition, stochasticity, etc., can be described exactly only through the rules of quantum mechanics [15]. Mathematically, a q-bit takes its values from a complex projective (Hilbert) space $\mathbb{C}^2$. Quantum states are invariant to multiplication by a scalar. Let $|0\rangle$, $|1\rangle$ be the base of this space. We assume that a quantum computer, like a conventional computer, is equipped with a discrete set of basic components, called the quantum circuits. Every quantum circuit is essentially a unitary transformation, which acts on a fixed number of q-bits. One of the fundamental principles of quantum mechanics asserts that the joint quantum state space of a system consisting of ℓ two-state systems is the tensor product of their component state spaces. Thus, the quantum state space of a system of ℓ q-bits is the projective Hilbert space $\mathbb{C}^{2^l}$. The set of base vectors of this state space can be parametrized by bit rows of length ℓ: $|b_1 b_2 \dots b_l\rangle$. Let us assume that classical data, a row i of length k, $k \le \ell$, is fed to the input of a quantum computer. In quantum computation, ℓ q-bits initially exist in the state $|i00\dots0\rangle$. An executable circuit is constructed from a finite number of quantum circuits acting on these q-bits. At the end of computation, the quantum computer passes to some state—a unit vector in the space $\mathbb{C}^{2^l}$. This state can be represented as

$$W = \sum_s \psi^{(s)} \,|\, s\rangle,$$

where summation with respect to s is taken over all binary rows of length ℓ, $\psi^{(s)} \in \mathbb{C}$, $\sum\limits_s |\psi^{(s)}|^2 = 1$. Here $\psi^{(s)}$ is called the probabilistic amplitude and W, the superposition of the base vectors $|s\rangle$. The Heisenberg uncertainty principle asserts that the state of a quantum system cannot be predicted exactly. Nevertheless, there are several possibilities of measuring all q-bits (or a subset of q-bits). The state space of our quantum system is Hilbertian and the concept of state measurement is equivalent to the scalar product in this Hilbert space with some given vector $\mathbf{V}$:

$$\langle \mathbf{V}, \mathbf{W} \rangle.$$

The projection of each of q-bits on the base $\{|0\rangle, |1\rangle\}$ is usually used in measurement. The result of this measurement is the computation result.

Let us examine the segment of the randomized stochastic approximation algorithm (3) (or (1), or (2)), that is used in computing the approximate value of the gradient vector $\widehat{\mathbf{g}}(\mathbf{x})$ function $f(\cdot) : \mathbb{R}^r \to \mathbb{R}$ at some point $\mathbf{x} \in \mathbb{R}^r$. Let the numbers in our quantum computer be expressed through p binary digits and $\ell = p \times r$ (in modern computers, $p = 16, 32, 64$). In the algorithm, let $\boldsymbol{\mathcal{K}}_n(\mathbf{x}) = \mathbf{x}$, $\beta = 2^{-j}$, $0 \le j < p$, and let the simultaneous test perturbation fed to the input of the algorithm be defined by a Bernoulli distribution. For any number $u \in \mathbb{R}$, let s_u denote its binary representation in the form of a bit array $s_u = \overline{b_u^{(1)} \dots b_u^{(p)}}$. Assume that there exists a quantum circuit that computes the value of the function $f(\mathbf{x})$. More exactly, there exists a unitary transformation $U_f : \mathbb{C}^{2^l} \to \mathbb{C}^{2^l}$ that for any $\mathbf{x} = \left(x^{(1)}, \dots, x^{(r)}\right)^{\mathrm{T}} \in \mathbb{R}^r$, associates the base element

$$|s_{x^{(1)}} \dots s_{x^{(r)}}\rangle = \left|b^{(1)}_{x^{(1)}} \dots b^{(p)}_{x^{(1)}} \dots b^{(1)}_{x^{(r)}} \dots b^{(p)}_{x^{(r)}}\right\rangle$$

with another base element

$$\left|s_{f(\mathbf{x})}00\dots0\right\rangle = \left|b^{(1)}_{f(\mathbf{x})} \dots b^{(p)}_{f(\mathbf{x})}00\dots0\right\rangle = U_f|s_{x^{(1)}} \dots s_{x^{(r)}}\Big\rangle.$$

Let the hypothetical quantum computer contain, at least, four ℓ q-bit registers: an input register $\mathcal{I}$ for transforming conventional data into quantum data, two working registers W_1 and W_2 for manipulating quantum data, and an output register Δ for storing the quantum value, projection onto which is the computation result in conventional form (measurement). The design of a quantum circuit for estimating the gradient vector of the function $f(\cdot)$ requires a few standard quantum (unitary) transformations that are applicable to register data. We assume that the transformation result is stored in the same register to which the transformation is applied.

Addition/Subtraction $U_{\pm X}$ of a vector with (from) another vector stored in the register X.

Rotation of the first q-bits $U_{R_{1,p+1,\dots,(r-1)p+1}}$ converts the state of q-bits $1, p+1, \dots, (r-1)p+1$ into $\frac{1}{\sqrt{2}}|0\rangle + \frac{1}{\sqrt{2}}|1\rangle$.

Shift by j q-bits U_{S_j} displaces the state vector by j q-bits, adding new bits $|0\rangle$.

The gradient vector of a function $f(\cdot)$ at a point x can be estimated by the following algorithm.

1. Zero the states of all registers $\mathcal{I}, W_1, W_2, \Delta$.
2. Feed a bit array s_x to the input $\mathcal{I}$ and transform the register Δ:

$$\Delta := U_{R_{1,p+1,\dots,(r-1)p+1}}\Delta.$$

3. Compute the value of the functions $f(\mathbf{x})$ and $f(\mathbf{x} + 2^{-j}\Delta)$

$$W_1 := U_f U_{+\mathcal{I}} W_1,$$
$$W_2 := U_f U_{+\mathcal{I}} U_{S_j} U_{+\Delta} W_2.$$

4. Compute the difference $W_2 := U_{-W_1} W_2$.
5. Sequentially $(i = 1, 2, \dots, r)$ measure the components of result

$$\widehat{g}^{\,(i)}(\mathbf{x}) = \langle \Delta, W_2 \rangle,\ W_2 := U_{S_p} W_2.$$

Computation result is determined by the formula for estimating the gradient vector of $f(\cdot)$:

$$\nabla f(\mathbf{x}) \approx \Delta \frac{f(\mathbf{x} + \beta\boldsymbol{\Delta}) - f(\mathbf{x})}{\beta}.$$

6. CONCLUSIONS

Unlike deterministic methods, stochastic optimization considerably widens the range of practical problems for which an exact optimal solution can be found. Stochastic optimization algorithms are effective for information network analysis, optimization via modeling, image processing, pattern recognition, neural network learning, and adaptive control. The role of stochastic optimization is expected to grow with the complication of modern systems in the same way as population growth and depletion of natural resources stimulate the use of more intensive technologies in those areas where they were not required earlier. The trend of modern development of computation techniques heralds that traditional deterministic algorithms will be supplanted by stochastic algorithms, because there already exist pilot quantum computers based on stochastic principles.

APPENDIX

Proof of Theorem 1. Let us consider algorithm (5). Theorem 1 for algorithms (2) and (3) are proved almost along similar lines and can be demonstrated by analogy. For sufficiently large n under which $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(f(\cdot)) \in \Theta_n$, using projection properties we easily obtain

$$\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right\|^2 \le \left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^* - \frac{\alpha_n}{\beta_n}\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)y_n\right\|^2.$$

Now applying the operation of conditional expectation for the σ-algebra $\mathcal{F}_{n-1}$, we obtain

$$\begin{aligned}\mathrm{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right\|^2 |\mathcal{F}_{n-1}\right\} \le \left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|^2 - 2\frac{\alpha_n}{\beta_n}\left\langle \widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*, \mathrm{E}\{\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)y_n \,|\, \mathcal{F}_{n-1}\}\right\rangle \\ +\frac{\alpha_n^2}{\beta_n^2}\mathrm{E}\left\{y_n^2 \left\|\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)\right\|^2 |\mathcal{F}_{n-1}\right\}. \qquad (6)\end{aligned}$$

By virtue of the mean value theorem, condition (SA.2) for the function $F(\cdot,\cdot)$ yields

$$|F(\mathbf{w},\mathbf{x}) - F(\mathbf{w},\boldsymbol{\theta}^*)| \le \frac{1}{2}\nabla_\theta F(\mathbf{w},\boldsymbol{\theta}^*)^2 + \left(A + \frac{1}{2}\right)\|\mathbf{x} - \boldsymbol{\theta}^*\|^2, \quad \mathbf{x} \in \mathbb{R}^r.$$

Since the function $F(\cdot,\boldsymbol{\theta})$ is uniformly bounded, using the notation

$$\nu_1 = \sup_{\mathbf{w}\in\mathbb{W}}\left(|F(\mathbf{w},\boldsymbol{\theta}^*)| + \frac{1}{2}\nabla_\theta F(\mathbf{w},\boldsymbol{\theta}^*)^2\right),$$

we find that

$$F(\mathbf{w},\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n\mathbf{x})^2 \le (\nu_1 + (2A+1)\left(\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\|^2 + \|\beta_n\mathbf{x}\|^2)\right)^2$$

uniformly for $w \in \mathbb{W}$. By condition (4), we have

$$\mathrm{E}\left\{v_n^2\|\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)\|^2|\mathcal{F}_{n-1}\right\} \le \sup_{\mathbf{x}}\boldsymbol{\mathcal{K}}_n(\mathbf{x})^2\xi_n^2.$$

Since the vector functions $\boldsymbol{\mathcal{K}}_n(\cdot)$ are bounded and their carriers are compact, the last two inequalities for the last term in the right side of (6) yield

$$\begin{aligned}\mathrm{E}\left\{\|y_n\|^2\|\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)\|^2|\mathcal{F}_{n-1}\right\} \le 2\mathrm{E}\left\{v_n^2\|\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)\|^2|\mathcal{F}_{n-1}\right\} \\ +2\iint F\left(\mathbf{w},\widehat{\boldsymbol{\theta}}_{n-1} + \beta_n\mathbf{x}\right)^2\|\boldsymbol{\mathcal{K}}_n(\mathbf{x})\|^2\mathrm{P}_n(d\mathbf{x})\mathrm{P}_w(d\mathbf{w}) \\ \le C_1 + C_2\left((d_n^2+1)\left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|^2 + \beta_n^2\right) + C_3\xi_n^2.\end{aligned}$$

Here and in what follows, $C_i,\ i = 1, 2, \ldots,$ are positive constants.

Now let us consider

$$\beta_n^{-1}\mathrm{E}\left\{y_n\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)|\mathcal{F}_{n-1}\right\} = \beta_n^{-1}\iint F\left(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n\mathbf{x}\right)\boldsymbol{\mathcal{K}}_n(\mathbf{x})\mathrm{P}_n(d\mathbf{x})\mathrm{P}_w(d\mathbf{w}) + \beta_n^{-1}\mathrm{E}\{v_n\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)|\mathcal{F}_{n-1}\}. \tag{7}$$

For the second term in the last relation, since v_n and Δ_n are independent, from (4) we obtain

$$\mathrm{E}\{v_n\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)\,|\,\mathcal{F}_{n-1}\} = \mathrm{E}\{v_n\,|\,\mathcal{F}_{n-1}\}\int\boldsymbol{\mathcal{K}}_n(\mathbf{x})\mathrm{P}_n(d\mathbf{x}) = 0.$$

Since the function $\nabla_\theta F(\cdot, \boldsymbol{\theta})$ is uniformly bounded, we have

$$\int\limits_{\mathbb{R}^p}\nabla_\theta F(\mathbf{w}, \mathbf{x})\mathrm{P}_w(d\mathbf{w}) = \nabla f(\mathbf{x}).$$

Using the last relation and condition (4), let us express the first term in (7) as

$$\beta_n^{-1}\iint F(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n\mathbf{x})\boldsymbol{\mathcal{K}}_n(\mathbf{x})\mathrm{P}_n(d\mathbf{x})\mathrm{P}_w(d\mathbf{w}) = \nabla f(\widehat{\boldsymbol{\theta}}_{n-1}) + \int\left(\beta_n^{-1}\int F(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1} + \beta_n\mathbf{x})\boldsymbol{\mathcal{K}}_n(\mathbf{x})\mathrm{P}_n(d\mathbf{x}) - \nabla_\theta F(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1})\right)\mathrm{P}_w(d\mathbf{w})$$

$$= \nabla f(\widehat{\boldsymbol{\theta}}_{n-1}) + \iint\boldsymbol{\mathcal{K}}_n(\mathbf{x})\mathbf{x}^{\mathrm{T}}\int\limits_0^1(\nabla_\theta F(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1} + t\beta_n\mathbf{x}) - \nabla_\theta F(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1}))dt\mathrm{P}_n(d\mathbf{x})\mathrm{P}_w(d\mathbf{w}).$$

Since conditions (SA.2) hold for any function $F(w, \cdot)$ and condition (4) holds for $\boldsymbol{\mathcal{K}}_n(\cdot)$, for the absolute value of the second term in the last equality we have

$$\left|\iint\boldsymbol{\mathcal{K}}_n(\mathbf{x})\mathbf{x}^{\mathrm{T}}\int\limits_0^1\left(\nabla_\theta F(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1} + t\beta_n\mathbf{x}) - \nabla_\theta F(\mathbf{w}, \widehat{\boldsymbol{\theta}}_{n-1})\right)dt\mathrm{P}_n(d\mathbf{x})\mathrm{P}_w(d\mathbf{w})\right| \le \iint\|\boldsymbol{\mathcal{K}}_n(\mathbf{x})\|\,\|\mathbf{x}\|A\|\beta_n\mathbf{x}\|\mathrm{P}_n(d\mathbf{x})\mathrm{P}_w(d\mathbf{w}) \le C_4\beta_n.$$

Consequently, for the second term in the right side of inequality (6) we obtain

$$-2\frac{\alpha_n}{\beta_n}\left\langle\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*, \mathrm{E}\{\boldsymbol{\mathcal{K}}_n(\boldsymbol{\Delta}_n)y_n\,|\,\mathcal{F}_{n-1}\}\right\rangle \le -2\alpha_n\left\langle\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*, \nabla f(\widehat{\boldsymbol{\theta}}_{n-1})\right\rangle + 2C_4\alpha_n\beta_n\left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|.$$

Substituting the expressions for the second and third terms in the right side into (6), we find that

$$\mathrm{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right\|^2\,|\,\mathcal{F}_{n-1}\right\} \le \left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|^2 - 2\alpha_n\left\langle\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*, \nabla f(\widehat{\boldsymbol{\theta}}_{n-1})\right\rangle + 2C_4\alpha_n\beta_n\left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\| + \frac{\alpha_n^2}{\beta_n^2}\left(C_1 + C_2\left((d_n^2+1)\left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|^2 + \beta_n^2\right) + C_3\xi_n^2\right).$$

Since condition (SA.1) holds for the function $f(\cdot)$ and the inequality

$$\left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\| \le \frac{\varepsilon^{-1}\beta_n + \varepsilon\beta_n^{-1}\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\|^2}{2}$$

holds for any $\varepsilon > 0$, we obtain

$$\mathrm{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right\|^2 |\mathcal{F}_{n-1}\right\} \le \left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|^2 \left(1 - (2\mu - \varepsilon C_4)\alpha_n + C_2\alpha_n^2\beta_n^{-2}(d_n^2 + 1)\right)$$
$$+\varepsilon^{-1}C_4\alpha_n\beta_n^2 + \frac{\alpha_n^2}{\beta_n^2}\left(C_1 + C_2\beta_n^2 + C_3\xi_n^2\right).$$

Choosing a small ε such that $\varepsilon C_4 < \mu$, a sufficiently large n, and applying the conditions of Theorem 1 for number sequences, let us transform the last inequality to the form

$$\mathrm{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right\|^2 |\mathcal{F}_{n-1}\right\} \le \left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|^2 (1 - C_5\alpha_n) + C_6\left(\alpha_n\beta_n^2 + \alpha_n^2\beta_n^{-2}(1 + \xi_n^2)\right).$$

Hence, by the conditions of Theorem 1, all conditions of the Robbins–Siegmund Lemma [16] that are necessary for the convergence of $\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}^*$ as $n \to \infty$ with probability 1 are satisfied. To prove the conditions of Theorem 1 for mean-square convergence, let us examine the unconditional mathematical expectation of both sides of the last inequality

$$\mathrm{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right\|^2\right\} \le \mathrm{E}\left\{\left\|\widehat{\boldsymbol{\theta}}_{n-1} - \boldsymbol{\theta}^*\right\|^2\right\}(1 - C_5\alpha_n) + C_6\left(\alpha_n\beta_n^2 + \alpha_n^2\beta_n^{-2}(1 + \sigma_n^2)\right).$$

The mean-square convergence of the sequence $\{\widehat{\boldsymbol{\theta}}_n\}$ to the point $\boldsymbol{\theta}^*$ is implied by Lemma 5 of [17].

This completes the proof of Theorem 1.

REFERENCES

1. Granichin, O.N., Stochastic Approximation with Input Perturbation for Dependent Observation Disturbances, *Vestn. Leningrad. Gos Univ.*, 1989, Ser. 1, no. 4, pp. 27–31.
2. Granichin, O.N., A Procedure of Stochastic Approximation with Input Perturbation, *Avtom. Telemekh.*, 1992, no. 2, pp. 97–104.
3. Granichin, O.N., Estimation of the Minimum Point for an Unknown Function with Dependent Background Noise, *Probl. Peredachi Inf.*, 1992, no. 2, pp. 16–20.
4. Polyak, B.T. and Tsybakov, A.B., On Stochastic Approximation with Arbitrary Noise (The KW Case), in *Topics in Nonparametric Estimation. Adv. in Soviet Mathematics. Am. Math. Soc.,* Khasminskii, R.Z., Ed., 1992, no. 12, pp. 107–113.
5. Polyak, B.T. and Tsybakov, A.B., Optimal Accuracy Orders of Stochastic Approximation Search Algorithms, *Probl. Peredachi Inf.*, 1990, no. 2, pp. 45–53.
6. Spall, J.C., A Stochastic Approximation Technique for Generating Maximum Likelihood Parameter Estimates, *Am. Control Conf.,* 1987, pp. 1161–1167.
7. Spall, J.C., Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, *IEEE Trans. Autom. Control*, 1992, vol. 37, pp. 332–341.
8. Alspector, J., Meir, R., Jayakumar, A., *et al.*, A Parallel Gradient Descent Method for Learning in Analog VLSI Neural Networks, in *Advances in Neural Information Processing Systems*, Hanson, S.J., Ed., San Mateo: Morgan Kaufmann, 1993, pp. 834–844.
9. Maeda, Y. and Kanata, Y., Learning Rules for Recurrent Neural Networks Using Perturbation and Their Application to Neuro-control, *Trans. IEE Japan*, 1993, vol. 113-C, pp. 402–408.
10. Chen, H.F., Duncan, T.E., and Pasik-Duncan, B., A Kiefer–Wolfowitz Algorithm with Randomized Differences, *IEEE Trans. Autom. Control*, 1999, vol. 44, no. 3, pp. 442–453.
11. Ljung, L. and Guo, L., The Role of Model Validation for Assessing the Size of the Unmodeled Dynamics, *IEEE Trans. Autom. Control*, 1997, vol. 42, no. 9, pp. 1230–1239.

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

13. Kushner, H.J. and Clark, D.S., *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Berlin: Springer-Verlag, 1978.

14. Shor, P.W., Quantum Computing, *9 Int. Math. Congress*, Berlin, 1998.

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

16. Robbins, H. and Siegmund, D., A Convergence Theorem for Nonnegative Almost Super-Martingales and Some Applications, in *Optimizing Methods in Statistics*, Rustagi, J.S., Ed., New York: Academic, 1971, pp. 233–257.

17. Polyak, B.T., *Vvedenie v optimizatsiyu* (Introduction to Optimization), Moscow: Nauka, 1983.

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