Yugoslav Journal of Operations Research
22 (2012), Number 1, 131-140
DOI:10.2298/YJOR101124003J

EFFICIENCY OF THE STOCHASTIC APPROXIMATION METHOD*

M. JAPUNDŽIĆ
Higher School of Professional Business Studies,
Novi Sad, Serbia
milos.japundzic@gmail.com

Received: November 2010 / Accepted: September 2011

Abstract: The practical aspect of the stochastic approximation method (SA) is studied. Specifically, we investigated the efficiency depending on the coefficients that generate the step length in optimization algorithm, as well as the efficiency depending on the type and the level of the corresponding noise. Efficiency is measured by the mean values of the objective function at the final estimates of the algorithm, over the specified number of replications. This paper provides suggestions how to choose already mentioned coefficients, in order to achieve better performance of the stochastic approximation algorithm.

Keywords: Stochastic approximation, step length, efficiency of the stochastic methods, noise.

MSC: 49K45, 62L20, 90C15

1. INTRODUCTION

There have been countless applications of the stochastic approximation method, in the period greater than a half century, since the seminal publication of Robbins and Monro [7] appeared. Some areas include neural network, simulation-based optimization, evolutionary algorithms, machine learning, experimental design, and signal processing applications such as noise cancellation and pattern recognition. This method is primarily used for solving systems of nonlinear equations in the presence of noisy measurements

$$g(\theta) = 0,\ \theta \in \Theta \subseteq \boldsymbol{R}^n \tag{1}$$

* Some results contained in this paper were first published in the author's MSc thesis [4].

where is $g(\theta) \in \boldsymbol{R}^n$. So, the problem of interest is a typical nonlinear system of n equations with n unknowns, based on noisy measurements of $g(\theta)$ in the form

$$Y(\theta) = g(\theta) + e(\theta), \tag{2}$$

where $e(\theta)$ represents the noise term.

The problem (1) is under certain assumptions equivalent to the problem of stochastic optimization

$$\min_{\theta \in \Theta} L(\theta) = E\left[y(\theta)\right]. \tag{3}$$

where $\Theta \subseteq \boldsymbol{R}^n$ is the domain of allowable values for a vector θ, $L(\theta)$ scalar function with n unknowns called objective function, $y(\theta)$ value of objective function L at vector θ in the presence of noise, and $E[y(\theta)]$ expected value. Namely, if we want to solve problem (3) for a differentiable function L, we can convert it into problem (1) choosing the gradient of objective function ($g(\theta) \equiv \partial L(\theta)/\partial\theta$) for function $g(\theta)$. Conversely, the problem (1) can be converted directly into an optimization problem by noting that θ with $g(\theta) = 0$ is equivalent to θ such that $\left\| g(\theta) \right\|$ is minimized for any vector norm $\| \ \|$. The choise of formulation typically depends on issues such as the basic applied problem structure and data format, the available and relevant search algorithms, the proclivities of the analyst and traditions of his/her field, and the available software.

Stochastic approximation algorithm is motivated by the deterministic steepest descent algorithm with the noisy measurement (2) replacing the exact root-finding function $g(\theta)$. In the case of unconstrained optimization the algorithm has the form

$$\theta_{k+1} = \theta_k - a_k Y_k(\theta_k), \quad k = 0,1,2,\ldots, \tag{4}$$

while in the case of constrained optimization that form is

$$\theta_{k+1} = \Psi_\Theta[\ \theta_k - a_k Y_k(\theta_k)], \quad k = 0,1,2,\ldots, \tag{5}$$

where Ψ_Θ is a user-defined mapping that projects any point out of the constraint domain Θ to a new point inside Θ. For both iterative rules, (4) and (5), holds $Y_k(\theta_k) = g(\theta_k) + e_k(\theta_k)$, where distribution of the noise $e_k(\theta_k)$ may vary from iteration to iteration ($g(\theta)$ is the gradient of objective function). An important special case is where $\{e_k\}_{k=0}^{\infty}$ is an independent and identically distributed (i.i.d.) sequence of mean-zero random vectors. In that case holds $e_k(\theta_k) = e(\theta_k)$ and $Y(\theta_k) = g(\theta_k) + e(\theta_k)$, for all $k = 0,1,2,\ldots$.

The basic results related to the stochastic approximation method can be found in [7] and [11], where convergence analysis of this method is presented. Contrary to it, in this paper we investigated the efficiency depending on the coefficients that generate the step length in optimization algorithm, as well as the efficiency depending on type and level of the corresponding noise.

The paper is organised as follows. In Section 2 we present a set of sufficient conditions for almost sure convergence of stochastic approximation method, with emphasis on the condition that refers to the step length. Section 3 contains analysis of choice of the coefficients that generate step length, where we proposed the way to choose these coefficients in order to achieve better performance of the algorithm. That section also contains numerical results that justify proposed choice of coefficients. All numerical results are obtained by using programming language Matlab.

2. CONVERGENCE OF STOCHASTIC APPROXIMATION

It is of interest for any search algorithm to know whether the iterate θ_k generated with stochastic approximation method converges to a solution θ^* as $k\to\infty$. That result guarantees that the iteration θ_k will fall into a small neighborhood of a solution θ^* after sufficient function evaluations. Many sufficient conditions were given for almost sure convergence of the SA recursions in (4) and (5). We shall present so called ″statistics″ conditions.

2.1. Convergence conditions

This subsection presents a set of sufficient conditions for almost sure convergence of the stochastic approximation iterations θ_k. These conditions are applicable if there is a unique root of problem (1). Hence, when used for optimization ($\partial L(\theta)/\partial\theta=0$), they could be applied if there are no local minima different from the (unique) global minimum. Note that these conditions are the sufficient ones, but many practical implementation of SA will produce satisfactory results even if one or more of the conditions are not satisfied.

″Statistics″ conditions:

(step length)

$$a_k > 0, \quad a_k \to 0, \quad \sum_{k=0}^{\infty} a_k = \infty, \quad \text{and} \quad \sum_{k=0}^{\infty} a_k^2 < \infty \ . \tag{C.1}$$

(search direction)

For some symmetric, positive definite matrix B and every $0< \eta <1$, (C.2)

$$\inf_{\eta<|\theta-\theta|<1/\eta} (\theta-\theta^*)^T B g(\theta) > 0$$

(mean-zero noise)

$$E[e_k(\theta)] = 0, \text{ for all } \theta \text{ and } k = 0,1,2,... \tag{C.3}$$

(growth and variance bounds)

$$\| g(\theta) \|^2 + E(\| e_k(\theta) \|^2) \le c(1+\| \theta \|^2), \tag{C.4}$$

for all θ and $k = 0,1,2,...$ and some $c>0$.

From the point of view of the user's input, condition C.1 is the most relevant one. This condition provides a careful balance in having the gain $\{a_k\}$ decay neither too fast nor too slow. In particular, the gain should approach zero sufficiently fast ($a_k\to 0$, $\sum_{k=0}^{\infty} a_k^2 < \infty$) to damp out the noise effects as the iterate gets near the solution θ^*, but it should also approach zero sufficiently slow ($\sum_{k=0}^{\infty} a_k = \infty$) to avoid premature (false) convergence of the algorithm. The choice of the gain sequence a_k is critical to the performance of stochastic approximation algorithm. The scaled harmonic sequence $a_k = a/(k+1)$, $a>0$, $k = 0,1,2,...,$ is the best-known example of a gain sequence that satisfies

condition C1. Usually, some numerical experimentations are required to choose the best value of the coefficient *a* that appears in the gain.

A common generalization of the harmonic sequence is $a_k = a/(k+1)^{\alpha}$ for strictly positive values *a* and α. From basic calculus, picking $1/2 < \alpha \leq 1$ yields to $\{a_k\}$ satisfying the conditions $\sum_{k=0}^{\infty} a_k = \infty$ and $\sum_{k=0}^{\infty} a_k^2 < \infty$ appearing in C.1.

When the desirability for a gain sequence that balances algorithm stability in the early iterations with nonnegligible step sizes in the later iterations is given, than the recommended gain form is

$$a_k = \frac{a}{(k+1+A)^{\alpha}}, \quad 1/2 < \alpha \leq 1, \tag{6}$$

where is $a>0$ and $A \geq 0$. Coefficient *A* is called stability constant, because it affects the stability of the algorithm.

The problem is how to choose the coefficients *a* and *A* in (6), to ensure the convergence of the SA. If we choose $A=0$, there are some potential problems depending on the size of the coefficient *a*. Choosing a large numerator *a*, in hope for producing nonnegligible step sizes after the algorithm has been running awhile, may cause unstable behavior in early iterations (when the denominator is still small). On the other hand, choosing a small *a*, can lead to a stable behavior in early iterations but sluggish performance in later iterations. For this reason, picking $A>0$ is usually recommended. A strictly positive *A* allows choice of a larger *a* without risking unstable behavior in early iterations. Then, in later iterations, the coefficient *A* in the denominator becomes negligible relative to the *k* while the relatively large *a* in the numerator helps maintain a nonnegligible step size. In [11] Spall recommended, as a reasonable choice for the stability constant, to pick *A* such that it is approximately 5 to 10 percent of the total number of allowed iterations in the search process.

2.2. Rate of convergence

However, convergence itself gives no information about the rate the iterates approach the solution. For that purpose we need the probability distribution of iterations θ_k, since the iterations generated by SA are random vectors. Knowledge of the distribution gives us a guidance to chose the a_k so as to minimize the likely deviation of θ_k from θ^*.

General results on the asymptotic distribution of the SA iterate θ_k are given in Fabian [2]. His work is a generalization of the first asymptotic distribution results for SA in Chung [1] and Sacks [8]. Fabian shows that, under appropriate regularity conditions,

$$k^{\frac{\alpha}{2}}(\theta_k - \theta^*) \overset{dist.}{\rightarrow} N(0,\Sigma), k \rightarrow \infty \tag{7}$$

where $\overset{dist.}{\rightarrow}$ denotes "converges in distribution", Σ is some covariance matrix that depends on the coefficients in the gain sequence a_k and on the Jacobian matrix of $g(\theta)$, and α governs the decay rate for the SA gain $\{a_k\}$. The intuitive interpretation of (7) is that iteration θ_k is approximately normally distributed with mean θ^* and covariance matrix Σ / k^{α} for *k* reasonably large.

Expresion (7) implies that the rate at which the iterate θ_k approaches θ^* is proportional, in a stochastic sense, to $k^{-\alpha/2}$ for large k. Under condition C.1 on the gain a_k for convergence of the iterate, the rate of convergence of θ_k to θ^* is maximized at $\alpha = 1$ when the gain has the standard form $a_k = a/(k+1)^{\alpha}$. That is, the maximum rate of convergence for the root-finding SA algorithm under the general conditions is $0(1/\sqrt{k})$ in an appropriate stochastic sense.

3. THE CHOICE OF COEFFICIENTS

Let us consider the standard stochastic optimization problem without constraints

$$\min_{\theta} L(\theta) = E\left[y(\theta)\right]. \tag{8}$$

Suppose that the gradient of the objective function $g(\theta)$ can only be measured in the presence of the noise $e(\theta)$. More specifically, suppose that measurements of $g(\theta)$ at any θ are available as $Y_k(\theta) = g(\theta) + e_k(\theta)$, $k = 0,1,2,...$ Estimation θ_k, which is close to the true solution θ^* of the problem (8), in most cases does not have to be the best in the sense of values of $L(\theta)$. This is the reason why efficiency is measured by the mean values of the objective function at the final estimations of solution. The true objective function values $L(\theta_k)$ are used in constructing all tables and figures. These values are not available to the algorithm, which uses only noisy measurements $Y(\theta)$ at the various values of θ.

In this section we shall analyse the choice of the coefficients a, A, and α appearing in the term

$$a_k = \frac{a}{(k+1+A)^{\alpha}}, \quad 1/2 < \alpha \le 1.$$

As it was already mentioned in Section 2, coefficient A is stability constant, while coefficient α regulates the decay rate of the gain $\{a_k\}$. The rate of convergence of θ_k to θ^* is maximized at $\alpha = 1$, but in practical problems it may not be the best to choose that value, because in practice, it is often (but not always as we shall see) preferable to have a slower decay rate. The intuitive reason for the desirability of $\alpha < 1$ is that a slower decay provides a larger step size in the iterations with large k, allowing the algorithm to move in bigger steps toward the solution. Intuitively, if the standard deviation of the noise is large, it is more difficult to converge to the solution. Acording to our numerical tests, in the case of larger deviation of the noise, smaller values of the coefficient a are desirable. It seems to be a reasonable choice, because smaller values of a could neutralize negative effects of the noise.

In order to verify the reported conclusions, a computer program is coded in Matlab to solve two standard test functions:

Test function 1:

$$L(\theta) = t_1^4 + t_1^2 + t_1 t_2 + t_2^2,$$

initial point $\theta_0 = [1,1]^T$,

optimal point $\theta^* = [0,0]^T$, $L(\theta^*) = 0$.

Test function 2 (Rozenbrock function):

$L(\theta) = 100(t_2 - t_1^2)^2 + (1 - t_1)^2$,

initial point $\theta_0 = [0,0]^T$,

optimal point $\theta^* = [1,1]^T$, $L(\theta^*) = 0$.

In the case of the test function 1, we suppose that the gradient of the objective function $g(\theta)$ can only be measured in the presence of the $\boldsymbol{N}(0, 0.1^2 I_2)$ noise (I_2 is a identity matrix of dimension 2). Table 1 shows the mean values of test function 1 at final estimates over $p = 10$ replications and $k = 10$ iterations. The pair of coefficients (A,a) represents the optimal choice of the coefficients, in the sense that their realization gives the smallest value of the mean values of the objective function. As we can see in Table 1, in this case ($k = 10$ iterations), better results are obtained by choosing $\alpha = 0.501$.

Table 1: Sample means for terminal values of the objective function in the case of $p = 10$ replications and $k = 10$ iterations

$\alpha = 0.501$			$\alpha = 1$		
A	a	Sample mean	A	a	Sample mean
0.5	0.35	$7.8055 * 10^{-4}$	0.5	0.5	$1.9022 * 10^{-3}$
0.6	0.36	$7.8962 * 10^{-4}$	0.6	0.53	$1.5971 * 10^{-3}$
0.7	0.37	$7.9857 * 10^{-4}$	0.7	0.56	$1.4094 * 10^{-3}$
0.8	0.38	$8.0704 * 10^{-4}$	0.8	0.59	$1.3097 * 10^{-3}$
0.9	0.39	$8.1592 * 10^{-4}$	0.9	0.62	$1.2746 * 10^{-3}$
1	0.4	$8.2643 * 10^{-4}$	1	0.65	$1.2855 * 10^{-3}$

The approximate optimal values of the coefficient a are chosen by trial and error over $k = 1000$ iterations. In the case of $\alpha = 0.501$, for the test function 1, the optimal choice of coefficient a is $0.085 \leq a \leq 0.092$, while in the case $\alpha = 1$, optimal choice is $1.5 \leq a \leq 2$. Figure 1 is created choosing $a = 0.09$, and $a = 1.8$, where the mean values of the test function 1 at the final estimates over p = 10 replications and k = 1000 iterations are presented. In contrast to the previous case of k = 10 iterations, here we can see that better results are obtained by choosing $\alpha = 1$. This appears to be a consequence of asymptotic theory when using 1000 measurements to estimate only two parameters.

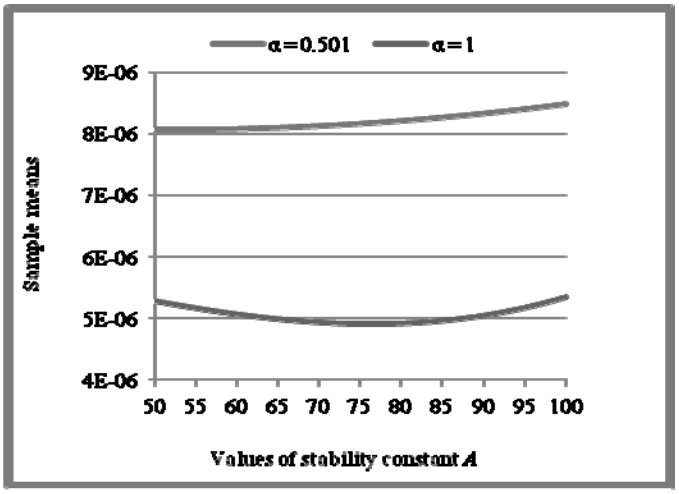

Figure 1: Sample means for terminal values of objective function in the case of p = 10 replications and k = 1000 iterations

To examine the effect of different levels of the random noise, 5 values of standard deviation $\sigma \in \{0.01, 0.1, 1, 5, 10\}$ were tested. We considered the noise with normal $\boldsymbol{N}(0, \sigma^2 I_2)$ distribution, as well as the noise with uniform $\boldsymbol{U}(c,d)$ distribution. For coefficients c and d we choose values $c = -\sigma\sqrt{3}$ and $d = \sigma\sqrt{3}$, in order to obtain mean 0 and variance σ^2 for uniform distribution. Tables 2-6 present the mean values of Rozenbrock function at the final estimates, over p = 20 replications, k = 1000 iterations, and selected $\alpha = 1$. The pair of coefficients (A,a) represent the optimal choice of coefficients, in the sense that their realization gives the smallest value of the mean values of the objective function.

Table 2: Sample means for standard deviation $\sigma = 0.01$

$\sigma = 0.01$				
A	a	$\boldsymbol{N}(0, \sigma^2 I_2)$	a	$\boldsymbol{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$
50	0.67	0.0403	0.67	0.0404
60	0.78	0.0330	0.78	0.0330
70	0.85	0.0299	0.85	0.0299
80	0.92	0.0276	0.92	0.0278
90	1.08	0.0249	1.08	0.0249
100	1.15	0.0238	1.15	0.0239

Table 3: Sample means for standard deviation $\sigma = 0.1$

$\sigma = 0.1$				
A	a	$\boldsymbol{N}(0, \sigma^2 I_2)$	a	$\boldsymbol{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$
50	0.66	0.0408	0.66	0.0410
60	0.76	0.0337	0.77	0.0335
70	0.83	0.0308	0.83	0.0310
80	0.97	0.0280	0.97	0.0282
90	1.08	0.0249	1.08	0.0251
100	1.13	0.0243	1.18	0.0244

Table 4: Sample means for standard deviation $\sigma = 1$

$\sigma = 1$				
A	a	$\boldsymbol{N}(0, \sigma^2 I_2)$	a	$\boldsymbol{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$
50	0.62	0.0436	0.62	0.0457
60	0.72	0.0370	0.73	0.0385
70	0.82	0.0333	0.84	0.0343
80	0.92	0.0304	0.94	0.0315
90	1.01	0.0286	1.04	0.0293
100	1.12	0.0272	1.14	0.0276

Table 5: Sample means for standard deviation $\sigma = 5$

$\sigma = 5$				
A	a	$\boldsymbol{N}(0, \sigma^2 I_2)$	a	$\boldsymbol{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$
50	0.53	0.0634	0.54	0.0759
60	0.63	0.0578	0.63	0.0696
70	0.7	0.0553	0.73	0.0643
80	0.77	0.0535	0.83	0.0614
90	0.84	0.0522	0.92	0.0596
100	0.89	0.0517	1.02	0.0583

Table 6: Sample means for standard deviation $\sigma = 10$

$\sigma = 10$				
A	a	$\boldsymbol{N}(0, \sigma^2 I_2)$	a	$\boldsymbol{U}(-\sigma\sqrt{3}, \sigma\sqrt{3})$
50	0.47	0.0944	0.48	0.1307
60	0.55	0.0909	0.55	0.1257
70	0.55	0.0898	0.62	0.1214
80	0.6	0.0883	0.67	0.1187
90	0.63	0.0876	0.73	0.1168
100	0.61	0.0873	0.79	0.1155

Analysing the data from tables 2-6 we can see that the type of the noise does not affect significantly the efficiency of the stochastic approximation method. Figure 2 and

Figure 3 show the values of coefficient *a* depending on the level of the noise. Also, Figure 2 and Figure 3 confirm that smaller values for the coefficient *a* are desirable in case of larger deviation of the noise.

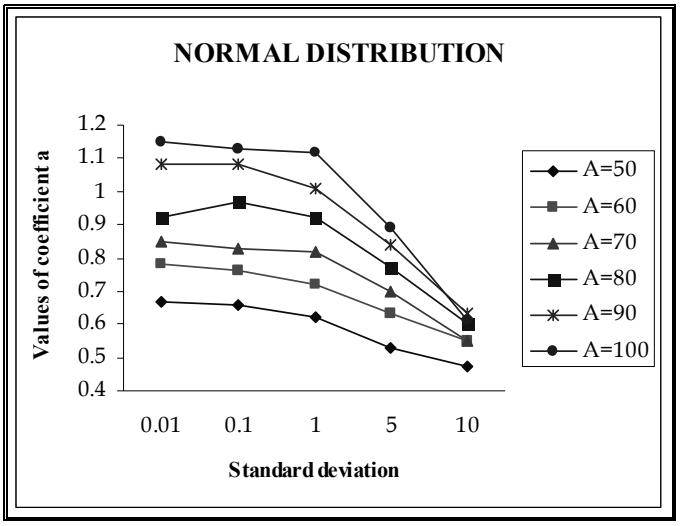

Figure 2: Values of coefficient *a* depending on the level of the noise

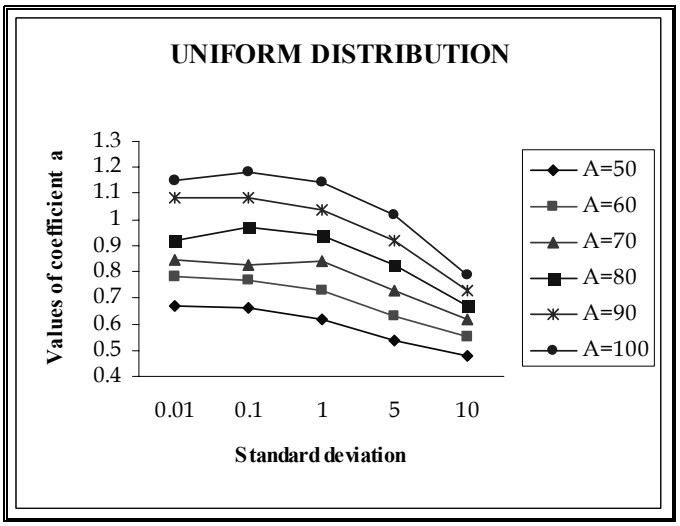

Figure 3: Values of coefficient *a* depending on the level of the noise

REFERENCES

[1] Chung, K.L., "On a stochastic approximation method", *Annals of Mathematical Statistics*, 25 (1954) 463-483.

[2] Fabian, V., "On asymptotic normality in stochastic approximation", *Annals of Mathematical Statistics*, 39 (1968) 1327-1332.

[3] Fabian, V., "Stochastic Approximation", in: *Optimizing Methods in Statistics*, Academic Pres, New York, 1971, 439-470.

[4] Japundžić, M., "Efficiency of the modifications of deterministic methods in solving the stochastic optimization problem", MSc Thesis (in Serbian language), University of Novi Sad, Faculty of Natural Sciences and Mathematics, October 2010.

[5] Kushner, H.J., Yin, G., *Stochastic Approximation and Recursive Algorithms and Applications*, Springer-Verlag, Second Edition, 2003.

[6] Nevel'son, M.B., Has'minskii, R.Z., *Stocshastic Approximation and Recursuve Estimation*, American Mathematical Society, 1973.

[7] Robbins, H., Monro, S., "A stochastic approximation method", *Annals of Mathematical Statistics*, 22 (1951) 400-407.

[8] Sacks, J., "Asymptotic distribution of stochastic approximation procedures", *Annals of Mathematical Statistics*, 29 (1958) 373-405.

[9] Spall, J.C., "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation", *IEEE Transactions on Automatic Control*, 37 (1992) 332-341.

[10] Spall, J.C., "Adaptive stochastic approximation by the simultaneous perturbation method", *IEEE Transactions on Automatic Control,* 45 (2000) 1839-1853.

[11] Spall, J.C., *Introduction to Stochastic Search and Optimization*, Wiley-Interscience, New Jersey, 2003, 95-125.