

Global Random Optimization by Simultaneous Perturbation Stochastic Approximation

John L. Maryak and Daniel C. Chin

Abstract—We examine the theoretical and numerical global convergence properties of a certain “gradient free” stochastic approximation algorithm called the “simultaneous perturbation stochastic approximation (SPSA)” that has performed well in complex optimization problems. We establish two theorems on the global convergence of SPSA, the first involving the well-known method of injected noise. The second theorem establishes conditions under which “basic” SPSA *without injected noise* can achieve convergence in probability to a global optimum, a result with important practical benefits.

Index Terms—Global convergence, simulated annealing, simultaneous perturbation stochastic approximation (SPSA), stochastic approximation (SA), stochastic optimization.

I. 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, $\theta^* \in \mathbf{R}^p$, 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 for the algorithm to 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

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

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 (e.g., [4], Chs. 1, 5, 6, and others, and [8], Chs. 6 and 7 and the references therein), and is known to converge to a local minimum of the loss function under various conditions.

Several authors (e.g., [3], [5], and [11]) 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., [3] and [5]. These results are based on the intuitive idea that promoting global convergence by the injection of extra noise terms into the recursion may allow the algorithm to escape θ -neighborhoods that produce local minimum points of $L(\theta)$, especially in the early iterations of the algorithm.

A somewhat different version of SA is obtained by using a “simultaneous perturbation” (SP) gradient approximation, as described in [10] for multivariable ($p > 1$) problems. Using SPSA often results in a recursion that is much more economical, in terms of loss-function evaluations, than the standard version of SA (see [10]). 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. Many studies (e.g., [1] and [10]) have shown SPSA to be very effective in complex optimization problems.

The main goal of this paper is to establish two theorems on the global convergence of SPSA. A considerable body of theory has been developed for SPSA (e.g., [1], [2], [9], and [10], and the references therein). This theory does not include global convergence results. As mentioned earlier, global convergence theory does exist for standard implementations of SA. However, because of the particular form of SPSA’s gradient approximation, the existing theory on global convergence of standard SA algorithms is not directly applicable to SPSA. In Section II 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 II) and Theorem 2 (Section III) are standard types of global convergence results (see, e.g., [3], [5], and [12]).

To overcome drawbacks associated with the noise-injection method (see Section III), we present in Section III a theorem showing that, under conditions different from those in Section II, the basic version of SPSA can perform as a global optimizer *without* the need for injected noise. Section IV is a summary. The proof of Theorem 2 is given in the Appendix.

II. SPSA WITH INJECTED NOISE AS A GLOBAL OPTIMIZER

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

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

where $\omega_k \in \mathbf{R}^p$ is independent identically distributed (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 “SP” gradient defined as follows:

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

where $c_k, \varepsilon_k^{(\pm)}$ are scalars, $\Delta_k \in \mathbf{R}^p$, and the inverse of a vector is defined to be the vector of inverses. This gradient definition follows that given in [10]. The ε_k terms represent (unknown) additive noise that may contaminate the loss function observation, c_k are parameters of the algorithm, the c_k sequence decreases to zero, and the Δ_{kl} components of Δ_k are chosen randomly according to the conditions in [10], usually (but not necessarily) from the Bernoulli (± 1) distribution.

Our theorem on global convergence of SPSA using injected noise is based on a result in [3]. In order to state the theorem, we need

Manuscript received February 20, 2003; revised August 26, 2005 and May 11, 2007. Recommended by Associate Editor I. Paschalidis. This work was supported in part by The Johns Hopkins University (JHU) Applied Physics Laboratory (APL) Independent Research and Development Program and in part by the U.S. Navy under Contract N00024-98-D-8124.

The authors are with the Applied Physics Laboratory, The Johns Hopkins University, Laurel, Maryland 20723-6099 USA (e-mail: jlmjlm@comcast.net; dcchin@comcast.net).

Digital Object Identifier 10.1109/TAC.2008.917738

to develop some notation (from [3]), starting with the definition of a key probability measure, P_η , used in hypothesis H7 next. Define P_η for any $\eta > 0$ by the Radon–Nikodym derivative: $dP_\eta(\theta)/d\theta = \exp(-2L(\theta)/\eta^2)/Z_\eta$, where $Z_\eta = \int_{\mathbf{R}^p} \exp(-2L(\theta)/\eta^2) d\theta$. Next, define an important constant, C_0 , for convergence theory as follows ([3]). For $t \in \mathbf{R}$ and $v_1, v_2 \in \mathbf{R}^p$, let $I(t, v_1, v_2) = \inf_\phi \frac{1}{2} \int_0^t |d\phi(s)/ds + g(\phi(s))|^2 ds$, where the inf is taken over all absolutely continuous functions $\phi: [0, \infty) \rightarrow \mathbf{R}^p$ such that $\phi(0) = v_1$ and $\phi(t) = v_2$, $g(\bullet)$ is the gradient, and $|\bullet|$ is the Euclidean norm. Let $V(v_1, v_2) = \lim_{t \rightarrow \infty} I(t, v_1, v_2)$, and $S_0 = \{\theta : g(\theta) = 0\}$. Then, $C_0 \equiv \frac{3}{2} \sup_{v_1, v_2 \in S_0} (V(v_1, v_2) - 2L(v_2))$. We will also need the following definition of tightness. If $\{X_k\}$ is a sequence of random p -dimensional vectors, then $\{X_k\}$ is *tight* if for any $\varepsilon > 0$, there exists a compact subset $K_\varepsilon \subset \mathbf{R}^p$ such that $P(X_k \in K_\varepsilon) > 1 - \varepsilon, \forall k > 0$. Finally, let $\zeta_k^* \equiv \hat{g}_k(\hat{\theta}_k) - g(\hat{\theta}_k)$ and let superscript prime ($'$) denote transpose.

The following are the hypotheses used in Theorem 1.

- H1. Let $\Delta_k \in \mathbf{R}^p$ be a vector of p mutually independent zero-mean random variables $\{\Delta_{k1}, \Delta_{k2}, \dots, \Delta_{kp}\}'$ such that $\{\Delta_k\}$ is a mutually independent sequence that is also independent of the sequences $\{\hat{\theta}_1, \dots, \hat{\theta}_{k-1}\}$, $\{\varepsilon_1^{(\pm)}, \dots, \varepsilon_{k-1}^{(\pm)}\}$, and $\{\omega_1, \dots, \omega_{k-1}\}$, and such that Δ_{ki} is symmetrically distributed about zero, $|\Delta_{ki}| \leq \alpha_1 < \infty$ a.s. and $E|\Delta_{ki}^{-2}| \leq \alpha_2 < \infty$, a.s. $\forall i, k$.
- H2. Let $\varepsilon_k^{(+)}$ and $\varepsilon_k^{(-)}$ represent random measurement noise terms that satisfy $E_k[(\varepsilon_k^{(+)} - \varepsilon_k^{(-)})|\Delta_k] = 0$ a.s. $\forall k$, where E_k denotes the conditional expectation given $\mathfrak{F}_k \equiv$ the sigma algebra induced by $\{\hat{\theta}_0, \omega_1, \dots, \omega_{k-1}, \Delta_1, \dots, \Delta_{k-1}, \varepsilon_1^{(\pm)}, \dots, \varepsilon_{k-1}^{(\pm)}\}$. The $\{\varepsilon_k^{(\pm)}\}$ sequences are not assumed independent. Assume that $E_k[(\varepsilon_k^{(\pm)})^2|\Delta_k] \leq \alpha_3 < \infty$ a.s. $\forall k$.
- H3. $L(\theta)$ is a thrice continuously differentiable map from \mathbf{R}^p into \mathbf{R} ; $L(\theta)$ attains the minimum value of zero; as $|\theta| \rightarrow \infty$, we have $L(\theta) \rightarrow \infty$ and $|g(\theta)| \rightarrow \infty$; $\inf(|g(\theta)|^2 - \text{Lap}(L(\theta))) > -\infty$ (Lap here is the Laplacian, i.e., the sum of the elements of the Hessian matrix $(\partial g(\theta)/\partial \theta')$ of $L(\theta)$); the individual elements of $L^{(3)}(\theta) \equiv \partial^3 L(\theta)/\partial \theta' \partial \theta' \partial \theta'$ satisfy $|L_{i_1 i_2 i_3}^{(3)}(\theta)| \leq \alpha_5 < \infty$.
- H4. The algorithm parameters have the form $a_k = a/k, c_k = c/k^\gamma$, for $k = 1, 2, \dots$, where $a, c > 0, q/a > C_0$, and $0 < \gamma < 1/2$.
- H5. $[(4p-4)/(4p-3)]^{1/2} < \lim_{|\theta| \rightarrow \infty} \inf(g(\theta)' \theta / (|g(\theta)| \|\theta\|)$.
- H6. Let $\{\omega_k\}$ be an i.i.d. $N(0, I)$ sequence, independent of the sequences $\{\hat{\theta}_k\}$, $\{\varepsilon_k^{(\pm)}\}$, and $\{\Delta_k\}$.
- H7. For any $\eta > 0, Z_\eta < \infty; P_\eta$ has a unique weak limit P as $\eta \rightarrow 0$.
- H8. The sequence $\{\hat{\theta}_k\}$ is tight.

Comments:

- 1) It is well known (e.g., [3]) that, under the above conditions, the measure P is concentrated on the set of global minima of $L(\theta)$.
- 2) Assumptions H3, H5, and H7 correspond to assumptions (A1) through (A3) of [3]; assumptions H4 and H8 supply hypotheses stated in [3, Th. 2]; and the definitions of a_k and q_k given in (2) correspond to those used in [3]. Since we show that assumption (A4) of [3] is satisfied by our algorithm, this allows us to use the conclusion of their Theorem 2.
- 3) Hypotheses H1 and H2 and the domain of γ given in H4 are commonly assumed for convergence results (e.g., [10]). Sufficient conditions for assumption H8 are given in [3, Th. 3].

We can now state our first theorem as follows:

Theorem 1: Under hypotheses H1 through H8, $\hat{\theta}_k$ [in (2)] converges in probability to the set of global minima of $L(\theta)$.

Proof: See [7], and the remark on convergence in probability in [3], p. 1003.

III. SPSA WITHOUT INJECTED NOISE AS A GLOBAL OPTIMIZER

The injection of noise into an algorithm, while providing for global optimization, introduces some difficulties such as the need for more “tuning” (i.e., selecting the coefficients) of the extra terms and retarded convergence in the vicinity of the solution, due to the continued addition of noise. Using results in [12], it can be shown ([6]) that the injection of noise has the potential to dramatically slow the rate of convergence of the SPSA algorithm.

The definition of SPSA gradient approximation suggests that SPSA might not need to use injected noise for global convergence. Although SPSA gradient approximation tends to work very well in an SA recursion, the SP gradient, evaluated at any single point in θ -space, obviously 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 set of conditions. The major difference in conditions is the addition of assumption J10 (which, although technical, is considered reasonable—see Note 2 next in this section) used in Theorem 2 next. Also, the condition H5 is replaced by J5, which is somewhat less restrictive.

In this section, we are working with the basic SPSA algorithm having the same form as (1)

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

where $\hat{g}_k(\bullet)$ is the SP approximate gradient defined in Section II, and now (obviously) no extra noise is injected into the algorithm. For use in the subsequent discussion, it will be convenient to define

$$\begin{aligned} b_k(\hat{\theta}_k) &\equiv E(\hat{g}_k(\hat{\theta}_k) - g(\hat{\theta}_k)|G_k), \text{ and } e_k(\hat{\theta}_k) \\ &\equiv \hat{g}_k(\hat{\theta}_k) - E(\hat{g}_k(\hat{\theta}_k)|G_k) \end{aligned}$$

where G_k denotes the sigma-algebra generated by $\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k\}$, which allows us to write (4) as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k [g(\hat{\theta}_k) + e_k(\hat{\theta}_k) + b_k(\hat{\theta}_k)]. \quad (5)$$

Another key element in the subsequent discussion is the ordinary differential equation (ODE)

$$\dot{\theta} = g(\theta) \quad (6)$$

which, in [6, Lemma 1] is shown to be the “limit mean ODE” for the algorithm (4).

Now we can state our assumptions for Theorem 2, as follows:

- J1. Let $\Delta_k \in \mathbf{R}^p$ be a vector of p mutually independent mean-zero random variables $[\Delta_{k1}, \Delta_{k2}, \dots, \Delta_{kp}]'$ such that $\{\Delta_k\}$ is a mutually independent sequence and Δ_k is independent of the sequences $\{\hat{\theta}_1, \dots, \hat{\theta}_{k-1}\}$ and $\{\varepsilon_1^{(\pm)}, \dots, \varepsilon_{k-1}^{(\pm)}\}$, and such that Δ_{ki} is symmetrically distributed about zero, $|\Delta_{ki}| \leq \alpha_1 < \infty$ a.s. and $E|\Delta_{ki}^{-2}| \leq \alpha_2 < \infty \forall i, k$.
- J2. Let $\int_0^x H_1(\psi_1(s), \psi_2(s)) ds = \limsup_{m, n} \frac{\Delta}{m} \log E \exp[\sum_{i=0}^{x/\Delta-1} \psi_1'(i\Delta) \sum_{j=i_m}^{im+m-1} b_{n+j}(\hat{\theta}_{n+j})]$. Let $\varepsilon_k^{(+)}$ and $\varepsilon_k^{(-)}$ represent random measurement noise terms that satisfy $E((\varepsilon_k^{(+)} - \varepsilon_k^{(-)})|G_k) = 0$ a.s. $\forall k$. The $\{\varepsilon_k^{(\pm)}\}$ sequences need not be assumed independent. Assume that $E((\varepsilon_k^{(\pm)})^2|G_k) \leq \alpha_3 < \infty$ a.s. $\forall k$.

- J3(a). $L(\theta)$ is thrice continuously differentiable in Θ , where $\Theta \subset \mathbf{R}^p$ denotes the θ -region (assumed to be open) under consideration, and the individual elements of the third derivative satisfy $|L_{i_1 i_2 i_3}^{(3)}(\theta)| \leq \alpha_5 < \infty$.
- (b). $|L(\theta)| \rightarrow \infty$ as $|\theta| \rightarrow \infty$. This is required by the theory in [5], and is usually satisfied even if Θ is a finite region.
- J4. The algorithm parameters satisfy the following: the gains $a_k > 0$, $a_k \rightarrow 0$ as $k \rightarrow \infty$, and $\sum_{k=1}^{\infty} a_k = \infty$ (unlike Theorem 1, we do not require $a_k = a/k$). The sequence $\{c_k\}$ is of form $c_k = c/k^\gamma$, where $c > 0$ and $0 < \gamma < 1/2$, and $\sum_{k=0}^{\infty} (a_k/c_k)^2 < \infty$.
- J5. The gradient $g(\theta)$ is Lipschitz continuous, and $|g(\theta)| < \infty$, $\forall \theta \in \Theta$.
- J6. The ODE (6) has a unique solution for each initial condition.
- J7. For the ODE (6), suppose that there exists a finite set of limit points in Θ of the ODE, and each limit point is contained in one of a collection of disjoint compact stable invariant sets (see [5]) K_1, K_2, \dots, K_m . These are closed sets containing all local (including global) minima of the loss function.
- J8. For any $\eta > 0$, $Z_\eta < \infty$; P_η has a unique weak limit P as $\eta \rightarrow 0$ (Z_η & P_η are defined in Section 2).
- J9. $E|\sum_{i=1}^k e_i(\hat{\theta}_i)| < \infty \forall k$.
- J10. For any asymptotically stable (in the sense of Liapunov) point, $\bar{\theta}$, of the ODE (6), there exists a neighborhood of the origin in \mathbf{R}^p such that the closure, Q_2 , of that neighborhood satisfies $\bar{\theta} + Q_2 \equiv \{\bar{\theta} + y : y \in Q_2\} \subset \Theta$. There is a neighborhood, Q_1 , of the origin in \mathbf{R}^p and a real-valued function $H_1(\psi_1, \psi_2)$, continuous in $Q_1 \times Q_2$, whose ψ_1 -derivative is continuous on Q_1 for each fixed $\psi_2 \in Q_2$, and such that the following limit holds. For any $\chi, \Delta > 0$, with χ being an integral multiple of Δ , and any functions $(\psi_1(\bullet), \psi_2(\bullet))$ taking values in $Q_1 \times Q_2$ and being constant on the intervals $[i\Delta, i\Delta + \Delta)$, $i\Delta < \chi$, we have

$$\int_0^\chi H_1(\psi_1(s), \psi_2(s)) ds = \limsup_{m,n} \frac{\Delta}{m} \log E \exp \left[\sum_{i=0}^{(\chi/\Delta)-1} \psi_1'(i\Delta) \sum_{j=im}^{im+m-1} b_{n+j}(\bar{\theta} + \psi_2(i\Delta)) \right] \quad (7)$$

Also, there is a real-valued function $H_2(\psi_3)$ that is continuous and differentiable in a neighborhood of the origin in \mathbf{R}^p , and such that

$$\int_0^\chi H_2(\psi_1(s)) ds = \limsup_{m,n} \frac{\Delta}{m} \log E \exp \times \left[\sum_{i=0}^{(\chi/\Delta)-1} \psi_1'(i\Delta) \sum_{j=im}^{im+m-1} e_{n+j}(\hat{\theta}_{n+j}) \right]. \quad (8)$$

A bit more notation is needed. Let $T > 0$ be interpreted such that $[0, T]$ is the total time period under consideration in the ODE (6). Let

$$\bar{H}(\psi_1, \psi_2) = 0.5[H_1(2\psi_1, \psi_2) + H_2(2\psi_1)],$$

$$\bar{L}(\beta, \psi_2) = \sup_{\psi_1} [\psi_1'(\beta - g(\psi_2)) - \bar{H}(\psi_1, \psi_2)].$$

If $\phi(\bullet)$ is a real-valued absolutely-continuous function on $[0, T]$ with $\phi(0) = x \in \mathbf{R}$, define the function

$S(T, \phi) = \int_0^T \bar{L}(\dot{\phi}(s), \phi(s)) ds$; otherwise define $S(T, \phi) = \infty$. $S(T, \phi)$ is the usual action functional of the theory of large deviations (adapted to our context). Define $t_n \equiv \sum_{i=0}^{n-1} a_i$, and

$t_k^n = \sum_{i=0}^{k-1} a_{n+i}$. Define $\{\hat{\theta}_k^n\}$ and $\theta^n(\bullet)$ by $\hat{\theta}_0^n = x \in \Theta$, $\hat{\theta}_{k+1}^n = \hat{\theta}_k^n - a_{n+k} \hat{g}_{n+k}(\hat{\theta}_k^n)$, and $\theta^n(t) = \hat{\theta}_k^n$ for $t \in [t_k^n, t_{k+1}^n)$.

Now we can state the last two assumptions for Theorem 2:

- J11. For each $\delta > 0$ and $i = 1, 2, \dots, m$, there is a ρ -neighborhood of K_i , denoted $N_\rho(K_i)$, and $\delta_\rho > 0$, $T_\rho < \infty$ such that, for each $x, y \in N_\rho(K_i)$, there is a path, $\phi(\bullet)$, with $\phi(0) = x$, $\phi(T_y) = y$, where $T_y \leq T_\rho$ and $S(T_\rho, \phi) \leq \delta$.
- J12. There is a sphere, D_1 , such that D_1 contains $\bigcup_i K_i$ in its interior, and the trajectories of $\theta^n(\bullet)$ stay in D_1 . All paths of the ODE (6) starting in D_1 stay in D_1 .

Note 1. Assumptions J1, J2, and J3(a) are from [10], and are used here to characterize the noise terms $b_k(\hat{\theta}_k)$ and $e_k(\hat{\theta}_k)$. Assumption J3(b) is used in [5, p. 178]. Assumption J4 expresses standard conditions on the algorithm parameters (see [10]), and implies hypothesis (A10.2) in [4, p. 174]. Assumptions J5 and J6 correspond to hypothesis (A10.1) in [4, p. 174]. Assumption J7 is from [5, p. 175]. Assumption J8 is a standard hypothesis (see [3]) used to establish the limiting distribution to which $\hat{\theta}_k$ will be shown to converge. Assumption J9 is used to establish the ‘‘mean’’ criterion for the martingale sequence mentioned in Note 2 next. Since the bound in J9 is not required to be uniform, the assumption is satisfied if $E|e_i(\hat{\theta}_i)| < \infty, \forall i$. Assumptions J11 and J12 are the ‘‘controllability’’ hypothesis A4.1 and the hypothesis A4.2, respectively, of [5, p. 176].

Note 2. Assumption J10 corresponds to hypotheses (A10.5) and (A10.6) in [4, pp. 179–181]. Although these hypotheses are standard ‘‘textbook’’ forms for this type of large deviation analysis, it is useful to note that they are reasonable in our setting. The first part [(7), involving noise terms $b_k(\hat{\theta}_k)$] of J10, is discussed in [4, p. 174], which states that the results of their Section 6.10 are valid if the noise terms (that they denote ξ_n) are bounded. So, our (7) hypothesis looks reasonable in light of the result [10] that the $b_k(\hat{\theta}_k)$ noise terms are $O(c_k^2)(c_k \rightarrow 0)$. The second part [(8), involving noise terms $e_k(\hat{\theta}_k)$] is justified by the discussion in [4, p. 174], which notes that the results in their Section 6.10 (used in the Lemma in the Appendix next) are valid if the noise terms they denote δM_n [corresponding to our noise terms $e_k(\hat{\theta}_k)$] satisfy the martingale difference property that we have established in [6, Lemma 2].

Now we can state our main theorem (for the proof, see the Appendix).

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

IV. SUMMARY

SPSA is an efficient gradient-free SA algorithm that has performed well on a variety of complex optimization problems. In Section II, we gave conditions under which (as with some standard SA algorithms) adding injected noise to the basic SPSA algorithm can result in a global optimizer. More significantly, in Section III and the Appendix, we established 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 much faster rate of convergence (no extra noise corrupting the algorithm in the vicinity of the solution). In numerical studies reported in [6], 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.

APPENDIX (LEMMA RELATED TO THEOREM 2 AND PROOF OF THEOREM 2)

In this Appendix, we are working with the basic SPSA algorithm as defined in (4): $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)$. We first establish an important Lemma that is needed in order to apply the results from [4] and [5] in the proof of Theorem 2. Let B_x be a set of continuous functions on $[0, T]$ taking values in Θ and with initial value x . Let B_x^0 denote the interior of B_x , and \bar{B}_x denote the closure (using L^1 norm).

Lemma. Under assumptions J4, J5, J6, and J10, we have

$$\begin{aligned} - \inf_{\phi \in B_x^0} S(T, \phi) &\leq \liminf_n \log P_x^n \{ \theta^n(\bullet) \in B_x \} \\ &\leq \limsup_n \log P_x^n \{ \theta^n(\bullet) \in B_x \} \\ &\leq - \inf_{\phi \in \bar{B}_x} S(T, \phi), \end{aligned} \quad (9)$$

where P_x^n denotes the probability under the condition that $\theta^n(0) = x$, and $S(T, \phi)$ is defined after hypothesis J10.

Proof: This result is a straightforward application of Theorems 10.1 and 10.4 in [4, p. 178] and [4, p. 181], respectively. The aforementioned fact that the ODE (6) is the “limit mean ODE” for algorithm (4) allows us to apply the analysis in [4] to our algorithm in (4). Note that our assumption J10 is a modified form of assumptions (A10.5) and (A10.6) in [4], using “equals” signs rather than inequalities. The two-sided inequality in (9) follows from J10 by essentially the same argument as in the proof of Theorem 10.1 in [4, p. 178], which uses an “equality” assumption [(A10.4), p. 174] to arrive at a two-sided large deviation result analogous to (9) given earlier. Q.E.D.

Proof of Theorem 2: This theorem follows immediately from a result in [5], once we establish that the injected noise in [5] can be replaced by the “effective noise” in the SPSA algorithm. We can write the SA algorithm in [5] (his (1.1)), taking $\sigma(X_n) = 1$ without loss in our notation as $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k [g(\hat{\theta}_k) + \zeta_k]$, where ζ_k is i.i.d. Gaussian (injected) noise. The key result for us is [5, Th. 2, p. 177]. In the proof of this Theorem 2, [5] uses the i.i.d. Gaussian assumption only to arrive at a large-deviation result exactly analogous to our aforesaid Lemma. The theorem [5, Th. 2] and its subsequent discussion are then based on the large-deviation result. Obviously, the SPSA algorithm without injected noise can be written as $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k [g(\hat{\theta}_k) + \zeta_k^*]$. Since we have established the aforesaid Lemma for SPSA, the results of [5] hold for the SPSA algorithm with its “effective” noise $\{\zeta_k^*\}$ replacing the $\{\zeta_k\}$ sequence used in [5].

Let us restate this fact, since it is a key step in our proof. Of course, our definitions and hypotheses were made specifically to reproduce the setting in [5] (see the Notes at the end of Section III). Inspection of the proofs involving [5, Th. 2] and its subsequent discussion show that the “i.i.d. Gaussian” assumption in [5] was used only to establish his large-deviation inequality (2.6). We have established in our Lemma the same large-deviation result with SPSA effective noise replacing the i.i.d. Gaussian injected noise. This means that the theorem [5, Th. 2] and its relevant subsequent discussion now hold true word-for-word for SPSA, since (of course, under our hypotheses) the nature of the noise is the only difference between our development and the development in [5].

In particular, the discussion in [5, pp. 178–179], of corollary results to his Theorem 2 is relevant to our Theorem 2 context (SPSA without injected noise). In a section of [5] on “The potential case,” the author notes that when $b(x, \xi) = \bar{b}(x)$ in his notation, the result he is discussing applies to his (1.1), which corresponds to our SPSA setup in (4). The relevant result in this discussion in [5] can be stated as: The difference between the measure of (his) X_n (which corresponds to our

$\hat{\theta}_k$) and the invariant measure (which we have denoted P_η) converges asymptotically ($n, k \rightarrow \infty, \eta \rightarrow 0$) to the zero measure weakly. It follows easily that, in the limit as $k \rightarrow \infty$, $\hat{\theta}_k$ is equivalent to P in the same sense as in [3, Th. 2] and (as in the proof of Theorem 1 earlier) the desired convergence in probability follows. Q.E.D.

ACKNOWLEDGMENT

We thank the reviewers for many insightful and helpful comments.

REFERENCES

- [1] D. C. Chin, “Comparative study of stochastic algorithms for system optimization based on gradient approximations,” *IEEE Trans. Syst., Man, Cybern., Part B, Cybern.*, vol. 27, no. 2, pp. 244–249, Apr. 1997.
- [2] J. Dippon and J. Renz, “Weighted means in stochastic approximation of minima,” *SIAM J. Control Optim.*, vol. 35, pp. 1811–1827, 1997.
- [3] S. B. Gelfand and S. K. Mitter, “Recursive stochastic algorithms for global optimization in \mathbf{R}^d ,” *SIAM J. Control Optim.*, vol. 29, pp. 999–1018, 1991.
- [4] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer, 1997.
- [5] H. J. Kushner, “Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo,” *SIAM J. Appl. Math.*, vol. 47, pp. 169–185, 1987.
- [6] J. L. Maryak and D. C. Chin, “Global random optimization by simultaneous perturbation stochastic approximation,” in *Proc. Amer. Control Conf.*, Arlington, VA, 2001, pp. 756–762.
- [7] J. L. Maryak and D. C. Chin, “Efficient global optimization using SPSA,” in *Proc. Amer. Control Conf.*, San Diego, CA, 1999, pp. 890–894.
- [8] J. C. Spall, *Introduction to Stochastic Search and Optimization*. Hoboken, NJ: Wiley, 2003.
- [9] J. C. Spall, “Adaptive stochastic approximation by the simultaneous perturbation method,” *IEEE Trans. Autom. Control*, vol. 45, no. 10, pp. 1839–1853, Oct. 2000.
- [10] J. C. Spall, “Multivariate stochastic approximation using a simultaneous perturbation gradient approximation,” *IEEE Trans. Autom. Control*, vol. 37, no. 3, pp. 332–341, Mar. 1992.
- [11] M. A. Styblinski and T. S. Tang, “Experiments in nonconvex optimization: Stochastic approximation with function smoothing and simulated annealing,” *Neural Netw.*, vol. 3, pp. 467–483, 1990.
- [12] G. Yin, “Rates of convergence for a class of global stochastic optimization algorithms,” *SIAM J. Optim.*, vol. 10, pp. 99–120, 1999.