Proceedings
of the
1996 Conference
on
Information Science and Systems

Volume II

Department of Electrical Engineering
Princeton University
Princeton, New Jersey 08544-5263

A Deterministic Analysis of Simultaneous Perturbation Stochastic Approximation

I-JENG WANG and EDWIN K.P. CHONG
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907-1285

Abstract

We study the convergence of the simultaneous perturbation stochastic approximation algorithm and establish deterministic necessary and sufficient conditions on both perturbation and noise sequences. We discuss the difference between the algorithm and a stochastic approximation with random directions, and propose ideas on further research in analysis and designs of these algorithms.

1. Introduction

The Kiefer-Wolfowitz (KW) algorithm is an application of stochastic approximation to local optimization based on a finite-difference estimate of the gradient. For an objective function with dimension p, the finite-difference estimation requires $2p$ observations at each iteration. This requirement usually results in unrealistic computational complexity when the dimension of the problem is high. In [9], Spall presents a KW type algorithm based on a "simultaneous perturbation" gradient approximation that requires only 2 observations at each iteration. It is suggested in [9] that the proposed algorithm can be significantly more efficient than the standard KW procedure. We refer to Spall's algorithm as the *simultaneous perturbation stochastic approximation* (SPSA) algorithm. In [1], Chen et al. propose a modification of the SPSA algorithm and prove its convergence under weaker conditions.

Although the convergence of Spall's algorithm has been established, it is not clear (at least intuitively) why random perturbations used in the algorithm would result in faster convergence. In both Spall's [9] and Chen's [1] results, conditions on random perturbations for convergence are stated in probabilistic settings. These stochastic conditions do not provide much insight into the essential properties of perturbations that contribute to the convergence and efficiency of the SPSA algorithm.

In this paper, we develop a deterministic framework for the analysis of the SPSA algorithm. We present five equivalent deterministic necessary and sufficient conditions on both the perturbation and noise for convergence of the SPSA algorithm. We believe that our sample-path characterization sheds light on what makes the SPSA algorithm effective.

This research was supported in part by the National Science Foundation through grants ECS-9410313 and ECS-9501652.

2. Stochastic Approximation with Simultaneous Perturbations

Consider the problem of recursively estimating the minimum of an objective function $L : \mathbb{R}^p \to \mathbb{R}$ based on noisy measurements of L. We assume that L satisfies the following conditions:

(A1) The gradient of L, denoted by $f = \nabla L$, exists and is uniformly continuous.

(A2) There exist $x^* \in \mathbb{R}^p$ such that

- $f(x^*) = 0$; and
- for all $\delta > 0$, there exists $h_\delta > 0$ such that $\|x - x^*\| \geq \delta$ implies $f(x)^T(x - x^*) \geq h_\delta \|x - x^*\|$.

We define the random perturbations as a sequence of vectors $d_n = [d_n^1, \cdots, d_n^p]^T$, $d_n \in \mathbb{R}^p$, $d_n^i \neq 0$. Next we define a sequence of vectors $\{r_n\}$ related to $\{d_n\}$ by $r_n = [\frac{1}{d_n^1}, \cdots, \frac{1}{d_n^p}]^T$. The SPSA algorithm is described by

$$x_{n+1} = x_n - a_n \frac{y_n^+ - y_n^-}{2c_n} r_n, \tag{1}$$

where a_n is the positive step-size satisfying the usual conditions

$$\lim_{n\to\infty} a_n = 0, \ \sum_{n=1}^{\infty} a_n = \infty; \tag{2}$$

$\{c_n\}$ is a positive sequence with $\lim_{n\to\infty} c_n = 0$; and y_n^+ and y_n^- are noisy measurements of the function L at perturbed points, defined by

$$\begin{aligned} y_n^+ &= L(x_n + c_n d_n) + e_n^+, \\ y_n^- &= L(x_n - c_n d_n) + e_n^-, \end{aligned}$$

with additive noise e_n^+ and e_n^-, respectively. For convenience, we write

$$f^d(x_n) = \frac{L(x_n + c_n d_n) - L(x_n - c_n d_n)}{2c_n} \tag{3}$$

as an approximation to the directional derivative, $d_n^T f(x_n)$.

For the purpose of analysis, we rewrite the algorithm (1) into a Robbins-Monro type of algorithm

$$\begin{aligned} x_{n+1} &= x_n - a_n(d_n^T f(x_n) - b_n)r_n + a_n \frac{e_n}{2c_n} r_n, \quad (4) \\ &= x_n - a_n f(x_n) + a_n b_n + a_n \tfrac{e_n}{2c_n} r_n \\ &\quad - a_n(r_n d_n^T - I) f(x_n), \quad (5) \end{aligned}$$

by defining

$$\begin{aligned} b_n &= d_n^T f(x_n) - f^d(x_n), \quad (6) \\ e_n &= e_n^- - e_n^+. \end{aligned}$$

The sequence $\{b_n\}$ represents the bias of the directional derivative approximation. The effective noise for the algorithm is the scaled difference between two measurement noise values, $\frac{e_n}{c_n} r_n$. Applying sample-path convergence results for Robbins-Monro algorithms [5, 10, 12], we can obtain sample-path conditions on both random perturbation $\{d_n\}$ and noise sequence $\{e_n\}$ for convergence of the algorithm (1). We first provide the necessary tools.

3. Convergence of Robbins-Monro Algorithms

We rely mainly on the following convergence theorem from [10] to derive conditions on perturbations and noise.

Theorem 1. *Consider the stochastic approximation algorithm*

$$x_{n+1} = x_n - a_n f(x_n) + a_n e_n + a_n \gamma_n, \quad (7)$$

where $\{x_n\}$, $\{e_n\}$, and $\{\gamma_n\}$ are sequences on $\mathbb{R}^p$, $f\colon \mathbb{R}^p \to \mathbb{R}^p$ satisfies Assumption (A2), $\{a_n\}$ is a sequence of positive real numbers satisfying (2), and $\lim_{n\to\infty} \gamma_n = 0$. Suppose that the sequence $\{f(x_n)\}$ is bounded. Then for any x_1 in $\mathbb{R}^p$, $\{x_n\}$ converges to x^* if and only if $\{e_n\}$ satisfies any of the following conditions:

(C1)

$$\lim_{n\to\infty} \left(\sup_{n \le k \le m(n,T)} \left\| \sum_{i=n}^{k} a_i e_i \right\| \right) = 0$$

for some $T > 0$, where
$m(n,T) \triangleq \max\{k : a_n + \cdots + a_k \le T\}$.

(C2)

$$\lim_{T\to 0} \frac{1}{T} \limsup_{n\to\infty} \left(\sup_{n \le k \le m(n,T)} \left\| \sum_{i=n}^{k} a_i e_i \right\| \right) = 0.$$

(C3) For any $\alpha, \beta > 0$, and any infinite sequence of nonoverlapping intervals $\{I_k\}$ on $\mathbb{N}$ there exists $K \in \mathbb{N}$ such that for all $k \ge K$,

$$\left\| \sum_{n\in I_k} a_n e_n \right\| < \alpha \sum_{n\in I_k} a_n + \beta.$$

(C4) There exist sequences $\{f_n\}$ and $\{g_n\}$ with $e_n = f_n + g_n$ for all n such that

$$\sum_{k=1}^{n} a_k f_k \text{ converges, and } \lim_{n\to\infty} g_n = 0.$$

The condition (C1) is the well-known Kushner and Clark condition [6]; the condition (C2) is a modification of Kushner and Clark's condition presented by Chen [2]; the condition (C3) is proposed by Kulkarni and Horn in [5]; while the decomposition condition (C4) has been widely applied in the literature [2, 3, 7, 8].

As shown in [11], all the four conditions described above are equivalent to convergence of a form of generalized average of $\{e_n\}$. We define precisely the generalized average of a sequence on $\mathbb{R}^p$.

Definition 1. The generalized average of a sequence $\{e_n\}$ on $\mathbb{R}^p$ with respect to a positive sequence $\{a_n\}$, $a_1 = 1$, $0 < a_n < 1$, $n \ge 2$, is a sequence $\{\bar{e}_n\}$ defined by

$$\bar{e}_n = \frac{1}{\beta_n} \sum_{k=1}^{n} \gamma_k e_k,$$

where

$$\begin{aligned} \beta_n &= \begin{cases} 1 & n = 1, \\ \prod_{k=2}^{n} \frac{1}{1-a_k} & \text{otherwise,} \end{cases} \\ \gamma_n &= a_n \beta_n. \end{aligned}$$

In this paper we consider only generalized averages with respect to the step-size sequence $\{a_n\}$. For simplicity, we assume in the sequel that the step-size $\{a_n\}$ also satisfies the condition in the above definition. Note that these additional assumptions on the step-size are only for convenience and are not crucial to our results. The convergence of the generalized average of $\{e_n\}$ is closely related to the convergence of stochastic approximation algorithms. We can establish the following equivalence [11]:

Lemma 1. *Each of the conditions on $\{e_n\}$ in Theorem 1 holds if and only if its generalized average $\left\{\frac{1}{\beta_n}\sum_{k=1}^{n} \gamma_k e_k\right\}$ converges to 0.*

Proof. See [11] for a proof of the equivalence between convergence of the generalized average and condition (C4). □

For ease of presentation, we will refer to the above generalized average condition as condition (C5). The following simple lemma concerning conditions (C1–5) will be used to prove our main result.

Lemma 2. *Let $\{x_n\}$ and $\{y_n\}$ be sequences on $\mathbb{R}^p$. Suppose that $\{x_n\}$ satisfies conditions (C1-5). Then $\{x_n + y_n\}$ satisfies conditions (C1-5) if and only if $\{y_n\}$ satisfies conditions (C1-5).*

With the above results we can derive necessary and sufficient conditions on both random perturbations d_n and noise e_n for convergence of the algorithm in (1).

4. Convergence of SPSA

We present a convergence theorem that provides a sample-path (deterministic) characterization for the random perturbations d_n. First, we show that a simple application of the mean value theorem yields the convergence of the finite-difference estimates of the directional derivatives. This result will be used to establish convergence of the algorithm (1).

Lemma 3. *Suppose that $L\colon \mathbb{R}^p \to \mathbb{R}$ satisfies Assumption (A1), $\{c_n\}$ converges to 0, and $\{d_n\}$ is bounded. Then the sequence $\{b_n\}$ defined by (6) converges to 0.*

Proof. By the Mean Value Theorem,

$$\begin{aligned} b_n &= d_n^T f(x_n) - \frac{L(x_n + c_n d_n) - L(x_n - c_n d_n)}{2c_n}, \\ &= d_n^T [f(x_n) - f(x_n + (2\lambda_n - 1)c_n d_n)], \end{aligned}$$

where $0 \le \lambda_n \le 1$ for all $n \in \mathbb{N}$. Since f is uniformly continuous, for any $\epsilon > 0$ there exists $\delta > 0$ such that $\|x - y\| < \delta$ implies $\|f(x) - f(y)\| < \frac{\epsilon}{\sup_n \|d_n\|}$. Furthermore, by the convergence of $\{c_n\}$ there exists $N \in \mathbb{N}$ such that $\|(2\lambda_n - 1)c_n d_n\| < \delta$ for all $n \ge N$. Hence, for all $n \ge N$

$$\begin{aligned} \|b_n\| &\le \|d_n\| \|f(x_n) - f(x_n + (2\lambda_n - 1)c_n d_n)\|, \\ &< \sup_n \|d_n\| \frac{\epsilon}{\sup_n \|d_n\|} = \epsilon. \end{aligned}$$

Therefore $\{b_n\}$ converges to 0. □

With the help of Theorem 1, Lemma 1, and Lemma 3, we establish a necessary and sufficient condition for convergence of the SPSA algorithm in the following theorem.

Theorem 2. *Suppose that the assumptions (A1-2) hold, and $\{d_n\}$, $\{r_n\}$, and $\{f(x_n)\}$ are bounded. Then, $\{x_n\}$ defined by (1) converges to x^* if and only if the sequences $\{(r_n d_n^T - I)f(x_n)\}$ and $\left\{\frac{e_n}{2c_n} r_n\right\}$ satisfy conditions (C1-5).*

Proof. ($\Longrightarrow$) Suppose that $\{x_n\}$ converges to x^*. Then $\{f(x_n)\}$ converges to $f(x^*) = 0$ by the continuity of f. Since $\{d_n\}$ and $\{r_n\}$ are bounded, $\|r_n d_n^T - I\|$ is bounded and $\{(r_n d_n^T - I)f(x_n)\}$ converges to 0. Thus $\{(r_n d_n^T - I)f(x_n)\}$ satisfies conditions (C1–5). By Theorem 1, $\left\{(r_n d_n^T - I)f(x_n) - \frac{e_n}{2c_n} r_n\right\}$ satisfies condition (C1–5). Therefore $\left\{\frac{e_n}{2c_n} r_n\right\}$ satisfies condition (C1–5) by Lemma 2.
($\Longleftarrow$) This follows directly from Lemma 2 and Theorem 1. □

In Theorem 2, the condition on the random perturbation d_n, although tight, is "coupled" with the function values $f(x_n)$ and cannot be verified directly. It seems possible to design an adaptive perturbation scheme based on the estimate of $f(x_n)$ so that the condition presented in Theorem 2 is satisfied. However this idea is not only difficult to justify theoretically but also hard to carry out due to the special structure of the matrix

$$r_n d_n^T - I = \begin{bmatrix} 0 & \frac{d_n^2}{d_n^1} & \cdots & \frac{d_n^p}{d_n^1} \\ \frac{d_n^1}{d_n^2} & 0 & \cdots & \frac{d_n^p}{d_n^2} \\ \vdots & \vdots & & \vdots \\ \frac{d_n^1}{d_n^p} & \frac{d_n^2}{d_n^p} & \cdots & 0 \end{bmatrix}. \tag{8}$$

We can see that it is difficult to scale the elements in the matrix according to $\{f(x_n)\}$. One solution to this is to establish probabilistic sufficient conditions on the perturbation to guarantee that the deterministic condition in Theorem 2 holds almost surely, as in [1, 9]. We present a general sufficient condition based on the martingale convergence theorem. In the following, we assume that $\{d_n\}$ and $\{e_n\}$ are random sequences.

Proposition 1. *Let $\mathcal{F}_n$ be the σ-algebra generated by $\{d_k\}_{k=1,\cdots,n}$ and $\{e_k\}_{k=1,\cdots,n}$. Assume that $\sum_{n=1}^{\infty} a_n^q < \infty$ for some $q > 1$, and $E\left(\frac{d_n^i}{d_n^j} | \mathcal{F}_{n-1}\right) = 0$ for $i \ne j$. Suppose that $\{d_n\}$, $\{r_n\}$, and $\{f(x_n)\}$ are bounded. Then $\{(r_n d_n^T - I)f(x_n)\}$ satisfies conditions (C1-5) almost surely.*

Proof. Let $z_n = a_n(r_n d_n^T - I)f(x_n)$, $z_n = [z_n^1, \cdots, z_n^p]^T$. Since

$$E(z_n|\mathcal{F}_{n-1}) = E[a_n(r_n d_n^T - I)f(x_n)|\mathcal{F}_{n-1}] = 0,$$

$\{\sum_{k=1}^n z_k\}$ is a martingale. Furthermore,

$$E(|z_n^i|^q) < \infty$$

for all $i \le p$ by the boundedness of $\{d_n\}$, $\{r_n\}$, $\{f(x_n)\}$, and $\sum_{n=1}^{\infty} a_n^p$. Hence by the L^q convergence theorem for martingales [4, (4.4), p.217], the sequence $\{\sum_{k=1}^n z_k\}$ converges almost surely. Therefore $\{(r_n d_n^T - I)f(x_n)\}$ satisfies condition (C4). □

In [1, 9], d_n is assumed to be a vector of p mutually independent random variables independent of $\mathcal{F}_{n-1}$. Under this assumption, it is clear that the condition in Proposition 1 can be satisfied by assuming either $E(d_n^i) = 0$, as in [9], or $E\left(\frac{1}{d_n^i}\right) = 0$, as in [1].

5. Some Remarks on Random Perturbation

As shown in [9], under appropriate assumptions, the SPSA algorithm exhibits better asymptotic behavior than the standard KW algorithm in the sense that the former attains a smaller asymptotic mean square error when both of them converge. However, it is still not fully understood why the random perturbation can improve the performance, and which type of distributions for the random perturbation are more desirable. We believe that these questions can be answered via a sample path analysis. Here we propose some possible approaches.

From (4), we can see that, besides the fact that the standard KW algorithm requires $2p$ observations for each iteration instead of 2 for the SPSA, the differences between the SPSA and the standard KW algorithm lie mainly in three terms: the bias term b_n, the noise term $\frac{e_n}{2c_n}r_n$, and the difference between the directional derivative and the gradient. It seems unlikely for the first two factors to contribute significantly to the performance improvement by the SPSA algorithm. Therefore it is important to study the difference between the following two algorithms:

$$x_{n+1} = x_n - a_n d_n^T f(x_n) r_n, \quad (9)$$
$$\tilde{x}_{n+1} = \tilde{x}_n - a_n f(\tilde{x}_n), \quad (10)$$

under the consideration that $\tilde{x}_n$ requires p times more observations than x_n. We do not have theoretical results at this stage. However, simulation results do not indicate any clear advantage of (9) over (10).

In [6], Kushner and Clark consider an algorithm similar to the SPSA algorithm, which they refer to as the random directions stochastic approximation (RDSA). The algorithm can be described by

$$\begin{aligned} x_{n+1} &= x_n - a_n \frac{y_n^+ - y_n^-}{2c_n} d_n \quad (11)\\ &= x_n - a_n v f(x_n) + a_n b_n + a_n \frac{e_n}{2c_n} d_n \\ &\quad - a_n (d_n d_n^T - vI) f(x_n), \quad (12) \end{aligned}$$

where y_n^+, y_n^-, b_n, and e_n are defined as for the SPSA algorithm, and $v > 0$ is an arbitrary constant real number. As can be seen from (1) and (11), the RDSA algorithm differs from the SPSA algorithm only in the direction it takes in each iteration. The RDSA algorithm "moves" in the direction d_n along which the directional derivative is estimated. Following the same arguments as in Section 4, we also obtain necessary and sufficient conditions for convergence of the RDSA algorithm.

Theorem 3. *Suppose that the assumptions (A1-2) hold, $\{d_n\}$ and $\{f(x_n)\}$ are bounded. Then, $\{x_n\}$ defined by (11) converges to x^* if and only if the sequences $\{(d_n d_n^T - vI)f(x_n)\}$ and $\left\{\frac{e_n}{2c_n} d_n\right\}$ satisfies conditions (C1-5).*

As in the case of the SPSA algorithm, it is difficult to design $\{d_n\}$ based on $\{f(x_n)\}$ to satisfy the condition for $\{(d_n d_n^T - vI)f(x_n)\}$ above, due to the structure of the matrix

$$d_n d_n^T - vI = \begin{bmatrix} (d_n^1)^2 - v & d_n^1 d_n^2 & \cdots & d_n^1 d_n^p \\ d_n^2 d_n^1 & (d_n^2)^2 - v & \cdots & d_n^2 d_n^p \\ \vdots & \vdots & & \vdots \\ d_n^p d_n^1 & d_n^p d_n^2 & \cdots & (d_n^p)^2 - v \end{bmatrix} \quad (13)$$

Although we are allowed to choose smaller d_n^i (the same is not true for SPSA since $\frac{1}{d_n^i}$ needs to be bounded), the diagonal terms always give a "weight" around the quantity v. Furthermore if we try to choose $v = v_n$ such that $v_n \approx (d_n^i)^2$, the resulting step-size $a_n v_n$ in (12) may decreases too fast and the algorithm may not converge. However, similar to the case for the SPSA algorithm, we can derive the following sufficient probabilistic condition on d_n. As before, we assume below that $\{d_n\}$ and $\{e_n\}$ are random sequences.

Proposition 2. *Let $\mathcal{F}_n$ be the σ-algebra generated by $\{d_k\}_{k=1,\cdots,n}$ and $\{e_k\}_{k=1,\cdots,n}$. Assume that $\sum_{n=1}^{\infty} a_n^q < \infty$ for some $q > 1$, $E[(d_n^i)^2 | \mathcal{F}_{n-1}] = v$, and $E(d_n^i d_n^j | \mathcal{F}_{n-1}) = 0$ for $i \neq j$. Suppose that $\{d_n\}$ and $\{f(x_n)\}$ are bounded. Then $\{(d_n d_n^T - vI)f(x_n)\}$ satisfies conditions (C1-5) almost surely.*

In [6], Kushner and Clark assume that $\{d_n\}$ is a sequence of independent vectors, each distributed uniformly over the surface of the unit p-sphere. They also state that the RDSA algorithm with this random direction distribution does not give better asymptotic performance than the standard KW algorithm. Together with the results in [9], this statement implies that the SPSA algorithm is asymptotically superior to the RDSA algorithm studied by Kushner and Clark. However, under the deterministic framework used in this paper, it is not clear whether the same is true for general choices of directions $\{d_n\}$. Furthermore, in the case where a Bernoulli-type distribution (which seems to be a good choice as illustrated by the simulation result in [9]) is used, the SPSA and RDSA algorithms are identical. To fully understand this issue, we believe that it is essential to compare the behaviors of the algorithm described by (9) and the algorithm described by

$$x_{n+1} = x_n - a_n d_n^T f(x_n) d_n. \quad (14)$$

Since the matrix $r_n d_n^T$ is positive semi-definite, $f(x_n)^T r_n d_n^T f(x_n) \geq 0$. Hence if $r_n^T f(x_n) \neq 0$ and $d_n^T f(x_n) \neq 0$, the directional derivatives $d_n^T f(x_n)$ and $r_n^T f(x_n)$ have the same sign. Therefore, it seems that the difference in performance between algorithms (9) and (14), if any, should be a result of the choice of $\{d_n\}$ (or $\{r_n\}$) but not of the difference between their structures. A more detailed analysis is needed to fully understand how to design the random direction to achieve better performance.

6. Conclusion

In this paper we presented a deterministic analysis for SPSA algorithms. We establish deterministic necessary and sufficient conditions on both perturbation and noise for convergence of the algorithm. The result illustrates that a sample-path analysis may provide better insight into stochastic problems. Further research is needed to fully understand why algorithms with random perturbations may outperform algorithms based on better gradient approximations.

References

[1] H. F. Chen, T. E. Duncan, and B. Pasik-Duncan, "A Kiefer-Wolfowitz algorithm with randomized differences," *manuscript*, 1996.

[2] H.-F. Chen, "Stochastic approximation and its new applications," in *Proceedings of 1994 Hong Kong International Workshop on New Directions of Control and Manufacturing*, pp. 2–12, 1994.

[3] H.-F. Chen and Y.-M. Zhu, "Stochastic approximation procedures with randomly varying truncations," *Scientia Sinica (Series A)*, vol. 29, no. 9, pp. 914–926, 1986.

[4] R. Durrett, *Probability: Thoery and Examples.* Wadsworth & Brooks/Cole, 1991.

[5] S. R. Kulkarni and C. Horn, "Necessary and sufficient conditions for convergence of stochastic approximation algorithms under arbitrary disturbances," in *Proceedings of the 34th IEEE Conference on Decision and Control*, pp. 3843–3848, 1995.

[6] H. K. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* New York: Springer, 1978.

[7] T. L. Lai, "Stochastic approximation and sequential search for optimum," in *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. LeCam and R. A. Olshen, eds.), vol. 2, (Monterey, CA), pp. 557–577, Wadsworth, 1985.

[8] L. Ljung, G. Pflug, and H. Walk, *Stochastic Approximation and Optimization of Random Systems.* Birkhäuser, 1992.

[9] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, Mar. 1992.

[10] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms," *Advances in Applied Probability*, to appear.

[11] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "Generalized averaging and stochastic approximation," *submitted to the 35th IEEE Conference on Decision and Control*, 1996.

[12] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "On equivalence of some noise conditions for stochastic approximation algorithms," in *Proceedings of the 34th IEEE Conference on Decision and Control*, pp. 3849–3853, 1995.