

Convergence of Simultaneous Perturbation Stochastic Approximation for Nondifferentiable Optimization

Ying He yhe@engr.colostate.edu
Electrical and Computer Engineering
Colorado State University
Fort Collins, CO 80523

Michael C. Fu mfu@isr.umd.edu
Steven I. Marcus marcus@isr.umd.edu
Institute for Systems Research
University of Maryland
College Park, MD 20742

Abstract

In this paper, we consider Simultaneous Perturbation Stochastic Approximation (SPSA) for function minimization. The standard assumption for convergence is that the function be three times differentiable, although weaker assumptions have been used for special cases. However, all work that we are aware of at least requires differentiability. In this paper, we relax the differentiability requirement and prove convergence using convex analysis.

Keywords

Simultaneous Perturbation Stochastic Approximation, Convex Analysis, Subgradient

I. INTRODUCTION

Simultaneous Perturbation Stochastic Approximation (SPSA), proposed by Spall [15], has been successfully applied to many optimization problems. Like other Kiefer-Wolfowitz-type stochastic approximation algorithms, such as the finite-difference based stochastic approximation algorithm, SPSA uses only objective function measurements. Furthermore, SPSA is especially efficient in high-dimensional problems in terms of providing a good solution for a relatively small number of measurements of the objective function [17].

Convergence of SPSA has been analyzed under various conditions. Much of the literature assumes the objective function be three times differentiable [15] [3] [5] [8] [10] [16] [18], though weaker assumptions are found as well, e.g. [1] [4] [12] [14] [19]. However, all of them require that the function be at least differentiable. Among the weakest assumptions on the objective function, Fu and Hill [4] assume that the function is differentiable and convex; Chen et al. [1] assume that the function is differentiable and the gradient satisfies a Lipschitz condition. In a semiconductor fab-level decision making problem [7], we found that the one-step cost function is continuous and convex with respect to the decision variables, but nondifferentiable, so that the problem of finding the one-step optimal action requires minimizing a continuous and convex function. So the question is: does the SPSA algorithm converge in this setting? The answer is affirmative, and the details will be presented.

Gerencsér et al. [6] have discussed non-smooth optimization. However, they approximate the non-smooth function by a smooth enough function, and then optimize the smooth function by SPSA. Thus, they take an indirect approach.

In this paper, we consider function minimization and show that the SPSA algorithm converges for nondifferentiable convex functions, which is especially important when the function is not differentiable at the minimizing point. First, similar to [19], we decompose the SPSA algorithm into four terms: a subgradient term, a bias term, a random direction noise term and an observation noise term. In our setting, the subgradient term replaces the gradient term in [19], since we assume that the function does not have to be differentiable. Hence, we need to show the asymptotic behavior of the algorithm follows a differential inclusion instead of an ordinary differentiable equation. Kushner and Yin

[9] state a theorem (Theorem 5.6.2) for convergence of a Kiefer-Wolfowitz algorithm in a nondifferentiable setting. However, this theorem is not general enough to cover our SPSA algorithm. We will prove a more general theorem to establish convergence of SPSA.

The general approach for proving convergence for these types of algorithms requires showing that the bias term vanishes asymptotically. In the differentiable case, a Taylor series expansion or the mean value theorem is used to establish this. These tools are not applicable in our more general setting, but we are able to use convex analysis for this task, which is one new contribution of this paper. For the random direction noise term, we use a similar argument as in [19] to show the noise goes to zero with probability 1 (w.p.1), except that now the term is a function of the subgradient instead of the gradient. For the observation noise term, the conditions for general Kiefer-Wolfowitz algorithms given in [9, pp.113-114] are used, and we also show it goes to zero w.p.1.

To be more specific, we want to minimize the function $E[F(\theta, \chi)] = f(\theta)$ over the parameter $\theta \in H \subset R^r$, where $f(\cdot)$ is continuous and convex, χ is a random vector and H is a convex and compact set. Let θ_k denote the k th estimate of the minimum, and let $\{\Delta_k\}$ be a random sequence of column random vectors with $\Delta_k = [\Delta_{k,1}, \dots, \Delta_{k,r}]^T$. $\Delta_1, \Delta_2, \dots$ are not necessary identically distributed. The two-sided SPSA algorithm to update θ_k is as follows:

$$\theta_{k+1} = \Pi_H \left(\theta_k - \alpha_k [\Delta_k^{-1}] \frac{F_k^+ - F_k^-}{2c_k} \right), \quad (1)$$

where Π_H denotes a projection onto the set H , $F_k^\pm$ are observations taken at parameter values $\theta_k \pm c_k \Delta_k$, c_k is a positive sequence converging to zero, α_k is the step size, and $[\Delta_k^{-1}]$ is defined as $[\Delta_k^{-1}] := [\Delta_{k,1}^{-1}, \dots, \Delta_{k,r}^{-1}]^T$.

Write the observation in the form

$$F_k^\pm = f(\theta_k \pm c_k \Delta_k) + \phi_k^\pm,$$

where $\phi_k^\pm$ are observation noises, and define

$$G_k := \frac{f(\theta_k + c_k \Delta_k) - f(\theta_k - c_k \Delta_k)}{2c_k}. \quad (2)$$

Then the algorithm (1) can be written as:

$$\theta_{k+1} = \Pi_H \left(\theta_k - \alpha_k [\Delta_k^{-1}] G_k + \alpha_k [\Delta_k^{-1}] \frac{\phi_k^- - \phi_k^+}{2c_k} \right). \quad (3)$$

The convergence of the SPSA algorithm (1) has been proved under various conditions. One of the weakest conditions on the objective function is that $f(\cdot)$ be differentiable and convex [4]. Under the differentiability condition, one generally invokes a Taylor series expansion or the mean value theorem to obtain $f(\theta_k \pm c_k \Delta_k) = f(\theta_k) \pm c_k \Delta_k^T \nabla f(\theta_k) + O(|c_k|^2 |\Delta_k|^2)$. Therefore, $G_k = \Delta_k^T \nabla f(\theta_k) + O(|c_k| |\Delta_k|^2)$, which means G_k can be approximated by $\Delta_k^T \nabla f(\theta_k)$. Then, suppose $H = R^r$, the algorithm (3) can be written as:

$$\begin{aligned} \theta_{k+1} = \theta_k - \alpha_k \nabla f(\theta_k) + \alpha_k (I - [\Delta_k^{-1}] \Delta_k^T) \nabla f(\theta_k) \\ + \alpha_k [\Delta_k^{-1}] \frac{\phi_k^- - \phi_k^+}{2c_k}, \end{aligned}$$

where a standard argument of the ODE method implies that the trajectory of θ_k follows the ODE

$$\dot{\theta} = -\nabla f(\theta).$$

In our context, however, we only assume that $f(\cdot)$ is continuous and convex – $\nabla f(\cdot)$ may not exist at some points, so a Taylor series expansion or the mean value theorem is not applicable. Instead, using convex analysis we show that G_k is close to the product of Δ_k^T and a *subgradient* of $f(\cdot)$.

II. SUBGRADIENT AND REFORMULATION OF THE SPSA ALGORITHM

First, we introduce some definitions and preliminary results on convex analysis, with more details in [11].

Let h be a real-valued convex function on R^r ; a vector $sg(x)$ is a *subgradient* of h at a point x if $h(z) \geq h(x) + (z - x)^T sg(x), \forall z$. The set of all subgradients of h at x is called the *subdifferential* of h at x and is denoted by $\partial h(x)$ [11, p.214]. If h is a convex function, the set $\partial h(x)$ is a convex set, which means that $\lambda z_1 + (1 - \lambda) z_2 \in \partial h(x)$ if $z_1 \in \partial h(x)$, $z_2 \in \partial h(x)$ and $0 \leq \lambda \leq 1$.

The *one-sided directional derivative* of h at x with respect to a vector y is defined to be the limit

$$h'(x; y) = \lim_{\lambda \downarrow 0} \frac{h(x + \lambda y) - h(x)}{\lambda}. \quad (4)$$

According to Theorem 23.1 in [11, p.213], if h is a convex function, $h'(x; y)$ exists for each y . Furthermore, according to Theorem 23.4 in [11, p.217], at each point x , the

subdifferential $\partial h(x)$ is a non-empty closed bounded convex set, and for each vector y the directional derivative $h'(x; y)$ is the maximum of the inner products $\langle sg(x), y \rangle$ as $sg(x)$ ranges over $\partial h(x)$. Denote the set of $sg(x)$ on which $h'(x; y)$ attains its maximum by $\partial h_y(x)$. Thus, for all $sg_y(x) \in \partial h_y(x)$ and $sg(x) \in \partial h(x)$,

$$h'(x; y) = y^T sg_y(x) \geq y^T sg(x).$$

Now let us discuss the relationship between G_k defined by (2) and subgradients.

Lemma 1: Consider the algorithm (1), assume $f(\cdot)$ is a continuous and convex function, $\lim_{k \rightarrow \infty} c_k = 0$, $\{\Delta_k\}$ has support on a finite discrete set. Then $\forall \varepsilon > 0$, $\exists \widetilde{sg}(\theta_k) \in \partial f(\theta_k)$ and finite K such that

$$|G_k - \Delta_k^T \widetilde{sg}(\theta_k)| < \varepsilon \quad \text{w.p.1}, \quad \forall k \geq K.$$

Proof: Since $f(\cdot)$ is a continuous and convex function, for fixed $\Delta_k = z$, both $f'(\theta_k; z)$ and $f'(\theta_k; -z)$ exist. By (4) and $\lim_{k \rightarrow \infty} c_k = 0$, $\forall \varepsilon > 0$, $\exists K_1(z), K_2(-z) < \infty$ s.t. $|f'(\theta_k; z) - \frac{f(\theta_k + c_k z) - f(\theta_k)}{c_k}| < \varepsilon, \forall k \geq K_1(z)$, and $|f'(\theta_k; -z) - \frac{f(\theta_k - c_k z) - f(\theta_k)}{c_k}| < \varepsilon, \forall k \geq K_2(-z)$. Let $K = \max_z \{K_1(z), K_2(-z)\}$. Since $\{\Delta_k\}$ has support on a finite discrete set, which implies it is bounded, K exists and is finite, $\forall k \geq K$

$$\begin{aligned} \left| f'(\theta_k; \Delta_k) - \frac{f(\theta_k + c_k \Delta_k) - f(\theta_k)}{c_k} \right| &< \varepsilon \quad \text{w.p.1}, \\ \left| f'(\theta_k; -\Delta_k) - \frac{f(\theta_k - c_k \Delta_k) - f(\theta_k)}{c_k} \right| &< \varepsilon \quad \text{w.p.1}. \end{aligned}$$

Since $G_k = (f(\theta_k + c_k \Delta_k) - f(\theta_k - c_k \Delta_k)) / (2c_k)$,

$$\left| G_k - \frac{1}{2}(f'(\theta_k; \Delta_k) - f'(\theta_k; -\Delta_k)) \right| < \varepsilon \quad \text{w.p.1}. \quad (5)$$

In addition, for $f'(\theta_k; \Delta_k)$, there exists $sg_{\Delta_k}(\theta_k) \in \partial f_{\Delta_k}(\theta_k)$ such that

$$f'(\theta_k; \Delta_k) = \Delta_k^T sg_{\Delta_k}(\theta_k). \quad (6)$$

Similarly, for $f'(\theta_k; -\Delta_k)$, there exists $sg_{-\Delta_k}(\theta_k) \in \partial f_{-\Delta_k}(\theta_k)$ such that

$$f'(\theta_k; -\Delta_k) = (-\Delta_k)^T sg_{-\Delta_k}(\theta_k). \quad (7)$$

Combining (5), (6) and (7), we conclude that $\forall \varepsilon > 0, \exists$ finite K and $sg_{\Delta_k}(\theta_k), sg_{-\Delta_k}(\theta_k) \in \partial f(\theta_k)$ such that

$$\left| G_k - \Delta_k^T \left(\frac{1}{2} sg_{\Delta_k}(\theta_k) + \frac{1}{2} sg_{-\Delta_k}(\theta_k) \right) \right| < \varepsilon \text{ w.p.1, } k \geq K. \quad (8)$$

Note that $\partial f(\theta_k)$ is a convex set, so $\widetilde{sg}(\theta_k) := \frac{1}{2} sg_{\Delta_k}(\theta_k) + \frac{1}{2} sg_{-\Delta_k}(\theta_k) \in \partial f(\theta_k)$, and $|G_k - \Delta_k^T \widetilde{sg}(\theta_k)| < \varepsilon$ w.p.1. $\square$

Define $\delta_k = \Delta_k^T \widetilde{sg}(\theta_k) - G_k$. The SPSA algorithm (3) can be decomposed as

$$\theta_{k+1} = \Pi_H \left(\begin{aligned} &\theta_k - \alpha_k \widetilde{sg}(\theta_k) + \alpha_k (I - [\Delta_k^{-1}] \Delta_k^T) \widetilde{sg}(\theta_k) \\ &+ \alpha_k [\Delta_k^{-1}] \delta_k + \alpha_k [\Delta_k^{-1}] \frac{\phi_k^- - \phi_k^+}{2c_k} \end{aligned} \right). \quad (9)$$

Suppose $H = R^r$, and if we can prove that the third, fourth, and fifth terms inside of the projection go to zero as k goes to infinity, the trajectory of θ_k would follow the differential inclusion [9, p.16]

$$\dot{\theta} \in -\partial f(\theta).$$

According to [11, p.264], the necessary and sufficient condition for a given x to belong to the minimum set of f (the set of points where the minimum of f is attained) is that $0 \in \partial f(x)$.

III. BASIC CONSTRAINED STOCHASTIC APPROXIMATION ALGORITHM

Kushner and Yin [9, p.124] state a theorem (Theorem 5.6.2) for convergence of a Kiefer-Wolfowitz algorithm in a nondifferentiable setting. However, this theorem is not general enough to cover the SPSA algorithm given by (9). So, we establish a more general theorem.

Note that the SPSA algorithm given by (9) is a special case of the stochastic approximation algorithm:

$$\theta_{k+1} = \Pi_H(\theta_k + \alpha_k \widetilde{sf}(\theta_k) + \alpha_k b_k + \alpha_k e_k) \quad (10)$$

$$= \theta_k + \alpha_k \widetilde{sf}(\theta_k) + \alpha_k b_k + \alpha_k e_k + \alpha_k Z_k, \quad (11)$$

where Z_k is the reflection term, b_k is the bias term, e_k is the noise term, and $\widetilde{sf}(\theta_k)$ can be any element of $-\partial f(\theta_k)$. Similar to (9), we need to show that b_k, e_k and Z_k go to zero.

As in [9, p.90], let $m(t)$ denote the unique value of k such that $t_k \leq t < t_{k+1}$ for $t \geq 0$, and set $m(t) = 0$ for $t < 0$, where the time scale t_k is defined as follows: $t_0 = 0$, $t_k = \sum_{i=0}^{k-1} \alpha_i$.

Define the shifted continuous-time interpolation $\theta^k(t)$ of θ_k as follows:

$$\begin{aligned} \theta^k(t) = \theta_k &+ \sum_{i=k}^{m(t+t_k)-1} \alpha_i \widetilde{sf}(\theta_i) + \sum_{i=k}^{m(t+t_k)-1} \alpha_i b_i \\ &+ \sum_{i=k}^{m(t+t_k)-1} \alpha_i e_i + \sum_{i=k}^{m(t+t_k)-1} \alpha_i Z_i. \end{aligned} \quad (12)$$

Define $B^k(t) = \sum_{i=k}^{m(t+t_k)-1} \alpha_i b_i$, and define $M^k(t)$ and $Z^k(t)$ similarly, with e_k and Z_k respectively in place of b_k . Since $\theta^k(t)$ is piecewise constant, we can rewrite (12) as

$$\theta^k(t) = \theta_k + \int_0^t \widetilde{sf}(\theta^k(s)) ds + B^k(t) + M^k(t) + Z^k(t) + \rho^k(t), \quad (13)$$

where

$$\rho^k(t) = \int_{t_{m(t+t_k)}}^t \widetilde{sf}(\theta^{m(t+t_k)}(s)) ds \leq \alpha_{m(t+t_k)} |\widetilde{sf}(\theta_{m(t+t_k)})|.$$

Note that $\rho^k(t)$ is due to the replacement of the first summation in (12) by an integral, and $\rho^k(t) = 0$ at the jump times t_k of the interpolated process, and $\rho^k(t) \rightarrow 0$, since α_k goes to zero as k goes to infinity.

We require the following conditions, similar to those of Theorem 5.3.1 in [9, pp.88-108]:

(A.1) $\alpha_k \rightarrow 0$, $\sum \alpha_k = \infty$.

(A.2) The feasible region H is a hyperrectangle. In other words, there are numbers $a_i < b_i$, $i = 1, \dots, r$, such that $H = \{x : a_i \leq x_i \leq b_i\}$.

(A.3) For some positive number T_1 ,

$$\lim_{k \rightarrow \infty} \sup_{|t| \leq T_1} |B^k(t)| = 0 \quad \text{w.p.1.}$$

(A.4) For some positive number T_2 ,

$$\lim_{k \rightarrow \infty} \sup_{|t| \leq T_2} |M^k(t)| = 0 \quad \text{w.p.1.}$$

For $x \in H$ satisfying (A.2), define the set $C(x)$ as follows. For $x \in H^0$, the interior of H , $C(x)$ contains only the zero element; for $x \in \partial H$, the boundary of H , let $C(x)$ be the infinite convex cone generated by the outer normals at x of the faces on which x lies [9, p.77].

Proposition 1: For the algorithm given by (11), where $\widetilde{sf}(\theta_k) \in \partial f(\theta_k)$, assume $\partial f(\theta)$ is bounded $\forall \theta \in H$ and (A.1)-(A.4) hold. Suppose that $f(\cdot)$ is continuous and convex, but not constant. Consider the differential inclusion

$$\dot{\theta} \in -\partial f(\theta) + z, \quad z(t) \in -C(\theta(t)), \quad (14)$$

and let S_H denote the set of stationary points of (14), i.e. points in H where $0 \in -\partial f(\theta) + z$. Then, $\{\theta_k(\omega)\}$ converges to a point in S_H , which attains the minimum of f .

Proof: see Appendix. $\square$

IV. CONVERGENCE OF THE SPSA ALGORITHM

We now use Proposition 1 to prove convergence of the SPSA algorithm given by (9), which we first rewrite as follows:

$$\begin{aligned} \theta_{k+1} = & \theta_k - \alpha_k \widetilde{sg}(\theta_k) + \alpha_k (I - [\Delta_k^{-1}] \Delta_k^T) \widetilde{sg}(\theta_k) \\ & + \alpha_k [\Delta_k^{-1}] \delta_k + \alpha_k [\Delta_k^{-1}] \frac{\phi_k^- - \phi_k^+}{2c_k} + \alpha_k \tilde{Z}_k, \end{aligned} \quad (15)$$

where $\tilde{Z}_k$ is the reflection term.

Note that the counterparts of b_k and e_k in (11) are $[\Delta_k^{-1}] \delta_k$ and $e_{r,k} + e_{o,k}$, respectively, where the latter quantity is decomposed into a random direction noise term $e_{r,k} := (I - [\Delta_k^{-1}] \Delta_k^T) \widetilde{sg}(\theta_k)$ and an observation noise term $e_{o,k} := \frac{[\Delta_k^{-1}]}{2c_k} (\phi_k^- - \phi_k^+)$.

Lemma 2: Assume that

$$E\left[\frac{\Delta_{k,i}}{\Delta_{k,j}} \mid \Delta_0, \dots, \Delta_{k-1}, \theta_0, \dots, \theta_{k-1}\right] = 0, i \neq j,$$

and $\{\widetilde{sg}(\theta_k)\}$, $\{\Delta_k\}$ and $\{[\Delta_k^{-1}]\}$ are bounded. Then, $e_{r,k} = (I - [\Delta_k^{-1}] \Delta_k^T) \widetilde{sg}(\theta_k)$ is a martingale difference.

Proof: By the boundedness assumption, each element of the sequence $\{\widetilde{sg}(\theta_k)\}$ is bounded by a finite B_k . Using a similar argument as in [19], define $M_k = \sum_{l=0}^k (I - [\Delta_l^{-1}] \Delta_l^T) \widetilde{sg}(\theta_l)$, so $e_{r,k} = M_k - M_{k-1}$.

$$\begin{aligned} E[M_{k+1} \mid M_l, l \leq k] &= E[M_k + (I - [\Delta_{k+1}^{-1}] \Delta_{k+1}^T) \widetilde{sg}(\theta_{k+1}) \mid M_l, l \leq k] \\ &= M_k + E[e_{r,k+1} \mid M_l, l \leq k], \end{aligned}$$

and the absolute value of $E[e_{r,k+1} \mid M_l, l \leq k]$

$$\begin{aligned} &|E[e_{r,k+1} \mid M_l, l \leq k]| \\ &\leq |B_k| \cdot |E[(I - [\Delta_{k+1}^{-1}] \Delta_{k+1}^T) \bar{\mathbf{1}} \mid M_l, l \leq k]| \\ &= |B_k| \cdot E[|\Lambda_k \bar{\mathbf{1}} \mid M_l, l \leq k]| \\ &= 0, \end{aligned}$$

where $\bar{\mathbf{1}}$ is a column vector with each element being 1, and

$$\Lambda_k = \begin{bmatrix} 0 & \frac{\Delta_{k+1,1}}{\Delta_{k+1,2}} & \cdots & \frac{\Delta_{k+1,1}}{\Delta_{k+1,r}} \\ \frac{\Delta_{k+1,2}}{\Delta_{k+1,1}} & 0 & \cdots & \frac{\Delta_{k+1,2}}{\Delta_{k+1,r}} \\ \vdots & \vdots & & \vdots \\ \frac{\Delta_{k+1,r}}{\Delta_{k+1,1}} & \cdots & \frac{\Delta_{k+1,r}}{\Delta_{k+1,2}} & 0 \end{bmatrix}.$$

$\{E|M_k|\}$ is bounded since $\{\widetilde{sg}(\theta_k)\}$, $\{\Delta_k\}$ and $\{[\Delta_k^{-1}]\}$ are bounded. Thus, $e_{r,k}$ is a martingale difference. $\square$

For the observation noise term $e_{o,k}$, we assume the following conditions:

(B.1) $\phi_k^- - \phi_k^+$ is a martingale difference.

(B.2) There is a $K < \infty$ such that for a small γ , all k , and each component $(\phi_k^- - \phi_k^+)_j$ of $(\phi_k^- - \phi_k^+)$,

$$E[e^{\gamma(\phi_k^- - \phi_k^+)_j} \mid \Delta_0, \dots, \Delta_{k-1}, \theta_0, \dots, \theta_{k-1}] \leq e^{\gamma^2 K/2}.$$

(B.3) For each $\mu > 0$, $\sum_k e^{-\mu c_k^2/\alpha_k} < \infty$.

(B.4) For some $T > 0$, there is a $c_1(T) < \infty$ such that for all k ,

$$\sup_{k \leq i < m(t_k + T)} \frac{\alpha_i/c_i^2}{\alpha_k/c_k^2} \leq c_1(T).$$

Examples of condition (B.2) and a discussion related to (B.3) and (B.4) can be found in [9, pp.110-112]. Note that the moment condition on the observation noises $\phi_k^\pm$ is not required in [1]. For other alternative noise conditions, see [2].

Lemma 3: Assume that (A.1) and (B.1) - (B.4) hold and $\{[\Delta_k^{-1}]\}$ is bounded. Then, for some positive T , $\lim_{k \rightarrow \infty} \sup_{|t| \leq T} |M^{o,k}(t)| = 0$ w.p.1, where $M^{o,k}(t) := \sum_{i=k}^{m(t+t_k)-1} \alpha_i e_{o,i}$.

Proof: Define $\tilde{M}^k(t) := \sum_{i=k}^{m(t+t_k)-1} \frac{\alpha_i}{2c_i} (\phi_i^- - \phi_i^+)$. By (A.1) and (B.1) - (B.4), there exists a positive T such that $\lim_{k \rightarrow \infty} \sup_{|t| \leq T} |\tilde{M}^k(t)| = 0$ w.p.1, following the same argument as the one in the proof of Theorem 5.3.2 and Theorem 5.3.3 in [9, pp.108-110].

$\{[\Delta_k^{-1}]\}$ is bounded, so $|M^{o,k}(t)| \leq D_k |\tilde{M}^k(t)|$, where D_k is the bound of $[\Delta_k^{-1}]$.

Thus, $\lim_{k \rightarrow \infty} \sup_{|t| \leq T_1} |M^{o,k}(t)| = 0$ w.p.1. $\square$

Proposition 2: Consider the SPSA algorithm (9), where $\widetilde{sg}(\theta_k) \in \partial f(\theta_k)$. Assume $\partial f(\theta_k)$ is bounded $\forall \theta_k \in H$, (A.1)-(A.2) and (B.1)-(B.4) hold, $\{\Delta_k\}$ has support on a finite discrete set, $\{[\Delta_k^{-1}]\}$ is bounded, and $E[\frac{\Delta_{k,i}}{\Delta_{k,j}} \mid \Delta_0, \dots, \Delta_{k-1}, \theta_0, \dots, \theta_{k-1}] = 0$ for $i \neq j$. Suppose

that $f(\cdot)$ is continuous and convex, but not constant. Consider the differential inclusion

$$\dot{\theta} \in -\partial f(\theta) + z, \quad z(t) \in -C(\theta(t)), \quad (16)$$

and let S_H denote the set of stationary points of (16), i.e. points in H where $0 \in -\partial f(\theta) + z$. Then, $\{\theta_k(\omega)\}$ converges to a point in S_H , which attains the minimum of f .

Proof: Since $\{[\Delta_k^{-1}]\}$ is bounded and $\lim_{k \rightarrow \infty} \delta_k = 0$ by Lemma 1, $\lim_{k \rightarrow \infty} [\Delta_k^{-1}] \delta_k = 0$ w.p. 1, which implies that (A.3) holds.

By Lemma 3, $\lim_{k \rightarrow \infty} \sup_{|t| \leq T} |M^{o,k}(t)| = 0$ w.p.1. So (A.4) holds for $e_{o,k}$.

Since $\{\Delta_k\}$ has support on a finite discrete set, $\{\widetilde{sg}(\theta_k)\}$ and $\{[\Delta_k^{-1}]\}$ are bounded, $E|e_{r,k}|^2 < \infty$. Using the martingale convergence theorem and Lemma 2, we get

$$\lim_{k \rightarrow \infty} \sup_{|t| \leq T} |M^{r,k}(t)| = 0 \quad w.p.1,$$

where $M^{r,k}(t) := \sum_{i=k}^{m(t+t_k)-1} \alpha_i e_{r,i}$. So (A.4) holds for $e_{r,k}$.

So all conditions of Proposition 1 are satisfied, and all its conclusions hold. $\square$

V. CONCLUSIONS

In this paper, we use convex analysis to establish convergence of constrained SPSA for the setting in which the objective function is not necessarily differentiable. As alluded to in the introduction, we were motivated to consider this setting by a capacity allocation problem in manufacturing, in which non-differentiability appeared for the case of period demands having discrete support rather than continuous. Similar phenomena arise in other contexts, e.g. in discrete event dynamic systems, as observed by Shapiro and Wardi [13]. Clearly, there are numerous avenues for further research in the non-differentiable setting. We believe the analysis can be extended to SPSA algorithms with nondifferentiable constraints, as well as to other (non-SPSA) stochastic approximation algorithms for nondifferentiable function optimization. For example, the same basic analysis could be used for RDSA (Random Directions Stochastic Approximation) algorithms as well. We specifically intend to consider global optimization of nondifferentiable functions along the line of [10]. On the more technical side, it would be desirable to weaken conditions such as (A.2) and (B.4), which are not required in [1].

Acknowledgements:

We thank the Associate Editor and three referees for their many comments and suggestions, which have led to an improved exposition. This work was supported in part by the National Science Foundation under Grants DMI-9713720 and DMI-9988867, by International SEMATECH and the Semiconductor Research Corporation under Grant NJ-877, by the Air Force Office of Scientific Research under Grant F496200110161, and by a fellowship from General Electric Corporate Research and Development through the Institute for Systems Research.

APPENDIX: PROOF OF PROPOSITION 1

Define $G^k(t) = \int_0^t \widetilde{s}f(\theta^k(s))ds$, and let B be the bound of $\widetilde{s}f(\theta^k(s))$, which exists due to the boundedness of $\partial f(\theta)$. Then, for each T and $\varepsilon > 0$, there is a δ which satisfies $0 < \delta < \varepsilon/B$, such that for all k , $\sup_{0 < t-s < \delta, |t| < T} |G^k(t) - G^k(s)| \leq \sup_{0 < t-s < \delta, |t| < T} \int_t^s \widetilde{s}f(\theta^k(u))du < \varepsilon$, which means $G^k(\cdot)$ is equicontinuous.

By (A.3) and (A.4), there is a null set O such that for sample points $\omega \notin O$, $M^k(\omega, \cdot)$ and $B^k(\omega, \cdot)$ go to zero uniformly on each bounded interval in $(-\infty, \infty)$ as $k \rightarrow \infty$. Hence, $M^k(\cdot)$ and $B^k(\cdot)$ are equicontinuous and their limits are zero.

By the same argument as in the proof of Theorem 5.2.1 in [9, pp.96-97], $Z^k(\cdot)$ is equicontinuous.

$\theta^k(\cdot)$ is equicontinuous since $M^k(\cdot)$, $B^k(\cdot)$, $G^k(\cdot)$ and $Z^k(\cdot)$ are equicontinuous.

Let $\omega \notin O$ and let k_j denote a subsequence such that

$$\{\theta^{k_j}(\omega, \cdot), G^{k_j}(\omega, \cdot)\}$$

converges, and denote the limit by $(\theta(\omega, \cdot), G(\omega, \cdot))$. The existence of such subsequences is guaranteed by Arzela-Ascoli theorem.

Since f is convex, according to Corollary 24.5.1 [11, p.234], $\forall \varepsilon > 0, \exists \delta > 0$, if $|\theta^{k_j}(\omega, s) - \theta(\omega, s)| < \delta$, then $-\partial f(\theta^{k_j}(\omega, s)) \subset N_\varepsilon(-\partial f(\theta(\omega, s)))$, where $N_\varepsilon(\cdot)$ means ε -neighborhood.

Furthermore, since $\lim_{j \rightarrow \infty} \theta^{k_j}(\omega, s) = \theta(\omega, s)$ for fixed ω and s , $\forall \varepsilon > 0, \exists$ finite J , if $j \geq J$, $-\partial f(\theta^{k_j}(\omega, s)) \subset N_\varepsilon(-\partial f(\theta(\omega, s)))$, i.e. for each $\widetilde{s}f(\theta^{k_j}(\omega, s)) \in -\partial f(\theta^{k_j}(\omega, s))$ and $\varepsilon > 0$, there is finite J and $\tilde{g}(\omega, s) \in -\partial f(\theta(\omega, s))$ such that if $j \geq J$, $|\widetilde{s}f(\theta^{k_j}(\omega, s)) -$

$$|\tilde{g}(\omega, s)| < \varepsilon.$$

Since $\tilde{s}f(\cdot)$ and $\tilde{g}(\omega, \cdot)$ are bounded functions on $[0, t]$, by the Lebesgue dominated convergence theorem,

$$\lim_{j \rightarrow \infty} \int_0^t \tilde{s}f(\theta^{k_j}(\omega, s))ds = \int_0^t \tilde{g}(\omega, s)ds,$$

which means $G(\omega, t) = \int_0^t \tilde{g}(\omega, s)ds$.

Thus, we can write $\theta(\omega, t) = \theta(\omega, 0) + \int_0^t \tilde{g}(\omega, s)ds + Z(\omega, t)$, where $\tilde{g}(\omega, s) \in -\partial f(\theta(\omega, s))$.

Using a similar argument as in [9, p.97], $Z(\omega, t) = \int z(\omega, s)ds$, where $z(\omega, s) \in -C(\theta(\omega, s))$ for almost all s .

Hence, the limit $\theta(\omega, \cdot)$ of any convergent subsequence satisfies the differential inclusion (14).

Note that S_H is the set of stationary points of (14) in H . Following a similar argument as in the proof of Theorem 5.2.1 [9, pp.95-99], we can show that $\{\theta_k(\omega)\}$ visits S_H infinite often, S_H is asymptotically stable in the sense of Lyapunov due to the convexity of f , and thus $\{\theta_k(\omega)\}$ converges to S_H w.p.1. Since f is a convex function and H is a non-empty convex set, by Theorem 27.4 in [11, pp.270-271], any point in S_H attains the minimum of f relative to H . $\square$

REFERENCES

- [1] H. F. Chen, T. E. Duncan, and B. Pasik-Duncan, "A Kiefer-Wolfowitz algorithm with randomized differences," *IEEE Transactions on Automatic Control*, vol. 44, no. 3, pp. 442–453, March 1999.
- [2] E. K. P. Chong, I.J. Wang, and S. R. Kulkarni, "Noise conditions for prespecified convergence rates of stochastic approximation algorithms," *IEEE Transactions on Information Theory*, vol. 45, no. 2, pp. 810–814, March 1999.
- [3] J. Dippon and J. Renz, "Weighted means in stochastic approximation of minima," *SIAM Journal on Control and Optimization*, vol. 35, no. 5, pp. 1811–1827, September 1997.
- [4] M. C. Fu and S. D. Hill, "Optimization of discrete event systems via simultaneous perturbation stochastic approximation," *IIE Transactions*, vol. 29, no. 3, pp. 233–243, March 1997.
- [5] L. Gerencsér, "Convergence rate of moments in stochastic approximation with simultaneous perturbation gradient approximation and resetting," *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 894–905, May 1999.
- [6] L. Gerencsér, G. Kozmann, and Z. Vágó, "SPSA for non-smooth optimization with application in ecg analysis," in *Proceedings of the IEEE Conference on Decision and Control*, Tampa, Florida, 1998, pp. 3907–3908.
- [7] Y. He, M. C. Fu, and S. I. Marcus, "A Markov decision process model for dynamic capacity allocation," in preparation.

- [8] N. L. Kleinman, J. C. Spall, and D. Q. Naiman, "Simulation-based optimization with stochastic approximation using common random numbers," *Management Science*, vol. 45, no. 11, pp. 1570–1578, November 1999.
- [9] H. J. Kushner and G. George Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 1997.
- [10] J. L. Maryak and D. C. Chin, "Stochastic approximation for global random optimization," in *Proceedings of the American Control Conference*, Chicago, IL, 2000, pp. 3294–3298.
- [11] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1970.
- [12] P. Sadegh, "Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation," *Automatica*, vol. 33, no. 5, pp. 889–892, May 1997.
- [13] A. Shapiro and Y. Wardi, "Nondifferentiability of steady-state function in discrete event dynamic systems," *IEEE Transactions on Automatic Control*, vol. 39, no. 8, pp. 1707–1711, August 1994.
- [14] J. C. Spall and J. A. Cristion, "Model-free control of nonlinear stochastic systems with discrete-time measurements," *IEEE Transactions on Automatic Control*, vol. 43, no. 9, pp. 1198–1210, September 1998.
- [15] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Transactions on Automatic Control*, vol. 37, no. 3, pp. 332–341, March 1992.
- [16] J. C. Spall, "A one-measurement form of simultaneous perturbation stochastic approximation," *Automatica*, vol. 33, no. 1, pp. 109–112, January 1997.
- [17] J. C. Spall, "An overview of the simultaneous perturbation method for efficient optimization," *Johns Hopkins APL Technical Digest*, vol. 19, no. 4, pp. 482–492, 1998.
- [18] J. C. Spall, "Adaptive stochastic approximation by the simultaneous perturbation method," *IEEE Transactions on Automatic Control*, vol. 45, no. 10, pp. 1839–1853, October 2000.
- [19] I. J. Wang and E. K. P. Chong, "A deterministic analysis of stochastic approximation with randomized directions," *IEEE Transactions on Automatic Control*, vol. 43, no. 12, pp. 1745–1749, December 1998.