Proceedings of the American Control Conference
Chicago, Illinois • June 2000

SPSA in noise free optimization [1]

László Gerencsér, Zsuzsanna Vágó
Computer and Automation Institute of the Hungarian Academy of Sciences,
H-1111 Budapest, Kende 13-17, Hungary, gerencser@sztaki.hu, vago@oplab.sztaki.hu

Abstract

The SPSA (simultaneous perturbation stochastic approximation) method for function minimization developed in [15] is analyzed for optimization problems without measurement noise. We prove the striking result that under appropriate technical conditions the estimator sequence converges to the optimum with geometric rate with probability 1. Numerical experiments support the conjecture that the top Lyapunov-exponent of defined in terms of the SPSA method is smaller than the Lyapunov-exponent of its deterministic counterpart. We conclude that randomization improves convergence rate while dramatically reducing the number of function evaluations.

Keywords: optimization; stochastic approximation; recursive estimation; Kiefer-Wolfowitz-methods; random products; Lyapunov-exponents.

1 Introduction

The aim of this paper is to analyze the convergence properties of the simultaneous perturbation stochastic approximation (SPSA) method for function minimization developed in [15] when the function values can be computed without measurement error. A basic feature of Spall's method is a new way of estimating the gradient using only two measurements at properly selected random parameter values. The application of SPSA is justified when the function evaluation is expensive.

The SPSA methods has been proposed in [15] where, under appropriate technical conditions, the almost sure convergence of the estimator process has been established. In the same paper asymptotic normality of a properly scaled estimation error process has been established. Similar results under weaker conditions have been obtained in [1]. A rate of convergence result for higher order moments of the estimation error has been given in [7]. A number of ideas related to SPSA methdos are given in [10]. For an up to date survey see [14].

SPSA for noise-free optimization was briefly considered in [7]. It was shown there that under suitable technical conditions the rate of convergence for the L_q-norms of the estimations error is $O(k^{-1/2})$, for any $q \geq 1$. In fact, in the noise-free case the SPSA procedure can be analyzed using results for Robbins-Monroe-type procedures. In particular, the asymptotic covariance of $k^{1/2}(\widehat{\theta}_k - \theta^*)$ can be determined using classical results of [11]. It is easy to see that, due to the multiplicative effect of the noise, this asymptotic covariance is equal to zero. Hence a convergence rate faster than $O(k^{-1/2})$ is expected. In fact, using the analysis of [7] in an inductive argument and exploiting the multiplicative nature of the noise it can be shown that the convergence rate is $O(k^{-m})$ for any finite m.

The question thus arises what is the actual rate of convergence of $\widehat{\theta}_k - \theta^*$ and what is the best choice of the perturbation size and the step-size. The main result of the paper is that fixed gain SPSA applied to noise-free optimization yields geometric rate of convergence almost surely, just like for *deterministic* gradient methods under appropriate conditions. In the terminology of optimization theory we would say that the convergence rate is linear indicating the error at step $k+1$ is less than constant multiple of an upper bound for the error at step k and this constant is strictly smaller than 1.

2 The problem formulation

The p-dimensional Euclidean-space will be denoted by $\mathbb{R}^p$. The Euclidean-norm of a vector x will be denoted by $|x|$. The operator norm of a matrix A will be denoted be $\|A\|$, i.e. $\|A\| = \sup_{x \neq 0} |Ax|/|x|$. We consider the following problem: minimize the function $L(\theta)$ defined for $\theta \epsilon \mathbb{R}^p$ under the following conditions:

Condition 2.1 *The function $L(.)$ is three-times continuously differentiable with respect to θ with bounded derivatives up to order three in any bounded domain. It is assumed that $L(.)$ has a unique minimizing value and it will be denoted by θ^*. We assume $\theta^* = 0$.*

A key assumption is that the computation of $L(.)$ is expensive and the gradient of $L(.)$ is not computable

[1] This work was supported by the National Research Foundation of Hungary (OTKA) under Grants no. T 032932 and T 029579

0-7803-5519-9/00 $10.00 © 2000 AACC

at all. Thus to minimize $L(.)$ we need a numerical procedure to estimate the gradient of $L(.)$ denoted by

$$G(\theta) = L_\theta(\theta). \tag{1}$$

Following [15] we consider random perturbations of the components of θ. For this we first consider a sequence of independent, identically distributed (i.i.d.) random variables Δ_{ki}, $k = 1, ...,$ $i = 1, ..., p$ defined over some probability space $(\Omega, \mathcal{F}, \mathcal{P})$ satisfying certain weak technical conditions. E.g. they may be chosen Bernoulli with

$$P(\Delta_{ki} = +1) = 1/2 \qquad P(\Delta_{ki} = -1) = 1/2.$$

Now let $0 < c_k \leq 1$ be a fixed sequence of positive numbers. A standard choice for c_k, proposed in [15] is $c_k = c/k^\gamma$ with some $\gamma > 0$. For any $\theta \epsilon \mathbb{R}^p$ we evaluate $L(.)$ at two randomly and symmetrically chosen points $\theta + c_k\Delta_k$ and $\theta - c_k\Delta_k$, respectively. Define the random vector

$$\Delta_k^{-1} = \left[\Delta_{k1}^{-1}, \ldots, \Delta_{kp}^{-1}\right]^T.$$

Then the estimator of the gradient is defined as

$$H(k,\theta) = \Delta_k^{-1}\frac{1}{2c_k}\Big(L(\theta + c_k\Delta_k) - L(\theta - c_k\Delta_k)\Big).$$

The SPSA procedure is then defined by

$$\widehat{\theta}_{k+1} = \widehat{\theta}_k - \frac{a}{k+1}H(k+1, \widehat{\theta}_k) \tag{2}$$

with $a > 0$ fixed. The convergence of this procedure under various conditions, in particular under the assumed presence of measurement noise, has been analyzed in in [15], [1], [7].

The analysis given in [7] is based on the adaptation of an ODE-method developed earlier in [5]. The characteristics of an ODE-method is that a segment of the piecewise linear trajectory defined by the estimator sequence $\widehat{\theta}_k$ is approximated by the solution trajectory of an associated ODE, which is defined by

$$\dot{y}_t = -\frac{a}{t}G(y_t) \qquad y_s = \xi \quad a > 0. \tag{3}$$

The convergence rate of the ODE to θ^* is $O(t^{-c})$ with some $c > 0$. A simple calculation gives that, assuming the validity of an ODE principle for a noise-free SPSA method, we should increase the stepsize from a/t. A radical choice is to consider a *fixed* stepsize and thus we consider the ODE

$$\dot{y}_t = -aG(y_t) \qquad y_s = \xi \quad a > 0. \tag{4}$$

Correspondingly, we will consider a fixed gain SPSA method defined the recursion

$$\widehat{\theta}_{k+1} = \widehat{\theta}_k - aH(k+1, \widehat{\theta}_k). \tag{5}$$

Fixed gain SPSA methods have been first considered in [8] in connection with discrete optimization. It has been shown there that choosing the perturbation size $c = a^{1/6}$, the error of the estimator is of the order of magnitude $O_M(a^{1/3})$. The notation $O_M(.)$ means that the $L_q(\Omega, \mathcal{F}, \mathcal{P})$-norm of the left hand side decreases with the rate given on the right hand side for any $q \geq 1$.

Condition 2.2 *Let $y(t, s, \xi)$ denote the unique solution of (4). Then with some $C_0, \alpha > 0$*

$$|y(t, s, \xi)| \leq C_0 e^{-a\alpha(t-s)}|\xi|$$

for every $\xi, t > s \geq 1$. Furthermore we have

$$\|\frac{\partial}{\partial \xi}y(t, s, \xi)\| \leq C_0 e^{-a\alpha(t-s)}. \tag{6}$$

The convergence properties of the proposed fixed gain SPSA method will be first established for quadratic functions. We have the following result:

Theorem 2.1 *Let L be a positive definite quadratic function, with Hessian-matrix A and let the smallest eigenvalue of A be α. Assume that the size of the perturbation, c is fixed. Then, for sufficiently small a there is a deterministic constant $\lambda < 0$ depending on a, such that for any initial condition θ_0 outside of a set of Lebesgue-measure zero we have*

$$\lim_{k\to\infty}\frac{1}{k}\log|\widehat{\theta}_k - \theta^*| = \lambda$$

with probability 1.

Remark: For an arbitrary initial condition we have $\leq$ in place of the equality. Thus the sequence of estimators $\widehat{\theta}_k$ converges to θ^* with geometric rate with probability 1. The deterministic constant λ is called the top *Lyapunov-exponent* of the problem. Comments on its magnitude will be given in the next section.

3 The proof

The proof of the theorem is surprisingly simple. First, it is easy to see that for quadratic functions

$$H(k,\theta) = \Delta_k^{-1}\Delta_k^T G(\theta).$$

But $G(\theta) = A(\theta - \theta^*)$, hence we get the following recursion for $\delta\theta_k = \theta - \theta^*$:

$$\delta\theta_{k+1} = (I - a\Delta_k^{-1}\Delta_k^T A)\delta\theta_k. \tag{7}$$

Now the sequence Δ_k is i.i.d., hence the matrix-valued process

$$A_k = (I - a\Delta_k^{-1}\Delta_k^T A)$$

is stationary and ergodic. Applying Oseledec's multiplicative ergodic theorem (cf. [12, 13]) the claim of the theorem follows immediately with some deterministic, not necessarily negative λ.

To establish that $\lambda < 0$ we apply [4] the conditions of which are easily verified. We also use an elementary device to couple continuous time and discrete time procedures. Since

$$\mathrm{E}\log(I - a\Delta_k^{-1}\Delta_k^T A)$$

is stable for sufficiently small a it follows that the random product $A_k A_{k-1} ... A_1$ converges to 0 at geometric rate with probability 1. Thus we must have $\lambda < 0$, as stated.

Remark: SPSA for noise-free quadratic problems is equivalent to the application of iterated random linear mappings (cf. (7)). Therefore we refer to this problem as the linear problem. In the general case of non-quadratic functions SPSA leads to the study of the effect of iterated random non-linear mappings. This case will be referred to as the non-linear problem.

The value of the top Lyapunov exponent λ is of great practical interest. A remarkable feature of λ can be easily established for the scalar case in a slightly different setting. Assume that ξ_k are positive i.i.d. non-constant random variables. Then the inequality

$$\mathrm{E}\log\xi < \log\mathrm{E}\xi$$

implies that the convergence rate of the randomized product is better than that of its deterministic counterpart. It is conjectured that a similar result holds for matrix products under appropriate conditions. The conjecture is supported by simulation results.

A theoretical expression for λ can be obtained as follows (cf. [3]). Let $Y_k = A_k ... A_1$ and define the normalized products $U_k = Y_k/||Y_k||$. Then it can be shown that the process (U_k, A_{k+1}) is asymptotically stationary. In our case the two components are also independent. Let μ denote the stationary distribution of U on the unit sphere of $p \times p$ matrices. Then we have

$$\lambda = \mathrm{E}_P \int \log||A_2 U_1||d\mu \tag{8}$$

where P denotes the probability distribution of A_2.

The invariant measure can be described by an integral equation, but this may be more useful for theoretical purposes than for practical computations.

Oseledec's theorem also implies that if the multiplicity of the top Lyapunov exponent is 1, then there is a p-dimensional *random* vector v_1 such that

$$\lim_k (\widehat{\theta}_k - \theta^*)/|\widehat{\theta}_k - \theta^*| = v_1.$$

Thus the direction of $(\widehat{\theta}_k - \theta^*)$ also converges just like in the case of the deterministic gradient method.

Finally for sufficiently small a Theorem 3 of [3] is applicable and we get that for some normalizing constants m, σ and for any pair of indices i, j

$$(\log Y_{k,ij} - nm)/(n^{1/2}\sigma) \to \mathcal{N}(0, \sigma^2)$$

in distribution.

4 The general case

In the general case a third order Taylor series expansion gives

$$H(k, \theta) = \Delta_k^{-1}\Delta_k^T G(\theta) + O(c_k^2).$$

Thus (5) defines a non-linear random mapping. The stability properties of stationary random mappings has been considered recently in [2]. However, these results are not directly applicable for non-stationary sequences of random mappings. Furthermore, even if we could ignore the presence of the residual term $O(c_k^2)$, the conditions of the cited paper are hard to verify. Namely, we have to consider the inequalities

$$|\Delta_k^{-1}\Delta_k^T G(\theta) - \Delta_k^{-1}\Delta_k^T G(\theta')| \leq L_k|\theta - \theta'|$$

where L_k is a random Lipschitz-constant, in some appropriately chosen metric $|.|$, depending only on Δ_k. Then we should check the condition

$$\mathrm{E}\log L_k < 0. \tag{9}$$

In this paper we choose to adapt the analysis of fixed gain recursive estimators given in [6]. By exploiting the multiplicative structure of the noise we can rewrite the analysis so that we arrive to inequalities which are very similar to inequalities given for the linear case in [4]. Thus the techniques of the two papers can be combined.

Since the analysis of [6] has been given primarily for continuous-time methods, we now switch to continuous-time SPSA procedures. Such procedures can be obtained by using a perturbation-process

$$\Delta(t) = (\Delta_1(t), ..., \Delta_p(t))^T$$

which is an L-mixing process such that for each t $\Delta_i(t)$ are i.i.d. Bernoulli random variables. E.g. a single component $\Delta_i(t)$ can be generated as the sign process of a Gaussian ARMA-process. Then the SPSA-method becomes

$$\frac{d}{dt}\widehat{\theta}(t) = -aH(t, \widehat{\theta}(t)). \tag{10}$$

In [6] it was assumed that $|H(k, \theta) - G(\theta)|$ is bounded. Thus the imposed stability of the associated ODE ensures that the estimator process lives in a bounded domain. In the case of fixed gain SPSA the random effect

is multiplicative and the above boundedness condition can not be guaranteed a priori. Therefore we have to modify the analysis of [6] to truncated procedures.

Let D_0 be a compact truncation domain. Then if $\widehat{\theta}_k \notin D_0$ then we reset its value the initial valued $\xi \in \text{int} D_0$. The choice of the set D_0 requires some a priori knowledge on the location of θ^*.

Unlike the decreasing gain case (cf. [5]) the probability of the resetting is typically not asymptotically vanishing. The first major step is to show that the conclusions of [6] remain valid for the truncated procedure. In particular, the tracking error $|\widehat{\theta}_t - y_t|$ is bounded by an L-mixing process the order of magnitude of which is $O_M(a^{1/2})$.

Now we give a short outline of the proof, indicating only the necessary changes to be made. First, assume that the residual term $O(c_k^2)$ is actually zero. Such situations do arise in connection with multivariable direct adaptive controller design (cf. [9]). Let the event that a resetting takes place in the interval $[nT, (n+1)T)$ be denoted by C_n. A key quantity that shows up in the ODE analysis is the normalized local tracking error η_n^* defined by

$$\sup_{\substack{nT \le t \le (n+1)T \\ \theta \in D_0}} |\int_{nT}^{t} \psi_N(s,\theta) \cdot \overline{H}(s, y(s,nT,\theta)) ds| / |\theta|$$

where $\psi_N(s,\theta) = (\partial/\partial\xi) y((n+1)T, s, y(s,nT,\theta))$ is the sensitivity matrix of the ODE and $\overline{H}(s, y(s,nT,\theta) = H(s, y(s,nT,\theta) - G(s,\theta)$. The new element here is the *normalization* by $|\theta|$. In analogy with Lemma 2.2 of [6] η_n^* can be analyzed. Furthermore, equation (2.11) of [6] can be rewritten as

$$\sup_{n \le t \le (n+1)T} |\widehat{\theta}_t - \overline{y}_t| \le c\eta^*(n) |\widehat{\theta}_{nT}| + R\chi_{C_n}$$

where χ_{C_n} is the characteristic function of the event C_n with some $R > 0$.

The above inequality reduces the non-linear problem to a linear one: except for the effect of resetting an identical inequality has been obtained as (2.6) in [4]. Thus the analysis is thus essentially reduced to the analysis of a truncated version of the linear algorithm given in [4], which can be carried out without much difficulty.

The handling of the effect of the residual term $O(c_k^2)$ is a routine exercise. Assuming $c_k = \beta^k$ with sufficiently small β, the exponential rate of convergence of the perturbed algorithm will still be ensured.

To enhance numerical stability we use Newton-type SPSA methods proposed in [17] (in condensed form in [16]) or higher order SPSA methods proposed in [7].

Finally we get the following non-linear extension of Theorem 2.1: for sufficiently small a we have

$$\limsup_{k \to \infty} \frac{1}{k} \log |\widehat{\theta}_k - \theta^*| = \lambda' < 0 \qquad (11)$$

with probability 1. We are unaware of any non-linear extension of Oseledec's theorem, thus we can not guarantee the existence of the limit on the left hand side Likewise, we are unaware of any non-linear extension of (8).

A natural idea is to consider the linear approximation of our procedure around θ^* and apply the theorem for the linear case. Unfortunately there is no guarantee that λ' is smaller than the top Lyapunov exponent of the linear problem and thus the effect of nonlinearity may stay dominant. We conclude that there seems to be no easy way to improve (11). On the other hand it would be of interest to find interesting classes of non-linear problems where $\lambda' < \lambda$. For such problems the asymptotic behaviour of the approximating linear procedure would become dominant and thus the linear theory would be applicable.

5 Simulation results

We have tested fixed gain SPSA for quadratic functions in $p = 20$ and $p = 50$ dimensions. In the gradient estimation we have chosen $c = 0.5$ and we have performed 1000 iterations. The purpose of the simulations was to provide empirical evidence for the viability of the method. In addition we have found the approximate value of the top Lyapunov-exponent for which no closed form theoretical expression exists in general. In the attached figures we plot

$$\hat{\lambda} = \frac{1}{k} \log |\widehat{\theta}_k - \theta^*|$$

against k. On Figures 1 and 2 we have two examples with nice stability properties with the following parameters:

$p = 20$, solid line: $a = 0.01$, dotted line: $a = 0.005$
$p = 50$, solid line: $a = 0.01$, dotted line: $a = 0.005$.

The remarkable thing is that the top Lyapunov-exponents are significantly larger in absolute value than what you would get for the deterministic algorithm. The latter would be roughly $-a$ for the present example.

If a is too small then, predictably, the top Lyapunov-exponent is very close to zero. On the other hand, if a is too large then the algorithm loses its stability, just like in the deterministic case. The stability of the empirical value of the Lyapunov exponent predicted by Oseledec's theorem is nicely demonstrated in all cases.

Acknowledgement

The first author expresses his thanks to James C. Spall and John L. Maryak for initiating and cooperating in this research and providing a number of useful comments on this particular paper and to Lorenzo Finesso, Francois Legland and Hakan Hjalmarssson for calling his attention to some useful references.

References

[1] H.F. Chen, T.E. Duncan, and B. Pasik-Duncan. A Kiefer-Wolfowitz algorithm with randomized differences. *IEEE Trans. Automat. Contr.*, 44:442–453, 1999.

[2] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41:45–76, 1999.

[3] H. Furstenberg and H.Kesten. Products of random matrices. *Ann. Math. Statist.*, 31:457–469, 1960.

[4] L. Gerencsér. Almost sure exponential stability of random linear differential equations. *Stochastics*, 36:411–416, 1991.

[5] L. Gerencsér. Rate of convergence of recursive estimators. *SIAM J. Control and Optimization*, 30(5):1200–1227, 1992.

[6] L. Gerencsér. On fixed gain recursive estimation processes. *J. of Mathematical Systems, Estimation and Control*, 6:355–358, 1996. Retreival code for full electronic manuscript: 56854.

[7] L. Gerencsér. Rate of convergence of moments for a simultaneuous perturbation stochastic approximation method for function minimization. *IEEE Trans. Automat. Contr.*, 44:894–906, 1999.

[8] L. Gerencsér, S. D. Hill, and Zs. Vágó. Optimization over discrete sets via SPSA. In *Proceedings of the 38-th Conference on Decision and Control, CDC'99*, pages 1791–1794. IEEE, 1999.

[9] H. Hjalmarsson. Efficient tuning of linear multivariable controllers using iterative feedback tuning. *Int. J. Adapt.Control Signal Process.*, 13:553–572, 1999.

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

[11] M.B. Nevel'son and R.Z. Has'minskii. *Stochastic Approximation and Recursive Estimation.* American Mathematical Soc., Providence RI, 1976.

[12] V. I. Oseledec. A multiplicative ergodic theorem. Lyapunov charasteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.*, 19:197–231, 1968.

[13] M. S. Ragunathan. A proof of oseledec's multiplicative ergodic theorem. *Israel Journal of Mathematics*, 42:356–362, 1979.

[14] J. C. Spall. Stochastic optimization and the simultaneous perturbation method. In P.A. Farrington, H.B. Nembhard, D. T. Sturrock, and G.W. Evans, editors, *Proc. of the 1999 Winter Simulation Conference, Phoenix, AZ, USA*, pages 101–109, 1999.

[15] J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Contr.*, 37:332–341, 1992.

[16] J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. In *Proceedings of the 1998 IEEE CDC*, pages 3872 – 3879, 1998.

[17] J.C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. on Automat. Contr.*, 45:in press, 2000.

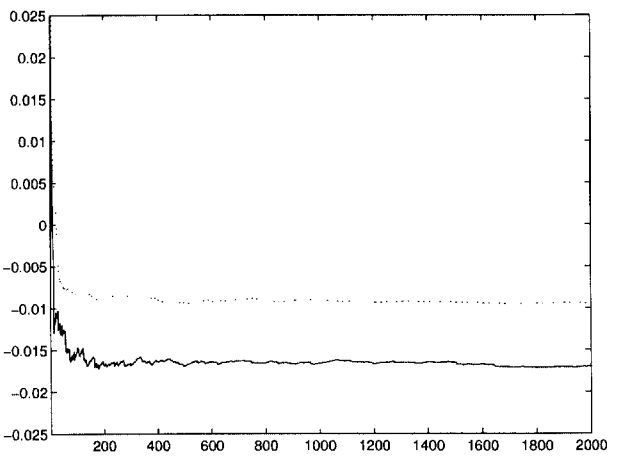

Figure 1: $p = 20$, $a = 0.01$ and $a = 0.005$, resp.

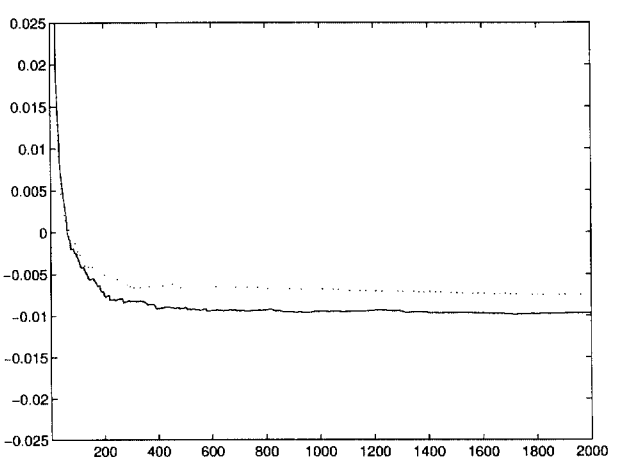

Figure 2: $p = 50$, $a = 0.01$ and $a = 0.005$, resp.