Accuracy for the SPSA algorithm with two measurements

Oleg N. Granichin
Saint Petersburg State University
Department of Mathematics
and Mechanics
28 Universitetsky pr.,
198504 Saint Petersburg
Russia
oleg_granichin@mail.ru

Alexander T. Vakhitov
Saint Petersburg State University
Department of Mathematics
and Mechanics
28 Universitetsky pr.,
198504 Saint Petersburg
Russia
av38@yandex.ru

Abstract: The case of SPSA algorithms with two trial simultaneous perturbations is discussed. The better asymptotic convergence rate of the algorithm estimates is proved under more wide assumption about the optimizing function. The Lyapunov function with the power from 1 to 2 is considered.

Key–Words: Stochastic optimization, SPSA, Arbitrary noise, Robustness, Rate of convergence

1 Introduction

Problem of function minimization is being solved in many applications. Sometimes the extremal values of a function can be found theoretically. In general, engineering systems have to deal with unknown functions, and it is only possible to measure its' values in some points.

Measurement always means a noise is present. Sometimes the algorithms which solve optimization problem precisely on the sheet of paper are not able to converge to the function minimum in the practice. Robustness of an algorithm is very useful.

The simultaneous perturbation stochastic approximation (SPSA) with one or two measurements on each iteration were begun to investigate since beginning of 90th years [1, 2, 3]. These algorithms are known for their good convergence properties in the case of measurements with "an almost arbitrary noise". The noise still needs only to be somehow bounded and does not depend on simultaneous perturbations, for example, brought into the system under experiment.

1.1 Previous Experience

This paper continues the investigations [4, 5, 6, 8]. There were discussed the common task and several proposed algorithms. We continue to determine the bounds of ability to apply the SPSA algorithm with two measurements. It is known this algorithm has the better convergence in practice, than the case with one trial simultaneous perturbation described in [8]. This paper proves the better asymptotic convergence of an algorithm under more wide assumption about the optimizing function, depending on the properties of bounding of additional noise in observations.

In the next two sections the main problem statement and the algorithm are discussed. Then the convergence theorem and it's proof are explained. In the conclusion we provide the discussion about achieved results.

2 The Problem Statement and Main Assumptions

Let $F(x,w):\ \mathbb{R}^q \times \mathbb{R}^p \to \mathbb{R}^1$ be the differentiable on the first argument function, $x_1, x_2, \dots$ be a sequence of arguments of F chosen during optimization procedure at each iteration $n = 1, 2, \dots$ (the design of an experiment), $\{w_n\}$ is uncontrollable sequence of random values from $\mathbb{R}^p$ with identical but unknown distribution $\mathrm{P}_w(\cdot)$ which has the finite support. The function $F(\cdot, w_n)$ can be observed with the added noise v_n:

$$y_n = F(x_n, w_n) + v_n \tag{1}$$

The problem is to minimize the function

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

based on observations $y_1, y_2, \dots$. This means to build the sequence of estimates $\{\hat{\theta}_n\}$ of unknown vector $\theta_\star$, which minimizes $f(x)$.

There are two important simpler forms of observation models:

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

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

which are included in (1).

We follow such notation: $\mathrm{E}\{\cdot\}$ is the expectation value, $\|\cdot\|_\rho$ is the norm in l_ρ and $\langle\cdot,\cdot\rangle$ ia the scalar product in $\mathbb{R}^q$.

Consider the Lyapunov function V:

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

where $\theta_\star$ is the minimum point of $f(x)$.

We make two assumptions about properties of functions $f(x)$ and $F(x,w)$ needed for proving consistency of the further algorithm:

A.1 Function $f(x)$ has a unique minimum and

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

with some constant $\mu > 0$.

A.2 For any w gradient of function $F(\cdot, w)$ satisfy:

$$\|\nabla_x F(x,w) - \nabla_x F(y,w)\|_{\frac{\rho}{\rho-1}} \leq M\|x-y\|_{\frac{\rho}{\rho-1}}$$

$\forall x, y \in \mathbb{R}^q$ with some constant $M > 0$.

3 Algorithm

Let the trial simultaneous perturbation Δ_n, $n = 1, 2, \ldots$, be a random sequence of zero-mean independent vectors from $\mathbb{R}^q$ with distributions $\mathrm{P}_n(\cdot)$, $n = 1, 2, \ldots,$, which have a uniformly bounded finite support and independent components. Consider sequences of real positive numbers $\{\alpha_n\}$ and $\{\beta_n\}$. To choose some initial vector $\hat{\theta}_0 \in \mathbb{R}^q$. In [5, 6, 7] the algorithm with two simultaneous perturbations was proposed for construction of sequences of measurement points and estimates:

$$\begin{cases} x_n^\pm = \hat{\theta}_{n-1} \pm \beta_n \Delta_n, \\ y_n^\pm = F(x_n^\pm, w_n^\pm) + v_n^\pm \ , \\ \hat{\theta}_n = \hat{\theta}_{n-1} - \alpha_n \Delta_n \frac{y_n^+ - y_n^-}{2\beta_n}. \end{cases} \qquad (2)$$

4 Convergence

Denote $\mathbb{W} = \mathrm{supp}(\mathrm{P}_w(\cdot)) \subset \mathbb{R}^p$ is the finite support of the distribution $\mathrm{P}_w(\cdot)$; $\mathcal{F}_n$ is a σ-algebra generated by $\hat{\theta}_0, \hat{\theta}_1, \ldots,$ $\hat{\theta}_n$, formed by an algorithm (2); $c_1 = \max_{w\in\mathbb{W}} \|\nabla_x F(\theta_*, w\|^\rho$; $c_2 = \max_{w^\pm\in\mathbb{W}} |F(\theta_*, w^+) - F(\theta_*, w^-)|^\rho$; $d_n = \alpha_n^\rho \beta_n^{-\rho}\rho$;

$$\gamma_n = \alpha_n(\mu - \beta_n \frac{\rho-1}{\rho} q^{\frac{\rho+1}{\rho}} M) + 2^{3\rho-2} c_1$$

$$\phi_n = \frac{1}{\rho}\alpha_n\beta_n q^{\frac{\rho+1}{\rho}} M + 2^{3\rho-2} c_1 \beta_n q^\rho + 2^{2\rho-2} c_2.$$

Theorem 1 . Let be $\rho \in (1, 2]$ and the next conditions are satisfied:
(A.1) for functions $f(x) = \mathrm{E}\{F(x,w)\}$;
(A.2) for functions $F(\cdot, w)$ $\forall w \in \mathbb{W}$;
functions $F(x,\cdot)$ and $\nabla_x F(x,\cdot)$ uniformly bounded on $\mathbb{W}$;
$\forall n \geq 1$ random values $v_1^\pm, \ldots, v_n^\pm$ and vectors $w_1^\pm, \ldots, w_{n-1}^\pm$ do not depend on $w_n^\pm, \Delta_n$, and random vectors $w_n^\pm$ do not depend on Δ_n;
$\mathrm{E}\{|v_n^+ - v_n^-|^\rho\} \leq \sigma_n^\rho$, $n = 1, 2, \ldots$; $\forall n$, $0 \leq \gamma_n \leq 1$, $\sum_n \gamma_n = \infty$, $\mu_n \to 0$ with $n \to \infty$, where

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

Then:
1) Sequence of estimations $\{\hat{\theta}_n\}$ generated by the algorithm (2) converges to $\theta_\star$ in meaning that:

$$\mathrm{E}\{V(\hat{\theta}_n)\} \to 0 \ \text{ as } \ n \to \infty;$$

2) if $\overline{\lim}_{n\to\infty} z_n \geq z > 1$, then

$$\mathrm{E}\{V(\hat{\theta}_n)\} = \mathcal{O}\left(\prod_{i=0}^{n-1}(1-\gamma_i)\right);$$

3) if $z_n \geq z > 1$ $\forall n$, then

$$\mathrm{E}\{V(\hat{\theta}_n)\} \leq (\mathrm{E}\{V(\hat{\theta}_0)\} + \frac{\mu_0}{z-1}) \prod_{i=0}^{n-1}(1-\gamma_i);$$

4) if, moreover,

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

then $\hat{\theta}_n \to \theta_\star$ with $n \to \infty$ with probability 1 and

$$\mathrm{P}\{V(\hat{\theta}_n) \leq \varepsilon, \ \forall n \geq n_0\} \geq 1 - \psi_{n_0}/\varepsilon, \qquad (3)$$

where $\psi_{n_0} = \mathrm{E}\{V(\hat{\theta}_{n_0})\} + \sum_{n=n_0}^\infty \phi_n + 2^\rho q^\rho d_n \sigma_n^\rho$.

Remark. Constant c_2 in basic case of $F(x,w) = f(x)$ can be 0.

5 Proof

We will bound the value of $V(\hat{\theta}_n)$ by the value of $V(\hat{\theta}_{n-1})$, what allows us to use Lemma from [9] to prove convergence as in the theorem statement.

Using the definition of algorithm (2) and properties of function $V(x)$, using mean-value theorem with some $t \in (0,1)$ we get

$$V(\hat{\theta}_n) \le V(\hat{\theta}_{n-1} - \alpha_n \Delta_n \frac{y_n^+ - y_n^-}{2\beta_n}) =$$

$$= V(\hat{\theta}_{n-1}) - \alpha_n \langle \nabla V(\hat{\theta}_{mid}), \Delta_n \frac{y_n^+ - y_n^-}{2\beta_n} \rangle =$$

$$= V(\hat{\theta}_{n-1}) - \frac{\alpha_n}{2\beta_n} \langle \nabla V(\hat{\theta}_{n-1} - t\alpha_n \Delta_n \frac{y_n^+ - y_n^-}{2\beta_n}), \Delta_n (y_n^+ - y_n^-) \rangle =$$

$$= V(\hat{\theta}_{n-1}) - \rho \frac{\alpha_n}{2\beta_n} (y_n^+ - y_n^-) \sum_{i=1}^{q} \mathrm{sign}_n^{(i)}(t) \Delta_n^{(i)} \times \times \left| \hat{\theta}_{n-1}^{(i)} - \theta_*^{(i)} - t\alpha_n \Delta_n^{(i)} \frac{y_n^+ - y_n^-}{2\beta_n} \right|^{\rho-1},$$

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

$$\hat{\theta}_{n-1}^{(i)} - \theta_*^{(i)} - t\alpha_n \Delta_n^{(i)} \frac{y_n^+ - y_n^-}{2\beta_n}.$$

Let $\widetilde{\mathrm{sign}}_{n-1}^{(i)} = 0$ or ± 1 depending on the sign of $\hat{\theta}_{n-1}^{(i)} - \theta_*^{(i)}$. Using inequality $-\mathrm{sign}(c-d)|c-d|^{\rho-1} b \le -\mathrm{sign}(c)|c|^{\rho-1} b + 2^{2-\rho}|d|^{\rho-1}|b|$ for $b, c, d \in \mathbb{R}$, we get:

$$V(\hat{\theta}_n) \le V(\hat{\theta}_{n-1}) - \rho\alpha_n \frac{y_n^+ - y_n^-}{2\beta_n} \times$$

$$\times \sum_{i=1}^{q} \widetilde{\mathrm{sign}}_{n-1}^{(i)} \left| \hat{\theta}_{n-1}^{(i)} - \theta_*^{(i)} \right|^{\rho-1} \Delta_n^{(i)} +$$

$$+ 2^{2-\rho} \rho \frac{\alpha_n}{2\beta_n} \sum_{i=1}^{q} \left| t\alpha_n \Delta_n^{(i)} \frac{y_n^+ - y_n^-}{2\beta_n} \right|^{\rho-1} \times$$

$$\times |\Delta_n^{(i)} (y_n^+ - y_n^-)| \le$$

$$\le V(\hat{\theta}_{n-1}) - \frac{\alpha_n}{2\beta_n} \langle \nabla V(\hat{\theta}_{n-1}), \Delta_n (y_n^+ - y_n^-) \rangle + \quad (4)$$

$$+ 2^{2-2\rho} c_n \|\Delta_n\|^\rho |y_n^+ - y_n^-|^\rho.$$

Using the model of observations (1) and the mean-value theorem for $F(\cdot, w_n)$, we derive with some $t' \in (0,1)$:

$$\Delta_n y_n = \Delta_n (F(\hat{\theta}_{n-1} + \beta_n \Delta_n, w_n) + v_n) =$$

$$= \Delta_n F(\hat{\theta}_{n-1}, w_n) + \Delta_n v_n +$$

$$+ \Delta_n \langle \nabla_x F(\hat{\theta}_{n-1} + t'\beta_n \Delta_n, w_n), \beta_n \Delta_n \rangle.$$

We apply the operation of conditional expectation by σ-algebra $\mathcal{F}_{n-1}$. Because of the independence of trial perturbations Δ_n from v_n and w_n we get

$$\mathrm{E}\{\Delta_n v_n | \mathcal{F}_{n-1}\} = \mathrm{E}\{\Delta_n | \mathcal{F}_{n-1}\} \mathrm{E}\{v_n | \mathcal{F}_{n-1}\} = 0,$$

$$\mathrm{E}\{\Delta_n F(\hat{\theta}_{n-1}, w_n) | \mathcal{F}_{n-1}\} =$$

$$= \mathrm{E}\{\Delta_n | \mathcal{F}_{n-1}\} \mathrm{E}\{F(\hat{\theta}_{n-1}, w_n) | \mathcal{F}_{n-1}\} = 0.$$

Consequently, for the second term of (4) we get:

$$-\alpha_n \mathrm{E}\{\langle \nabla V(\hat{\theta}_{n-1}), \Delta_n \frac{y_n^+ - y_n^-}{2\beta_n} \rangle | \mathcal{F}_{n-1}\} =$$

$$= -\alpha_n \langle \nabla V(\hat{\theta}_{n-1}), \mathrm{E}\{\Delta_n \frac{y_n^+ - y_n^-}{2\beta_n} | \mathcal{F}_{n-1}\} \rangle =$$

$$= -\frac{\alpha_n}{2\beta_n} \langle \nabla V(\hat{\theta}_{n-1}), \mathrm{E}\{\Delta_n \langle (\nabla_x F(\hat{\theta}_{n-1} +$$

$$+ t'\beta_n \Delta_n, w_n^+) + \nabla_x F(\hat{\theta}_{n-1} - t''\beta_n \Delta_n, w_n^-)),$$

$$\beta_n \Delta_n \rangle | \mathcal{F}_{n-1}\} \rangle \le -\frac{\alpha_n}{2} \langle \nabla V(\hat{\theta}_{n-1}),$$

$$\mathrm{E}\{\Delta_n \langle \nabla_x F(\hat{\theta}_{n-1}, w_n^+) + F(\hat{\theta}_{n-1}, w_n^-), \Delta_n \rangle | \mathcal{F}_{n-1}\} \rangle +$$

$$+ \frac{\alpha_n}{2} | \langle \nabla V(\hat{\theta}_{n-1}), \mathrm{E}\{\Delta_n \langle \nabla_x F(\hat{\theta}_{n-1} + t'\beta_n \Delta_n, w_n) +$$

$$+ \nabla_x F(\hat{\theta}_{n-1} - t''\beta_n \Delta_n, w_n) -$$

$$- \nabla_x F(\hat{\theta}_{n-1}, w_n^+) - \nabla_x F(\hat{\theta}_{n-1}, w_n^-), \Delta_n \rangle | \mathcal{F}_{n-1}\} \rangle |$$

From uniform boundness of function $\nabla_x F(\cdot, w_n)$, Hoelder inequality [10] and conditions (A.1) and (A.2), using Yung inequality [10]: $a^{1/r} b^{1/s} \le \frac{1}{r} a + \frac{1}{s} b$, $r > 1$, $a, b > 0$, $\frac{1}{r} + \frac{1}{s} = 1$, we derive

$$-\frac{\alpha_n}{2\beta_n} \mathrm{E}\{\langle \nabla V(\hat{\theta}_{n-1}), \Delta_n (y_n^+ - y_n^-) \rangle | \mathcal{F}_{n-1}\} \le$$

$$\le -\alpha_n \langle \nabla V(\hat{\theta}_{n-1}), \nabla f(\hat{\theta}_{n-1}) \rangle +$$

$$+ \frac{\alpha_n}{2} V(\hat{\theta}_{n-1})^{\frac{\rho-1}{\rho}} q^{1/\rho} \times$$

$$\times |\mathrm{E}\{\langle \nabla_x F(\hat{\theta}_{n-1} + t'\beta_n \Delta_n, w_n^+) - \nabla_x F(\hat{\theta}_{n-1}, w_n^+) +$$

$$\nabla_x F(\hat{\theta}_{n-1} - t''\beta_n \Delta_n, w_n^-) -$$

$$- \nabla_x F(\hat{\theta}_{n-1}, w_n^-), \Delta_n \rangle | \mathcal{F}_{n-1}\}| \le$$

$$\leq -\alpha_n \mu V(\hat{\theta}_{n-1}) + \alpha_n V(\hat{\theta}_{n-1})^{\frac{\rho-1}{\rho}} q^{\frac{2}{\rho}} \times$$

$$\times M \|\beta_n \Delta_n\|_{\frac{\rho}{\rho-1}} \leq -\alpha_n \mu V(\hat{\theta}_{n-1}) +$$

$$+\alpha_n \left(\frac{\rho-1}{\rho} V(\hat{\theta}_{n-1}) + \frac{1}{\rho} \right) q^{\frac{\rho+1}{\rho}} M \beta_n \leq$$

$$\leq -\alpha_n (\mu - \beta_n \frac{\rho-1}{\rho} q^{\frac{\rho+1}{\rho}} M) V(\hat{\theta}_{n-1}) + \frac{1}{\rho} \alpha_n \beta_n q^{\frac{\rho+1}{\rho}} M.$$

Let's bound the third term in the right side of inequality (4). First, for some point x_m, from the segment between $\hat{\theta}_{n-1} + \beta_n \Delta_n$ and θ_* from mean-value theorem, using Hoelder inequality, conditions (A.2) and inequality $(\frac{a+b+c+d}{4})^\rho \leq \frac{1}{4}(a^\rho + b^\rho + c^\rho + d^\rho)$, we get:

$$|y_n^+ - y_n^-|^\rho = |F(x_n^+, w_n^+) - F(\theta_*, w_n^+) + F(\theta_*, w_n^+) -$$

$$-F(\theta_*, w_n^-) + F(\theta_*, w_n^-) - F(x_n^-, w_n^-) + v_n^+ - v_n^-|^\rho \leq$$

$$\leq 2^{2\rho-2} \|\nabla_x F(x_m^+, w_n^+)(\hat{\theta}_{n-1} + \beta_n \Delta_n - \theta_*)\|^\rho +$$

$$+2^{2\rho-2} \|\nabla_x F(x_m^-, w_n^-)(\hat{\theta}_{n-1} - \beta_n \Delta_n - \theta_*)\|^\rho +$$

$$+2^{2\rho-2} |F(\theta_*, w_n^+) - F(\theta_*, w_n^-)|^\rho + 2^{2\rho-2} |v_n^+ - v_n^-|^\rho \leq$$

$$\leq 2^{3\rho-2} c_1 (V(\theta_{n-1}) + \beta_n q^\rho) +$$

$$+2^{2\rho-2} c_2 + 2^{2\rho-2} |v_n^+ - v_n^-|^\rho.$$

For the conditional expectation of the third term in (4) without coefficient $2^{2-\rho} c_n$, using independence of Δ_n and v_n we get:

$$\mathrm{E}\{\|\Delta_n\|^\rho |y_n^+ - y_n^-|^\rho | \mathcal{F}_{n-1}\} \leq$$

$$\leq \mathrm{E}\{\|\Delta_n\|^\rho | \mathcal{F}_{n-1}\} (2^{3\rho-2} c_1 (V(\theta_{n-1}) + \beta_n q^\rho) +$$

$$+2^{2\rho-2} c_2 + 2^{2\rho-2} \mathrm{E}\{|v_n^+ - v_n^-|^\rho | \mathcal{F}_{n-1}\})$$

Using the notation defined and the estimations found, inequality (4) can be rewritten as:

$$V(\hat{\theta}_n) \leq V(\hat{\theta}_{n-1})(1 - \gamma_n) + \phi_n +$$

$$+2^\rho q^\rho d_n \mathrm{E}\{|v_n^+ - v_n^-|^\rho | \mathcal{F}_{n-1}\}).$$

Applying expectation, the following is achieved:

$$\mathrm{E}\{V(\hat{\theta}_n)\} \leq \mathrm{E}\{V(\hat{\theta}_{n-1})\}(1 - \gamma_n) + \phi_n +$$

$$+2^\rho d_n \sigma_n^\rho$$

Next, we use the Lemma from [9], page 90. The statements of this theorem imply from the corresponding statements of [9].

6 Conclusions

Convergence, as we have seen in the theorem formulation, depends on values of some coefficients, and inequalities, which need to be satisfied. After this inequalities are written once, they need to be simplified and then the importance of each of them should be analyzed.

References:

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

[2] B.T. Polyak and A.B. Tsybakov, Optimal Orders of Accuracy for Search Algorithms of Stochastic Optimization, Problems Inform. Transmission, vol.26, 1990, 2, pp.126–133.

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

[4] O.N. Granichin, Estimation of parameters of linear regression under arbitrary disturbances, Automat. Remote Contr., 2002, 1, pp. 30-41.

[5] O.N. Granichin, Randomized algorithms of stochastic approximation under arbitrary noise, Automat. Remote Contr., 2002, 2, pp. 44-55.

[6] O.N. Granichin, Optimal convergence rate of randomized algorithms of stochastic approximation under arbitrary noise, Automat. Remote Contr., 2003, 2, pp. 88-99.

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

[8] O.N. Granichin, S.S. Sysoev and A.T. Vakhitov, Precision of estimation of randomized algorithm for stochastic approximation, Automat. Remote Contr.,to appear, 2006.

[9] B.T. Polyak, Convergence and rate of convergence in iterative stochastic processes.I.The general case, Automat. Remote Contr. 12, 1976,pp. 83-94.

[10] V.A. Zorich, Mathematical Analysis, MCNMO, Moscow, 2001.