RANDOMIZED ALGORITHMS OF OPTIMIZATION AND THEIR IMPLEMENTATION ON QUANTUM COMPUTERS

Oleg Granichin[1]

Abstract — **New algorithms of the simultaneously perturbation stochastic approximation (SPSA) are considered under "almost" arbitrary noise in measurements of an unknown multi-variable minimized function. The function is measured sequently not at a point of the previous estimate but at estimate's slightly excited position for all vector components simultaneously. The offered algorithms use only one or two values of noisy measurements of the unknown function per each iteration. They give not only consistent estimates of unknown parameters but also they have so simple form that these algorithms "are naturally included" into implementation on new quantum electronic device for calculation of approximate value of the gradient vector of multi-variable function.**

Index Terms— **Multi-dimensional optimization, parameter estimation, randomized algorithm, SPSA, simultaneous perturbation, stochastic approximation, random direction, linear regression, filtering, prediction, quantum computing.**

I. Introduction

Recently electronics development has closely approached to creating intelligent control devices. Even now we have a real possibility of effective use of new algorithms of mathematical theories of optimization, optimal and adaptive control, identification of dynamic systems unknown parameters and design of experiments for practice problems solving. Earlier for some problems we could only try to get the most probable set that contains a vector of unknown parameters of the functioning dynamic system (or plant). But now the opportunity to get more precise solutions for them appears.

The exact solution of any problem can be found in the case when there is a precise formulation. In our real existing world all connections and relationships are so difficult and many–sided that it is impossible to give a strict mathematical description for many phenomena. Typically theoretic approach is to choose a mathematical model close to a real process and to include different noises (disturbances) into it. Noises represent some kind of "roughness" of the mathematical model from the one side and are characteristics of outside uncontrolled perturbations of a plant or a system from the other. During the last 50 years in majority of mathematical researches some "useful" probabilistic properties are arrogated to "noises" so that it would be easy to develop algorithms for task solving and then to prove mathematically the consistency of algorithms based on these properties. For instance, the most frequent assumptions are measurements noises independence and

[1]O.N.Granichin is with the Department of Mathematics and Mechanics, St. Petersburg State University (e-mail: Oleg_granitchin@mail.ru)

zero–mean. In the engineering the validity of the algorithms that are based on "least–squares method" or on "maximum likelihood method" is mostly grounded at probabilistic noise properties assumptions without enough basis. As a matter of fact it is inexpedient to use these algorithms in the conditions of a possible enemy counteraction. It is well known to specialists in the theory of the unknown parameters identification that if the "noise" is a deterministic unknown function (an enemy "jam" a signal) or the observation noise is a probabilistic "dependent" sequence, then the getting decisions is wrong. Then some theorists say that observation sequence is "degenerate" (not rich) and the solutions of such kind of problems are not studied at all. Another relevant problem is insufficient "variety" of the observations sequence. For example, the main purpose of adaptive control synthesis is a minimization of system state vector deviation from the specific trajectory that frequently turns to the "degenerate" sequence of observations. It causes complexity of the identification problem, for successful realization of which a "variety" of observations should be supplied.

At the case of the "bad" observation noise the ground of a new approach to solving of problems of estimation and optimization is the using of *trial simultaneous perturbations (disturbances)*. For the purpose of "enriching" information in the observation channel sometimes there is a possibility to include a new simultaneous perturbation with well–known probabilistic properties into the input system channel. In several problems the measurable random process that is already presented in a system is a such simultaneous perturbation. In control systems it is natural to add the trial simultaneous perturbations (actions) through a control channel. In other cases the simultaneous perturbation role can be played by randomization of an experiment design. Frequently it is possible to apply the methods which already became conventional for updated system with the simultaneous perturbation in order to research the convergence of new algorithms and their fields of applicability. Sometimes this updated system is the rewritten old one in the other form. One of the remarkable characteristics of such type of algorithms is a convergence under the "almost" arbitrary noise. A considerable restriction for using these algorithms is an assumption of "weak correlation" or "independence" of the measurement noise and the simultaneous perturbation which is added into the system, while there are no other assumptions about measurement noise properties. This restriction is natural in the case when the noise is generated from either an "unknown but bounded" deterministic function or by an enemy, who does not know real values and probabilistic properties of our simultaneous perturbation.

It is enough strangely that for a long time it was not the understanding that searching algorithms with sequential estimate $\{\theta_n\}$ changing in an axis direction of some random centered vector Δ_n (mean value is zero)

$$\theta_n = \theta_{n-1} - \Delta_n Y_n,$$

can converge to a true parameters vector $\theta^\star$ in the conditions when the observations Y_n are done with an "almost" arbitrary noise and Y_n are determined by previous estimate θ_{n-1} and the vector Δ_n of simultaneous perturbation. In this paper the algorithms of such type are named *"randomized algorithms of an estimation"* because the substantiation of

their convergence under "almost" arbitrary noise essentially uses a "stochastic" nature of the simultaneous perturbation. In the nearest future the caution attitude to stochastic algorithms and their outcomes will be significantly changed. The modern generation of computers will be replaced with a new kind of quantum computers. These new computers are stochastic systems, because their work is based on Heisenberg uncertainty principle of quantum mechanics. Probably, taking into account capabilities of quantum parallelism, the randomized algorithms of the optimization and estimation will lie down naturally into the basis of the future quantum computing devices.

A. Preview Example

Let's consider a problem of the detecting of the scalar "useful" signal φ_n, which is known and can be present or be absent in an observation channel. Suppose that $\{\varphi_n\}$ is the realization of some sequence of independent identically distributed random values with the mean value M_φ and the finite dispersion σ_φ^2. At each time moment n the measurement y_n is made with additive bounded noise v_n. Assume that cases $\{\theta^\star = 1\}$ and $\{\theta^\star = 0\}$ correspond to the situation when a signal is present or absent in the receiver. We can write the sequence of equations

$$y_n = \varphi_n \theta^\star + v_n, \; n = 1, 2, \ldots.$$

The our objective is to determine the value of $\theta^\star$ based on available input-output measurements y_i, φ_i, $i \leq N$.

A classical approach to a solving of this problem is the consideration of stochastic problem setting where noises are generated by the sequence of random variables with known probabilistic properties. For instance consider the sequence of estimates $\{\theta_n\}$ which are generated by the least mean squares method

$$\hat{\theta}_n = \frac{\sum_{k=1}^n \varphi_k y_k}{\sum_{k=1}^n \varphi_k^2}.$$

It is well-known that in the case of independent identically distributed random noises v_n the sequence $\{\theta_n\}$ converges with the probability 1 to the true value $\theta^\star$ when random sequences $\{\varphi_n\}$ and $\{v_n\}$ are independent, $\{v_n\}$ is bounded in the mean squares sense and the mean value of v_n is equal zero ($M_v = 0$). In the case $M_v \neq 0$ it converges to the value

$$\theta^\star + \frac{M_v M_\varphi}{\sigma_\varphi^2}.$$

When the mean value M_v is unknown and absolute value of the second item is more the 1 the previous estimates does not help us to solve the problem.

An alternative method is the membership set approach which based on *a priori* information about the level C_v of measurement noises:

$$|v_n| \leq C_v, \; n = 1, 2, \ldots.$$

Then at the time moment N the membership set

$$\Theta_N = \cap_{n=1}^N \{\theta \in \mathrm{R} : \; |y_n - \varphi_n \theta| \leq C_v\}$$

is the set of all possible values of parameter that are consistent with the observation scheme. But it is not possible to show the convergence of the sets sequence $\{\Theta_N\}$ to the point $\theta^\star$ as $N \to \infty$ under some general assumptions. An advantage of this approach is that it doesn't require any specific conditions on the noise types. The disadvantage is, however, that an accuracy of the estimation directly depends on a noise level. The quality of estimates isn't so good when a noise level is high. This approach leads to the pessimistic answer for the question about possibility to precisely identify the unknown value of parameter $\theta^\star$. The main purpose of this contribution is to show that in the sufficiently general cases this pessimistic answer not close the possibility to solve the identification problem.

How to solve the initial problem? Denote $\Delta_n = \varphi_n - M_\varphi$, $n = 1, 2, \ldots$ are centered inputs. Suppose that $\mathrm{E}\{|\Delta_n|^4\} < \infty$, $\Delta_1 \neq 0$ and sequences $\{\Delta_n\}$ and $\{v_n\}$ are independent (or $\{v_n\}$ is formed by an "unknown but bounded" deterministic function). Let's both parts of equation for y_n are multiplied by Δ_n. We derive

$$\Delta_n y_n = \Delta_n^2 \theta^\star + \Delta_n M_\varphi \theta^\star + \Delta_n v_n$$

and

$$\frac{1}{n}\sum_{k=1}^{n} \Delta_k y_k = \frac{1}{n}\sum_{k=1}^{n} \Delta_k^2 \theta^\star + \frac{1}{n}\sum_{k=1}^{n} \Delta_k M_\varphi \theta^\star + \frac{1}{n}\sum_{k=1}^{n} \Delta_k v_k, \; n = 1, 2, \ldots.$$

The first and second terms in the right part converge to $\sigma_\varphi^2 \theta$ and zero with the probability 1 as $n \to \infty$. It is possible to show that the last term converges to zero too. Hence the sequence $\{\hat{\theta}_n\}$, $n = 1, 2, \ldots$, formed by the rules

$$\hat{\theta}_n = \frac{\sum_{k=1}^{n} \Delta_k y_k}{\sum_{k=1}^{n} \Delta_k^2}, \; n = 1, 2, \ldots$$

converges to $\theta^\star$ with the probability 1. So we can get a precise solution of the initial problem for sufficiently large N. In the recurrent form the last method can be rewritten as randomized algorithm which is above.

B. Earlier Works

The idea of using random inputs to eliminate bias was put forward by Fisher [1] as the randomized principle in the design of experiments. The problem of linear regression parameters estimation was considered in [2–7] at the case of nonstandard assumptions about the observation noises. Some generalization of the usual LR problem setting were discussed in the [2] and [3]. There were supposed that the vector of unknown parameters can be time-varying and the algorithms for the estimation of the mean value of parameters were offered. In the partial case of centered random input signals and time invariant regression parameters the linear regression problems were considered by Plyak and Goldenshluger [4] with an arbitrary noise. But the algorithms which were suggested in [4] doesn't achieve an optimal rate of convergence in the general case. The possibility to get strongly consistent parameter estimates was also discussed in [5] when the noises are bounded and deterministic and the input sequences are suitably chosen.

As an important partial case of the randomized estimation algorithms there are algorithms of the simultaneously perturbation stochastic approximation (SPSA) [8–12]. The main features of the SPSA algorithms are the following: at each algorithm iteration we need only one or two values of an unknown minimized function measurements and the unknown function is measured not at a point of the previous estimate but at estimate's slightly excited position for all unknown vector components simultaneously. The convergence analysis of SPSA algorithm under the "almost" arbitrary noise was originated in [8]. In Polyak and Tsybakov [9] the algorithms of the same type were considered in "good" observation noise conditions, and their asymptotic optimality among a wide class of recurrent algorithms was proved. The term "SPSA algorithm" was offered in the [11] by Spall. He showed that in multi-dimensional case the essential reduction of measurements quantity at each iteration, in comparison with a classical Kiefer-Wolfowitz procedure of stochastic approximation, does not increase amount of iterations, which are necessary for obtaining the same accuracy of the estimation. Historically many authors (see [13, 14]) have been using another term, random direction stochastic approximation (RDSA) for denomination of the algorithms being looked at in the papers mentioned above. For the "training" of neuron networks SPSA algorithms were also offered in the [15] and [16]. The competence of SPSA algorithms under the "almost" arbitrary noise were considered in [7, 17–21].

The filtering problems with random inputs in an observation channel were discussed in Zhang [22, 23] for linear systems with non-Gaussian disturbances and in [7, 24] with almost arbitrary noise.

One of the possible way of a representation of the SPSA algorithm as quantum computing circuit is given in [20].

C. Contents of This Paper

The paper is organized as follows. In the next section, we state the problem of a minimization of an unknow function and main assumptions about the functions. We consider the multi-variable objective function like an average risk and we suppose that it is possible to measure values of the integrand loss function with the "almost" arbitrary noise only. This approach is more general than it usually studied. In Section III, we offer three types of SPSA algorithms. There is shown that the sequence of received estimates converges to the true value of the unknown parameters with the probability 1 and in the mean-square sense. In Section IV, we study the question how to increase the mean-square rate of convergence. In Section V, we formulate the linear regression problem and the main assumptions about inputs (regressors) and noises. Further we offer the randomized stochastic approximation (RSA) and mean squares (RMS) algorithms for estimation of the vector of the unknown LR parameters. There is shown that the sequences of generated estimates converges to the true value of the unknown parameters almost sure and in the mean-square sense. In Section VI, we study the filtering problem at the case of random inputs (regressors) in an observation channel and mixed type of uncertainties. The process, which is being filtered, is generated from a white noise sequence through a stable linear filter. The observation noises are formed from the values of an "unknown but

bounded" deterministic function or at the case of random observation noise we assume that it is bounded in the mean square sense and independent with inputs. There is offered to use the randomized least mean squares (RLMS) algorithm for the prediction. The upper boundary of the mean value of the prediction errors is established. It can be sufficiently small under appropriate choice of the probabilistic distribution of regressors. In Section VII, we give examples and numerical simulations to demonstrate the performance of our schemes. Under various types of observation noises the typical behavior of the offered estimates compares with the one of the standard LMS or KF estimates. The numerical results indicate two facts when observation noises don't satisfy any "good" statistical properties. Firstly, our estimation and filtering schemes outperform standard algorithms. Secondly, it is possible to achieve the value of averaged errors much less then the observation noise level. Section VIII is devoted to the representation of SPSA algorithms on a quantum computing device. In Appendix, we go through the proofs of the main results.

II. Problem Statement and Main Assumptions

Let $F(w,\theta):\ \mathbb{R}^p\times\mathbb{R}^d\to\mathbb{R}^1$ be a continuously differentiatable on the second argument of the function, $x_1,x_2\ldots$ is an observation plan, a sequence of points selected by an experimenter where the observation value of the unknown function $F(w_n,\cdot)$ is accessible at each iteration $n=1,2,\ldots$ with additive noise v_n

$$y_n=F(w_n,x_n)+v_n.$$

Here $\{w_n\}$ is an uncontrollable sequence of random vectors $w_n\in\mathbb{R}^p$ with an identical unknown distribution $\mathrm{P}_w(\cdot)$ which has a finite support.

Problem statement. It is required to find the unknown vector $\theta^\star$ which minimizes a function

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

by using the observations $y_1,y_2\ldots$. Usually the problem of function $f(\cdot)$ minimization is considered when using more simple observation model

$$y_n=f(x_n)+v_n,$$

which is easily included in above. Generalization made in the problem statement is dictated, as a minimum, by tendency to take into account the case of multiplicative noise in observations

$$y_n=w_nf(x_n)+v_n.$$

This case is included in the general scheme with the function $F(w,x)=wf(x)$.

We adopt the following notations. $\mathrm{E}\{\cdot\}$ is used to denote the expectation of a random variable. The Euclidean norm of a vector x in $\mathbb{R}^d$ is denoted by $\|x\|$. The scalar product in R^d is denoted by $\langle\cdot,\cdot\rangle$. I is d-dimensional identity matrix. The trace of a matrix A

is denoted by Tr[A]. A > 0 means that A is a positive definite matrix. The maximum (minimum) eigenvalue of A is denoted by $\lambda_{\max}(\mathrm{A})$ ($\lambda_{\min}(\mathrm{A})$), and the Euclidean norm of A is defined as its maximum singular value, i.e.

$$\|\mathrm{A}\| = \sqrt{\lambda_{\max}(\mathrm{A}\mathrm{A}^{\mathrm{T}})}.$$

Let's formulate the main assumptions.

(A.1) Function $f(\cdot)$ has a unique root in R^d at some point $\theta^\star = \theta^\star(f(\cdot))$ and

$$\langle x - \theta^\star(f(\cdot)), \nabla f(x)\rangle \geq \mu\|x - \theta^\star(f(\cdot))\|^2, \ \ \forall x \in \mathrm{R}^d$$

with some constant $\mu > 0$.

(A.2) the gradient of a $f(\cdot)$ Lipschitz condition holds

$$\|\nabla f(x) - \nabla f(\theta)\| \leq A\|x - \theta\|, \ \forall x, \theta \in \mathrm{R}^d$$

with some constant $A > \mu$.

(A.3) Function $f(\cdot) \in C^\ell$ is ℓ–times continuously differentiable and for all its partial derivatives up to the order ℓ the Holder condition of order ρ ($0 < \rho \leq 1$), holds on R^d so that

$$|f(x) - \sum_{|\bar{l}|\leq\ell} \frac{1}{\bar{l}!} D^{\bar{l}} f(\theta)(x-\theta)^{\bar{l}}| \leq M\|x-\theta\|^\gamma,$$

where $\gamma = \ell + \rho \geq 2$, M — some constant,
$\bar{l} = (l^{(1)}, \ldots, l^{(d)})^{\mathrm{T}} \in \mathbb{N}^d$ is a multi-index, $l^{(i)} \geq 0$, $i = 1, \ldots, d$,
$\mid \bar{l} \mid = l^{(1)} + \ldots + l^{(r)}, \bar{l}! = l^{(1)}! \cdots l^{(d)}!$,
$x \in \mathbb{R}^d, x^{\bar{l}} = (x^{(1)})^{l^{(1)}} \cdots (x^{(d)})^{l^{(d)}}, D^{\bar{l}} = \partial^{|\bar{l}|}/(\partial x^{(1)})^{l^{(1)}} \cdots (\partial x^{(d)})^{l^{(d)}}$.
If $\gamma = 2$ then $M = A/2$ in the condition (A.1) .

III. Basic Algorithms and Trial Simultaneous Perturbations

Let $\{\Delta_n\}_{n=1,2,\ldots}$ is an observable sequence of independent random vectors. $\{\Delta_n\}$ is called *trial simultaneous perturbation.* $\Delta_n \in \mathrm{R}^d$ and its distribution function is $\mathrm{P}_n(\cdot)$.

Let $\theta_0 \in \Theta$ be some initial vector and $\{\alpha_n\}$, $\{\beta_n\}$ — some numerical sequences tending to zero. Consider three algorithms for constructing the plan of experiments $\{x_n\}$ and the sequence of estimates $\{\theta_n\}$. The first one is

$$\begin{cases} x_n = \hat{\theta}_{n-1} + \beta_n \Delta_n, \ y_n = F(w_n, x_n) + v_n \ , \\ \\ \hat{\theta}_n = \hat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n} \mathcal{K}_n(\Delta_n) y_n, \end{cases} \tag{1}$$

that uses one measurement per each iteration. Next algorithms, the "smoothed" versions of the Kiefer-Wolfowitz procedure, use two measurements

$$\begin{cases} x_{2n} = \hat{\theta}_{n-1} + \beta_n \Delta_n, \; x_{2n-1} = \hat{\theta}_{n-1} - \beta_n \Delta_n, \\ \hat{\theta}_n = \hat{\theta}_{n-1} - \frac{\alpha_n}{2\beta_n} \mathcal{K}_n(\Delta_n)(y_{2n} - y_{2n-1}), \end{cases} \tag{2}$$

$$\begin{cases} x_{2n} = \hat{\theta}_{n-1} + \beta_n \Delta_n, \qquad x_{2n-1} = \hat{\theta}_{n-1}, \\ \hat{\theta}_n = \hat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n} \mathcal{K}_n(\Delta_n)(y_{2n} - y_{2n-1}). \end{cases} \tag{3}$$

In all algorithms there are some smoothed vector-functions (kernels) $\mathcal{K}_n(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, $n = 1, 2, \ldots$ with the finite support: $\mathcal{K}_n(x) = 0$ for $\|x\| \geq C_\Delta$, where C_Δ is some constant. These functions and the trial simultaneous perturbation distribution functions $\mathrm{P}_n(\cdot)$ satisfy the following conditions

$$\int \mathcal{K}_n(x) \mathrm{P}_n(dx) = 0, \quad \int \mathcal{K}_n(x) x^{\mathrm{T}} \mathrm{P}_n(dx) = \mathrm{I}, \tag{4}$$

$$\sup_n \int \|\mathcal{K}_n(x)\|^2 \mathrm{P}_n(dx) < \infty, \quad n = 1, 2, \ldots.$$

The first time the algorithm (2) was proposed by Kushner and Clark in the book [13] for the case of uniform distributed trial perturbation and functions $\mathcal{K}_n(x) = x$. In the [8] the algorithm (1) was considered with the same kernel function but with more general type of the trial distribution under "almost" arbitrary measurement noise. Both algorithms (1) and (2) were proposed by Polyak and Tsybakov in the [9] with kernel vector-functions $\mathcal{K}_n(\cdot)$ of the more general type and the uniform distributed trial simultaneous perturbation. In the [11] Spall began to consider the algorithm (2) with general type of the trial simultaneous perturbation distribution and the vector-functions

$$\mathcal{K}_n(\Delta_n) = \begin{pmatrix} \frac{1}{\Delta_n^{(1)}} \\ \frac{1}{\Delta_n^{(2)}} \\ \vdots \\ \frac{1}{\Delta_n^{(d)}} \end{pmatrix}.$$

In the [19] was offered the algorithm (3) with the same vector-functions $\mathcal{K}_n(\cdot)$ and the same type of the trial simultaneous perturbation distribution

For some technical reasons we modify the algorithm (1)

$$\begin{cases} x_n = \hat{\theta}_{n-1} + \beta_n \Delta_n, \; y_n = F(w_n, x_n) + v_n \;, \\ \hat{\theta}_n = \mathcal{P}_{\Theta_n}(\hat{\theta}_{n-1} - \frac{\alpha_n}{\beta_n} \mathcal{K}_n(\Delta_n) y_n), \end{cases} \tag{5}$$

where $\{\Theta_n\}$ is a sequence of convex closed sets $\Theta_n \subset \mathrm{R}^d$, which contain the point $\theta_\star$ for all sufficiently large $n \geq 1$, $\mathcal{P}_{\Theta_n}$ is a projector on Θ_n. Let $d_n = \mathrm{diam}\Theta_n$ be Euclidean diameter of the set Θ_n. If we know beforehand the convex closed set Θ, which contains the point $\theta^\star$ then $\Theta_n = \Theta$. In the contra case the sequence $\{d_n\}$ can be infinitely increasing.

Denote $\mathbb{W} = \mathrm{supp}(\mathrm{P}_w(\cdot)) \subset \mathbb{R}^p$; $\mathcal{F}_{n-1}$ — σ-algebra generated by $\hat{\theta}_0, \hat{\theta}_1, \ldots, \hat{\theta}_{n-1}$, which are formed by algorithm (5) (or (2), or (3)); for algorithms (2) or (3)

$$\bar{v}_n = v_{2n} - v_{2n-1},\ \bar{w}_n = \begin{pmatrix} w_{2n} \\ w_{2n-1} \end{pmatrix},\ d_n = 1,$$

for the algorithm (5)

$$\bar{v}_n = v_n,\ \bar{w}_n = w_n.$$

Theorem 1 If the condition (A.1) *is held for the function* $f(\theta) = \mathrm{E}\{F(w,\theta)\}$*;*
(A.2) *for functions* $F(w,\cdot)$ $\forall w \in \mathbb{W}$*;*
(4) *for functions* $\mathcal{K}_n(\cdot)$ *and* $\mathrm{P}_n(\cdot)$, $n = 1, 2, \ldots$*;*
$\forall \theta \in \mathbb{R}^d$ *functions* $F(\cdot,\theta)$ *and* $\nabla_\theta F(\cdot,\theta)$ *uniformly bounded on* $\mathbb{W}$*;*
$\forall n \geq 1$ *random values* $\bar{v}_1, \ldots, \bar{v}_n$ *and vectors* $\bar{w}_1, \ldots, \bar{w}_{n-1}$ *don't depend on* $\bar{w}_n, \Delta_n$*, random vector* $\bar{w}_n$ *doesn't depend on* Δ_n*;*
$\mathrm{E}\{\bar{v}_n^2\} \leq \sigma_n^2$, $n = 1, 2, \ldots$.

If $\sum_n \alpha_n = \infty$ *and* $\alpha_n \to 0$, $\beta_n \to 0$, $\alpha_n^2 \beta_n^{-2}(1 + d_n^2 + \sigma_n^2) \to 0$ *as* $n \to \infty$,

then *the sequence of estimates* $\{\hat{\theta}_n\}$*, generated by the algorithm (5) (or (2), or (3)), converges to the point* $\theta^\star$ *in the mean-square sense* $\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2\} \to 0$ *as* $n \to \infty$.

Moreover, if $\sum_n \alpha_n \beta_n^2 < \infty$ *and with the probability 1*

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

then $\hat{\theta}_n \to \theta^\star$ *as* $n \to \infty$ *with the probability 1.*

The proof of Theorem 1 is given in Appendix.

Remark 1. For the function $F(w,x) = wf(x)$ conditions (A.1,2) of Theorem 1 are held when they are satisfied for the function $f(x)$.

Remark 2. Under fulfillment of Theorem 1 conditions, measurement noises v_n are "almost" arbitrary one in some sense. They can be either the nonrandom "unknown but bounded" deterministic sequence or the realization of some stochastic process with any kind of internal dependences. In particular, it is not necessary to assume anything about the dependence between v_n and σ-algebra $\mathcal{F}_{n-1}$.

Remark 3. The condition of independence of measurement noise and trial simultaneous perturbation can be loosed. As in [18] it is enough to require the tending to zero the conditional correlation between v_n and $\mathcal{K}_n(\Delta_n)$ as $n \to \infty$ with the rate being not less than

$$\|\mathrm{E}\{v_n \mathcal{K}_n(\Delta_n) | \mathcal{F}_{n-1}\}\| = \mathcal{O}(\frac{\alpha_n}{\beta_n})$$

with the probability 1.

IV. Rate of Convergence

Let's consider a particular structure of the trial simultaneous perturbation distribution functions and some special kind of vector functions $\mathcal{K}_n(\cdot),\ n = 1, 2, \ldots$ which satisfy the conditions (4). By using these functions, the algorithm (2)(or (3), or (5)) achieves asymptotically optimum mean–square rate of convergence $\mathcal{O}(n^{-\frac{\gamma-1}{\gamma}})$ as in the [9], provided the researched function satisfies to the condition of a smoothness (A.3).

Let independent random vectors $\{\Delta_n\}$ be distributed identically at $n = 1, 2, \ldots$, all components of the vectors Δ_n are independent among themselves and have symmetrical as to zero identical scalar distribution function P_Δ distinct from zero only within some interval $[-C_\Delta, C_\Delta] \subset \mathrm{R},\ C_\Delta > 0$.

Further we shall consider vector functions $\mathcal{K}_n(\cdot)$ independent from n

$$\mathcal{K}_n(x) = \mathcal{K}(x) = (\mathcal{K}^{(1)}(x), \ldots, \mathcal{K}^{(d)}(x))^{\mathrm{T}},\ x \in \mathrm{R}^d,\ n = 1, 2, \ldots,$$

components of which $\mathcal{K}^{(i)}(\cdot),\ i = 1, \ldots, d$ are calculated using the formulas

$$\mathcal{K}^{(i)}(x) = K_0(x^{(i)}) \prod_{j \neq i} K_1(x^{(j)}),\ i, j = 1, \ldots, d,\ x \in \mathbb{R}^d. \tag{6}$$

determined by two scalar bounded functions $K_0(\cdot)$ and $K_1(\cdot)$ (kernels) with the finite support $[-C_\Delta, C_\Delta]$, satisfying the conditions

$$\begin{gathered} \int u K_0(u) \mathrm{P}_\Delta(du) = 1, \int u^k K_0(u) \mathrm{P}_\Delta(du) = 0,\ k = 0, 2, \ldots, \ell, \\ \\ \int K_1(u) \mathrm{P}_\Delta(du) = 1,\ \int u^k K_1(u) \mathrm{P}_\Delta(du) = 0,\ k = 1, \ldots, \ell - 1. \end{gathered} \tag{7}$$

In particular, for the scalar case ($d = 1$) the definition of $\mathcal{K}(\cdot)$ is $\mathcal{K}(x) = K_0(x)$.

It is simple to be convinced that functions $\mathcal{K}(\cdot)$ together with the distribution function of a trial simultaneous perturbation

$$\mathrm{P}_n(x) = \mathrm{P}(x) = \prod_{i=1}^{d} \mathrm{P}_\Delta(x^{(i)})$$

satisfies the conditions (4) while fulfilling the conditions (7).

Let's indicate some possible ways of the construction of kernels functions $K_0(\cdot)$ and $K_1(\cdot)$ which satisfy the conditions (7). Let $\{p_m(\cdot)\}_{m=0}^{\ell}$ is some system of polynomials on an interval $[-C_\Delta, C_\Delta]$, orthogonal relative to a measure generated by distribution $\mathrm{P}_\Delta(\cdot)$. Let's take

$$K_0(u) = \sum_{m=0}^{\ell} a_m p_m(u)\ , \qquad a_m = p'_m(0) / \int_{-C_\Delta}^{C_\Delta} p_m^2(u) \mathrm{P}_\Delta(du),$$

$$K_1(u) = \sum_{m=0}^{\ell-1} b_m p_m(u)\ , \qquad b_m = p_m(0)\ /\ \int_{-C_\Delta}^{C_\Delta} p_m^2(u) \mathrm{P}_\Delta(du)$$

and check that so defined functions $K_0(\cdot)$ and $K_1(\cdot)$ satisfy the conditions (7). Really, the function u^k can be presented as

$$u^k = \sum_{j=0}^{k} c_j p_j(u),$$

where c_j are numeric coefficients. Therefore, for the kernel $K_0(\cdot)$ which is constructed by orthogonal polynomials, we get

$$\int u^k K_0(u) \mathrm{P}_\Delta(du) = \sum_{m=0}^{\ell} \sum_{j=0}^{k} \int_{-C_\Delta}^{C_\Delta} a_m c_j p_m(u) p_j(u) \mathrm{P}_\Delta(du) = \sum_{j=0}^{k} c_j p'_j(0) = \delta_{k,1},$$

after taking a derivative from the formula of the u^k expansion and then putting $u = 0$. Here $\delta_{i,j} = 1$ if $i = j$, and $\delta_{i,j} = 0$ otherwise. The kernel $K_1(\cdot)$ satisfies to conditions (7) too. This fact is proved more simply.

For the construction of kernels $K_0(\cdot)$ and $K_1(\cdot)$ on an interval $[-1/2, 1/2]$ it was offered in [9] to use orthogonal Legendre's polynomials $p_m(\cdot), m = 0, 1, \ldots, \ell$ with the uniform probabilistic distribution of trial perturbation components on an interval $[-1/2, 1/2]$. In this case for initial values of $\ell = 1, 2$ (i.e. $2 \le \gamma \le 3$) we have
$K_0(u) = 12u, K_1(u) = 1,$
for next values of $\ell = 3, 4$ (i.e. $3 < \gamma \le 5$)
$K_0(u) = 5u(15 - 84u^2), \quad K_1(u) = 9/4 - 15u^2.$
Note, that by definitions $K_0(u) = 0$ and $K_1(u) = 0$ as $|u| > 1/2$.

In many practical cases the possibility of using more general (than in [9]) consideration of the trial simultaneous perturbation distributions and the other orthogonal polynomials set is caused by the problem statement itself, which sometimes contains a type of the trial perturbation distribution P_Δ or suggests more convenient one. Solving some problems it is possible to use trial simultaneous perturbation distributions belonging to some narrow fixed class only.

The following theorem gives sufficient conditions where the asymptotic rate of convergence of algorithm (2) (or (5)) is optimal.

Theorem 2 Let be $\alpha_n = \alpha\, n^{-1},\ \beta_n = \beta\, n^{-\frac{1}{2\gamma}},\ \beta > 0.$
If the following conditions are fulfilled:
(A.1,3) *under* $\gamma \ge 2,\ \alpha\beta > \frac{\gamma-1}{2\mu\gamma}$ *for the function* $f(\theta) = \mathrm{E}\{F(w,\theta)\}$;
(A.2) *for functions* $F(w,\cdot)\ \forall w \in \mathbb{W}$;
(7) *for functions* $K_0(\cdot), K_1(\cdot)$ *and* $\mathrm{P}_\Delta(\cdot)$;
$\forall \theta \in \mathbb{R}^d$ *functions* $F(\cdot,\theta)$ *and* $\nabla_\theta F(\cdot,\theta)$ *uniformly bounded on* $\mathbb{W}$;
$d_n n^{-1+\frac{1}{2\gamma}} \to 0$ *as* $n \to \infty$;
$\forall n \ge 1$ *random vectors* $\bar{w}_n, \Delta_n$ *don't depend on* $\bar{v}_1, \ldots, \bar{v}_n, \bar{w}_1, \ldots, \bar{w}_{n-1}$, *random vector* Δ_n

doesn't depend on $\bar{w}_n$*;*
$\mathrm{E}\{(v_{2n}-v_{2n-1})^2/2\} \le \sigma_2^2$, $(\mathrm{E}\{v_n^2\} \le \sigma_1^2)$;

then *for the sequence of estimates* $\{\theta_n\}$*, generated by the algorithm (2) (or (5)), the mean-square convergence rate is*

$$\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2\} = \mathcal{O}(n^{-\frac{\gamma-1}{\gamma}})$$

asymptotically when $n \to \infty$*.*

The proof of Theorem 2 is also given in Appendix.

Remark 4. Define $\chi = 1$ when $\{v_n\}$ are independent and mean–zero, $\chi = 2$ in other cases. The constants $\hat{K} = \int \|\mathcal{K}(x)\|^2 \mathrm{P}(dx)$ and $\bar{K} = \int \|x\|^\gamma \|\mathcal{K}(x)\| \mathrm{P}(dx)$ are bounded because vector-functions $\mathcal{K}(\cdot)$ are bounded and $supp(\mathrm{P}_\Delta(\cdot))$ is finite. At the case $M > 0$ in the proof of Theorem 2 the best value of parameters and quantitative evaluations of the asymptotic convergence rate will be established for the algorithms (2) and (5)

$$\alpha^\star = 1/(\mu\beta^\star),\ \beta^\star = (2\chi(\nu_i + \sigma_i^2/i)\hat{K})^{\frac{1}{2\gamma}}(\sqrt{\gamma(\gamma-1)}M\bar{K})^{-\frac{1}{\gamma}},$$
$$\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2\} \le n^{-\frac{\gamma-1}{\gamma}}\kappa_i \hat{K}^{\frac{\gamma-1}{\gamma}}\bar{K}^{\frac{2}{\gamma}} + o(n^{-\frac{\gamma-1}{\gamma}}),$$
$$\kappa_i = \gamma^{\frac{1+\gamma}{\gamma}}(\gamma-1)^{\frac{1-\gamma}{\gamma}}\mu^{-2}(\chi(\nu_i + \sigma_i^2/i))^{\frac{\gamma-1}{\gamma}}M^{\frac{2}{\gamma}},\ \ i = 1, 2.$$

Here $\nu_1 = \sup_{w\in\mathbb{W}}\Big(F(w,\theta^\star) + \frac{1}{2}(\nabla_\theta F(w,\theta^\star))^2\Big)^2$, corresponds to the algorithm (5); and for the algorithm (2) $\nu_2 =$

$$= \sup_{w_1,w_2\in\mathbb{W}}\Big(2|F(w_1,\theta^\star) - F(w_2,\theta^\star)| + (\nabla_\theta F(w_1,\theta^\star))^2 + (\nabla_\theta F(w_2,\theta^\star))^2\Big)^2/8.$$

In the case of $F(w,x) = f(x)$, the iteration asymptotic convergence rate is always better for the sequence of estimates formed by the algorithm (2), which using two observation, than for the algorithm (5), which can be clearly seen from views κ_1 and κ_2. The algorithm (2) advantage stops being indisputable if we try to compare the algorithms' behavior taking into account the number of measurements (twice as much measurements needs to be made at each iteration when using the algorithm (2)). Comparing values κ_1 and $2\kappa_2$, it is easily convincing that if $2^{\frac{1}{\gamma-1}}\sigma_2^2 - \sigma_1^2 > \nu_1 - 2^{\frac{1}{\gamma-1}}\nu_2$ then the asymptotic rate of convergence that takes into account the number of measurements is better for the algorithm (5) than for (2) .

Let's consider the scalar case, $d = 1$, and $F(w,x) = f(x)$, $\gamma = 2$, the trial perturbation $\{\Delta_n\}$ formed by uniformly distributed independent random variables lying on an interval $[-\frac{1}{2}, \frac{1}{2}]$, $\mathcal{K}(x) = 12x$, $|x| \le 1/2$, measurement noises $\{v_n\}$ are random independent zero–mean values,$\{v_n\} : \mathrm{E}\{v_n^2\} \le \sigma_v^2$, For algorithm (2) we have

$$\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2\} \le \frac{9A\sigma_v}{4\sqrt{3}\mu^2}n^{-1/2} + o(n^{-1/2}),\ \alpha^\star = \frac{1}{\mu\beta^\star},\ \beta^\star = \frac{4\sqrt{\sigma_v}}{\sqrt{A}\sqrt[4]{3}},$$

and for (5) $\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2\} \le 4,5\sqrt{f(\theta^\star)^2 + \sigma_v^2}/(\sqrt{6}\mu^2)\, n^{-1/2} + o(n^{-1/2})$. Hence, if $f(\theta^\star)^2 < \sigma_v^2$, then it is more preferable to use the algorithm (5).

V. LR Parameter Estimation

Consider a linear regression model

$$y_n \;=\; \varphi_n^{\mathrm{T}}\,\theta_n^\star \;+\; v_n, \quad \theta_n^\star = \theta^\star + w_n, \quad n = 0, 1, \ldots \tag{8}$$

Here $y_n \in \mathrm{R}^1$ is an output of the observation made at time n, input $\varphi_n \in \mathrm{R}^d$ is a vector that is known at time n, and $v_n \in \mathrm{R}^1$, $w_n \in \mathrm{R}^d$ represent the noises (disturbances). The unknown parameters vector $\theta^\star$ is to be estimated from the observations y_i, φ_i, $i \le n$.

Let $\mathcal{F}_n$ be the σ-algebra generated by $\{\varphi_0, \ldots, \varphi_n,\, w_0, \ldots, w_n,\, v_1, \ldots, v_n\}$, $\hat{\mathcal{F}}_n$ be the σ-algebra generated by $\{\varphi_0, \ldots, \varphi_n,\, w_0, \ldots, w_n,\, v_0, \ldots, v_{n+1}\}$, and $\tilde{\mathcal{F}}_n$ be the σ-algebra generated by $\{\varphi_0, \ldots, \varphi_n,\, w_0, \ldots, w_{n+1},\, v_0, \ldots, v_{n+1}\}$, $\mathcal{F}_{n-1} \subset \hat{\mathcal{F}}_{n-1} \subset \tilde{\mathcal{F}}_{n-1} \subset \mathcal{F}_n$.

We make the following Assumptions.

(A) The inputs $\{\varphi_n\}_{n\ge 0}$ form a sequence of independent identically distributed random vectors with known bounded mean values $\|\mathrm{E}\{\varphi_n\}\| = M_\varphi < \infty$ and φ_n is independent of $\tilde{\mathcal{F}}_{n-1}$. The random vectors $\Delta_n = \varphi_n - \mathrm{E}\{\varphi_n\}$ have a symmetric distribution function $\mathrm{P}(\cdot)$ (i.e. $\mathrm{P}(\Omega) = \mathrm{P}(-\Omega)$ for any Borel set $\Omega \subset \mathbb{R}^d$), $\mathrm{E}\{\Delta_n \Delta_n^{\mathrm{T}}\} = \mathrm{B} > 0$, $\mathrm{E}\{\|\Delta_n\|^4\} \le M_4 < \infty$.

(B) $\forall n$ w_n is independent of $\hat{\mathcal{F}}_{n-1}$ and $\mathrm{E}\{w_n\} = 0$. The noises $\{v_n\}_{n\ge 1}$ and $\{w_n\}_{n\ge 1}$ satisfy one of the conditions:

(i) $\qquad \mathrm{E}\{\, v_n^2 | \mathcal{F}_{n-1}\} \le \sigma_v^2 < \infty$, a.s., $\mathrm{E}\{\|w_n\|^2\} \le \sigma_w^2 < \infty$,

(ii) $\qquad \mathrm{E}\{v_n^2\} \le \sigma_v^2 < \infty, \qquad \mathrm{E}\{w_n w_n^{\mathrm{T}}\} \le \mathrm{Q}_w < \infty$,

where σ_v, σ_w are some constants, Q_w is a symmetric matrix.

Note that standard assumptions about observation noise $\{v_n\}$ in LR parameter estimation with random inputs are somewhat different [25]. It is assumed, in particular, that $\mathrm{E}\{v_n\} = 0$, and $\{v_n\}$ is a sequence of independent identically distributed random variables, and $\{v_n\}$ are independent of $\{\varphi_n\}$.

Let's study the randomized stochastic approximation estimator (RSA) for the model (8)

$$\hat{\theta}_n = \hat{\theta}_{n-1} - \alpha_n \Gamma \Delta_n (\varphi_n^{\mathrm{T}} \hat{\theta}_{n-1} - y_n), \quad n = 1, 2, \ldots, \tag{9}$$

where $\alpha_n \ge 0$ is a nonrandom step–size and Γ is a positive definite symmetric matrix. We also suppose that the initial value $\hat{\theta}_0$ is an arbitrary nonrandom vector in $\mathbb{R}^d$.

Theorem 3 Let Assumption (A) *be fulfilled and the sequence* $\{\alpha_n\}$ *satisfy*

$$\sum_{n=1}^{\infty} \alpha_n = \infty, \quad \alpha_n \to 0 \quad \text{as} \quad n \to \infty. \tag{10}$$

If Assumption (Bi) *holds and* $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$ *then for the estimates generated by the algorithm (9) we have* $\hat{\theta}_n \to \theta^\star$ *a.s. as* $n \to \infty$.

If Assumption (Bii) *holds then* $\mathrm{E}\{(\hat{\theta}_n - \theta^\star)(\hat{\theta}_n - \theta^\star)^{\mathrm{T}}\} \to 0$ *as* $n \to \infty$.

In the case when $\Gamma = \mathrm{B}^{-1}$ the result of Theorem 3 follows from Theorem 1 immediately with

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

The algorithm (9) is equal to the (2).

The following theorem establishes the rate of convergence for the algorithm (9).

Theorem 4 Let Assumptions (A) *and* (Bii) *be fulfilled,* $\alpha_n = n^{-1}$, *and* $-\Gamma\mathrm{B} + \frac{1}{2}\mathrm{I}$ *be a Hurwitz matrix, i.e. all its eigenvalues lie in the left half–plane.*

Then for the algorithm (9) we have

$$\mathrm{E}\{(\hat{\theta}_n - \theta^\star)(\hat{\theta}_n - \theta^\star)^{\mathrm{T}}\} \le n^{-1}\,\mathrm{S} \;+\; o(n^{-1}), \tag{11}$$

where S *is a solution of matrix equation*

$$\Gamma\,\mathrm{B}\,\mathrm{S} \;+\; \mathrm{S}\,\mathrm{B}\,\Gamma \;-\; \mathrm{S} \;=\; \Gamma\mathrm{R}\Gamma. \tag{12}$$

Here $\mathrm{R} = (\sigma_v^2(1 + M_\varphi^2\rho) + M_\varphi^2\mathrm{Tr}[\mathrm{Q}_w])\mathrm{B} + \mathrm{E}\{\Delta_n\Delta_n^{\mathrm{T}}\mathrm{Q}_w\Delta_n\Delta_n^{\mathrm{T}}\}$ *and* $\rho > 0$ *is any small positive constant.*

The proof of Theorem 4 is given in Appendix.

Equation (12) can be explicitly solved in the case when $\Gamma = \mathrm{B}^{-1}$, $\mathrm{Tr}[\mathrm{Q}_w] = 0$ and $M_\varphi = 0$. For the algorithm

$$\hat{\theta}_n \;=\; \hat{\theta}_{n-1} \;-\; (n\mathrm{B})^{-1}\varphi_n(\varphi_n^{\mathrm{T}}\hat{\theta}_{n-1} - y_n), \tag{13}$$

we have

$$\mathrm{E}\{(\hat{\theta}_n - \theta^\star)(\hat{\theta}_n - \theta^\star)^{\mathrm{T}}\} \le n^{-1}\sigma_v^2\mathrm{B}^{-1} \;+\; o(n^{-1}).$$

The same rate of convergence holds for the algorithm (13) when v_n are independent zero–mean random variables, see [25]. Moreover, it was shown in the [25] that this choice of α_n and Γ is an optimal for algorithms of similar kinds.

Remark 5: In Theorem 4 statement in the case of equalities in Assumption (Bii) the equality also holds in the equation (11) for the rate of convergence.

Now we consider the randomized least squares estimator (RMS) for the regression model (8):

$$\begin{cases} \hat{\theta}_n = \hat{\theta}_{n-1} - \Gamma_n\Delta_n(\varphi_n^{\mathrm{T}}\hat{\theta}_{n-1} - y_n), \\ \\ \Gamma_n = \Gamma_{n-1} - \Gamma_{n-1}\Delta_n\Delta_n^{\mathrm{T}}\Gamma_{n-1}/(1 + \Delta_n^{\mathrm{T}}\Gamma_{n-1}\Delta_n),\ \Gamma_0 = \gamma_0^{-1}\mathrm{I}, \end{cases} \tag{14}$$

where $\gamma_0 > 0$ is a small positive number (regularization parameter, see [26, 27]). We again assume that the initial value $\hat{\theta}_0$ is an arbitrary nonrandom vector in $\mathbb{R}^d$.

Theorem 5 Let Assumption (A) *be fulfilled.*

If Assumption (Bi) *holds then for the algorithm (14)* $\hat{\theta}_n \to \theta^\star$ *a.s. as* $n \to \infty$.

If $|v_n| \le C_v$, $\|w_n\| \le C_w$, $\|\Delta_n\| \le C_\Delta$ *a.s. then for the algorithm (14)* $\mathrm{E}\{(\hat{\theta}_n - \theta^\star)(\hat{\theta}_n - \theta^\star)^{\mathrm{T}}\} \to 0$ as $n \to \infty$. *Here* C_v, C_w, $C_\Delta < \infty$ *are some constants.*

The proof of Theorem 5 is also given in Appendix.

VI. Prediction of the Signal (Filtering)

We'll study the special case of the problem of filtering when the observations are related by equations

$$y_n = \varphi_n^{\mathrm{T}}\theta_n^\star + v_n,\ n = 0, 1, \ldots \tag{15}$$

Here $y_n \in \mathbb{R}^1$ is an observation made at time n and φ_n is a d-dimensional vector that is known at time n, $v_n \in \mathbb{R}^1$ represents an observation noise, and the vector signal process $\{\theta_n^\star\}$, $\theta_n^\star \in \mathbb{R}^d$ is generated from a white noise sequence through a stable linear filter

$$\theta_{n+1}^\star = \mathrm{A}\theta_n^\star + w_{n+1},\ \theta_0^\star \in \mathbb{R}^r, \tag{16}$$

where A is a known stable matrix (i.e. $\|\mathrm{A}\| < 1$), $\{w_n\}$ and $\{v_n\}$ satisfy conditions (Bii).

It is the objective of the problem of the prediction for one step to estimate the vector $\theta_{n+1}^\star$ from measurements y_i, φ_i, $i \le n$. Let $\hat{\theta}_{n+1}$ be a current estimate of the vector $\theta_{n+1}^\star$. The quality of prediction (the filtering performance) is determined by the mean value of square of the prediction error

$$\mathrm{E}\{\|\hat{\theta}_{n+1} - \theta_{n+1}^\star\|^2\}.$$

Usually in the filtering problem statement deterministic sequence of vectors $\{\varphi_n\}$ is considered. Here we will suppose that vectors φ_n, $n = 1, 2, \ldots$ are random and satisfy Assumption (A).

The randomized least mean squares algorithm (RLMS) is defined recursively by

$$\hat{\theta}_{n+1} = \mathrm{A}\hat{\theta}_n - \alpha\mathrm{A}\Gamma\Delta_n(\varphi_n^{\mathrm{T}}\hat{\theta}_n - y_n),\ n = 0, 1, \ldots, \tag{17}$$

where $\alpha > 0$ is a step–size and Γ is a positive definite symmetric matrix. We suppose that the initial value $\hat{\theta}_0$ is some vector in $\mathbb{R}^d$.

The prediction errors satisfy by the following equation which can be obtained by substituting (15) and (16) into (17):

$$\hat{\theta}_{n+1} - \theta_{n+1}^\star = \mathrm{A}(\mathrm{I} - \alpha\Gamma\Delta_n\Delta_n^{\mathrm{T}})(\hat{\theta}_n - \theta_n^\star) - \alpha\mathrm{A}\Gamma\Delta_n(\mathrm{E}\{\varphi_n\}^{\mathrm{T}}(\hat{\theta}_n - \theta_n^\star) - v_n) - w_{n+1}.$$

If Assumptions (A) and (Bii) hold then successively taking the conditional expectations with respect to σ-algebra $\mathcal{F}_n$ and $\tilde{\mathcal{F}}_{n-1}$ we conclude that

$$\mathrm{E}\{\|\hat{\theta}_{n+1} - \theta_{n+1}^\star\|^2|\tilde{\mathcal{F}}_{n-1}\} \le (1 - 2\alpha\lambda_{\min}(\mathrm{B}\Gamma) + \alpha^2\|\Gamma\|^2 M_4^4)\|\mathrm{A}\|^2\|\hat{\theta}_n - \theta_n^\star\|^2 +$$

$$+ \alpha^2(\mathrm{E}\{\varphi_n\}^{\mathrm{T}}(\hat{\theta}_n - \theta_n^\star) - v_n)^2\|\Gamma\|^2\mathrm{Tr}[\mathrm{B}] + \mathrm{Tr}[\mathrm{Q}_w].$$

Further, taking the unconditional expectation from the last inequality we obtain for the mean value of the square of prediction errors with any $\rho > 0$

$$\mathrm{E}\{\|\hat{\theta}_{n+1} - \theta_{n+1}^\star\|^2\} \le b(\alpha, \rho)\mathrm{E}\{\|\hat{\theta}_n - \theta_n^\star\|^2\} + \alpha^2(1 + M_\varphi\rho)\|\Gamma\|^2\mathrm{Tr}[\mathrm{B}]\sigma_v^2 + \mathrm{Tr}[\mathrm{Q}_w],$$

where

$$b(\alpha,\rho) = (1 - 2\alpha\lambda_{\min}(\mathrm{B}\Gamma) + \alpha^2\|\Gamma\|^2 M_4)\|\mathrm{A}\|^2 + \alpha^2(M_\varphi + \frac{1}{\rho})M_\varphi\|\Gamma\|^2\mathrm{Tr}[\mathrm{B}]. \qquad (18)$$

The next result follows directly from the last inequality.

Theorem 6: Let Assumption (A) and (Bii) be fulfilled. *Then* the mean value of the square of prediction errors of the estimates $\{\hat{\theta}_n\}$ generated from the algorithms (17) $\forall\rho > 0$ and sufficiently small α: $b(\alpha,\rho) < 1$, satisfy the next inequality

$$\mathrm{E}\{\|\hat{\theta}_{n+1} - \theta^\star_{n+1}\|^2\} \le \frac{\mathrm{Tr}[\mathrm{Q}_w] + \alpha^2(1 + M_\varphi\rho)\|\Gamma\|^2\mathrm{Tr}[\mathrm{B}]\sigma_v^2}{1 - b(\alpha,\rho)} + b(\alpha,\rho)^n\mathrm{E}\{\|\hat{\theta}_0 - \theta^\star_0\|^2\},$$

where the constant $b(\alpha,\rho)$ is determining from (18).

Let's suppose that $\Gamma = \mathrm{B}^{-1}$, $\|\mathrm{A}\|^{-2} = 1 + \mathcal{O}(\alpha^3)$ $\mathrm{E}\{\|\hat{\theta}_0 - \theta^\star_0\|^2\} = 0$. Denote

$$r(\rho) = \frac{M_4 + (M_\varphi + 1/\rho)M_\varphi\mathrm{Tr}[\mathrm{B}]}{2\lambda^2_{\min}(\mathrm{B})}.$$

In this case for the sufficiently small α result of Theorem 6 leads to the inequality

$$\mathrm{E}\{\|\hat{\theta}_{n+1} - \theta^\star_{n+1}\|^2\} \le D(\alpha,\rho) + \mathcal{O}(\alpha^2),$$

where

$$D(\alpha,\rho) = \frac{1}{2}\mathrm{Tr}[\mathrm{Q}_w]\Big(\frac{1}{\alpha} + r(\rho) + \left(r(\rho)^2 + \frac{(1 + M_\varphi\rho)\mathrm{Tr}[\mathrm{B}]\sigma_v^2}{\mathrm{Tr}[\mathrm{Q}_w]\lambda^2_{\min}(\mathrm{B})}\right)\alpha\Big). \qquad (19)$$

From the last expression the trade–offs between the filtering ability and noise sensitivity are clearly visible. If $M_\varphi = 0$ then the similar result one can get directly from the [28] (theorem 4, p. 764) for the tracking problem solving.

Minimizing the expression for $D(\alpha,\rho)$ with respect to α one obtain

$$\alpha^\star = \left(r(\rho^\star)^2 + \frac{(1 + M_\varphi\rho^\star)\mathrm{Tr}[\mathrm{B}]\sigma_v^2}{\mathrm{Tr}[\mathrm{Q}_w]\lambda^2_{\min}(\mathrm{B})}\right)^{-\frac{1}{2}},$$

where $\rho^\star$ is the minimum point of the function

$$\bar{D}(\rho) = \frac{1}{2}\mathrm{Tr}[\mathrm{Q}_w]\left(r(\rho) + 2\sqrt{r(\rho)^2 + (1 + M_\varphi\rho)\mathrm{Tr}[\mathrm{B}]\sigma_v^2\lambda^{-2}_{\min}(\mathrm{B})/\mathrm{Tr}[\mathrm{Q}_w]}\right). \qquad (20)$$

If $M_\varphi = 0$ then the function $D(\alpha,\rho)$ is independent of ρ. Hence we get

$$\alpha^\star = 2\lambda^2_{\min}(\mathrm{B})\sqrt{\frac{\mathrm{Tr}[\mathrm{Q}_w]}{M_4^2\mathrm{Tr}[\mathrm{Q}_w] + 2\lambda^2_{\min}(\mathrm{B})\mathrm{Tr}[\mathrm{B}]\sigma_v^2}}$$

and

$$\bar{D}^\star = \frac{\mathrm{Tr}[\mathrm{Q}_w]}{4\lambda^2_{\min}(\mathrm{B})}\left(M_4 + 2\sqrt{M_4^2 + 2\lambda^2_{\min}(\mathrm{B})\mathrm{Tr}[\mathrm{B}]\sigma_v^2/\mathrm{Tr}[\mathrm{Q}_w]}\right).$$

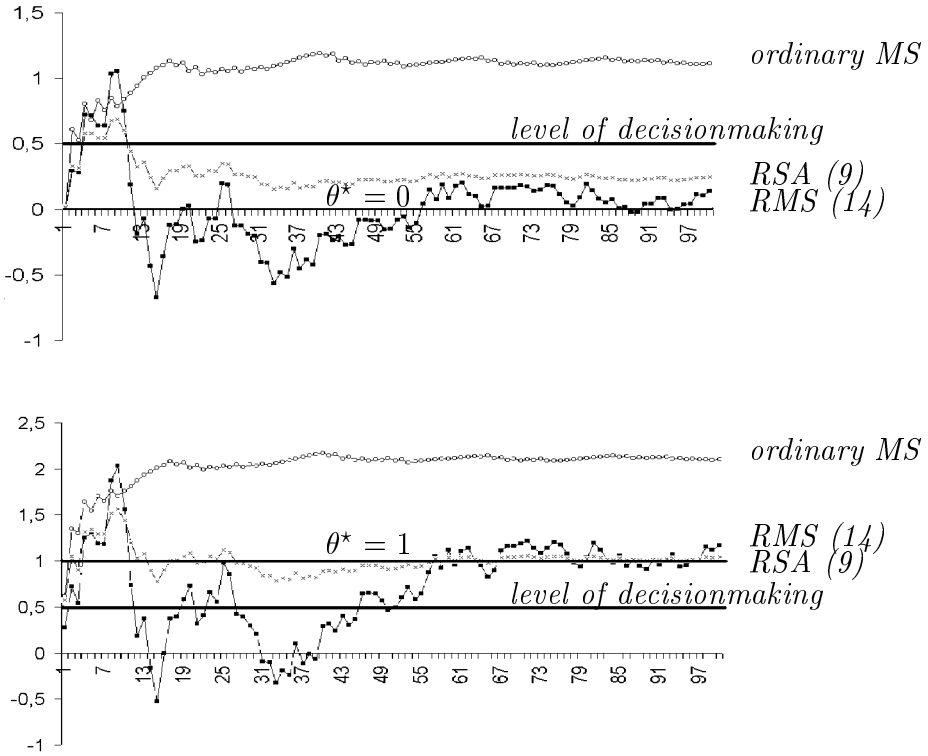

Fig. 1. Detecting useful signal

In the case when regressor's vectors can be arbitrary chosen from d–dimensional cube $[-1,1]^d$ it is easy to conclude from the last equation that random vectors with the probabilistic distribution of Bernoulli (± 1) are more appropriate as regressors.

Let $d = 1$, $\sigma_w^2 = \mathrm{Tr}[\mathrm{Q}_w] << \sigma_v^2$ and $\{\varphi_n\}$ be a scalar independent Bernoulli process ($\mathrm{Prob}\{\varphi_n = \bar{\varphi}\} = \mathrm{Prob}\{\varphi_n = -\bar{\varphi}\} = 1/2$). We have $\alpha^\star \mathrm{A}\Gamma \approx \sigma_w/|\bar{\varphi}|\sigma_v$. Note that this value equals approximately the limit value of Kalman coefficient for the optimal Kalman filter when noises $\{v_n\}$ are independent and equal $\pm\sigma_v$ with the identical probability.

VII. Experimental Results

A. Signal detecting

Let's consider a problem of the detecting of a known signal φ_n which was stated in introduction. At each time moment n the measurement y_n is made with additive bounded noise v_n:

$$y_n = \varphi_n \theta^\star + v_n, \; n = 1, 2, \ldots.$$

The case $\{\theta^\star = 1\}$ corresponds to the situation when a signal is present in the receiver and $\{\theta^\star = 0\}$ is absent. The our objective is to determine the value of $\theta^\star$ based on available input-output measurements y_i, φ_i, $i \leq N$.

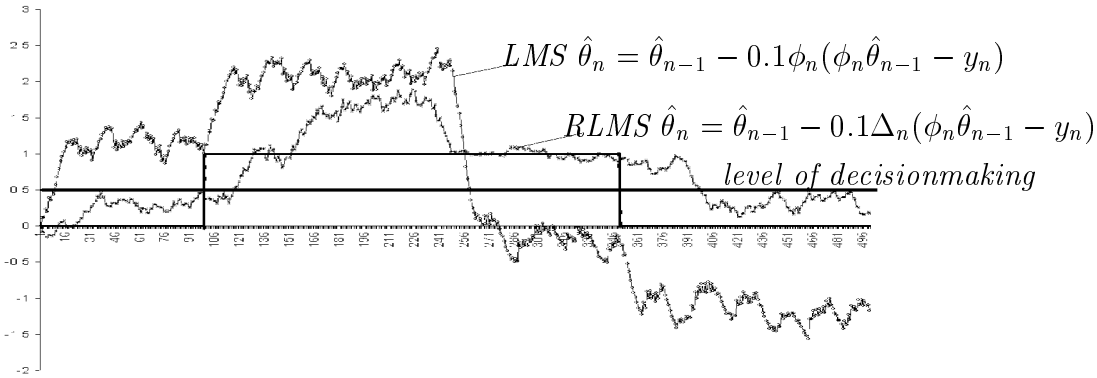

Fig.2 Tracking useful signal

At the simulation on the computer the useful signal $\{\varphi_n\}$ was selected as uniformly distributed on an interval $[0.5, 1.5]$ and was also observed on a hum of noises, which were determined from an "unknown but bounded" deterministic function, $|v_n| \leq C_v = 2$. Fig. 1 shows the typical trajectories of estimates which were generated by the three algorithms in the two cases when useful signal was (or wasn't) present in the receiver. Some deterministic sequence with average value more than +1 (displaced from zero) was used as an observation noise in the simulation on computer. The level of observation noise is so high that estimates generated from the ordinary mean squares (MS) algorithm exceed the decisionmaking level almost always without the dependence from presence or absence of a signal, while the algorithms (9)and (14) give the correct answers after 50 iterations.

In the following example the useful signals $\{\varphi_n\}$ and observation noises $\{v_n\}$ satisfy above conditions but useful signals "are actuated" in an observation channel temporarily, though its value are accessible to the experimenter during the all observation period. The problem is to design a rule, on which at each moment one could answer the question: does the useful signal acts in the observation channel or there is just noise being registered. The observation noises were formed by some deterministic sequence which average value is zero. But first half of its values are the positive numbers with average value more than +1 and last half are the negative numbers with average value less than −1. Comparison RLMS with the ordinary least mean squares algorithm (LMS) shows that RLMS tracking the changes of useful signal parameters more precisely, it gives only 13% of incorrect answers, see Fig.2.

B. Filtering

Let's consider the simple scalar case: $d = 1$. Investigated process $\{\theta_n^\star\}$ is generated through a stable linear filter (16)

$$\theta_{n+1}^\star = 0.9999\theta_n^\star + w_{n+1},\ n = 0, 1, \ldots$$

with A = 0.9999 and $\theta_0^\star = 0$ from the values of an independent process $\{w_n\}$ uniformly distributed on interval $[-\frac{1}{3}, \frac{1}{3}]$: $\mathrm{E}\{w_n\} = 0$, $\mathrm{E}\{w_n^2\} = \frac{2}{81}$. At each time moment n one can measure the values φ_n and y_n which are related with $\theta_n^\star$ by the equation (15), where v_n represent immeasurable bounded noise (disturbance): $|v_n| \leq 2$.

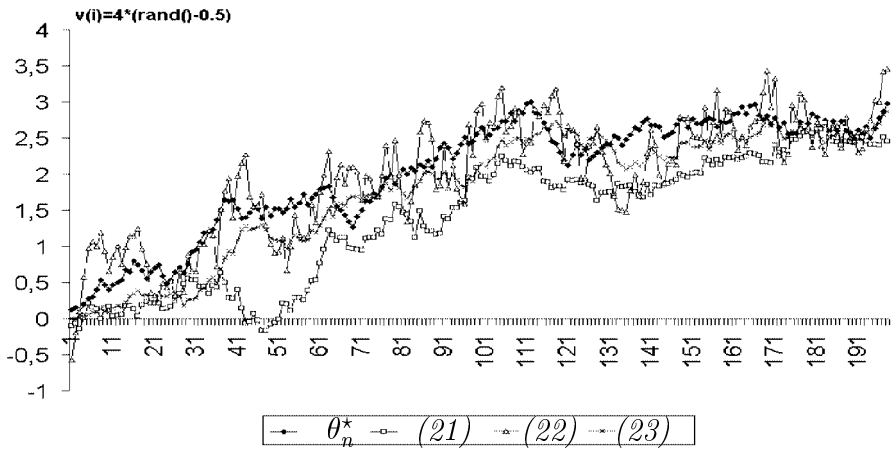

Fig. 3. Estimates of filtering at white noises case

At the computer simulation the random values $\{\varphi_n\}$ was selected as uniformly distributed on an interval $[0.5, 1.5]$. Observations of the process $\{\theta_n^\star\}$ were made on the time interval from $n = 1$ to 199. A quality of the prediction was determined by

$$\tilde{D}(\{\hat{\theta}_n\}) = \frac{1}{199}\sum_{n=1}^{199} \|\hat{\theta}_n - \theta_n^\star\|^2.$$

Minimizing function $\bar{D}(\rho)$ (see (20)) with respect to ρ one can obtain $\rho^\star = 0.269$. Hence for the RLMS algorithm (15) an optimal step–size $\alpha^\star$ is 11.3808, $\Gamma = 1/48$, and the correspondence value of $D(\alpha^\star, \rho^\star)$ is 1.3699 that less than $\sigma_v^2 = 4$.

Under different kind of observation noises we compare the performances of three trajectories of prediction estimates which were generated by the RLMS

$$\hat{\theta}_{n+1} = 0.9999(\hat{\theta}_n - 0.2371(\varphi_n - 1.0)(\varphi_n\hat{\theta}_n - y_n)), \tag{21}$$

by the ordinary least mean squares algorithm (LMS)

$$\hat{\theta}_{n+1} = 0.9999(\hat{\theta}_n - 0.2371\varphi_n(\varphi_n\hat{\theta}_n - y_n)), \tag{22}$$

and by the Kalman filter (KF)

$$\hat{\theta}_{n+1} = 0.9999\hat{\theta}_n - k_n\varphi_n(\varphi_n\hat{\theta}_n - y_n), \tag{23}$$

$$k_n = \frac{0.9999\gamma_{n-1}}{\frac{16}{3} + \gamma_{n-1}\varphi_n^2},\ \gamma_n = \gamma_{n-1}0.9999^2 - \frac{\varphi_n^2\gamma_{n-1}^2}{\frac{16}{3} + \gamma_{n-1}\varphi_n^2} + \frac{2}{81},\ \gamma_0 = 0.$$

Fig. 3 and 4 are showing in the typical cases of them behaviors.

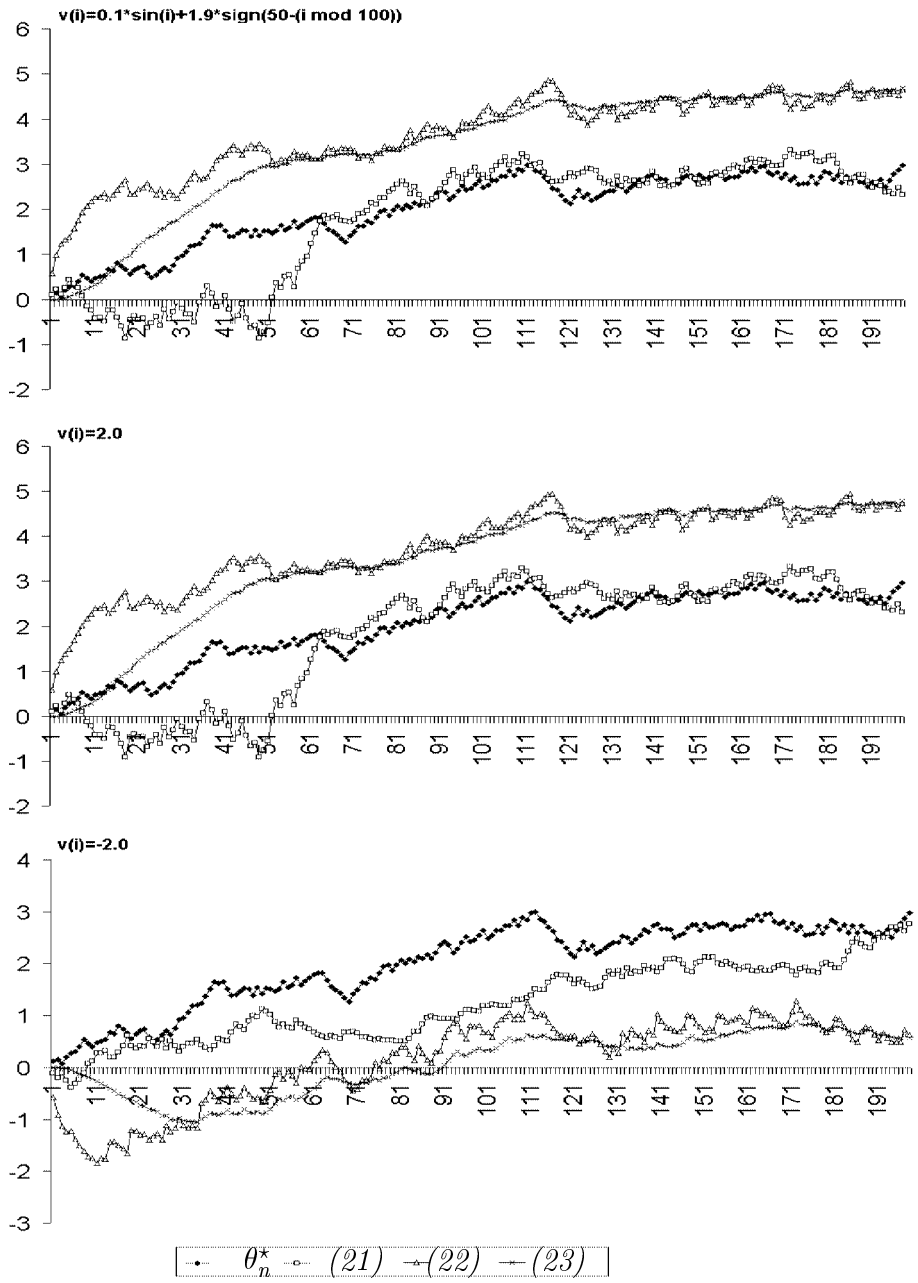

Fig. 4. Estimates of filtering at "unknown but bounded" non–random noises

The numerical results are summarized in Table I. It is well–known that KF estimates (23) give an optimal performance index when $\{v_n\}$ and $\{w_n\}$ are Gaussian white noises, behavior of the LMS estimates (22) is sufficiently good when $\{v_n\}$ and $\{w_n\}$ are centered independent random processes. Therefore behavior of the estimates (22),(23) is sufficiently good in the first case though the level of the observation noises $\{v_n\}$ is high (see Fig. 3). In the second case when the observation noises $\{v_n\}$ is zero–mean but non–regular and in the case of unknown constant noise the mean values of the prediction errors of algorithms (22) and (23) are equal approximately to the square of the level of observation noises (see Fig. 4). But the performance indexes of the RLMS estimates are approximately identical in all considered cases and they are in 5-7 times less than the square of the level of the observation noises.

TABLE I
AVERAGED ERRORS OF VARIOUS ALGORITHMS

	$\tilde{D}((21))$	$\tilde{D}((22))$	$\tilde{D}((23))$
$v_n = 4.0 * (rand() - 0.5)$	0.5309	0.1803	0.1256
$v_n = 0.1 * sin(n) + 1.9 * sign(50 - n \bmod 100)$	0.5700	2.8254	2.2640
$v_n = 2.0$	0.5954	3.1387	2.5335
$v_n = -2.0$	0.7826	3.4989	3.9582

VIII. QUANTUM COMPUTER AND RANDOMIZED ALGORITHMS

During the last couple of years the area of the randomized stochastic approximation algorithms has been constantly getting wider. The algorithms' implementation simplicity allows to use them not only in special computing devices but also in the design of a classical type electronic devices with an immediate use of "simultaneous perturbation" principal [29]. Effectiveness of a method is explained from two basic moments:

- Only one or two measurements of function values are required for the calculation of the approximate value of a multivariable function gradient.
- Algorithms have robustness qualities of a very high degree in a sense that the estimation algorithms convergence is proved under the "almost" arbitrary noises in function measurements.

The main problem of a convergence substantiation which arises in practical use of SPSA algorithms is the way of generation of trial simultaneous perturbations. They are

to be independent from noises in the observation channel as well as the components of a trial simultaneous perturbation vector should be independent between themselves. The efficiency of offered algorithms reduces when we use a classical computer[2] that sequentially executes elementary operations one by one (in comparison with theoretical results). The title "simultaneous perturbation" itself has an insistent demand for practical use to be "parallel". At the same time the problem of a parallel calculations organization on computers of classical type is difficult when it deals with the large dimensionality of a function arguments vector (five, tens, ..., thousands).

In the given section, the model of "hypothetical" quantum computer will be considered. Offered above algorithms can be realized most efficiently on this new kind of computers. In other words, it will be considered below is an example of design of an electronic quantum device that calculates for "one step" the "good" approximation of the gradient vector of an unknown multivariable function with rather high degree of an accuracy. The word "hypothetical" is consciously quoted since after the P. Shor report [30] on the Berlin mathematical congress in 1998 many serious authors began to write about quantum computers as an engineering of the nearest future. In August 2000 the first practical successive result was declared (Reuters, "IBM Says It Develops Most Advanced Quantum Computer").

A. The Quantum Circuit Model

Firstly let's describe the mathematical model of quantum computer producing calculation on determined circuits. There have been a considerable number of important developments in the field of an extension of classical information–theoretic concepts to a quantum–mechanical setting (see [31],[32]). The classical computer treats bits, receiving values from the set $\{0, 1\}$. It is equipped with a final set of schemes, which can be applied to the sets of bits. The quantum computer treats the quantum bits (or qubits), representing typically a two–state microscopic system, possibly an atom or nuclear spin or polarized photon, the behavior of which (e.g., entanglement, interference, superposition, stochasticity, ...) can be accurately explained using the rules of quantum theory [33] only. Mathematically, a state (more precisely, a pure state) of a qubit is a unit vector in the complex space $\mathbb{C}^2$ with inner product. The quantum states are invariant concerning multiplication by scalar value. Let's denote base vectors of this space $|0\rangle$, $|1\rangle$. Assume that the quantum computer is equipped with a discrete set of fundamental components, called by quantum schemes. Each quantum scheme is an unitary transformation, which acts on a fixed number of qubits. One of the fundamental principles of a quantum computer model is that the joint quantum state space of a system, consisting from k two–state systems, is the tensor products of their individual Hilbert spaces. Thus, the quantum state space of k qubits systems is the complex projective space $\mathbb{C}^{2^k}$. The basis vectors set of this state space can be parameterized by bits lines of length k

$$|b_1 b_2 \dots b_k\rangle = |b_1\rangle \otimes |b_2\rangle \otimes \dots \otimes |b_k\rangle.$$

[2]By term "classical" we mean "nonquantum" throughout this paper.

Let's assume, that the "classical" information, bits line of the length l, $l \leq k$, is a quantum computer input. Originally in the quantum calculation the last qubits with number $i > l$ get the value of $|0\rangle$. The executed circuit is created from the final quantity of quantum schemes operating with these qubits. At the end of calculation the quantum computer comes into some state, which is a unit vector in the space $\mathbb{C}^{2^k}$. This state can be presented as

$$W = \sum_s \psi_s |s\rangle,$$

where the summation by s is done for all binary lines of length k, $\psi_s \in \mathbb{C}$, $\sum_s |\psi_s|^2 = 1$. Sizes $|\psi_s|$ are named as probabilistic amplitudes, and W is named as a superposition of the basis vectors $|s\rangle$. The quantum mechanics uncertainty Heisenberg principle states that it is impossible to measure the quantum system received state precisely. However some capabilities to execute the measurement for all qubits (or subset of qubits) exist. The quantum system state space is a Hilbert space. The state measurement concept is equivalent to a scalar product in this Hilbert space with some specific vector V

$$\langle V, W\rangle \ (= \langle V|W\rangle).$$

Usually the projection on some basis state is used for measuring. The measurement result is an outcome of calculation.

B. Quantum Circuit for the Approximation of a Function Gradient

Let's consider the part of the SPSA algorithm (1) with Bernoulli trial simultaneous perturbation and $\mathcal{K}_n(x) = x$, which calculates the approximation of a gradient vector $\nabla f(X) \approx (\hat{g}^{(1)}(X), \hat{g}^{(2)}(X), \ldots, \hat{g}^{(d)}(X))^T$ of a function $f(\cdot) : \mathrm{R}^d \to \mathrm{R}$ at a point $X \in \mathrm{R}^d$. Suppose that p binary digit is used for representation of numbers in our computer (in modern computers, most frequently $p = 16, 32, 64$) and $k = p \times d$. For all $x \in \mathrm{R}$ binary representation of x in the form of the bits line is denoted as $s_x = \overline{b_x^{(1)} \ldots b_x^{(p)}}$. Let's assume that we have some quantum circuit that calculates function $f(X)$ values. To be more precise, it is possible to consider that the unitary transformation is given: $U_f : \mathbb{C}^{2^k} \to \mathbb{C}^{2^k}$. It maps one basis element

$$|s_{x^{(1)}} \ldots s_{x^{(d)}}\rangle = |b_{x^{(1)}}^{(1)} \ldots b_{x^{(1)}}^{(p)} \ldots b_{x^{(d)}}^{(1)} \ldots b_{x^{(d)}}^{(p)}\rangle$$

to another basis element

$$|s_{f(X)} 00 \ldots 0\rangle = |b_{f(X)}^{(1)} \ldots b_{f(X)}^{(p)} 00 \ldots 0\rangle = U_f |s_{x^{(1)}} \ldots s_{x^{(d)}}\rangle$$

for all $X = (x^{(1)}, \ldots, x^{(d)})^{\mathrm{T}} \in \mathrm{R}^d$

Assume that our "hypothetical" quantum calculator contains at least three k–qubits registers: *input* $\mathcal{I}$, transferring the "classical" input data to the "quantum"; *worker* W, permitting to manipulate the "quantum" data; and *simultaneous perturbation* Δ. For the quantum circuit design, which realizes the approximate calculation of a gradient vector

of a function $f(\cdot)$, several "standard" quantum (unitary) transformations are required. They can be applied to the data stored in registers. The outcome of transformation is saved in the register to which transformation was applied.

Sum U_{+R}. It adds to the vector another one which is stored in register R.

Turn of the first qubits $U_{R_{1,p+1,\ldots,(d-1)p+1}}$. It transforms the state of qubits numbered as $1, p+1, \ldots, (d-1)p+1$ to the state $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$.

Shift by j *qubits* U_{S_j}. It shifts state vector by j qubits, adding new $|0\rangle$.

For the approximate calculation of a function $f(\cdot)$ gradient at the point X the following algorithm can be used.

1. To send zero to all three registers $\mathcal{I}, W, \Delta$.
2. To submit the line of bits s_X to an input register $\mathcal{I}$. To transform the register Δ :
$$\mathcal{I} := |s_x\rangle,\ \Delta := U_{R_{1,p+1,\ldots,(d-1)p+1}}\Delta.$$
3. To calculate function's value $f(X + 2^{-j}\Delta),\ 0 \le j \le p-1$ in register W
$$W := U_f U_{+\mathcal{I}} U_{S_j} U_{+\Delta} W,$$
4. To measure the outcome of calculations
$$\hat{g}^{(i)}(X) = \langle U_{S_{-(i-1)p}}\Delta, W\rangle\ (= \langle U_{S_{-(i-1)p}}\Delta | U_f U_{+\mathcal{I}} U_{S_j} U_{+\Delta} | W\rangle),\ i = 1, 2, \ldots, d.$$

The last expressions are equivalent to the function $f(\cdot)$ gradient approximation expression
$$\nabla f(X) \approx \Delta f(X + 2^{-j}\Delta).$$

It is not so hard to show that the approximate accuracy represents smallness size $o(2^{-j})$. If it was known that the function $f(\cdot)$ has continuous partial derivatives up to the order $\ell > 3$ inclusively then it is possible to make measurements (step 4) not with vector Δ but with some extent transformation $U_{\mathcal{K}}\Delta$ by the analogy to Theorem 2. Probably, an order of approximation error will decrease.

C. One Approach of Creating the Intelligent System

Let's look at some achievements in the contemporary mathematics and computer engineering of the past century which allow us to look with a great optimism at the solution of the problem in the future. The first computers were specialized for solution of the concrete tasks. It was based both on the economic reasons and on the weak development of electronic components. For a short time the engineering progress has allowed to develop the universal computer and theoretical development has made it possible to effectively solve many of the tasks set. The final conception of the universal programming language

appeared in the middle of 70th and during the 80-90th years the programming technology has been on a top of its development. However we are still left with considerable problems tied with the machine's ability to make a quick and effective decision in tasks with a large data dimension (matrix transformation, large number factorization, Fourier or Laplace transformation, calculation of function convolution and so on). A lot of them form a minimal base of any artificial intelligence system.

Today the most practical way is a creation of fast computing devices for solving concrete tasks based on a possibility of the development of "parallel" fastest algorithms for processing the large sizes of data. Two directions are the most perspective. The first one (almost developed) is to create the classical electronic devices as VLSIC using the high-speed technologies. Such devices will allow in the real time to convert matrixes of the large dimensionalities which is necessary in many tasks of recognition. Second and even more perspective direction in the development of computing devices in a future is the use of "quantum computers". Channels for data transfer in such computer will be more informative. For implementation of a fixed structure of parallel calculations the special set of atoms (molecule) will be used. Probably in a future it will be possible to attain the highest speed of calculations by using the equivalence of the descriptions of many phenomenons in micro and macro world.

Despite of wide use of specialized parallel processors many reseachers agree that we still don't have the common methodology for a development of appropriate algorithms and devices. Probably we move to the necessity to comprehensively analyze a lot of artificial intelligence tasks. We need to somehow classify their settings and possible algorithms for their solution and to create a convenient formal logical means for the description of the tasks, algorithms and data.

Let's take a look at one of the possible solutions of creating some kind of an intelligent system. The term "intelligent" means here the ability of a system to adapt to the real world conditions by making the effective choice of the task to solve in the current moment. Such system could consist of two parts : internal one and ensuring interaction with the external world. Set of sensor controls and power installations controlling various organs of the device will link it with the external world. The internal part will match the following conditions:

- for each considered particular task the system has the particular device which is able to optimize the solving of this task by an appropriate choosing of the system's parameter from some final set,
- information from external world must be delivered to all such devices simultaneously.

By this conditions the system is able to interpret and to process the data received from external world through different devices simultaneously (in parallel).

We can imagine many ways of realization of such system. Further we'll discuss one of them. Suppose that we want to have a system able to solve any problem of some final set:

$$\{P_1, P_2, \ldots, P_m\}.$$

Here the m is some integer (may be equal $1, 2, 3$ or sufficiently large $100000, 1000000$). For instance our "hypothetical" device can be presented as some big complicated molecule consisting of smaller parts (particular devices)

$$\{D_1, D_2, \ldots, D_m\}$$

curled up in spiral or ball one by another. It is possible to consider each of them as some "quantum computer". Such form of spatial representation allows to suppose that some biological or physical influence acts to the all of particular devices simultaneously. We'll assume that the behavior of each of units is determined by their internal structure, input flow of data and some parameters $\theta^{(i)}$ from some set $\Theta^{(i)}$ with final dimension

$$D_i = D_i(\theta^{(i)}),\ \theta^{(i)} \in \Theta^{(i)},\ i = 1, 2, \ldots, m.$$

Suppose that each of units is equipped with special register which indicate the performance index of data w processing

$$f_i = f_i(w, \theta^{(i)}),\ i = 1, 2, \ldots, m.$$

The parameters of data processing can be passed to devices among with the input data. The objective of complete system is to make choice of the unit with higher performance index and to give them the foreground process. The general system's parameter is denoted by

$$\theta = \begin{pmatrix} \theta^{(1)} \\ \theta^{(2)} \\ \vdots \\ \theta^{(m)} \end{pmatrix},\ \theta \in \Theta = \Theta^{(1)} \otimes \cdots \otimes \Theta^{(m)}.$$

Let's define the function

$$F(w, \theta) = -\sum_{i=1}^{m} f_i(w, \theta^{(i)}).$$

This scheme satisfies both of above conditions.

Here we came to the concept of the "informational resonance". If a partial performance index f_i of some units "i" tends to +infinity then we'll say that unit "i" has a resonance. After the data is delivered from all the sensors to the all devices, these devices begin to work and solve each its own task. There can be three different situations:

- only one device has resonance,
- several devices have resonance,
- none of the devices has resonance.

In the first case the resonant device accepts control on the system. In the second case the system should determine which device to use. It can be done by choosing the device with the greatest value of f_i . The third case means that none of the devices has resonance with the data but we need to choose the strong rule of a system behavior. Assume that this rule is determined by setting the system parameter $\theta \in \Theta \subset \mathrm{R}^d$. We have a problem of the stochastic multi-dimensional optimization. There is an effective SPSA algorithm (1) for solving this problem. We need to include the quantum circuit of them into our intelligent system. And in the third case it would be chosen as foreground process.

IX. Conclusion

The stochastic optimization considerably expands the range of practical problems, for which it is possible to find the precisely optimum solution regarding standard deterministic methods. The stochastic optimization algorithms effectively allow to decide problems in such areas as information network analysis, based simulation optimization, a pattern recognition and classification, neuron networks training, image processing and a non-linear control. It is expected that the stochastic optimization role will continue to grow together with thickening of modern systems so as the population increment and the natural recourse exhaustion initiates using more heavy technologies in spheres where they were unnecessary before. Since the first quantum computers working on stochastic principles appeared already, the logic of a modern computers development also lead to replacement of conventional deterministic algorithms by stochastic ones.

Appendix

In this appendix we give the proofs of Theorems 1,2,4 and 5.

A. Proof of Theorem 1

Firstly let's consider the algorithm (5). Using properties of projection, at rather large n, when $\theta^\star = \theta^\star(f(\cdot)) \in \Theta_n$, it is easy to get the following inequality

$$\|\hat{\theta}_n - \theta^\star\|^2 \leq \|\hat{\theta}_{n-1} - \theta^\star - \frac{\alpha_n}{\beta_n}\mathcal{K}_n(\Delta_n)y_n\|^2 \ .$$

Applying the operation of conditional expectation by σ–algebra $\mathcal{F}_{n-1}$ to the last inequality, we have

$$\begin{aligned} &\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2|\mathcal{F}_{n-1}\} \leq \|\hat{\theta}_{n-1} - \theta^\star\|^2 - \\ &-2\frac{\alpha_n}{\beta_n}\langle\hat{\theta}_{n-1} - \theta^\star \ , \mathrm{E}\{\mathcal{K}_n(\Delta_n)y_n|\mathcal{F}_{n-1}\}\rangle + \\ &+\frac{\alpha_n^2}{\beta_n^2}\mathrm{E}\{y_n^2\|\mathcal{K}_n(\Delta_n)\|^2|\mathcal{F}_{n-1}\}. \end{aligned} \tag{24}$$

By virtue of Theorem about mean value from the condition (A.2) for the function $F(\cdot,\cdot)$ it follows

$$|F(w,x)-F(w,\theta^\star)| \le \frac{1}{2}\nabla_\theta F(w,\theta^\star)^2 + (A+\frac{1}{2})\|x-\theta^\star\|^2,\ x\in\mathbb{R}^d.$$

From here by virtue of the function $F(\cdot,\theta)$ uniform boundedness it is obtained

$$F(w,\hat\theta_{n-1}+\beta_n x)^2 \le (\nu_1 + (2A+1)(\|\hat\theta_{n-1}-\theta^\star\|^2 + \|\beta_n x\|^2))^2$$

uniformly on $w\in\mathbb{W}$. By virtue of the condition (4) we get

$$\mathrm{E}\{v_n^2\|\mathcal{K}_n(\Delta_n)\|^2|\mathcal{F}_{n-1}\} \le \sup_x \mathcal{K}_n(x)^2\xi_n^2.$$

For the last term in the right part (24) from the last two inequalities, taking into account boundedness of vector functions $\mathcal{K}_n(\cdot)$ and compactness of their support, we have

$$\mathrm{E}\{\|y_n\|^2\|\mathcal{K}_n(\Delta_n)\|^2|\mathcal{F}_{n-1}\} \ \le\ 2\mathrm{E}\{v_n^2\|\mathcal{K}_n(\Delta_n)\|^2|\mathcal{F}_{n-1}\} +$$

$$+\,2\int\int F(w,\hat\theta_{n-1}+\beta_n x)^2\|\mathcal{K}_n(x)\|^2\mathrm{P}_n(dx)\mathrm{P}_w(dw)\ \le$$

$$\le\ C_1 + C_2((d_n^2+1)\|\hat\theta_{n-1}-\theta^\star\|^2+\beta_n^2) + C_3\xi_n^2.$$

Here and below $C_i,\ i=1,2,\ldots$ are designated as some positive constants.

Further we shall consider

$$\beta_n^{-1}\mathrm{E}\{y_n\mathcal{K}_n(\Delta_n)|\mathcal{F}_{n-1}\}\ =$$

$$=\ \beta_n^{-1}\int\int F(w,\hat\theta_{n-1}+\beta_n x)\mathcal{K}_n(x)\mathrm{P}_n(dx)\mathrm{P}_w(dw)+ \tag{25}$$

$$+\beta_n^{-1}\mathrm{E}\{v_n\mathcal{K}_n(\Delta_n)|\mathcal{F}_{n-1}\}.$$

For the second term by virtue of condition (4) and independence of v_n and Δ_n the following could be obtained

$$\mathrm{E}\{v_n\mathcal{K}_n(\Delta_n)|\mathcal{F}_{n-1}\} = \mathrm{E}\{v_n|\mathcal{F}_{n-1}\}\int\mathcal{K}_n(x)\mathrm{P}_n(dx) = 0.$$

The function $\nabla_\theta F(\cdot,\theta)$ is uniform bounded, which implies that

$$\int_{\mathbb{R}^p}\nabla_\theta F(w,x)\mathrm{P}_w(dw) = \nabla f(x).$$

Using the condition (4), we shall transform the first term in right part the (25) to the following

$$\beta_n^{-1}\int\int F(w,\hat\theta_{n-1}+\beta_n x)\mathcal{K}_n(x)\mathrm{P}_n(dx)\mathrm{P}_w(dw) = \nabla f(\hat\theta_{n-1}) +$$

$$\int\Big(\beta_n^{-1}\int F(w,\hat\theta_{n-1}+\beta_n x)\mathcal{K}_n(x)\mathrm{P}_n(dx)-\nabla_\theta F(w,\hat\theta_{n-1})\Big)\mathrm{P}_w(dw)=\nabla f(\hat\theta_{n-1})+$$

$$+\int\int\mathcal{K}_n(x)x^{\mathrm{T}}\int_0^1(\nabla_\theta F(w,\hat\theta_{n-1}+t\beta_n x)-\nabla_\theta F(w,\hat\theta_{n-1}))dt\mathrm{P}_n(dx)\mathrm{P}_w(dw).$$

For an absolute value of the second term in the last equality we have

$$\Big|\int\int\mathcal{K}_n(x)x^{\mathrm{T}}\int_0^1(\nabla_\theta F(w,\hat\theta_{n-1}+t\beta_n x)-\nabla_\theta F(w,\hat\theta_{n-1}))dt\mathrm{P}_n(dx)\mathrm{P}_w(dw)\Big|\leq$$

$$\leq\int\int\|\mathcal{K}_n(x)\|\|x\|A\|\beta_n x\|\mathrm{P}_n(dx)\mathrm{P}_w(dw)\ \leq\ C_4\beta_n$$

by the fulfillment of conditions (4) for $\mathcal{K}_n(\cdot)$ and (A.2) for any function $F(w,\cdot)$. Hence, for the second term in the right part of inequality (24) we get

$$-2\frac{\alpha_n}{\beta_n}\langle\hat\theta_{n-1}-\theta^\star,\mathrm{E}\{\mathcal{K}_n(\Delta_n)y_n|\mathcal{F}_{n-1}\}\rangle\leq-2\alpha_n\langle\hat\theta_{n-1}-\theta^\star,\nabla f(\hat\theta_{n-1})\rangle+$$

$$+2C_4\alpha_n\beta_n\|\hat\theta_{n-1}-\theta^\star\|.$$

Using assessments obtained above for the second and third terms of (24), we obtain

$$\mathrm{E}\{\|\hat\theta_n-\theta^\star\|^2|\mathcal{F}_{n-1}\}\leq\|\hat\theta_{n-1}-\theta^\star\|^2-2\alpha_n\langle\hat\theta_{n-1}-\theta^\star,\nabla f(\hat\theta_{n-1})\rangle+$$

$$+2C_4\alpha_n\beta_n\|\hat\theta_{n-1}-\theta^\star\|+\frac{\alpha_n^2}{\beta_n^2}\Big(C_1+C_2((d_n^2+1)\|\hat\theta_{n-1}-\theta^\star\|^2+\beta_n^2)+C_3\xi_n^2\Big).$$

Using the condition (A.1) for function $f(\cdot)$ and inequality

$$\|\hat\theta_{n-1}-\theta^\star\|\leq\frac{\varepsilon^{-1}\beta_n+\varepsilon\beta_n^{-1}\|\hat\theta_{n-1}-\theta^\star\|^2}{2},$$

which is true at any $\varepsilon>0$, we get

$$\mathrm{E}\{\|\hat\theta_n-\theta^\star\|^2|\mathcal{F}_{n-1}\}\leq\|\hat\theta_{n-1}-\theta^\star\|^2\Big(1-(2\mu-\varepsilon C_4)\alpha_n+C_2\alpha_n^2\beta_n^{-2}(d_n^2+1)\Big)+$$

$$+\varepsilon^{-1}C_4\alpha_n\beta_n^2+\frac{\alpha_n^2}{\beta_n^2}\Big(C_1+C_2\beta_n^2+C_3\xi_n^2\Big).$$

Let's select the ε so small that $\varepsilon C_4<\mu$ and let n be rather great. Using the conditions of Theorem 1 for the numeric sequences, we shall transform the last inequality into the form

$$\mathrm{E}\{\|\hat\theta_n-\theta^\star\|^2|\mathcal{F}_{n-1}\}\leq\|\hat\theta_{n-1}-\theta^\star\|^2(1-C_5\alpha_n)+C_6(\alpha_n\beta_n^2+\alpha_n^2\beta_n^{-2}(1+\xi_n^2)).$$

From here by virtue of Theorem 1 conditions $\sum_n\alpha_n=\infty$ and $\sum_n\frac{\alpha_n^2}{\beta_n^2}(1+\xi_n^2)<\infty$ a.s. it is shown that all conditions of Robbins–Siegmund Lemma [34] necessary for the convergence with the probability 1 are held and $\theta_n\to\theta^\star$ as $n\to\infty$. For the mean-square convergence

proof in an appropriate Theorem 1 conditions we shall take unconditional expectation from both parts of the last inequality

$$\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2\} \le \mathrm{E}\{\|\hat{\theta}_{n-1} - \theta^\star\|^2\}(1 - C_5\alpha_n) + C_6(\alpha_n\beta_n^2 + \alpha_n^2\beta_n^{-2}(1+\sigma_n^2)).$$

The convergence of estimates $\{\theta_n\}$ sequence to the point $\theta^\star$ in a mean-square sense follows from [35](ch.2,s.2).

The proof for the algorithm (2) is a little bit different.

B. Proof of Theorem 2.

The proof scheme in many respects repeats the proof of Theorem 1. At first let's consider the algorithm (5) . Using properties of projection and applying the operation of conditional expectation by σ–algebra $\mathcal{F}_{n-1}$, for rather large n, at which $\theta^\star \in \Theta_n$, we get

$$\mathrm{E}\{\|\hat{\theta}_n - \theta^\star\|^2|\mathcal{F}_{n-1}\} \le \|\hat{\theta}_{n-1} - \theta^\star\|^2 - \tag{26}$$

$$-2\alpha n^{-1+\frac{1}{2\gamma}}(\hat{\theta}_{n-1} - \theta^\star, \mathrm{E}\{y_n\mathcal{K}(\Delta_n)|\mathcal{F}_{n-1}\}) + \alpha^2 n^{-2+\frac{1}{\gamma}}\mathrm{E}\{y_n^2\|\mathcal{K}(\Delta_n)\|^2|\mathcal{F}_{n-1}\}.$$

By virtue of (6) and condition (7) we have $\int \mathcal{K}(x)\mathrm{P}(dx) = 0$. Thus, by the independence of Δ_n and v_n we derive $\mathrm{E}\{v_n\mathcal{K}(\Delta_n)|\mathcal{F}_{n-1}\} = 0$, and, hence,

$$\mathrm{E}\{y_n\mathcal{K}(\Delta_n)|\mathcal{F}_{n-1}\} = \int\int F(w, \hat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}x)\mathcal{K}(x)\mathrm{P}(dx)\mathrm{P}_w(dw).$$

Note that by virtue of (6) and (7) we obtain

$$\beta^{-1}n^{\frac{1}{2\gamma}}\int\sum_{|\bar{l}|\le\ell}\frac{1}{\bar{l}!}D^{\bar{l}}f(\hat{\theta}_{n-1})\beta^{|\bar{l}|}n^{-\frac{|\bar{l}|}{2\gamma}}x^{\bar{l}}\mathcal{K}(x)\mathrm{P}(dx) \;=\; \nabla f(\hat{\theta}_{n-1}).$$

From the definition of the function $f(\cdot)$ one can get

$$\beta^{-1}n^{\frac{1}{2\gamma}}\mathrm{E}\{y_n\mathcal{K}(\Delta_n)|\mathcal{F}_{n-1}\} = \nabla f(\hat{\theta}_{n-1}) +$$

$$+\beta^{-1}n^{\frac{1}{2\gamma}}\int(f(\hat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}x) - \sum_{|\bar{l}|\le\ell}\frac{1}{\bar{l}!}D^{\bar{l}}f(\hat{\theta}_{n-1})\beta^{|\bar{l}|}n^{-\frac{|\bar{l}|}{2\gamma}}x^{\bar{l}})\mathcal{K}(x)\mathrm{P}(dx).$$

Fulfillment of the condition (A.3) implies the inequality

$$\Big|\int(f(\hat{\theta}_{n-1} + \beta_n x) - \sum_{|\bar{l}|\le\ell}\frac{1}{\bar{l}!}D^{\bar{l}}f(\hat{\theta}_{n-1})\beta^{|\bar{l}|}n^{-\frac{|\bar{l}|}{2\gamma}}x^{\bar{l}})\mathcal{K}(x)\mathrm{P}(dx)\Big| \le$$

$$\le\; M\int\|x\beta n^{-\frac{1}{2\gamma}}\|^\gamma\|\mathcal{K}(x)\|\mathrm{P}(dx) \le M\bar{K}\beta^\gamma n^{-\frac{1}{2}}.$$

Using obtained above assessments for the second and third terms in the right part (26) from the fulfillment of a condition (A.1) for a function $f(\cdot)$ and inequality

$$\|\hat{\theta}_{n-1}-\theta^\star\| \le \frac{\varepsilon^{-1}n^{-\frac{\gamma-1}{2\gamma}}M\bar{K}\beta^{\gamma-1}+\varepsilon(n^{-\frac{\gamma-1}{2\gamma}}M\bar{K}\beta^{\gamma-1})^{-1}\|\theta_{n-1}-\theta^\star\|^2}{2},$$

fair at any $\varepsilon>0$ we derive

$$\mathrm{E}\{\|\hat{\theta}_n-\theta^\star\|^2|\mathcal{F}_{n-1}\} \le \|\hat{\theta}_{n-1}-\theta^\star\|^2-2\alpha\beta n^{-1}(\hat{\theta}_{n-1}-\theta^\star,\nabla f(\hat{\theta}_{n-1}))+$$

$$+2\alpha\beta n^{-1-\frac{\gamma-1}{2\gamma}}M\bar{K}\beta^{\gamma-1}\|\hat{\theta}_{n-1}-\theta^\star\|+\alpha^2n^{-2+\frac{1}{\gamma}}\mathrm{E}\{y_n^2\|\mathrm{K}(\Delta_n)\|^2|\mathcal{F}_{n-1}\} \le$$

$$\le \|\hat{\theta}_{n-1}-\theta^\star\|^2(1-\alpha\beta(2\mu-\varepsilon)n^{-1})+n^{-2+\frac{1}{\gamma}}(\alpha\beta^{2\gamma-1}\varepsilon^{-1}M^2\bar{K}^2+$$

$$+\alpha^2\hat{K}\chi(\int\int F(w,\hat{\theta}_{n-1}+\beta n^{-\frac{1}{2\gamma}}x)^2\mathrm{P}(dx)\mathrm{P}_w(dw)+\mathrm{E}\{v_n^2|\mathcal{F}_{n-1}\})).$$

As in the proof of Theorem 1 from condition (A.2) we obtain

$$|F(w,\hat{\theta}_{n-1}+\beta n^{-\frac{1}{2\gamma}}x)| \le \sqrt{\nu_1}+(2A+1)(\|\hat{\theta}_{n-1}-\theta^\star\|^2+\|\beta n^{-\frac{1}{2\gamma}}x\|^2)$$

uniformly on $w\in\mathbb{W}$. Taking into account the unconditional expectation we conclude

$$\mathrm{E}\{\|\hat{\theta}_n-\theta^\star\|^2\} \le \mathrm{E}\{\|\hat{\theta}_{n-1}-\theta^\star\|^2\}(1-\psi n^{-1}+o(n^{-1}))+C_1n^{\frac{1}{\gamma}-2}+o(n^{\frac{1}{\gamma}-2}),$$

where $\psi=\alpha\beta(2\mu-\varepsilon)$, $C_1=\alpha\beta^{2\gamma-1}\varepsilon^{-1}M^2\bar{K}^2+\alpha^2\hat{K}\chi(\nu_1+\sigma_1^2)$. By lemma 9 from [35] if $\psi>(\gamma-1)/\gamma$ then for arbitrary $\varepsilon>0$ we have

$$n^{1-\frac{1}{\gamma}}\mathrm{E}\{\|\hat{\theta}_n-\theta^\star\|^2\} \le C_1(\alpha\beta(2\mu-\varepsilon)-\frac{\gamma-1}{\gamma})^{-1}+o(1). \tag{27}$$

The inequality $\psi>(\gamma-1)/\gamma$ is equvivalent to the condition $2\mu\alpha\beta>(\gamma-1)/\gamma$.

The proof for the algorithm (2) is different in some details. In particulary it is nessesary to use the inequality

$$\frac{1}{2}(F(w_1,\theta+x)-F(w_2,\theta-x))^2 \le \nu_2+2(2A+1)^2(\|x\|^2+\|\theta-\theta^\star\|^2)^2,$$

which is held uniformly on $w\in\mathbb{W}$. In a result of the proof is deduced inequality which is similar to (27) with the constant $C_2=\alpha\beta^{2\gamma-1}\varepsilon^{-1}M^2\bar{K}^2+\alpha^2\hat{K}\chi(\nu_2+\sigma_2^2/2)$ instead the C_1.

If we want to compare the constants C_1 and C_2 then we can to use the formula $C_1=C_2+\hat{K}\alpha^2\chi(\nu_1+\sigma_1^2-\sigma_2^2/2-\nu_2)$.

The right part of the (27) (or of the similar inequality for the algorithm (2)) is a function of α, β and ε. Optimizing by this parameters we can to derive the optimal values $\alpha^\star$, $\beta^\star$ $\varepsilon^\star=2\mu/\gamma$.

The proof of Theorem 2 and the second remark to it is completed.

C. Proof of Theorem 4

Let $\eta_n = \hat{\theta}_{n-1} - \theta^\star_n$, $\xi_n = v_n - \mathrm{E}\{\varphi_n\}^{\mathrm{T}}\eta_n$, $\mathrm{D}_n = (\hat{\theta}_n - \theta^\star)(\hat{\theta}_n - \theta^\star)^{\mathrm{T}}$.

Rewrite the matrix equation (12) as

$$(\Gamma\,\mathrm{B} - \frac{1}{2}\mathrm{I})\mathrm{S} + \mathrm{S}(\mathrm{B}\,\Gamma - \frac{1}{2}\mathrm{I}) = \Gamma\mathrm{R}\Gamma.$$

Since $-\Gamma\,\mathrm{B} + \frac{1}{2}\mathrm{I}$ is a Hurwitz matrix, by Lyapunov's Lemma there exists a positive definite matrix S which is a solution of matrix equation (12).

From (8) and (9) taking the conditional expectation sequentially with respect to σ-algebra $\tilde{\mathcal{F}}_{n-1}$ and $\hat{\mathcal{F}}_{n-1}$ we conclude in view of assumption (A) and the second part of assumption (Bii) that

$$\mathrm{E}\{\mathrm{D}_n|\hat{\mathcal{F}}_{n-1}\} \le \mathrm{D}_{n-1} - \alpha_n(\Gamma\,\mathrm{B}\mathrm{D}_{n-1} + \mathrm{D}^{\mathrm{T}}_{n-1}\mathrm{B}\Gamma) + \alpha_n^2\Big(\|\mathrm{D}_{n-1}\|\mathrm{E}\{\|\Delta_n\|^4\}\Gamma^2 +$$

$$+\Gamma\Big((v_n^2(1+M_\varphi^2\rho) + \|\mathrm{D}_{n-1}\|\rho^{-1} + M_\varphi^2\mathrm{Tr}[\mathrm{Q}_w])\mathrm{B} + \mathrm{E}\{\Delta_n\Delta_n^{\mathrm{T}}\mathrm{Q}_w\Delta_n\Delta_n^{\mathrm{T}}\}\Big)\Gamma\Big).$$

Now, taking the unconditional expectation and making use the first part of Assumption (Bii), one can obtain for matrix $\mathrm{V}_n := \mathrm{E}\{\mathrm{D}_n\}$

$$\mathrm{V}_n \le \mathrm{V}_{n-1} - \alpha_n(\Gamma\mathrm{B}\mathrm{V}_{n-1} + \mathrm{V}_{n-1}\mathrm{B}\Gamma) + \alpha_n^2\Gamma\mathrm{R}\Gamma + \alpha_n^2\mathcal{O}(\|\mathrm{V}_{n-1}\|),$$

where R is defined by Theorem 2. This implies $\mathrm{V}_n \to 0$ as $n \to \infty$ by Lemma 3 from [36].

From the last inequality we have

$$\mathrm{V}_n \le \mathrm{V}_{n-1} - n^{-1}\Gamma\mathrm{B}\mathrm{V}_{n-1} - n^{-1}\mathrm{V}_{n-1}\mathrm{B}\Gamma + n^{-2}\Gamma\mathrm{R}\Gamma + n^{-2}\mathcal{O}(\|\mathrm{V}_{n-1}\|).$$

Denote $\mathrm{W}_n = n\mathrm{V}_n - \mathrm{S}$. Then in view of Theorem 2 conditions we obtain

$$\mathrm{W}_n \le \mathrm{W}_{n-1} - (n-1)^{-1}(\Gamma\mathrm{B} - \frac{1}{2}\mathrm{I})\mathrm{W}_{n-1} - (n-1)^{-1}\mathrm{W}_{n-1}(\mathrm{B}\Gamma - \frac{1}{2}\mathrm{I}) + n^{-2}\mathcal{O}(\|\mathrm{W}_{n-1}\|).$$

Hence applying again Lemma 3 of [36] we get $\mathrm{W}_n \to 0$ as $n \to \infty$ and thus Theorem 4 is proved.

D. Proof of Theorem 5

The following auxiliary results from the [4] will be used in the proof of Theorem 5.

Lemma 1: Under the conditions of Theorem 5 the next facts hold:

(a) $\sum\limits_{n=1}^{\infty} \Delta_n^{\mathrm{T}}\,\Gamma_n^2\,\Delta_n < \infty$ a.s. and $\sum_{n=1}^{\infty} \|\Delta_n\|^4\,\lambda_{max}^2(\Gamma_n) < \infty$ a.s.

(b) $\sum\limits_{n=1}^{\infty} \Delta_n^{\mathrm{T}}\,\Gamma_n\,\Delta_n = \infty$ a.s.

Substituting (14) into (8) we have

$$\|\hat{\theta}_n - \theta^\star\|^2 = ((\hat{\theta}_{n-1} - \theta^\star)^{\mathrm{T}} - \eta_n^{\mathrm{T}}\Delta_n\Delta_n^{\mathrm{T}}\Gamma_n)((\hat{\theta}_{n-1} - \theta^\star) - \Gamma_n\Delta_n\Delta_n^{\mathrm{T}}\eta_n) + \xi_n^2\Delta_n^{\mathrm{T}}\Gamma_n^2\Delta_n +$$

$$+((\hat{\theta}_{n-1}-\theta^\star)^{\mathrm{T}}-\eta_n^{\mathrm{T}}\Delta_n\Delta_n^{\mathrm{T}}\Gamma_n)\xi_n\Gamma_n\Delta_n+\xi_n\Delta_n^{\mathrm{T}}\Gamma_n((\hat{\theta}_{n-1}-\theta^\star)-\Gamma_n\Delta_n\Delta_n^{\mathrm{T}}\eta_n).$$

Taking the conditional expectation with respect to σ-algebra $\tilde{\mathcal{F}}_{n-1}$ we obtain by an Assumption (A)

$$\mathrm{E}\{\|\hat{\theta}_n-\theta^\star\|^2|\tilde{\mathcal{F}}_{n-1}\}\le\|\hat{\theta}_{n-1}-\theta^\star\|^2+\mathrm{E}\{(v_n-\mathrm{E}\{\varphi_n\}^{\mathrm{T}}\eta_n)^2\,|\tilde{\mathcal{F}}_{n-1}\}\mathrm{E}\{\Delta_n^{\mathrm{T}}\Gamma_n^2\Delta_n|\tilde{\mathcal{F}}_{n-1}\}-$$
$$-\eta_n^{\mathrm{T}}\mathrm{E}\{\Delta_n\Delta_n^{\mathrm{T}}\Gamma_n|\tilde{\mathcal{F}}_{n-1}\}(\hat{\theta}_{n-1}-\theta^\star)-(\hat{\theta}_{n-1}-\theta^\star)^{\mathrm{T}}\mathrm{E}\{\Gamma_n\Delta_n\Delta_n^{\mathrm{T}}|\tilde{\mathcal{F}}_{n-1}\}\eta_n+$$
$$+\eta_n^{\mathrm{T}}\mathrm{E}\{\Delta_n\Delta_n^{\mathrm{T}}\Gamma_n^2\Delta_n\Delta_n^{\mathrm{T}}|\tilde{\mathcal{F}}_{n-1}\}\eta_n.$$

Hence averaging sequentially over w_n and v_n yields

$$\mathrm{E}\{\|\hat{\theta}_n-\theta^\star\|^2|\mathcal{F}_{n-1}\}\;\le\;(1+\mathrm{E}\{2M_\varphi^2\Delta_n^{\mathrm{T}}\Gamma_n^2\Delta_n+\|\Delta_n\|^4\,\lambda_{max}^2(\Gamma_n)|\mathcal{F}_{n-1}\})\|\hat{\theta}_{n-1}-\theta^\star\|^2-$$
$$-(\hat{\theta}_{n-1}-\theta^\star)^{\mathrm{T}}\mathrm{E}\{(\Gamma_n\Delta_n\Delta_n^{\mathrm{T}}+\Delta_n\Delta_n^{\mathrm{T}}\Gamma_n)|\mathcal{F}_{n-1}\}(\hat{\theta}_{n-1}-\theta^\star)+\sigma_w^2\mathrm{E}\{\|\Delta_n\|^4\,\lambda_{max}^2(\Gamma_n)|\mathcal{F}_{n-1}\}+$$
$$+(2\sigma_v^2+M_\varphi^2\sigma_w^2)\mathrm{E}\{\Delta_n^{\mathrm{T}}\Gamma_n^2\Delta_n|\mathcal{F}_{n-1}\}.$$

Further, applying the Robbins-Siegmund Lemma from [34] and taking into account (a) we conclude that sequence $\{\|\hat{\theta}_n-\theta^\star\|^2\}$ has a finite limit a.s. and

$$\sum_{n=1}^{\infty}(\hat{\theta}_{n-1}-\theta^\star)^{\mathrm{T}}\mathrm{E}\{\Gamma_n\Delta_n\Delta_n^{\mathrm{T}}+\Delta_n\Delta_n^{\mathrm{T}}\Gamma_n|\hat{\mathcal{F}}_{n-1}\}(\hat{\theta}_{n-1}-\theta^\star)<\infty.$$

By (b) $\sum_{n=1}^{\infty}\mathrm{E}\{\Gamma_n\,\Delta_n\,\Delta_n^{\mathrm{T}}|\hat{\mathcal{F}}_{n-1}\}=\infty$, therefore $\|\hat{\theta}_n-\theta^\star\|^2\to 0$ a.s. as $n\to\infty$. The proof of the first part of Theorem 5 is completed.

It follows from (14) that

$$\hat{\theta}_n=\Gamma_n\sum_{k=1}^{n}\Delta_k(y_k-\mathrm{E}\{\varphi_k^{\mathrm{T}}\}\theta_{n-1}),\;\Gamma_n=(\sum_{k=1}^{n}\Delta_k\Delta_k^{\mathrm{T}}+\gamma_0\mathrm{I})^{-1}.$$

Denote $\epsilon_k=\xi_k+\Delta_k^{\mathrm{T}}(\theta_k^\star-\theta^\star)$. We have

$$\mathrm{D}_n=\Gamma_n\Big(\gamma_0^2\theta^\star\theta^{\star\mathrm{T}}+\sum_{k=1}^{n}\Delta_k\Delta_k^{\mathrm{T}}\epsilon_k^2-\gamma_0\theta^\star\sum_{i=1}^{n}\Delta_i^{\mathrm{T}}\epsilon_i-\gamma_0\sum_{i=1}^{n}\Delta_i\epsilon_i\theta^{\star\mathrm{T}}+\sum_{\substack{i,j=1\\ i\neq j}}^{n}\Delta_i\Delta_j^{\mathrm{T}}\epsilon_i\epsilon_j\Big)\Gamma_n.$$

As in the [4] it is possible to show $\mathrm{E}\{\Gamma_n\Delta_i\Delta_j^{\mathrm{T}}\epsilon_i\epsilon_j\Gamma_n\}=0$ for any $i\neq j$, $i,j=1,\ldots,n$ and $\mathrm{E}\{\Gamma_n\gamma_0\theta^\star\Delta_i^{\mathrm{T}}\epsilon_i\Gamma_n\}=0$ for $i=1,\ldots,n$. Now we obtain by boundedness of v_n, w_n and Δ_n for sufficiently large n

$$\mathrm{E}\{\mathrm{D}_n\}=\mathrm{E}\{\Gamma_n(\gamma_0^2\theta^\star\theta^{\star\mathrm{T}}+\sum_{k=1}^{n}\Delta_k\Delta_k^{\mathrm{T}}\epsilon_k^2)\Gamma_n\}\le\hat{C}\mathrm{E}\{\Gamma_n\}$$

with some constant $\hat{C}$. Since $\sum_{k=1}^{n}\Delta_k\Delta_k^{\mathrm{T}}\to\infty$ a.s. as $n\to\infty$ and $\|\Gamma_n\|\le\gamma_0^{-1}$ we obtain by the Lebesgue dominated convergence Theorem that $\mathrm{E}\{\mathrm{D}_n\}\to 0$ as $n\to\infty$. This completes the proof of Theorem 4.

REFERENCES

[1] R.A. Fisher, *The Design of Experiments.* Oliver and Boyd, Edinburgh, 1935.

[2] O.N. Granichin, "A stochastic approximation algorithm with perturbations in the input for identification of static nonstationary plant," *Vestnik Leningr. Univ., Math.*, vol. 21, pp. 92–93, 1988.

[3] O.N. Granichin, "Unknown Function Minimum Point Estimation under Dependent Noise," *Problems Inform. Transmission*, v. 28, No. 2, pp.16–20, 1992.

[4] A.V. Goldenshluger and B.T. Polyak, "Estimation of regression parameters with arbitrary noise," *Mathematical Methods of Statistics*, vol. 2, pp. 18–29, 1993.

[5] L. Ljung and L. Guo, "The Role of Model Validation for Assessing the Size of the Unmodeled Dynamics," *IEEE Trans. Automat. Contr.*, vol. 42, pp.1230–1239, Sept. 1997.

[6] O.N. Granichin, "Estimating the parameters of linear regression in an arbitrary noise," *Automat. and Remote Contr.*, v. 63, pp. 25–35, 2002.

[7] O.N. Granichin and B.T. Polyak *Randomized Algorithms of an Estimation and Optimization Under Almost Arbitrary Noises.* M.: Nauka, 2003.

[8] O.N. Granichin, "A Stochastic Recursive Procedure with Dependent Noises in the Observation that Uses Sample Perturbations in the Input," *Vestnik Leningrad Univ., Math.*, vol. 22, No. 1(4), pp.27–31, 1989.

[9] B.T. Polyak and A.B. Tsybakov, "Optimal Orders of Accuracy for Search Algorithms of Stochastic Optimization," *Problems Inform. Transmission*, vol. 26, No. 2, pp.126–133, 1990.

[10] O.N. Granichin, "Stochastic Approximation with Sample Perturbations in the Input," *Automat. Remote Control*, vol. 53, No. 2, pp.232–237, 1992.

[11] J.C. Spall, "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Trans. Automat. Contr.,* vol. 37, pp. 332–341, 1992.

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

[13] H.J. Kushner and D.S. Clark, *Stochastic Approximation Methods for Constrain and Unconstraint System.* Berlin, Germany: Springer-Verlag, 1978.

[14] H.J Kushner. and G.G. Yin, *Stochastic Approximation Algorithms and Applications.* New York: Springer–Verlag, 1997.

[15] J. Alspector, R. Meir, A. Jayakumar and D. Lippe, "A Parallel Gradient Descent Method for Learning in Analog VLSI Neural Networks," in S.J. Hanson, J.D. Cowan and C. Lee *Advances in Neural Information Processing Systems.* San Mateo, CA: Morgan Kaufmann Publishers, Inc. pp.834–844, 1993.

[16] Y. Maeda, and Y. Kanata, "Learning Rules for Recurrent Neural Networks Using Perturbation and their Application to Neuron-control," *Trans. of IEE of Japan*, vol. 113-C, pp. 402–408, 1993.

[17] B.T. Polyak and A.B. Tsybakov, "On stochastic approximation with arbitrary noise (the KW case)," in Topics in Nonparametric Estimation, R.Z. Khasminskii ed., *Advances in Soviet Mathematics*, Amer. Math. Soc., Providence, No. 12, pp. 107–113, 1992.

[18] O.N. Granichin, "Stochastic Approximation under Dependent Noises, Detecting Signals and Adaptive Control," in *Approximation, Probability, and Related Fields*, pp. 247–271, Plenum, 1994.

[19] H.F. Chen, T.E. Duncan and B. Pasik-Duncan, "A Kiefer-Wolfowitz Algorithm with Randomized Differences," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 442–453, Mar. 1999.

[20] O.N. Granichin, "Randomized algorithms for stochastic approximation under arbitrary disturbances," *Automat. and Remote Contr.*, v. 63, pp. 209–219, 2002.

[21] O.N. Granichin, "Optimal convergence rate of the randomized algorithms of stochastic approximation in arbitrary noise," *Automat. and Remote Contr.*, v. 64, pp. 252–262, 2003.

[22] Q. Zhang, "Optimal filtering of discrete-time hybrid systems," *J. Optim. Theory App.*, vol. 100, pp. 123–144, 1999.

[23] Q. Zhang, "Hybrid filtering for linear systems with non-Gaussian disturbances," *IEEE Trans. Automat. Contr.*, vol. 45, pp. 50–61, 2000.

[24] O.N. Granichin, "Nonminimax filtering in unknown irregular constrained observation noise," *Automat. and Remote Contr.*, v. 63, pp. 1482–1488, 2002.

[25] B.T. Polyak and Ya.Z. Tsypkin, "Adaptive estimation algorithms (convergence, optimality, stability)," *Automat. Remote Contr.*, vol. 40, pp. 378–389, 1979.

[26] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, MA - London, 1983.

[27] P.C. Young, *Recursive Estimation and Time-Series Analysis. An Introduction.* Berlin–Heidelberg: Springer, 1984.

[28] L. Guo, L. Ljung and G-J. Wang, "Necessary and Sufficient Conditions for Stability of LMS," *IEEE Trans. on Automat. Contr.*, vol. 42, pp. 761–770, 1997.

[29] Y. Maeda, A. Nakazawa and K. Yakichi, "Hardware Implementation of a Pulse Density Neural Network Using Simultaneous Perturbation Learning Rule," *Analog Integrated Circuits and Signal Processing*, vol. 18, pp. 153–162, 1999.

[30] P.W. Shor, "Quantum computing," *Proc. of the 9th Int. Math. Congress,* Berlin, 1998, www.math.nine.edu/documenta/xvol-icm/Fields/Fields.html

[31] C.H. Bennet and P.W. Shor, "Quantum information theory," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2724–2742, Oct. 1998.

[32] J. Preskill "Quantum Information and Computation." [Online]. Available: http://www.theory.caltech.edu/˜ preskill/ph229

[33] A. Peres, *Quantum Theory: Concepts and Methods.* Dordrecht, The Netherlands: Kluwer, 1995.

[34] H. Robbins and D. Siegmund, "A convergence theorem for nonnegative almost supermartingales and some applications," in *Optimizing Methods in Statistics*, J.S. Rustagi ed., Academic Press, New York, pp. 233–257, 1971.

[35] B.T. Polyak, *Introduction to Optimization.* New York: Optim.Software, 1987.

[36] B.T. Polyak, "Convergence and rate of convergence of recursive stochastic algorithms. Linear case," *Automat. Remote Contr.*, vol. 38, pp. 537–542, 1977.