2014 European Control Conference (ECC)
June 24-27, 2014. Strasbourg, France

Finite Difference and Simultaneous Perturbation Stochastic Approximation with Fixed Step Sizes in Case of Multiplicative Noises

Alexander Vakhitov[1]

Abstract— **Simultaneous perturbation stochastic approximation method was shown to be superior over finite difference (Kiefer-Wolfowitz) method in case of unknown but bounded additive measurement noise. This paper is devoted to analysis of the behaviour of these methods in case of multiplicative noise and fixed step sizes. It gives theoretical bounds for the mean squared error and variance after finite number of iiterations for finite difference and simultaneous perturbation methods. The multiplicative noise is present in cost functions in many different fields, and ability to cope with them is a good side of for an optimization method. Fixed step size algorithms are easy to implement and analyze as well as can be used in nonstationary optimization problems. The simulation includes the case when the algorithms' parameters are chosen as theoretically optimal and the case when they are chosen as practically giving the best results after finite number of iterations. Comparative analysis shows similar performance of both methods in terms of mean squared error and slightly better performance of SPSA in terms of variance. Simulation results are provided to illustrate the theoretical contributions.**

I. Introduction

Stochastic approximation procedures are widely used in different fields such as control, machine learning, signal processing etc. Since the first publication of Robbins and Monro [1], a lot of research effort was devoted to study iterative procedures with noisy inputs aiming at finding roots of functions or their extremal points. Finite-difference method for finding a minimum using stochastic approximation was proposed by Kiefer and Wolfowitz in 1952 [2]. However, not all the theoretical results are ready to be applied in computers and technology due to the fact that most of SA procedures rely on absolutely diverging and diminishing sequences of step sizes which cannot be implemented in computers with finite number of registers. The algorithms with fixed step sizes are easier to implement, more robust, however they require new theoretical approaches.

In the late 1980s simultaneous perturbation stochastic approximation-type procedures were proposed by several authors [3], [4]. This paper is devoted to comparison of so called finite difference (FD) and simultaneous perturbation (SP) stochastic approximation (SA) procedures [3], [4] in case of finite step size and multiplicative noises giving theoretical bounds for estimation errors and estimates variances on finite horizon analogously to what was done for diminishing step sizes in [3]. To our knowledge, this paper for the first time gives bounds for the estimates' variance for SPSA-type algorihm with finite step size.

[1]Alexander Vakhitov is with Faculty of Mathematics and Mechanics, St. Petersburg State University, 198504 Universitetsky pr., 28, Stary Peterhof, St. Petersburg, Russia a.vakhitov@spbu.ru

Granichin proved that the SPSA method converges in case of arbitrary but bounded additive noise opposite to FDSA which make biased gradient estimates in this case [4]. Polyak and Tsybakov [5] proved asymptotically optimal convergence rate of SPSA-type procedures in a class of zero order optimization algorithms minimizing cost functions measured with additive noises. SPSA became a popular and used technique in the field of control (especially model-free control) as well as statistics, game theory [6].

Application of SPSA-type procedures to global optimization was studied in [7] and [8]. It was shown that due to smoothing properties of randomized observations it is possible to minimize functions with plenty of minima like Griewank function, and some theoretical conditions on convergence in average were given.

Finite-step SPSA-type procedures were applied to a problem of minimum tracking when a cost function changes in time and the minimum estimate needs to be continuously re-adjusted, like in extremal control setting [9]. This type of problems prohibits the use of diminishing step-size sequences. The results from this paper can be generalized to be used in tracking problems.

The importance of a case when measurements of function values are done with multiplicative noises can be illustrated as follows. If we consider a strongly convex function $f(x)$ then its values give us knowledge about the distance from some point x in space to a minimum point. If we have some measurement noise, and the noise amplitude grows if we go further and further from the minimum point, then we can formulate this noise as multiplicative:

$$y = wf(x) = f(x) + (w-1)f(x),$$

where we denote measurement as y, point as x, multiplicative noise as w, $Ew = 1$, and the "additive" noise component growing with distance is therefore $\varepsilon = (w-1)f(x)$. Similar effect appears in non-linear least squares problems where we need to fit functions $\phi_i(x)$ to measurements d_i corrupted by noise v_i. In this case cost function takes the form such as

$$y = \sum_i (d_i - \phi_i(x) + v_i)^2 =$$

$$= \sum_i \{(d_i - \phi_i(x))^2 + 2v_i(d_i - \phi_i(x)) + v_i^2\}.$$

From this formula we see that noise has not additive, but multiplicative nature here. Comparative study of the SPSA and FDSA algorithms in case of multiplicative noises is interesting also because it is well known [4] that these algorithms differ in working with additive noises, however

their performance in case of multiplicative noise was up to now unknown.

II. Problem statement

The problem is to find an estimate of the point of minimum of the function $f(x)$ using measurements made at points $x_n \in \mathbb{R}^q$, $n \in \mathbb{N}$ freely chosen by the optimization algorithm. These measurements are corrupted by additive noise v_n and multiplicative niose w_n which has expected value equal to 1 and finite variance σ_w^2:

$$y_n = w_n f(x_n) + v_n.$$

We denote as θ_* the minimum point of f.

We will make the following assumptions:

Assumption 1. Strong convexity of $f(x)$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2,$$

where $\mu > 0$ is constant of strong convexity .

Assumption 2. Gradient of $f(x)$ satisfies Lipshitz property:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|,\ L > 0$$

Assumption 3. Function value is bounded at minimum point:

$$|f(\theta_*)| < f_* < \infty.$$

Assumption 4. Random perturbation Δ_n used in SPSA algorithm is a Bernoulli vector with i.i.d. components which take values +1 or -1 with equal probability.

Assumption 5. Additive noise

$$E|v_n^+ - v_n^-| < \sigma_v^1,\ E|v_n^+ - v_n^-|^2 < \sigma_v^2,$$

where v_n^+, v_n^- are consequent realizations of the additive noise. This noise can be just bounded but not random.

Assumption 6. Multiplicative noise w_n has bounded first and second moments and known expected value:

$$E|w_n| < \sigma_w^1,\ E|w_n|^2 < \sigma_w^2, Ew_n = 1,$$

w_n are i.i.d. for $n \in \mathbb{N}$.

Assumption 7. The domain of possible argument values is a bounded convex set Ω, and exists a ball containing it of radius $R < \infty$ with a center in θ_*:

$$x \in \Omega \implies \|x - \theta_*\| < R$$

III. Algorithms

In this paper we compare two zero-order optimization algorithms. First is finite difference stochastic approximation proposed by Kiefer and Wolfowitz [2] with fixed step size:

$$\hat{\theta}_n = \hat{\theta}_{n-1} - \alpha k_n(\hat{\theta}_{n-1}, \beta), \tag{1}$$

where

$$k_n^{(i)}(\hat{\theta}_n, \beta) = \frac{1}{2\beta}(w_n^{i+} f(\hat{\theta}_n + \beta e_i) - w_n^{i-} f(\hat{\theta}_n - \beta e_i) + \\ + v_n^{i+} - v_n^{i-}),\ i = 1 \dots q,$$

$\beta > 0$ is the trial step size, e_i is the i-th canonical basis element.

Second algorithm is two measurements SPSA-type algorithm with fixed step size:

$$\hat{\theta}_n = \hat{\theta}_{n-1} - \alpha g_n(\hat{\theta}_{n-1}, \beta, \Delta_n), \tag{2}$$

where

$$g_n(\hat{\theta}_n, \beta, \Delta_n) = \frac{\Delta_n}{2\beta}(w_n^1 f(\hat{\theta}_n + \beta\Delta_n) - w_n^2 f(\hat{\theta}_n - \beta\Delta_n) + \\ + v_n^+ - v_n^-),$$

and Δ_n is decsribed in assumption 4.

For the both algorithms, $\hat{\theta}_0$ is chosen arbitrarily. In the algorithms' comparison (following [3]) we fix the number of measurements made by the algorithm during the run.

Denote as $E_n\{\cdot\}$ the expectation conditioned on the past observations:

$$E_n\{\cdot\} = E\{\cdot | y_{n-1}, y_{n-2}, \dots\}.$$

The measurement point and the estimate belong to the ball around the minimum point which is defined as $\|x - \theta_*\| < R$.

IV. Convergence and Variance

Theorem 1. Mean squared error for FDSA with fixed step size. Denote

$$k = 1 - 2\alpha\mu + \Big(\Big(q\frac{L^2R^2 + 4f_*L}{8\beta^2} + \frac{(q+2)L^2}{4} + \frac{q}{2}\Big)\sigma_w^2 + L^2\Big)\alpha^2,$$

$$h = \alpha^2\Big(q\beta L^2 + \frac{\sqrt{q}\sigma_v^1 L}{\beta}\Big) + \alpha\sqrt{q}L,$$

$$l = \alpha^2\Big(q\frac{2f_*^2\sigma_w^2 + \sigma_v^2}{4\beta^2} + \frac{q\sigma_w^2 + 2q^2}{8}\beta^2L^2 + \frac{qf_*\sigma_w^2 L}{2} + \\ + \frac{q^{3/2}\sigma_v^1 L}{2}\Big)$$

For the algorithm (1) in the assumptions stated above if $k + \varepsilon/2 \in (0,1)$ for some sufficiently small $\varepsilon > 0$ the estimates $\hat{\theta}_n$ satisfy the following inequalities:

$$E\|\hat{\theta}_n - \theta_*\|^2 \leq (k + \varepsilon/2)^n \|\hat{\theta}_0 - \theta_*\|^2 + \\ + \frac{(h^2/(2\varepsilon) + l)(1 - (k + \varepsilon/2)^n)}{1 - k - \varepsilon/2},$$

$$\limsup_{n\to\infty} (E\|\hat{\theta}_n - \theta_*\|^2)^{1/2} \leq \frac{h}{2(1-k)}(1 + \sqrt{1 + 4l(1-k)h^{-2}}).$$

Proof.

$$\|\hat{\theta}_{n+1} - \theta_*\|^2 = \|\hat{\theta}_n - \theta_*\|^2 - 2\alpha\langle\hat{\theta}_n - \theta_*, k_n(\hat{\theta}_n, \beta)\rangle + \\ + \alpha^2\|k_n(\hat{\theta}_n, \beta)\|^2.$$

- $\|k_n(\hat{\theta}_n, \beta)\|^2$:

$$\|k_n(\hat{\theta}_n, \beta)\|^2 = (2\beta)^{-2}\{\Big((w_{i,n}^{(1)} - 1)f(\hat{\theta}_n + \beta e_i)\Big)_{i=1}^q + \\ + \Big((1 - w_{i,n}^{(2)})f(\hat{\theta}_n - \beta e_i)\Big)_{i=1}^q + \\ + \Big(f(\hat{\theta}_n + \beta e_i) - f(\hat{\theta}_n - \beta e_i)\Big)_{i=1}^q +$$

$$+v_n^+ - v_n^-\}^2$$

$$E_n\{\|k_n(\hat{\theta}_n,\beta)\|^2\} = (2\beta)^{-2}\Big(E_n\{\sigma_w^2\sum_{i=1}^{q}(f^2(\hat{\theta}_n+\beta e_i)+$$

$$+f^2(\hat{\theta}_n-\beta e_i)) + \|(f(\hat{\theta}_n+\beta e_i)-f(\hat{\theta}_n-\beta e_i))_{i=1}^{q}\|^2+$$

$$+q\sigma_v^2+2\sigma_v^1\sqrt{q}\|(f(\hat{\theta}_n+\beta e_i)-f(\hat{\theta}_n-\beta e_i))_{i=1}^{q}\|\Big).$$

$$E_n(2\beta)^{-2}\sigma_w^2\sum_{i=1}^{q} f^2(\hat{\theta}_n+\beta e_i)+f^2(\hat{\theta}_n-\beta e_i) \leq$$

$$\leq \Big(q\frac{L^2R^2+4f_*L}{8\beta^2}+\frac{(q+2)L^2}{4}\Big)\sigma_w^2\|\hat{\theta}_n-\theta_*\|^2+ \quad (3)$$

$$+\frac{qf_*^2\sigma_w^2}{2\beta^2}+\frac{qf_*\sigma_w^2L}{2}+\frac{q\beta^2\sigma_w^2L^2}{8}$$

$$(2\beta)^{-2}\|\Big(f(\hat{\theta}_n+\beta e_i)-f(\hat{\theta}_n-\beta e_i)\Big)_{i=1}^{q}\| \leq$$

$$\leq \frac{L}{2\beta}\|\hat{\theta}_n-\theta_*\|+\frac{qL}{4}.$$

$$(2\beta)^{-2}\|\Big(f(\hat{\theta}_n+\beta e_i)-f(\hat{\theta}_n-\beta e_i)\Big)_{i=1}^{q}\|^2 \leq$$

$$\leq L^2\|\hat{\theta}_n-\theta\|^2+q\beta L^2\|\hat{\theta}_n-\theta_*\|+\frac{q^2\beta^2L^2}{4}.$$

$$E_n\{\|k_n(\hat{\theta}_n,\beta)\|^2\} \leq \Big(\Big(q\frac{L^2R^2+4f_*L}{8\beta^2}+\frac{(q+2)L^2}{4}\Big)\sigma_w^2+$$

$$+L^2\Big)\|\hat{\theta}_n-\theta_*\|^2+\Big(q\beta L^2+\frac{\sqrt{q}\sigma_v^1L}{\beta}\Big)\|\hat{\theta}_n-\theta_*\|+$$

$$+q\frac{2f_*^2\sigma_w^2+\sigma_v^2}{4\beta^2}+\frac{q\sigma_w^2+2q^2}{8}\beta^2L^2+\frac{qf_*\sigma_w^2L}{2}+\frac{q^{3/2}\sigma_v^1L}{2}.$$

- $-\langle\hat{\theta}_n-\theta_*,k_n(\hat{\theta}_n,\beta)\rangle$:

$$-\langle\hat{\theta}_n-\theta_*,k_n(\hat{\theta}_n,\beta)\rangle \leq -\mu\|\hat{\theta}_n-\theta_*\|^2+\frac{L\beta}{2}\sqrt{q}\|\hat{\theta}_n-\theta_*\|.$$

We have finally proved the following bound:

$$E_n\|\hat{\theta}_n-\theta_*\|^2 \leq k\|\hat{\theta}_{n-1}-\theta_*\|^2+h\|\hat{\theta}_n-\theta_*\|+l.$$

For some sufficiently small $\varepsilon>0$ such that $k+\varepsilon/2<1$,

$$E_n\|\hat{\theta}_n-\theta_*\|^2 \leq (k+\varepsilon/2)\|\hat{\theta}_{n-1}-\theta_*\|^2+l+h^2/(2\varepsilon^2).$$

Let us denote $e_n=(E\|\hat{\theta}_n-\theta_*\|^2)^{1/2}$.

If we take the unconditional expectation on the both sides of the last equation, in the same way as in [9], we get

$$e_n^2 \leq (k+\varepsilon/2)^n e_0^2+\frac{(h^2/(2\varepsilon)+l)(1-(k+\varepsilon/2)^n)}{1-k-\varepsilon/2},$$

$$\limsup_{n\to\infty} e_n \leq \frac{h}{2(1-k)}(1+\sqrt{1+4l(1-k)h^{-2}}).$$

QED

Theorem 2. Variance for FDSA with fixed step size. Denote

$$k=1-2\alpha\mu+\alpha^2\Big(q\frac{L^2R^2+4f_*L}{8\beta^2}+\frac{L^2}{2\beta}+\frac{q}{2}\Big)\sigma_w^2,$$

$$h=\alpha\beta\sqrt{q}L,$$

$$l=\alpha^2(\frac{2qf_*^2\sigma_w^2+q\sigma_v^2}{4\beta^2}+\frac{qf_*\sigma_w^2L}{2}+\frac{q\beta^2\sigma_w^2L^2}{8}).$$

For the algorithm (1) in the assumptions stated above if $k+\varepsilon/2\in(0,1)$ for some sufficiently small $\varepsilon>0$ the variance $E\|\hat{\theta}_n-E\hat{\theta}_n\|^2$ of estimates $\hat{\theta}_n$ satisfies the following inequalities :

$$E\|\hat{\theta}_n-E\hat{\theta}_n\|^2 \leq \frac{(h^2/(2\varepsilon)+l)(1-(k+\varepsilon/2)^n)}{1-k-\varepsilon/2},$$

$$\limsup_{n\to\infty}(E\|\hat{\theta}_n-E\hat{\theta}_n\|^2)^{1/2} \leq \frac{h}{2(1-k)}(1+$$

$$+\sqrt{1+4l(1-k)h^{-2}}).$$

Proof.

$$E\|\hat{\theta}_{n+1}-E\hat{\theta}_{n+1}\|^2=\|\hat{\theta}_n-E\hat{\theta}_n\|^2-2\alpha E\langle\hat{\theta}_n-E\hat{\theta}_n,$$

$$k_n(\hat{\theta}_n,\beta)-Ek_n(\hat{\theta}_n,\beta)\rangle+\alpha^2E\|k_n(\hat{\theta}_n,\beta)-Ek_n(\hat{\theta}_n,\beta)\|^2. \quad (4)$$

$$E\langle\hat{\theta}_n-E\hat{\theta}_n,k_n(\hat{\theta}_n,\beta)-Ek_n(\hat{\theta}_n,\beta)\rangle \geq \mu E\|\hat{\theta}_n-E\hat{\theta}_n\|^2-$$

$$-\sqrt{q}(\frac{\beta L}{2}+\sigma_v^1)E\|\hat{\theta}_n-E\hat{\theta}_n\|.$$

The third term of (4) can be bounded as:

$$\alpha^2E\|k_n(\hat{\theta}_n,\beta)-Ek_n(\hat{\theta}_n,\beta)\|^2 \leq \alpha^2\Big(q\frac{L^2R^2+4f_*L}{8\beta^2}+\frac{L^2}{2\beta}+$$

$$+\frac{q}{2}\Big)\sigma_w^2E\|\hat{\theta}_n-\theta_*\|^2+\frac{\sqrt{q}\sigma_v^1\alpha^2L}{\beta}E\|\hat{\theta}_n-\theta_*\|+$$

$$+\alpha^2L^2E\|\hat{\theta}_n-E\hat{\theta}_n\|^2+2\alpha^2L(\frac{\sqrt{q}\sigma_v^1}{\beta}+\sqrt{q}\beta L)E\|\hat{\theta}_n-E\hat{\theta}_n\|+$$

$$+\alpha^2(\frac{\sigma_v^2+2qf_*^2\sigma_w^2}{4\beta^2}+\frac{q\beta^2\sigma_w^2L^2}{8}+\frac{qL^2}{4}+\frac{qf_*\sigma_w^2L}{2}+\frac{q^{3/2}\sigma_v^1L}{2}).$$

In the following, we will use as $e_n=\sqrt{E\|\hat{\theta}_n-\theta_*\|^2}$ the bounds give by the theorem 1. The variance can be bounded as:

$$E\|\hat{\theta}_{n+1}-E\theta_{n+1}\|^2 \leq (1-2\alpha\mu+\alpha^2L^2)E\|\hat{\theta}_n-E\hat{\theta}_n\|^2+$$

$$+2\alpha^2L(\frac{\sqrt{q}\sigma_v^1}{\beta}+\sqrt{q}\beta L)E\|\hat{\theta}_n-E\hat{\theta}_n\|+$$

$$+\alpha^2\Big(q\frac{L^2R^2+4f_*L}{8\beta^2}+\frac{L^2}{2\beta}+\frac{q}{2}\Big)\sigma_w^2e_n^2+\frac{\sqrt{q}\sigma_v^1\alpha^2L}{\beta}e_n+$$

$$+\alpha^2(\frac{\sigma_v^2+2qf_*^2\sigma_w^2}{4\beta^2}+\frac{q\beta^2\sigma_w^2L^2}{8}+\frac{qL^2}{4}+\frac{qf_*\sigma_w^2L}{2}+\frac{q^{3/2}\sigma_v^1L}{2}).$$

In the following, we use the inequality under the integral

$$E\,h\|\hat{\theta}_n - E\theta_n\| \le E\frac{\varepsilon}{2}\|\hat{\theta}_n - E\theta_n\|^2 + \frac{h^2}{2\varepsilon}.$$

$$E\|\hat{\theta}_{n+1} - E\theta_{n+1}\|^2 \le kE\|\hat{\theta}_n - E\theta_n\|^2 + hE\|\hat{\theta}_n - E\theta_n\| +$$

$$+l \le (k+\varepsilon/2)E\|\hat{\theta}_n - E\theta_n\|^2 + l + \frac{h^2}{2\varepsilon}.$$

Denote $\delta_n = \sqrt{E\|\hat{\theta}_n - E\hat{\theta}_n\|^2}$. Iterating this inequality for $n-1, n-2, \ldots, 0$ and using the fact that $\delta_0 = 0$ we get the inequality and the asymptotic bound from the theorem statement.

QED

Theorem 3. Mean squared error for SPSA with fixed step size.

Denote

$$k = 1 - 2\alpha\mu + \alpha^2\Big(\sigma_w^2\big(\frac{\delta_2(L^2R^2+4f_*L)}{8\beta^2} + \frac{2\delta_2L^2+\delta_4L^2}{4}\big) +$$

$$+L^2\delta_4\Big), \quad h = \alpha^2\big(\beta\delta_5L^2 + \frac{L\sigma_v^1\delta_3}{\beta}\big) + \alpha\beta\delta_2L,$$

$$l = \alpha^2\big(\frac{\delta_2(2\sigma_w^2f_*^2+\sigma_v^2)}{4\beta^2} + \frac{\delta_4(\sigma_w^2f_*L+\sigma_v^1L)}{2} +$$

$$+\frac{\beta^2\delta_6L^2(\sigma_w^2+2)}{8}\big).$$

For the algorithm (2) in the assumptions stated above if $k+\varepsilon/2 \in (0,1)$ for some sufficiently small $\varepsilon > 0$ the estimates $\hat{\theta}_n$ satisfy the following inequalities:

$$E\|\hat{\theta}_n - \theta_*\|^2 \le (k+\varepsilon/2)^n\|\hat{\theta}_0 - \theta_*\|^2 +$$

$$+\frac{(h^2/(2\varepsilon)+l)(1-(k+\varepsilon/2)^n)}{1-k-\varepsilon/2},$$

$$\limsup_{n\to\infty}(E\|\hat{\theta}_n - \theta_*\|^2)^{1/2} \le \frac{h}{2(1-k)}\big(1+\sqrt{1+4l(1-k)h^{-2}}\big).$$

Proof.

$$\|\hat{\theta}_{n+1} - \theta_*\|^2 \le \|\hat{\theta}_n - \theta_*\|^2 - 2\alpha\langle\hat{\theta}_n - \theta_*, g_n(\hat{\theta}_n,\beta,\Delta_n)\rangle + \quad (5)$$

$$+\alpha^2\|g_n(\hat{\theta}_n,\beta,\Delta_n)\|^2.$$

Let us bound the terms one by one. $\|g_n(\hat{\theta}_n,\beta,\Delta_n)\|^2$:

$$\|g_n(\hat{\theta}_n,\beta,\Delta_n)\|^2 = \|\Delta_n\|^2(2\beta)^{-2}\Big((w_n^{(1)}-1)f(\hat{\theta}_n+\beta\Delta_n) +$$

$$+(1-w_n^{(2)})f(\hat{\theta}_n-\beta\Delta_n) +$$

$$+(v_n^1 - v_n^2) + (f(\hat{\theta}_n+\beta\Delta_n) - f(\hat{\theta}_n-\beta\Delta_n))\Big)^2$$

$$E_n(2\beta)^{-2}\sigma_w^2\|\Delta_n\|^2(f^2(\hat{\theta}_n+\beta\Delta_n) + f^2(\hat{\theta}_n-\beta\Delta_n)) \le \quad (6)$$

$$\|\hat{\theta}_n - \theta\|^2\sigma_w^2\big(\frac{\delta_2(L^2R^2+4f_*L)}{8\beta^2} + \frac{2\delta_2L^2+\delta_4L^2}{4}\big) +$$

$$+\sigma_w^2\big(\frac{\delta_2f_*^2}{2\beta^2} + \frac{\delta_4f_*L}{2} + \frac{\beta^2\delta_6L^2}{8}\big).$$

These bounds lead to the following bound for the conditional expectation of the squared pseudogradient in (5):

$$E_n\|g_n(\hat{\theta}_n,\beta,\Delta_n)\|^2 \le \|\hat{\theta}_n - \theta\|^2\Big(\sigma_w^2\big(\frac{\delta_2(L^2R^2+4f_*L)}{8\beta^2} +$$

$$+\frac{2\delta_2L^2+\delta_4L^2}{4}\big) + L^2\delta_4\Big) + \|\hat{\theta}_n - \theta_*\|\big(\beta\delta_5L^2 + \frac{L\sigma_v^1\delta_3}{\beta}\big) +$$

$$+\frac{\delta_2(2\sigma_w^2f_*^2+\sigma_v^2)}{4\beta^2} + \frac{\delta_4(\sigma_w^2f_*L+\sigma_v^1L)}{2} + \frac{\beta^2\delta_6L^2(\sigma_w^2+2)}{8}.$$

$-2\langle\hat{\theta}_n - \theta_*, g_n(\hat{\theta}_n,\beta,\Delta_n)\rangle$:

For the second term of (5) we get:

$$E_n\{-2\alpha\langle\hat{\theta}_n - \theta_*, g_n(\hat{\theta}_n,\beta,\Delta_n)\rangle\} \le -2\alpha\mu\|\hat{\theta}_n - \theta_*\|^2 +$$

$$+\frac{L}{2}\alpha\beta\delta_2\|\hat{\theta}_n - \theta_*\|.$$

If we take the unconditional expectation of the both sides of (5), we get

$$e_n^2 \le ke_{n-1}^2 + he_{n-1} + l,$$

where $e_n = E\|\hat{\theta}_n - \theta_*\|^2$ and k,h,l are as defined in the theorem. Analogously to previous theorems, we finish the proof.

Theorem 4. Variance for SPSA with fixed step size. Denote

$$k = 1 - 2\alpha\mu + \alpha^2L^2, \quad h = 2*\alpha^2*q^{3/2}\beta L^2 + \alpha\beta\delta_2L,$$

$$l = e_n^2\alpha^2\Big(\sigma_w^2\big(\frac{\delta_2(L^2R^2+4f_*L)}{8\beta^2} + \frac{2\delta_2L^2+\delta_4L^2}{4}\big) +$$

$$+(q-1)L^2\Big) + \alpha^2\{\big(\frac{\delta_2(2\sigma_w^2\ f_*^2+\sigma_v^2)}{4\beta^2} + \frac{\delta_4L(\sigma_w^2f_*+\sigma_v^1)}{2} +$$

$$+\frac{\beta^2\delta_6L^2(\sigma_w^2+2)}{8}\big)\}.$$

For the algorithm (2) in the assumptions stated above if $k+\varepsilon/2 \in (0,1)$ for some sufficiently small $\varepsilon > 0$ the variance $E\|\hat{\theta}_n - E\hat{\theta}_n\|^2$ of estimates $\hat{\theta}_n$ satisfies the following inequalities :

$$E\|\hat{\theta}_n - E\hat{\theta}_n\|^2 \le \frac{(h^2/(2\varepsilon)+l)(1-(k+\varepsilon/2)^n)}{1-k-\varepsilon/2},$$

$$\limsup_{n\to\infty}(E\|\hat{\theta}_n - E\hat{\theta}_n\|^2)^{1/2} \le \frac{h}{2(1-k)}\big(1+$$

$$+\sqrt{1+4l(1-k)h^{-2}}\big).$$

Proof.

$$E\|\hat{\theta}_{n+1} - E\hat{\theta}_{n+1}\|^2 = E\|\hat{\theta}_n - \theta\|^2 - 2\langle\hat{\theta}_n - E\hat{\theta}_n,$$

$$g_n(\hat{\theta}_n,\beta,\Delta_n) - Eg_n(\hat{\theta}_n,\beta,\Delta_n)\rangle +$$

$$+E\|g_n(\hat{\theta}_n,\beta,\Delta_n) - Eg_n(\hat{\theta}_n,\beta,\Delta_n)\|^2. \quad (7)$$

- $E\langle \hat{\theta}_n - E\hat{\theta}_n, g_n(\hat{\theta}_n,\beta,\Delta_n) - Eg_n(\hat{\theta}_n,\beta,\Delta_n)\rangle$:

$$-2\alpha E\langle \hat{\theta}_n - E\hat{\theta}_n, g_n(\hat{\theta}_n,\beta,\Delta_n) - Eg_n(\hat{\theta}_n,\beta,\Delta_n)\rangle \leq$$
$$\leq -2\alpha\mu E\|\hat{\theta}_n - E\hat{\theta}_n\|^2 + \alpha\beta\delta_2 L E\|\hat{\theta}_n - E\hat{\theta}_n\|.$$

- $E\|g_n(\hat{\theta}_n,\beta,\Delta_n) - Eg_n(\hat{\theta}_n,\beta,\Delta_n)\|^2$:
For the whole third term of (7),

$$E\|g_n(\hat{\theta}_n,\beta,\Delta_n) - Eg_n(\hat{\theta}_n,\beta,\Delta_n)\|^2 \leq E\|\hat{\theta}_n - E\hat{\theta}_n\|^2 L^2 +$$
$$+E\|\hat{\theta}_n - E\hat{\theta}_n\| 2\delta_3\beta L^2 + E\|\hat{\theta}_n - \theta_*\|^2\Big(\sigma_w^2\cdot$$
$$(\frac{\delta_2(L^2R^2+4f_*L)}{8\beta^2} + \frac{2\delta_2L^2+\delta_4L^2}{4}) + (q-1)L^2\Big) +$$
$$+\sigma_w^2(\frac{\delta_2 f_*^2}{2\beta^2} + \frac{\delta_4 f_* L}{2} + \frac{\beta^2\delta_6 L^2}{8}) + \frac{\delta_2}{4\beta^2}\sigma_v^2 + \frac{\delta_6\beta^2L^2}{4} +$$
$$+\frac{\delta_4\sigma_v^1 L}{2}.$$

$$E\|\hat{\theta}_{n+1} - E\theta_{n+1}\|^2 \leq (1 - 2\alpha\mu + \alpha^2L^2)E\|\hat{\theta}_n - E\hat{\theta}_n\|^2 +$$
$$+E\|\hat{\theta}_n - E\hat{\theta}_n\|(2\alpha^2\delta_3\beta L^2 + \alpha\beta\delta_2 L) +$$
$$+E\|\hat{\theta}_n - \theta_*\|^2\alpha^2\Big(\sigma_w^2(\frac{\delta_2(L^2R^2+4f_*L)}{8\beta^2} + \frac{2\delta_2L^2+\delta_4L^2}{4}) +$$
$$+(q-1)L^2\Big) + \alpha^2\{(\frac{\delta_2(2\sigma_w^2 f_*^2 + \sigma_v^2)}{4\beta^2} + \frac{\delta_4 L(\sigma_w^2 f_* + \sigma_v^1)}{2} +$$
$$+\frac{\beta^2\delta_6 L^2(\sigma_w^2+2)}{8})\}.$$

In the following, we use the inequality under the integral

$$E\,h\|\hat{\theta}_n - E\theta_n\| \leq E\frac{\varepsilon}{2}\|\hat{\theta}_n - E\theta_n\|^2 + \frac{h^2}{2\varepsilon}.$$

and in analogous to theorem 2 way we get the inequality and the asymptotic bound from the theorem statement. QED

V. Simulations

We want to compare the performance of the algorithms at some test problem. We run experiment with

$$y_n = (1+\xi_n)\|x\|^2,$$

where ξ_n is Gaussian with $\sigma = 3$, $\hat{\theta}_0 = (1,\ldots,1)^T$, dimension $q = 10$, domain is a ball with radius $R = 10$. In this problem, the standard deviation of multiplicative noise is 3 times more than the scale of expected function, so th eamount of noise is high. To analyze only case with multiplicative noise, we do not use any additive noise here. We allow each algorithm to do $2qx100 = 2000$ measurements.

The minimal expected bound on the error of FDSA according to theorem 1 is achieved when $\alpha_{FD} = 0.0041$, $\beta_{FD} = 5$. The predicted theoretically by theorem 1 expected square of the norm of esimation error $E\|\hat{\theta}_{100} - \theta_*\|^2$ is shown at the fig. 1 as well as error averaged from 1000 runs of the algorithm with the same parameters.

The minimal expected bound on the error of SPSA according to theorem 3 is achieved when $\alpha_{SP} = 0.00036$, $\beta_{SP} = 1.5$. The predicted theoretically by theorem 3 expected square of the norm of esimation error $E\|\hat{\theta}_{1000} - \theta_*\|^2$ is shown at the fig. 2 as well as error averaged from 1000 runs of the algorithm with the same parameters.

At the fig. 3 you see comparison of FDSA and SPSA with theoretically optimal parameters on the longer run with 10 000 measurements. SPSA variance is slightly lower, maybe that is because the theoretically optimal parameters for both methods were chosen on a finite grid, and in case of SPSA the value was closer to optimal one. We conclude that although FDSA bounds are tighter both for expected error norm squared and for variance, the parameters found by optimizing theoretical bounds for the algorithms are giving slightly better results when using SPSA comparing to FDSA.

At the fig. 4 you see the performance of both algorithms with practically optimal parameters. Theoretical bounds are not available for these cases since the theorem conditions are violated ($k > 1$). We see that on the long run algorithms give equal results both in expected error norm and in variance, although the variance of SPSA estimates is lower in the middle of the graph. This effect may be explained from comparative analysis of the theoretical bounds and emphasizes the importance of variance bounds in addition to mean squared error bounds.

VI. Conclusions

In this paper we have shown the theoretical bounds for mean-squared estimation error and variance of estimates for the FDSA and SPSA algorithms with fixed step sizes in case of multiplicative noise. We have illustrated the bounds with numerical simulations and we have noted that the eperiments show similar performance of the algorithms in this setting.

In future we are planning to improve the precision of the bounds and generalize the results given here to a case of nonstationary optimization with minimum point slowly moving in time. We will perform asymptotic analysis of the theorem bounds and compare the algorithms' asymptotic performance.

References

[1] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[2] Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

[3] James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *Automatic Control, IEEE Transactions on*, 37(3):332–341, 1992.

[4] Oleg Granichin. Procedure of stochastic approximation with disturbances at the input. *Automation and Remote Control*, 53(1):232–237, 1992.

[5] Boris Teodorovich Polyak and Aleksandr Borisovich Tsybakov. Optimal order of accuracy of search algorithms in stochastic optimization. *Problemy Peredachi Informatsii*, 26(2):45–53, 1990.

[6] James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. Wiley. com, 2005.

[7] Ivan Minin and Alexander Vakhitov. Randomized smoothing for near-convex functions in context of image processing. In *American Control Conference (ACC), 2012*, pages 833–838. IEEE, 2012.

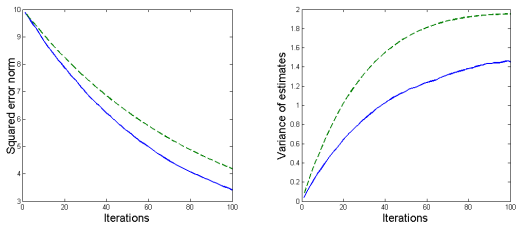

Fig. 1. Performance of FDSA algorithm with theoretically best parameters. Left: squared estimation error bound from theorem 1 (dashed) and averaged over 1000 runs error (solid). Right: variance bound from theorem 2 (dashed) and actual variance estimated using 1000 trial runs (solid)

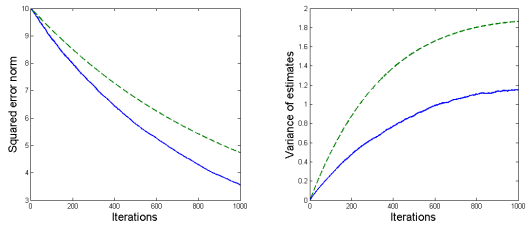

Fig. 2. Performance of SPSA algorithm with theoretically best parameters. Left: squared estimation error bound from theorem 3 (dashed) and averaged over 1000 runs error (solid). Right: variance bound from theorem 4 (dashed) and actual variance estimated using 1000 trial runs (solid)

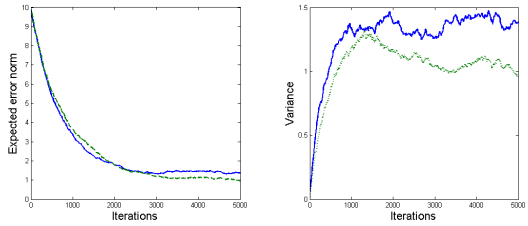

Fig. 3. Performance of SPSA and FDSA algorithm with theoretically best parameters. Left: squared estimation errors averaged over 100 runs error for FDSA (solid) and SPSA (dashed). Right: variance estimated using 100 trial runs for FDSA (solid) and SPSA (dashed). The iterations are counted in SPSA scale, he methods are synchronized by measurements (the estimated number x is the estimated having 2x measurements)

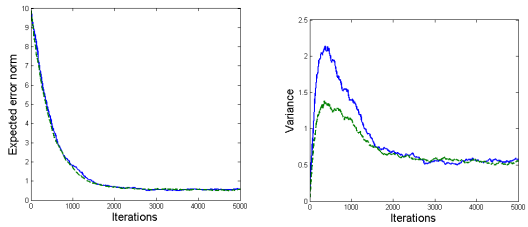

Fig. 4. Performance of SPSA and FDSA algorithm with practically best parameters. Left: squared estimation errors averaged over 100 runs error for FDSA (solid) and SPSA (dashed). Right: variance estimated using 100 trial runs for FDSA (solid) and SPSA (dashed). The iterations are counted in SPSA scale, he methods are synchronized by measurements (the estimated number x is the estimated having 2x measurements)

[8] John L Maryak and Daniel C Chin. Global random optimization by simultaneous perturbation stochastic approximation. In *American Control Conference, 2001.*, volume 2, pages 756–762. IEEE, 2001.

[9] Oleg Granichin, Lev Gurevich, and Alexander Vakhitov. Discrete-time minimum tracking based on stochastic approximation algorithm with randomized differences. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 5763–5767. IEEE, 2009.