Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

GLOBAL RANDOM OPTIMIZATION BY SIMULTANEOUS PERTURBATION STOCHASTIC APPROXIMATION

John L. Maryak
Daniel C. Chin

The Johns Hopkins University
Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel, MD 20723-6099, U.S.A.

ABSTRACT

A desire with iterative optimization techniques is that the algorithm reach the global optimum rather than get stranded at a local optimum value. Here, we examine the global convergence properties of a "gradient free" stochastic approximation algorithm called "SPSA," that has performed well in complex optimization problems. We establish two theorems on the global convergence of SPSA. The first provides conditions under which SPSA will converge in probability to a global optimum using the well-known method of injected noise. In the second theorem, we show that, under different conditions, "basic" SPSA *without injected noise* can achieve convergence in probability to a global optimum. This latter result can have important benefits in the setup (tuning) and performance of the algorithm. The discussion is supported by numerical studies showing favorable comparisons of SPSA to simulated annealing and genetic algorithms.

1 INTRODUCTION

A problem of great practical importance is the problem of stochastic optimization, which may be stated as the problem of finding a minimum point, $\theta^* \in R^p$, of a real-valued function $L(\theta)$, called the "loss function," that is observed in the presence of noise. Many approaches have been devised for numerous applications over the long history of this problem. A common desire in many applications is that the algorithm reach the global minimum rather than get stranded at a local minimum value. In this paper, we consider the popular stochastic optimization technique of stochastic approximation (SA), in particular, the form that may be called "gradient-free" SA. This refers to the case where the gradient, $g(\theta) = \partial L(\theta)/\partial\theta$, of the loss function is not readily available or not directly measured (even with noise). This is a common occurrence, for example, in complex systems where the exact functional relationship between the loss function value and the parameters, θ, is not known and the loss function is evaluated by measurements on the system (or by other means, such as simulation). In such cases, one uses instead an approximation to $g(\theta)$ (the well-known form of SA called the Kiefer-Wolfowitz type is an example).

The usual form of this type of SA recursion is:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k), \tag{1}$$

where $\hat{g}_k(\theta)$ is an approximation (at the k^{th} step of the recursion) of the gradient $g(\theta)$, and $\{a_k\}$ is a sequence of positive scalars that decreases to zero (in the standard implementation) and satisfies other properties. This form of SA has been extensively studied, and is known to converge to a local minimum of the loss function under various conditions.

Several authors (e.g., Chin (1994), Gelfand and Mitter (1991), Kushner (1987), and Styblinski and Tang (1990)) have examined the problem of *global* optimization using various forms of gradient-free SA. The usual version of this algorithm is based on using the standard "finite difference" gradient approximation for $\hat{g}_k(\theta)$. It is known that carefully injecting noise into the recursion based on this standard gradient can result in an algorithm that converges (in some sense) to the global minimum. For a discussion of the conditions, results, and proofs, see, e.g., Fang et al. (1997), Gelfand and Mitter (1991), and Kushner (1987). The amplitude of the injected noise is decreased over time (a process called "annealing"), so that the algorithm can finally converge when it reaches the neighborhood of the global minimum point.

A somewhat different version of SA is obtained by using a "simultaneous perturbation" gradient approximation, as described in Spall (1992) for multivariable ($p > 1$) problems. The gradient approximation in simultaneous-perturbation SA (SPSA) is much faster to compute than the finite-difference approximation in multivariable problems. More significantly, using SPSA often results in a recursion that is much more economical, in terms of loss-function evaluations, than the standard version of SA. The loss function evaluations can be the most expensive part of an optimization, especially if computing the loss function requires making measurements on the physical system. Several studies (e.g., Spall (1992), Chin (1997)) have shown SPSA to be very effective in complex optimization problems. A considerable body of theory has been developed for SPSA (Spall (1992), Chin (1997), Dippon and Renz (1997), Spall (2000), and the references therein), but, because of the special form of its gradient approximation, existing theory on global convergence of standard SA algorithms is not directly applicable to SPSA. In Section 2 of this paper, we present a theorem showing that SPSA can achieve global convergence (in probability) by the technique of injecting noise. The "convergence in probability" results of our Theorem 1 (Section 2) and Theorem 2 (Section 3) are standard types of global convergence results. Several authors have shown or discussed global convergence in probability or in distribution (Chiang *et al.* (1987), Gelfand and Mitter (1991), Gelfand and Mitter (1993), Geman and Geman (1984), Fang *et al.* (1997), Hajek (1988), Kushner (1987), Yakowitz *et al.* (2000), and Yin (1999)). Stronger "almost sure" global convergence results seem only to be available by using generally infeasible exhaustive search (Dippon and Fabian (1994)) or random search methods (Yakowitz (1993)), or for cases of optimization in a discrete (θ -) space (Alrefaei and Andradottir (1999)).

The method of injection of noise into the recursions has proven useful, but naturally results in a relative slowing of the rate of convergence of the algorithm (e.g., Yin (1999)) due to the continued injection of noise when the recursion is near a global solution. In addition, the implementation of the extra noise terms adds to the complexity of setting up the algorithm. In Section 3, we present a theorem showing that, under different (more demanding) conditions, the basic version of SPSA can perform as a global optimizer *without* the need for injected noise. Section 4 contains numerical studies demonstrating SPSA's performance compared to two other popular strategies for global optimization, namely, simulated annealing and genetic algorithms; and Section 5 is a summary.

2 SPSA WITH INJECTED NOISE AS A GLOBAL OPTIMIZER

Our first theorem applies to the following algorithm, which is the basic SPSA recursion indicated in equation (1), modified by the addition of extra noise terms:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k) + q_k \omega_k , \tag{2}$$

where $\omega_k \in R^p$ is i.i.d. $N(0,I)$ injected noise, $a_k = a/k$, $q_k^2 = q/k \log\log k$, $a > 0$, $q > 0$, and $\hat{g}_k(\bullet)$ is the "simultaneous perturbation" gradient defined as follows:

$$\hat{g}_k(\theta) \equiv (2c_k\Delta_k)^{-1}[L(\theta + c_k\Delta_k) - L(\theta - c_k\Delta_k) + \varepsilon_k^{(+)} - \varepsilon_k^{(-)}], \tag{3}$$

where $c_k, \varepsilon_k^{(\pm)}$ are scalars, $\Delta_k \in R^p$, and the inverse of a vector is defined to be the vector of inverses. This gradient definition follows that given in Spall (1992). The ε_k terms represent (unknown) additive noise that may contaminate the loss function observation, the c_k sequence is deterministic and chosen to decrease to zero, and the Δ_{kl} components are chosen randomly according to the conditions in Spall (1992), usually (but not necessarily) from the Bernoulli (± 1) distribution. (Uniformly or normally distributed perturbations are *not* allowed by the regularity conditions.)

Our theorem on global convergence of SPSA using injected noise is based on a result in Gelfand and Mitter (1991). The statements of the hypotheses, the associated definitions, and the proofs for the two theorems are quite long and involved. These will not be given here, but are available in the references quoted below. We can now state our first theorem as follows:

Theorem 1: Under hypotheses H1 through H9 in Maryak and Chin (2001), $\hat{\theta}_k$ in algorithm (2) converges in probability to the set of global minima of $L(\theta)$.

Proof: See Maryak and Chin (1999), and the remark on convergence in probability in Gelfand and Mitter (1991), p. 1003.

3 SPSA WITHOUT INJECTED NOISE AS A GLOBAL OPTIMIZER

As indicated in the introduction above, the injection of noise into an algorithm, while providing for global optimization, introduces some difficulties such as the need for more "tuning" of the extra terms and retarded convergence in the vicinity of the solution, due to the continued addition of noise. This effect on the rate of convergence of an algorithm using injected noise is technically subtle, but may have an important influence on the algorithm's performance. In particular, Yin (1999) shows that an algorithm of the form (2) converges at a rate proportional to $\sqrt{\log\log(k + const)}$, while the nominal local convergence rate for an algorithm *without* injected noise is $k^{1/3}$, i.e., $k^{1/3}(\hat{\theta}_k - \theta^*)$ converges in distribution (Spall (1992)). These rates indicate a significant difference in performance between the two algorithms.

A certain characteristic of the SPSA gradient approximation led us to question whether SPSA needed to use injected noise for global convergence. Although this gradient approximation tends to work very well in an SA recursion, the SPSA gradient, evaluated at any single point in θ-space, tends to be less accurate than the standard finite-difference gradient approximation evaluated at θ. So, one is led to consider whether the *effective* noise introduced (automatically) into the recursion by this inaccuracy is sufficient to provide for global convergence *without* a further injection of additive noise. It turns out that *basic* SPSA (i.e., *without* injected noise) does indeed achieve the same type of global convergence as in Theorem 1, but under a different, and more difficult to check, set of conditions.

In this Section, we designate Kushner (1987) as K87, and Kushner and Yin (1997) as KY97. Here we are working with the basic SPSA algorithm having the same form as equation (1):

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k), \tag{4}$$

where $\hat{g}_k(\bullet)$ is the simultaneous-perturbation approximate gradient defined in Section 2, and now (obviously) no extra noise is injected into the algorithm

Now we can state our main theorem:

Theorem 2. Under assumptions J1 through J12 in Maryak and Chin (2001), $\hat{\theta}_k$ in algorithm (4) converges in probability to the set of global minima of $L(\theta)$.

The idea of the proof is as follows (see Maryak and Chin (2001) for the details). This theorem follows from results (in a different context) in K87 for an algorithm $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k[g(\hat{\theta}_k) + \zeta_k]$, where ζ_k is i.i.d. Gaussian (injected) noise. In order to prove our Theorem 2, we start by writing the SPSA recursion as $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k[g(\hat{\theta}_k) + \zeta_k^*]$, where $\zeta_k^* \equiv \hat{g}_k(\hat{\theta}_k) - g(\hat{\theta}_k)$ is the "effective noise" introduced by the inaccuracy of the SPSA gradient approximation. So, our algorithm has the same form as that in K87. However, since ζ_k^* is not i.i.d. Gaussian, we cannot use K87's result directly. Instead, we use material in Kushner and Yin (1997) to establish a key "large deviation" result related to our algorithm (4), which allows the result in K87 to be used with ζ_k^* replacing the ζ_k in his algorithm.

4 NUMERICAL STUDIES: SPSA WITHOUT INJECTED NOISE

4.1 Two-Dimensional Problem

A study was done to compare the performance of SPSA to a recently published application of the popular genetic algorithm (GA). The loss function is the well-known Griewank function (see Haataja (1999)) defined for a two-dimensional $\theta = (t_1, t_2)'$, by:

$$L(\theta) = \cos(t_1 - 100)\cos[(t_2 - 100)/\sqrt{2}] \\ - [(t_1 - 100)^2 + (t_2 - 100)^2]/4000 - 1,$$

which has thousands of local minima in the vicinity of a single global minimum at $\theta = (100{,}100)'$ at which $L(\theta) = 0$. Haataja (1999) describes the application of a GA to this function (actually, to find the *maximum* of $-L(\theta)$) based on noise-free evaluations of $L(\theta)$ (i.e., $\varepsilon_k = 0$). This study achieved a success rate of 66% (see Haataja's Table 1.3, p.16) in 50 independent trials of the GA, using 300 generations and 9000 $L(\theta)$ evaluations in each run of the GA. Haataja's definition of a successful solution is a reported solution where the norm of the solution minus the correct value, θ^*, is less than 0.2, and the value of the loss function at the reported solution is within 0.01 of the correct value of zero. We examined the performance of basic SPSA (without adding injected noise) on this problem, using $a_k = a/(A+k)^{\alpha}$, with $A = 60,\ a = 100$ and $\alpha = .602$, a slowly decreasing gain sequence of a form that has been used in many applications (see Spall (1998)). For the gradient approximation (equation (3)), we chose each component of Δ_k to be an independent sample from a Bernoulli (±1) distribution, and $c_k = c/k^{\gamma}$, with $c = 10$ and $\gamma = .101$. Since we used the exact loss function, the ε_k noise terms were zero. We ran SPSA, allowing 3000 function evaluations in each of 50 runs, and starting the algorithm (each time) at a point randomly chosen in the domain $[-200, 400] \times [-200, 400]$. Haataja's θ-domain was also constrained to lie in a box, but the dimensions of the box were not specified. Hence we chose a domain that is a cube centered at the global minimum, in which there are many local minima of $L(\theta)$ (as seen in Haataja's (1999) Figure 1.1). SPSA successfully located the global minimum in all 50 runs (100% success rate).

4.2 Ten-Dimensional Problem

For a more ambitious test of the global performance of SPSA, we applied SPSA to a loss function given in Example 6 of Styblinski and Tang (1990), which we will designate for convenience as ST90. The loss function is:

$$L(\theta) = (2p)^{-1}\sum_{i=1}^{p} t_i^2 - 4p\prod_{i=1}^{p}\cos(t_i) ,$$

where $p = 10$ and $\theta = (t_1,...,t_p)'$. This function has the global minimum value of -40 at the origin, and a large number of local minima. As in the two-dimensional study above, we used the exact loss function. Our goal is to compare the performance of SPSA without injected noise with simulated annealing and with a GA.

For the simulated annealing algorithm, we use the results reported in ST90. They used an advanced form of simulated annealing called fast simulated annealing (FSA). According to ST90, FSA has proven to be much more efficient than classical simulated annealing due to using Cauchy (rather than Gaussian) sampling and using a fast (inversely linear in time) cooling scheme. For more details on FSA, see ST90. The results of their application of FSA to the above $L(\theta)$ are given in Table 1 below (FSA values taken from Table 10 of ST90). Table 1 shows the results of 10 independent runs of each algorithm. In each case (each run of each algorithm), the best value of $L(\theta)$ found by the algorithm is shown. In their study, although FSA was allowed to use 50,000 function evaluations for each of the runs, the algorithm showed very limited success in locating the global minimum. It should be noted that the main purpose of the ST90 paper was to examine a relatively new algorithm, stochastic approximation combined with convolution smoothing. This algorithm, which they call SAS, was much more effective than FSA, yielding results between those shown in Table 1 for GA and SPSA.

For the genetic algorithm (GA), we implemented a GA using the popular features of elitism (elite members of the old population pass unchanged into the new population), tournament selection (tournament size = 2), and real-number encoding (see Mitchell (1996), pp. 168, 170, and 157, respectively). After considerable experimentation, we found the following settings for the GA algorithm to provide the best performance on this problem. The population size was 100, the number of elite members (those carried forward unchanged) in each generation was 10, the crossover rate was 0.8, and mutation was accomplished by adding a Gaussian random variable with mean zero and standard deviation 0.01 to each component of the offspring. The original population of 100 (10-dimensional) θ-vectors was created by uniformly randomly generating points in the 10-dimensional hypercube centered at the origin, with edges of length 6 (so, all components had absolute value less than or equal to 3 radians). We also constrained all component values in subsequent generations to be less than or equal to 3 in absolute value. All runs of the GA algorithm reported here used 50,000 evaluations of the loss function. The results of the 10 independent runs of GA are shown in Table 1. Although the algorithm did reasonably well in getting close to the minimum loss value of –40, it only found the global minimum in one of the 10 runs (run #8). In the other nine cases, a few (typically two or four) of the components were trapped in a local minimum (around $\pm\pi$ radians), while the rest of the components (approximately) achieved the correct value of zero.

We examined the performance of basic SPSA (without adding injected noise), using the algorithm parameters defined in Subsection 4.1 with $A = 60$, $a = 1$, $\alpha = .602$, $c = 2$, and $\gamma = .101$. We started θ randomly within the same domain in which we chose initial values for the GA algorithm, and we did not constrain the search space for SPSA as we did for GA (the initialization and search space for FSA were not reported in ST90). We ran 10 Monte Carlo trials (i.e., randomly varying the choices of Δ_k). The results are tabulated in Table 1. The results of these numerical studies show a strong performance of the basic SPSA algorithm in difficult global optimization problems.

Table 1: Best Loss Function Value in Each of 10 Independent Runs of Three Algorithms

Run	SPSA	GA	FSA
1	–40.0	–36.6	–24.9
2	–40.0	–38.3	–15.5
3	–40.0	–38.3	–29.0
4	–40.0	–36.6	–32.1
5	–40.0	–35.0	–30.2
6	–40.0	–38.3	–30.1
7	–40.0	–36.6	–27.9
8	–40.0	–40.0	–20.9
9	–40.0	–36.6	–28.5
10	–40.0	–38.3	–34.6
Average Value	–40.0	–37.5	–27.4
Number of Function Evaluations	2,500	50,000	50,000

5 SUMMARY

SPSA is an efficient gradient-free SA algorithm that has performed well on a variety of complex optimization problems. We showed in Section 2 that, as with some standard SA algorithms, adding injected noise to the basic SPSA algorithm can result in a global optimizer. More significantly, in Section 3, we showed that, under certain conditions, the basic SPSA recursion can achieve global convergence *without the need for injected noise*. The use of basic SPSA as a global optimizer can ease the implementation of the global optimizer (no need to tune the injected noise) and result in a significantly faster rate of convergence (no extra noise corrupting the algorithm in the vicinity of the solution). In the numerical studies, we found significantly better performance of SPSA as a global optimizer than for the popular simulated annealing and genetic algorithm methods, which are often recommended for global optimization. In particular, in the case of a 10-dimensional optimization parameter (θ), the fast simulated annealing and genetic algorithms generally failed to find the global solution.

ACKNOWLEDGMENTS

This work was partially supported by the JHU/APL Independent Research and Development Program and U.S. Navy contract N00024-98-D-8124.

REFERENCES

Alrefaei, M.H. and Andradottir, S. (1999), "A Simulated Annealing Algorithm with Constant Temperature for Discrete Stochastic Optimization," *Management Science*, **45**, pp. 748-764.

Chaing T-S., Hwang, C-R., and Sheu, S-J. (1987), "Diffusion for Global Optimization in R^n," *SIAM J. Control Optim.*, **25**, pp. 737-753.

Chin, D.C. (1997), "Comparative Study of Stochastic Algorithms for System Optimization Based on Gradient Approximations," *IEEE Trans. Systems, Man, and Cybernetics – Part B: Cybernetics*, **27**, pp. 244-249.

Chin, D.C. (1994), "A More Efficient Global Optimization Algorithm Based on Styblinski and Tang," *Neural Networks*, **7**, pp. 573-574.

Dippon, J. and Fabian, V. (1994), "Stochastic Approximation of Global Minimum Points," *J. Statistical Planning and Inference*, **41**, pp. 327-347.

Dippon, J. and Renz, J. (1997), "Weighted Means in Stochastic Approximation of Minima," *SIAM J. Control Optim.,* **35**, pp. 1811-1827.

Fang, H., Gong, G., and Qian, M. (1997), "Annealing of Iterative Stochastic Schemes," *SIAM J. Control Optim.*, **35**, pp.1886-1907.

Gelfand, S.B. and Mitter, S.K. (1991), "Recursive Stochastic Algorithms for Global Optimization in R^d," *SIAM J. Control Optim.*, **29**, pp. 999-1018.

Gelfand, S.B. and Mitter, S.K. (1993), "Metropolis-Type Annealing Algorithms for Global Optimization in R^d," *SIAM J. Control Optim.*, **31**, pp. 110-131.

Geman S. and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Trans. Pattern Analysis and Machine Intelligence*, **PAMI-6**, pp. 721-741.

Haataja, J. (1999), "Using Genetic Algorithms for Optimization: Technology Transfer in Action," Chapter 1, pp. 3 – 22, in *Evolutionary Algorithms in Engineering and Computer Science,* Edited by K. Miettinen, M.M. Makela, P. Neittaanmaki, and J. Periaux, Wiley, Chichester.

Hajek, (1988), "Cooling Schedules for Optimal Annealing," *Mathematics of Operations Research*, **13**, pp. 311-329.

Kushner, H.J. and Yin, G.G. (1997), *Stochastic Approximation Algorithms and Applications*, Springer, New York.

Kushner, H.J. (1987), "Asymptotic Global Behavior for Stochastic Approximation and Diffusions with Slowly Decreasing Noise Effects: Global Minimization Via Monte Carlo," *SIAM J. Appl. Math.*, **47**, pp. 169-185.

Maryak, J.L. and Chin, D.C. (1999), "Efficient Global Optimization Using SPSA," *Proc. Amer. Control Conf.*, San Diego, June 2-4, pp. 890-894.

Maryak, J.L. and Chin, D.C. (2001), "Global Random Optimization by Simultaneous Perturbation Stochastic Approximation," *Proc. Amer. Control Conf.*, Arlington, VA, June 25-27, pp. 756-762.

Mitchell, M. (1996), *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, Mass.

Spall, J.C. (2000), "Adaptive Stochastic Approximation by the Simultaneous Perturbation Method," *IEEE Trans. Automat. Control,* **45**, pp. 1839-1853.

Spall, J.C. (1998), "Implementation of the Simultaneous Perturbation Algorithm for Stochastic Optimization," *IEEE Trans. Aerospace and Electronic Systems,* **34**, pp. 817-823.

Spall, J.C. (1992), "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Trans. Automat. Control*, **37**, pp. 332-341.

Styblinski, M.A. and Tang, T.-S. (1990), "Experiments in Nonconvex Optimization: Stochastic Approximation with Function Smoothing and Simulated Annealing," *Neural Networks*, **3**, pp. 467-483.

Yakowitz, S. (1993), "A Globally Convergent Stochastic Approximation," *SIAM J. Control Optim.* **31** pp. 30-40.

Yakowitz, S., L'Ecuyer, P., and Vazquez-Abad, F. (2000), "Global Stochastic Optimization with Low-Dispersion Point Sets," *Operations Research*, **48**, pp. 939-950.

Yin, G. (1999), "Rates of Convergence for a Class of Global Stochastic Optimization Algorithms," *SIAM J. Optim.*, **10**, pp. 99-120.

AUTHOR BIOGRAPHIES

JOHN L. MARYAK is a specialist in mathematics and statistics as applied to systems modeling and estimation problems. Dr. Maryak has worked since 1977 at the Johns Hopkins University, Applied Physics Laboratory (JHU/APL) on diverse tasks involving the analysis and performance assessment of complex military systems. Dr. Maryak has published a number of papers on Bayesian statistical methods, mathematical modeling, Kalman filtering, and maximum likelihood estimation. He is a member of the American Statistical Association (ASA), the Institute of Electrical and Electronics Engineers (IEEE), and Sigma Xi. His email address is `john.maryak@ jhuapl.edu`.

DANIEL C. CHIN is a mathematician at the Johns Hopkins University Applied Physics Laboratory and has experienced in stochastic approximation, Bayesian analysis, statistical estimation and simulation, and image data processing. His accomplishments include the identification of buried objects (metallic or non metallic) with induced electric current, control strategies and simulation studies for system-wide traffic-adaptive control, and an optimal representation of the Kalman filter process and its square root formulation. Mr. Chin is a member of the IEEE society, Sigma Xi, and the team that won the Hart Prize for most outstanding IR&D project at the Johns Hopkins University Applied Physics Lab in 1990. His email address is `daniel.chin@jhuapl.edu`.