2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Discrete Simultaneous Perturbation Stochastic Approximation on Loss Function with Noisy Measurements

Qi Wang and James C. Spall

Abstract— **Consider the stochastic optimization of a loss function defined on *p*-dimensional grid of points in Euclidean space. We introduce the middle point discrete simultaneous perturbation stochastic approximation (DSPSA) algorithm for such discrete problems and show that convergence to the minimum is achieved. Consistent with other stochastic approximation methods, this method formally accommodates noisy measurements of the loss function.**

Keywords— **Stochastic optimization; recursive estimation; SPSA; noisy data; discrete optimization.**

I. Introduction

THE optimization of real-world stochastic systems typically involves the use of a mathematical algorithm that iteratively seeks out the solution. It is often the case that the domain of optimization is discrete. Resource allocation, for instance, involves the distribution of discrete amount of some resource to a finite number of users in the face of uncertainty; other problems of interest within this framework include weapons assignment, plant location, network resource and experimental design. This paper introduces a method for stochastic discrete optimization that is based on stochastic approximation techniques customarily used in continuous optimization problems.

Many methods have been proposed to deal with discrete optimization problems. These methods include random search [2], simulated annealing [1], stochastic comparison [7], ordinal optimization [11], nested partitions [18]. Recently Hannah and Powell [8] propose an algorithm for one-stage stochastic combinatorial optimization problems, based on evolutionary policy iteration. Li et al. [13] introduce a method based on random search in the most promising area proposed in [12]. And Sklenar [19] considers an exhaustive local search method which is designed explicitly for noisy loss.

The aim here is to present an alternative method that can fully use the information of the structure of objective functions (e.g. "gradient") and potentially involve fewer function measurements. The simultaneous perturbation stochastic approximation (SPSA) algorithm [20, 21] was developed for continuous optimization problems of high dimension and where the loss function is expensive to evaluate. SPSA is a popular algorithm that creates gradient-type information from only two noisy function measurements in each iteration. The increase in efficiency over the finite difference stochastic approximation method, for example, has been shown to be a factor equal to the dimension of the problem [20]. Spall [20] has considered the convergence of SPSA for three times differentiable functions, whereas He et al. [9] have analyzed the convergence for nondifferentiable, but continuous optimization. Also Yousefian et al. [23] have discussed a local randomized smoothing technique for convex nondifferentiable continuous stochastic optimization. We want to use a similar idea of SPSA for the discrete case. Because the usual notion of a gradient does not apply in discrete problems, it is not obvious that the convergence properties demonstrated for SPSA hold for the discrete case. Hill et al. [10] considers a discrete form of SPSA and develops preliminary results associated with convergence for a separable discrete loss function under special conditions. However, this algorithm can be shown to not converge to the optimal solution in simple examples. We introduce a different form of discrete algorithm that applies to a broader range of problems while potentially retaining the essential efficiency advantages of standard SPSA.

In particular, we introduce a middle point discrete simultaneous perturbation stochastic approximation (DSPSA) algorithm that applies in a class of discrete problems. As in conventional SPSA, the method needs only two noisy measurements of the loss function at each iteration. Although a full convergence and convergence rate analysis has not yet been conducted, we show conditions for almost sure convergence of the algorithm to the true parameter value.

The paper is organized as the follows. In Section II, we motivate the general approach by considering the case of one dimensional θ, and describe the basic DSPSA algorithm for general $p \geq 1$. In Section III, we show that the algorithm converges to the optimal solution for some class of function under some conditions. In Section IV, we show how this algorithm compares with the localized random search method in two examples. In Section V, we conclude with a discussion.

Qi Wang is with the Department of Applied Mathematics and Statistics of the Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: qwang29@ jhu.edu).

James C. Spall is with The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099 USA and with the Department of Applied Mathematics and Statistics of the Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: james.spall@ jhuapl.edu).

II. Problem Formulation

A. Motivation: One Dimension Case

Let us first consider one dimensional discrete function L: $\mathbb{Z} \to \mathbb{R}$, where $\mathbb{Z}$ denotes the set of integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$. We want to find the minimal solution of the loss function L. Let the noisy measurement of the loss function be y, where $y = L + \varepsilon$ and ε indicates the noise. Fig. 1 shows an example of a discrete function in one dimension with a line connecting the neighboring integer points. The line $\bar{L}$ can be regarded as a continuous extension of L, but $\bar{L}$ is a nondifferentiable function at the integer points. For a point $\theta \in \mathbb{R} \setminus \mathbb{Z}$, the gradient is

$$\begin{aligned} g(\theta) &= L(\lceil \theta \rceil) - L(\lfloor \theta \rfloor) \\ &= L\left(\pi(\theta) + \frac{1}{2}\right) - L\left(\pi(\theta) - \frac{1}{2}\right) \\ &= \frac{L\left(\pi(\theta) + \frac{1}{2}\Delta\right) - L\left(\pi(\theta) - \frac{1}{2}\Delta\right)}{\Delta}, \end{aligned}$$

where $\lfloor . \rfloor$ is the floor function, $\lceil . \rceil$ is the ceiling function, and $\pi(\theta) = \left(2\lfloor \theta \rfloor + 1\right)/2$ is the middle point between $\lfloor \theta \rfloor$ and $\lceil \theta \rceil$, and Δ is a Bernoulli random variable taking the values ± 1. Actually $\pi(\theta) = n/2$ and n is an odd number, so $\pi(\theta) \pm \Delta/2$ must be integers. We can see that $g(\theta)$ is also well defined at integer points θ, and it is a subgradient (a vector $\boldsymbol{\gamma}$ is a subgradient of $L(\cdot)$ at $\boldsymbol{\theta}$ if $L(\boldsymbol{\mu}) - L(\boldsymbol{\theta}) \geq \boldsymbol{\gamma}^T (\boldsymbol{\mu} - \boldsymbol{\theta})$ for all $\boldsymbol{\mu} \in \mathbb{R}^p$) at θ ($\pi(\theta)$ is now the middle point between $\theta = \lfloor \theta \rfloor$ and $\theta + 1$). Then the estimated gradient for noisy function is

$$\hat{g}(\theta) = \frac{y\left(\pi(\theta) + \frac{1}{2}\Delta\right) - y\left(\pi(\theta) - \frac{1}{2}\Delta\right)}{\Delta}.$$

Ref. [9] has shown that SPSA method still converges for nondifferentiable functions when the functions are continuous and convex and the domains are convex and compact sets.

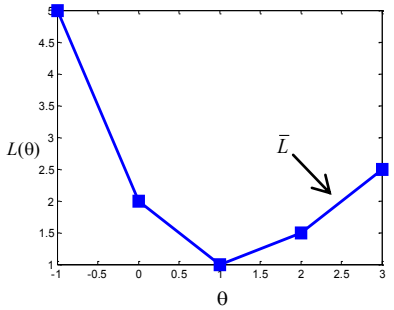

Fig. 1. Example of strictly discrete convex function and $\bar{L}$ is a continuous extension

B. Basic Algorithm of DSPSA

Motivated by the special example shown above, we will consider the case when $\boldsymbol{\theta}$ is p-dimensional, $p = 1,2,3,\ldots$. We have the general basic algorithm as below for function $y = L + \varepsilon$, where L: $\mathbb{Z}^p \to \mathbb{R}$ and ε is noise.

The basic algorithm is:

Step0: Pick an initial guess $\hat{\boldsymbol{\theta}}_0$.

Step1: Generate $\boldsymbol{\Delta}_k = [\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp}]^T$, where the Δ_{ki} are independent Bernoulli random variables taking the values ± 1 with probability $1/2$.

Step2: $\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) = \left(2\lfloor \hat{\boldsymbol{\theta}}_k \rfloor + \mathbf{1}_p\right)/2$, where $\mathbf{1}_p$ is a p-dimensional vector with all components equal unity and $\lfloor \hat{\boldsymbol{\theta}}_k \rfloor = \left[\lfloor \hat{\theta}_{k1} \rfloor, \ldots, \lfloor \hat{\theta}_{kp} \rfloor\right]^T$.

Step3: Evaluate y at $\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\Delta}_k/2$ and $\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{\Delta}_k/2$, form the estimate of $\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k)$,

$$\hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) = \left[y\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{2}\boldsymbol{\Delta}_k\right) - y\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\boldsymbol{\Delta}_k\right)\right] \boldsymbol{\Delta}_k^{-1},$$

where $\boldsymbol{\Delta}_k^{-1} = \left[\Delta_{k1}^{-1}, \ldots, \Delta_{kp}^{-1}\right]^T$.

Step4: Update the estimate according to the recursion

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

In the theoretical analysis below, we make use of the following mean gradient-like expression centered at $\boldsymbol{\pi}(\boldsymbol{\theta})$:

$$\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\boldsymbol{\theta})) \equiv E\left\{ \left[L\left(\boldsymbol{\pi}(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\Delta}\right) - L\left(\boldsymbol{\pi}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\Delta}\right)\right] \boldsymbol{\Delta}^{-1} \middle| \boldsymbol{\theta} \right\},$$

where $\boldsymbol{\Delta}$ is p-dimensional vector that has the same definition as $\boldsymbol{\Delta}_k$ mentioned above and $\boldsymbol{\theta}$ may be a random variable in some cases. If each direction is chosen equally, then

$$\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\boldsymbol{\theta})) = \frac{1}{2^p} \sum\nolimits_{\Delta} \left[L\left(\boldsymbol{\pi}(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{\Delta}\right) - L\left(\boldsymbol{\pi}(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\Delta}\right)\right] \boldsymbol{\Delta}^{-1},$$

where $\sum_{\Delta}$ indicates the summation over all possible directions $\boldsymbol{\Delta}$. Note that $\boldsymbol{\Delta}_k^{-1} = \boldsymbol{\Delta}_k$ and $\boldsymbol{\Delta}^{-1} = \boldsymbol{\Delta}$ in the Bernoulli ± 1 case; we use $\boldsymbol{\Delta}_k^{-1}$ to accommodate future extension to perturbation distributions other than Bernoulli ± 1.

III. CONVERGENCE PROPERTIES

We now present an almost sure (a.s.) convergence result for $\hat{\boldsymbol{\theta}}_k$. First we introduce some definitions that are used in

the proof to follow. For any point $\boldsymbol{\theta}$, we denote the set of middle points of all unit hypercubes containing $\boldsymbol{\theta}$ as $\mathcal{M}_\theta$. If $\boldsymbol{\theta}$ lies strictly inside one unit hypercube, $\mathcal{M}_\theta$ contains one point. But if $\boldsymbol{\theta}$ lies on the boundary, $\mathcal{M}_\theta$ contains multiple points. For any point $\boldsymbol{m}_\theta$ in $\mathcal{M}_\theta$, we have $|m_{\theta i} - t_i| \le 1/2$ for $i = 1,...,p$, where $\boldsymbol{\theta} = [t_1, t_2, ..., t_p]^T$ and $m_{\theta i}$ is the ith component of $\boldsymbol{m}_\theta$. Furthermore let $\Im_k = \{\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_k\}$.

Theorem 1. Assume L is a bounded function on $\mathbb{Z}^p$, and it has unique minimal point $\boldsymbol{\theta}^*$. Assume also (i) $a_k > 0$, $\lim_{k\to\infty} a_k = 0$, $\sum_{k=0}^{\infty} a_k = \infty$ and $\sum_{k=0}^{\infty} a_k^2 < \infty$; (ii) the components of $\boldsymbol{\Delta}_k$ are independently Bernoulli ± 1 distributed; (iii) For all k, $E[(\varepsilon_k^+ - \varepsilon_k^-) \mid \Im_k, \boldsymbol{\Delta}_k] = 0$ a.s. and the variance of $\varepsilon^{\pm}$ is uniformly bounded; (iv) $\sup_{k\ge 0} \| \hat{\boldsymbol{\theta}}_k \| < \infty$ a.s.; and (v) $-\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta)^T (\boldsymbol{\theta} - \boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{m}_\theta \in \mathcal{M}_\theta$ and all $\boldsymbol{\theta} \in \mathbb{R}^p \backslash \{\boldsymbol{\theta}^*\}$. Then $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^*$ a.s.

Proof. By the algorithm, we have

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{k+1} &= \hat{\boldsymbol{\theta}}_k - a_k \hat{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) \\ &= \hat{\boldsymbol{\theta}}_k - a_k \left[y\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{2}\boldsymbol{\Delta}_k\right) - y\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\boldsymbol{\Delta}_k\right)\right] \boldsymbol{\Delta}_k^{-1} \\ &= \hat{\boldsymbol{\theta}}_k - a_k \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{2}\boldsymbol{\Delta}_k\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\boldsymbol{\Delta}_k\right) + \varepsilon_k^+ - \varepsilon_k^- \right] \boldsymbol{\Delta}_k^{-1}. \quad (1)\end{aligned}$$

By conditions (i), (ii), (iv) and boundedness of L, we have

$$\lim_{k\to\infty} a_k \left(L(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + 1/2\boldsymbol{\Delta}_k) - L(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - 1/2\boldsymbol{\Delta}_k)\right) \boldsymbol{\Delta}_k^{-1} = \mathbf{0} \text{ a.s.} \quad (2)$$

Also suppose the variance of $\varepsilon_k^{\pm}$ are $(\sigma_k^{\pm})^2$. Then by Chebyshev's inequality and (iii),

$$\lim_{m\to\infty} \mathrm{P}\left\{ \left| a_k \varepsilon_k^{\pm} \right| > \eta \text{ for some } k \ge m \right\} \le \lim_{m\to\infty} \sum_{k=m}^{\infty} a_k^2 \frac{(\sigma_k^{\pm})^2}{\eta^2} \to 0,$$

implying by [4, Theorem 4.1.] that

$$\lim_{k\to\infty} a_k \{\varepsilon_k^+ - \varepsilon_k^-\} \boldsymbol{\Delta}_k^{-1} = \mathbf{0} \text{ a.s.} \quad (3)$$

Through (1), we have the relationship that $\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k = -a_k \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\Delta}_k/2\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \boldsymbol{\Delta}_k/2\right) + \boldsymbol{\varepsilon}_k^+ - \boldsymbol{\varepsilon}_k^- \right] \boldsymbol{\Delta}_k^{-1}$, and by the results of (2) and (3), we get $\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k \to \mathbf{0}$ a.s. Hence there exists $\Omega_1 \subseteq \Omega$ such that $\hat{\boldsymbol{\theta}}_{k+1}(\omega) - \hat{\boldsymbol{\theta}}_k(\omega) \to \mathbf{0}$ and $P(\Omega_1) = 1$. By condition (iv), $\{\hat{\boldsymbol{\theta}}_k(\omega)\}$ is a bounded sequence for any $\omega \in \Omega_1$. Then there exists a subsequence $\{\hat{\boldsymbol{\theta}}_{k_s}(\omega)\}$ and point $\boldsymbol{\theta}'(\omega)$ such that $\{\hat{\boldsymbol{\theta}}_{k_s}(\omega)\} \to \boldsymbol{\theta}'(\omega)$.

In addition, we can rewrite (1) as

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{k+1} &= \hat{\boldsymbol{\theta}}_k - a_k \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k)) \\ &\quad + a_k \left\{ \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k)) - \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{2}\boldsymbol{\Delta}_k\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\boldsymbol{\Delta}_k\right)\right] \boldsymbol{\Delta}_k^{-1} \right\} \\ &\quad - a_k \left(\varepsilon_k^+ - \varepsilon_k^-\right) \boldsymbol{\Delta}_k^{-1}.\end{aligned}$$

By the definition of $\bar{\boldsymbol{g}}(\cdot)$, we have $\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k)) = E\left\{ \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{2}\boldsymbol{\Delta}_k\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\boldsymbol{\Delta}_k\right)\right] \boldsymbol{\Delta}_k^{-1} \middle| \Im_k \right\}$

$= \frac{1}{2^p} \sum_{\Delta_k} \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{2}\boldsymbol{\Delta}_k\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\boldsymbol{\Delta}_k\right)\right] \boldsymbol{\Delta}_k^{-1}$. Let $b_k = L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{2}\boldsymbol{\Delta}_k\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_k) - \frac{1}{2}\boldsymbol{\Delta}_k\right)$, then for all $i < j$ we have $E\left[\left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - b_i \boldsymbol{\Delta}_i^{-1} \right)^T \left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_j)) - b_j \boldsymbol{\Delta}_j^{-1} \right) \right] =$

$E\left\{ E\left[\left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - b_i \boldsymbol{\Delta}_i^{-1} \right)^T \left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_j)) - b_j \boldsymbol{\Delta}_j^{-1} \right) \middle| \Im_j \right] \right\} =$

$E\left\{ \left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - b_i \boldsymbol{\Delta}_i^{-1} \right)^T E\left[\left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_j)) - b_j \boldsymbol{\Delta}_j^{-1} \right) \middle| \Im_j \right] \right\} = 0$. Then due to conditions (i) (ii) and (iv), for any k we have

$$E\left[\left\| \sum_{i=k}^{\infty} a_i \left\{ \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) + \frac{\boldsymbol{\Delta}_i}{2}\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) - \frac{\boldsymbol{\Delta}_i}{2}\right)\right] \boldsymbol{\Delta}_i^{-1} \right\} \right\|^2 \right] =$$

$$\sum_{i=k}^{\infty} a_i^2 E\left(\left\| \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) + \frac{\boldsymbol{\Delta}_i}{2}\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) - \frac{\boldsymbol{\Delta}_i}{2}\right)\right] \boldsymbol{\Delta}_i^{-1} \right\|^2 \right) < \infty. \quad (4)$$

Similarly, due to conditions (i) (ii) and (iii), we have

$$E\left[\left\| \sum_{i=k}^{\infty} a_i (\varepsilon_i^+ - \varepsilon_i^-) \boldsymbol{\Delta}_i^{-1} \right\|^2 \right] = \sum_{i=k}^{\infty} a_i^2 \mathrm{E}\left(\left\| (\varepsilon_i^+ - \varepsilon_i^-) \boldsymbol{\Delta}_i^{-1} \right\|^2 \right) < \infty. \quad (5)$$

Since for all k, $\left\{ \sum_{i=k}^{n} a_i \left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - \left(L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) + \boldsymbol{\Delta}_i/2\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) - \boldsymbol{\Delta}_i/2\right)\right) \boldsymbol{\Delta}_i^{-1} \right) \right\}_n$ and $\left\{ \sum_{i=k}^{n} a_i (\varepsilon_i^+ - \varepsilon_i^-) \boldsymbol{\Delta}_i^{-1} \right\}_n$ are martingales, then by (4), (5) and [3, Theorem 35.5], we know for all k, $\sum_{i=k}^{\infty} a_i \left(\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - \left(L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) + \boldsymbol{\Delta}_i/2\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) - \boldsymbol{\Delta}_i/2\right)\right) \boldsymbol{\Delta}_i^{-1} \right)$ exists and $\sum_{i=k}^{\infty} a_i (\varepsilon_i^+ - \varepsilon_i^-) \boldsymbol{\Delta}_i^{-1}$ exists. Let $M_k = \sum_{i=k}^{\infty} a_i (\varepsilon_i^+ - \varepsilon_i^-) \boldsymbol{\Delta}_i^{-1}$ and $N_k = \sum_{i=k}^{\infty} a_i \left\{ \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i)) - \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) + \boldsymbol{\Delta}_i/2\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i) - \boldsymbol{\Delta}_i/2\right)\right] \boldsymbol{\Delta}_i^{-1} \right\}$, then $\{M_k\}$ and $\{N_k\}$ are reverse martingales ([2, *p*.472]), and by [2, Theorem 35.8], there exist random variables M and N, such that $M_k \to M$ a.s. and $N_k \to N$ a.s. Furthermore

due to (4) and (5), we have $\lim_{k\to\infty} E\left[\|M_k\|^2\right] = 0$, and $\lim_{k\to\infty} E\left[\|N_k\|^2\right] = 0$. Also $\lim_{k\to\infty} E[M_k] = 0$ and $\lim_{k\to\infty} E[N_k] = 0$. Then $M = 0$ a.s., which indicates $M_k \to 0$ a.s. and $N = 0$ a.s. which indicates $N_k \to 0$ a.s. Then there exists $\Omega_2 \subseteq \Omega$ and $\Omega_3 \subseteq \Omega$ such that $P(\Omega_2) = 1$, $P(\Omega_3) = 1$ and such that for any $\omega \in \Omega_2$, $\sum_{i=k}^{\infty} a_i(\varepsilon_i^+(\omega) - \varepsilon_i^-(\omega))\Delta_i^{-1} \to 0$ and for any $\omega \in \Omega_3$, $\sum_{i=k}^{\infty} a_i \left\{\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega))) - \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega)) + \boldsymbol{\Delta}_i/2\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega)) - \boldsymbol{\Delta}_i/2\right)\right]\boldsymbol{\Delta}_i^{-1}\right\}$ $\to 0$. Let $\Omega_4 = \Omega_1 \cap \Omega_2 \cap \Omega_3$ with $P(\Omega_4)=1$. Then for any $\omega \in \Omega_4$ we have

$$\begin{aligned}\boldsymbol{\theta}'(\omega) &= \hat{\boldsymbol{\theta}}_{k_s}(\omega) - \sum_{i=k_s}^{\infty} a_i \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega))) \\ &\quad + \sum_{i=k_s}^{\infty} a_i \left\{\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega))) - \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega)) + \frac{1}{2}\boldsymbol{\Delta}_i\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega)) - \frac{1}{2}\boldsymbol{\Delta}_i\right)\right]\boldsymbol{\Delta}_i^{-1}\right\} \\ &\quad - \sum_{i=k_s}^{\infty} a_i \left(\varepsilon_i^+(\omega) - \varepsilon_i^-(\omega)\right)\boldsymbol{\Delta}_i^{-1},\end{aligned}$$

implying $\sum_{i=k_s}^{\infty} a_i \left(\varepsilon_i^+(\omega) - \varepsilon_i^-(\omega)\right)\boldsymbol{\Delta}_i^{-1} \to \mathbf{0}$ and $\sum_{i=k_s}^{\infty} a_i \left\{\bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega))) - \left[L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega)) + \boldsymbol{\Delta}_i/2\right) - L\left(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega)) - \boldsymbol{\Delta}_i/2\right)\right]\boldsymbol{\Delta}_i^{-1}\right\}$ $\to \mathbf{0}$ as $s \to \infty$. In addition, we know $\{\hat{\boldsymbol{\theta}}_{k_s}(\omega)\} \to \boldsymbol{\theta}'(\omega)$, indicating that

$$\sum_{i=k_s}^{\infty} a_i \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega))) \to \mathbf{0} \text{ as } s \to \infty. \tag{6}$$

Because $\hat{\boldsymbol{\theta}}_{k_s}(\omega) \to \boldsymbol{\theta}'(\omega)$, then for any $\delta > 0$, there exists $S > 0$ such that when $s > S$, $\left\|\hat{\boldsymbol{\theta}}_{k_s}(\omega) - \boldsymbol{\theta}'(\omega)\right\| < \delta$. Thus there exists S', when $s > S'$, all $\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_{k_s}(\omega)) \in \mathcal{M}_{\theta'}$. We now show $\boldsymbol{\theta}'(\omega)$ is the optimal point. By way of contradiction, suppose $\boldsymbol{\theta}'(\omega)$ is not the optimal solution. Then by condition (v), we have $-\bar{\boldsymbol{g}}(\boldsymbol{m}_{\theta'})^T(\boldsymbol{\theta}'(\omega) - \boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{m}_{\theta'} \in \mathcal{M}_{\theta'}$, which is a contradiction of $\sum_{i=k_s}^{\infty} a_i \bar{\boldsymbol{g}}(\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_i(\omega))) \to \mathbf{0}$ when $s > S'$. Then for all $\omega \in \Omega_4$ the limiting point of the sequence $\{\hat{\boldsymbol{\theta}}_k(\omega)\}$ is unique, which is equal to $\boldsymbol{\theta}^*$. Thus $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^*$ a.s.

Comment 1: The inner product condition (v) is a natural extension of the standard inner product condition for continuous problem (e.g. [22, p.106]), which includes convex function as a special case.

Comment 2: Actually some people have considered the discrete convexity. Miller [14] is a forerunner in the early 1970s in the area of discrete convex function. Ref. [14] has introduced the definition of discrete convex function and showed that the local optimal points for discrete convex function are also global optimal solutions. There are other definitions of discrete convex functions [5][15][16][6], but [17] shows that Miller's discrete convexity contains the other classes of discrete convexity. Note that Miller's definition does not include all functions satisfying condition (v), and condition (v) does not include all functions satisfy Miller's definition of discrete convexity. However, for $p = 1$, discrete convex functions satisfying Miller's definition also satisfy (v).

The corollaries below give two common functions satisfying condition (v). Even though we describe the functions in continuous form, for DSPSA we only use their values at multivariate integer points. Strictly convex separable functions mentioned in corollary 1 are discussed in [10].

Corollary 1. Strictly convex separable functions with minimal value at multivariate integer point satisfy the condition (v) in Theorem 1.

Proof. A separable function can be written as $\bar{L}(\boldsymbol{\theta}) = \sum_{i=1}^{p} \bar{L}_i(t_i)$, where $\boldsymbol{\theta} = [t_1, ..., t_p]^T$. And L is a discrete function has same values with $\bar{L}$ at multivariate integer points. Suppose the unique minimal point of $\bar{L}$ is $\boldsymbol{\theta}^*$, and $\boldsymbol{\theta}^*$ is a multivariate integer point with $\boldsymbol{\theta}^* = [t_1^*, ..., t_p^*]^T$. Then $\boldsymbol{\theta}^*$ is also the optimal point of L. Because it is strictly convex, then for all $\boldsymbol{\theta} \in \mathbb{R}^p \backslash \{\boldsymbol{\theta}^*\}$ and any subgradient $\partial\bar{L}_i(t_i)$, we have $\partial\bar{L}_i(t_i)(t_i^* - t_i) < 0$ for i=1,…, p. Moreover, for any $\boldsymbol{m}_\theta \in \mathcal{M}_\theta$, $\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta) = \frac{1}{2^p}\sum_{\boldsymbol{\Delta}}\left[L\left(\boldsymbol{m}_\theta + \boldsymbol{\Delta}/2\right) - L\left(\boldsymbol{m}_\theta - \boldsymbol{\Delta}/2\right)\right]\boldsymbol{\Delta}^{-1} = \frac{1}{2^p}\sum_{\boldsymbol{\Delta}}\left[\sum_{i=1}^{p}\left(L_i(m_{\theta i} + \Delta_i/2) - L_i(m_{\theta i} - \Delta_i/2)\right)\right]\boldsymbol{\Delta}^{-1} = \sum_{i=1}^{p}\frac{1}{2^p}\sum_{\boldsymbol{\Delta}}\left(L_i(m_{\theta i} + \Delta_i/2) - L_i(m_{\theta i} - \Delta_i/2)\right)\boldsymbol{\Delta}^{-1} = \sum_{i=1}^{p}\left(L_i(m_{\theta i} + 1/2) - L_i(m_{\theta i} - 1/2)\right)\boldsymbol{e}_i$. Then we have $\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta)^T(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \sum_{i=1}^{p}\left(L_i(m_{\theta i} + 1/2) - L_i(m_{\theta i} - 1/2)\right)(t_i - t_i^*)$. Because the minimal point is a multivariate integer point, then $\bar{L}_i(m_{\theta i} + 1/2) - \bar{L}_i(m_{\theta i} - 1/2)$ has the same sign with one of the subgradient of $\bar{L}_i$ at t_i, indicating that $\left(\bar{L}_i(m_{\theta i} + 1/2) - \bar{L}_i(m_{\theta i} - 1/2)\right)(t_i^* - t_i) < 0$ for all i=1,…,p.

Thus $-\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)<0$ for all $\boldsymbol{m}_\theta \in \mathcal{M}_\theta$ and all $\boldsymbol{\theta} \in \mathbb{R}^p \backslash \{\boldsymbol{\theta}^*\}$. Q.E.D.

Corollary 2. $\bar{L}$ is a strictly convex piecewise linear function with minimal value at a multivariate integer point and it is linear in each unit hypercube, then L satisfy the condition (v) in Theorem 1.

Proof. L is a discrete function that has same values with $\bar{L}$ at multivariate integer points. Since $\bar{L}$ is strictly convex function, then for all $\boldsymbol{\theta} \in \mathbb{R}^p \backslash \{\boldsymbol{\theta}^*\}$, and for any subgradient $\partial\bar{L}(\boldsymbol{\theta})$, we have $\partial\bar{L}(\boldsymbol{\theta})^T(\boldsymbol{\theta}^*-\boldsymbol{\theta})<0$. Furthermore for any $\boldsymbol{m}_\theta \in \mathcal{M}_\theta$, $\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta)=\frac{1}{2^p}\sum_\Delta\left[L(\boldsymbol{m}_\theta+\boldsymbol{\Delta}/2)-L(\boldsymbol{m}_\theta-\boldsymbol{\Delta}/2)\right]\boldsymbol{\Delta}^{-1}$ $=\frac{1}{2^p}\sum_\Delta\left[\bar{L}(\boldsymbol{m}_\theta+\boldsymbol{\Delta}/2)-\bar{L}(\boldsymbol{m}_\theta-\boldsymbol{\Delta}/2)\right]\boldsymbol{\Delta}^{-1}=\frac{1}{2^p}\sum_\Delta\left(\nabla\bar{L}(\boldsymbol{m}_\theta)^T\boldsymbol{\Delta}\right)\boldsymbol{\Delta}^{-1}$, where ∇ is the notation of gradient. Thus $\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta)^T(\boldsymbol{\theta}^*-\boldsymbol{\theta}) = \frac{1}{2^p}\sum_\Delta\left(\nabla\bar{L}(\boldsymbol{m}_\theta)^T\boldsymbol{\Delta}\right)\boldsymbol{\Delta}^{-T}(\boldsymbol{\theta}^*-\boldsymbol{\theta}) = \nabla L(\boldsymbol{m}_\theta)^T\frac{1}{2^p}\sum_\Delta\boldsymbol{\Delta}\boldsymbol{\Delta}^{-T}(\boldsymbol{\theta}^*-\boldsymbol{\theta}) = \nabla\bar{L}(\boldsymbol{m}_\theta)^T(\boldsymbol{\theta}^*-\boldsymbol{\theta})$. In addition for any $\boldsymbol{m}_\theta \in \mathcal{M}_\theta$, there will be one subgradient $\partial\bar{L}(\boldsymbol{\theta})$ at point $\boldsymbol{\theta}$, such that $\nabla\bar{L}(\boldsymbol{m}_\boldsymbol{\theta})=\partial\bar{L}(\boldsymbol{\theta})$. Then $\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta)^T(\boldsymbol{\theta}^*-\boldsymbol{\theta}) = \partial\bar{L}(\boldsymbol{\theta})^T(\boldsymbol{\theta}^*-\boldsymbol{\theta}) < 0$, which indicates $-\bar{\boldsymbol{g}}(\boldsymbol{m}_\theta)^T(\boldsymbol{\theta}-\boldsymbol{\theta}^*)<0$, for all $\boldsymbol{m}_\theta \in \mathcal{M}_\theta$ and all $\boldsymbol{\theta} \in \mathbb{R}^p \backslash \{\boldsymbol{\theta}^*\}$. Q.E.D.

IV. Comparision with Localized Random Search Method

Let us now compare the performance of DSPSA and the localized random search method for two loss functions. The first function considered here is a separable function $\sum_{i=1}^p t_i^2$. The second one is a skewed quartic loss function which is mentioned in [22, Ex 6.6]: $L(\boldsymbol{\theta}) = \boldsymbol{\theta}^T B^T B\boldsymbol{\theta} + 0.1\sum_{i=1}^p (B\boldsymbol{\theta})_i^3 + 0.01\sum_{i=1}^p (B\boldsymbol{\theta})_i^4$, where pB is an upper triangular matrix of 1's. Even though the skewed quartic loss function does not satisfy condition (v), we will see that DSPSA still works for this loss function. We consider the high-dimensional case for both functions, where p = 200, and the measurement noise ε is i.i.d $N(0,1)$. Since the localized random search method is more efficient in noise-free cases than in noisy cases, then we will consider both the noise-free situation and noisy situation. The localized random search method is described in [22, Sections 2.2−2.3], which consider both noise-free loss functions and noisy loss measurements, where a threshold parameter τ_k is involved. We will restrict the random search to the closest neighbor points, and all these points are chosen with equal probability. Here for DSPSA, let $a_k = a/(k+1+A)^\alpha$, a = 0.06 (for separable); a = 0.01 (for skewed quartic), A = 100, $\alpha = 0.602$. For the localized random search method, we choose $\tau_k = 2$ for the noisy case after several tuning. The initial guess is set to be $10\times\mathbf{1}_{200}$ in all runs. Fig. 2 and 3 show the performance of both methods under noise-free and noisy situations for separable function. And Fig. 4 and 5 show the performance of both methods for a skewed quartic function. We can see that DSPSA does better than the random search method for these two examples.

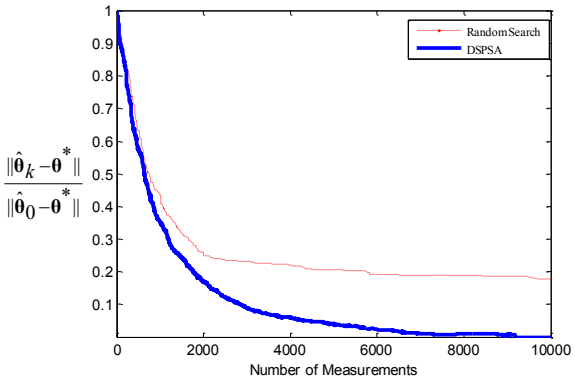

Fig. 2. Performance of localized random search method and DSPSA under noise-free situation for separable function.

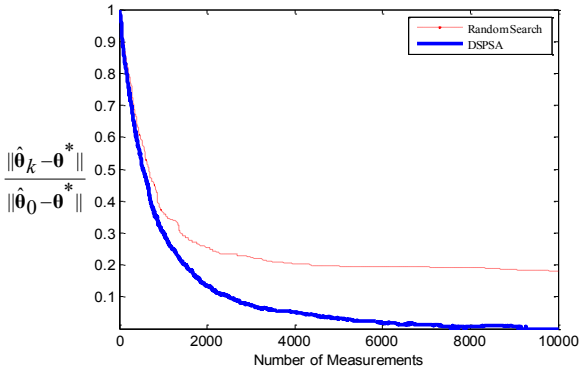

Fig. 3. Performance of localized random search method and DSPSA with noisy measurements for separable function.

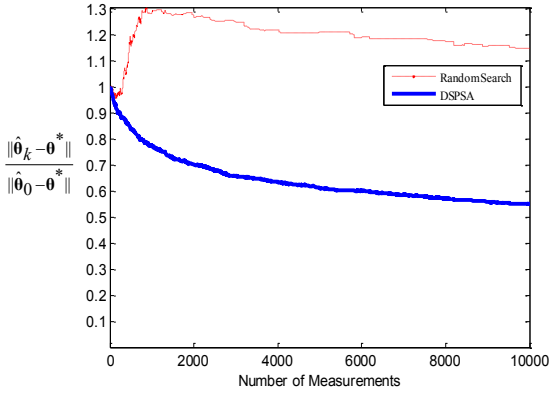

Fig. 4. Performance of localized random search method and DSPSA under noise-free situation for skewed quartic function.

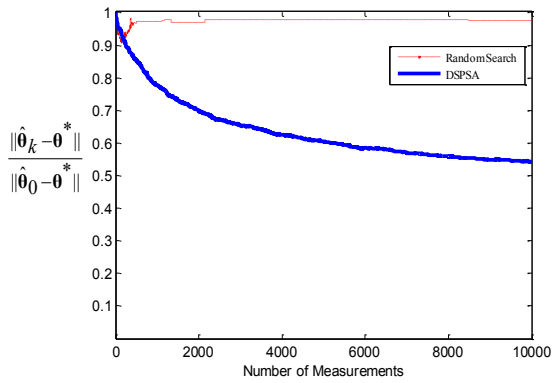

Fig. 5. Performance of localized random search method and DSPSA with noisy measurements for skewed quartic function.

V. Conclusion

In this paper, we introduced a discrete SPSA algorithm, and presented some preliminary convergence analysis. A preliminary numerical study shows that DSPSA works well on high-dimensional problems with or without noise in the loss measurements. As part of future work, we plan to formally study the convergence rate of the DSPSA and consider non-Bernoulli random variables for the perturbation vectors.

Also we intend to compare DSPSA with other popular discrete optimization algorithms, including those designed explicitly for handling noisy loss measurements (e.g. [7][13][19]). Two important practical problems of interest that involve stochastic discrete optimization are resource allocation, where a finite amount of a valuable commodity must be optimally allocated, and experimental design, where it is necessary to choose the best subset of input combinations from a large number of possible input combinations in a full-factorial design (e.g., [24]). We intend to explore the application of DSPSA to these or other problems.

Acknowledgment

This work was supported in part by the JHU/APL IRAD Program.

References

[1] M. H. Alrefaei, S. Andradóttir, "A Simulated Annealing Algorithm with Constant Temperature for Discrete Stochastic Optimization," *Management Sci.,* vol. 45, No.5, May 1999, pp. 748−764.

[2] S. Andradóttir, "A Method for Discrete Stochastic Optimization," *Management Sci.,* vol. 41, No.12, December 1995, pp. 1946−1961.

[3] P. Billingsley, *Probability and Measure*, Wiley-Interscience, Third Edition, 1995

[4] K. L. Chung, *A Course in Probability Theory,* Academic Press, Third Edition, 2001.

[5] P. Favati, F. Tardella, "Convexity in Nonlinear Integer Programming," *Ricerca Operativa ,* vol. 53, 1990, pp. 3−44.

[6] S. Fujishige, K. Murota, "Notes on L-/M-convex Functions and the Separation Theorems," *Mathematical Programming,* vol. 88, 2000, pp. 129−146.

[7] W. B. Gong, Y. C. Ho, W. Zhai, "Stochastic Comparison Algorithm for Discrete Optimization with Estimation," *SIAM J. Optim.,* vol. 10, No.2, 2000, pp. 384−404.

[8] L. A. Hannah and W.B. Powell, "Evolutionary Policy Iteration Under a Sampling Regime for Stochastic Combinatorial Optimization," *IEEE Transactions on Automatic Control,* vol. 55, No.5, May 2010, pp. 1254−1257.

[9] Y. He, M. C. Fu, and S. I. Marcus, "Convergence of Simultaneous Perturbation Stochastic Approximation for Nondifferentiable Optimization," *IEEE Transactions on Automatic Control,* vol. 48, No.8, August 2003, pp. 1459−1463.

[10] S. D. Hill, L. Gerencsér and Z. Vágó, "Stochastic Approximation on Discrete Sets Using Simultaneous Difference Approximations," *Proceeding of the 2004 American Control Conference*, Boston, MA, June 30−July 2, 2004, pp. 2795−2798.

[11] Y. C. Ho, Q. C. Zhao, and Q. S. Jia, *Ordinal Optimization: Soft Optimization for Hard Problems.* Springer, New York, NY, 2007.

[12] L. J. Hong and B. L. Nelson, "Discrete Optimization via Simulation Using COMPASS," *Oper. Res.,* vol. 54, No.1, 2006, pp. 115−129.

[13] J. Li, A. Sava, and X. Xie, "Simulation-Based Discrete Optimization of Stochastic Discrete Event Systems Subject to Non Closed-Form Constraints," *IEEE Transactions on Automatic Control,* vol. 54, No.12, December 2009, pp. 2900−2904.

[14] B. L. Miller, "On Minimizing Nonseparable Function Defined on the Integer with an Inventory Application," *SIAM Journal on Applied Mathematics,* vol. 21, No.1, July 1971, pp. 166−185.

[15] K. Murota, "Discrete Convex Analysis," *Mathematical Programming,* vol. 83, 1998, pp. 313−371.

[16] K. Murota, A. Shioura, "M-convex Function on Generalized Polymatroid," *Mathematics of Operations. Research,* vol. 24, 1999, pp. 95−105.

[17] K. Murota, A. Shioura, "Relationship of M-/L- Convex Function with Discrete Convex Functions by Miller and Favati-Tardella," *Discrete Applied Mathematics*, vol. 115, 2001, pp. 151−176.

[18] L. Shi and S. Olafsson, "Nested Partitions Method for Global Optimization," *Oper. Res.,* vol. 48, No.3, 2000, pp. 390−407.

[19] J. Sklenar, P. Popela "Integer Simulation Based Optimization by Local Search," *Procedia Computer Science,* vol. 1, 2010, pp. 1341−1348.

[20] J. C. Spall, "Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation," *IEEE Transactions on Automatic Control,* vol. 37, No.3, March 1992, pp. 332−341.

[21] J. C. Spall, "An Overview of the Simultaneous Perturbation Method for Efficient Optimization," *Johns Hopkins APL Technical Digest,* vol. 19, No.4, 1998, pp. 482−492.

[22] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control.* Wiley, Hoboken, NJ, 2003.

[23] F. Yousefian, A. Nedić, and U. V. Shanbhag, "Convex Nondifferentiable Stochastic Optimization: A Local Randomized Smooting Technique," *Proceedings of the American Control Conference*, Baltimore, MD, June 30–July 2, 2010, pp. 4875−4880.

[24] J. C. Spall, "Factorial Design for Choosing Input Values in Experimentation: Generating Informative Data for System Identification," *IEEE Control Systems Magazine*, vol. 30, no. 5, October 2010, pp. 38−53.