Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

AN APPROXIMATE ANNEALING SEARCH ALGORITHM TO GLOBAL OPTIMIZATION AND ITS CONNECTION TO STOCHASTIC APPROXIMATION

Jiaqiao Hu
Ping Hu

Department of Applied Mathematics and Statistics
State University of New York at Stony Brook
Stony Brook, NY 11794, USA

ABSTRACT

The Annealing Adaptive Search (AAS) algorithm searches the feasible region of an optimization problem by generating candidate solutions from a sequence of Boltzmann distributions. However, the difficulty of sampling from a Boltzmann distribution at each iteration of the algorithm limits its applications to practical problems. To address this difficulty, we propose an approximation of AAS, called Model-based Annealing Random Search (MARS), that samples solutions from a sequence of surrogate distributions that iteratively approximate the target Boltzmann distributions. We present the global convergence properties of MARS by exploiting its connection to the stochastic approximation method and report on numerical results.

1 INTRODUCTION

Random search methods have been recognized as a class of useful and effective tools for optimization of complex global optimization problems with little structure. Over the past few decades, various random search algorithms have been developed, ranging from the classical methods such as simulated annealing (Kirkpatrick, Gelatt, and Vecchi 1983), pure/adaptive random search (Zabinsky and Smith 1992), tabu search (Glover 1990), and genetic algorithms (Goldberg 1989), to the more recent ant colony optimization (Dorigo and Gambardella 1997), nested partitions method (Shi and Ólafsson 2000), estimation of distribution algorithms (Larranaga and Lozano 2002), cross-entropy (CE) as in the work (Rubinstein and Kroese 2004), and model reference adaptive search (MRAS) (Hu, Fu, and Marcus 2007), to name just a few. Because only the function values rather than structural information of the objective function such as continuity and differentiability are required, these methods are robust, easy to implement, and can be applied to a broad class of optimization problems.

The Annealing Adaptive Search (AAS) was first proposed in Romeijn and Smith (1994a) as an idealistic model to understand the behavior of simulated annealing. The algorithm samples candidate solutions according to a sequence of Boltzmann distributions parameterized by time-dependent temperatures. For a class of nonlinear optimization problems, AAS has the promising property that its computational complexity increases at most linearly with the problem dimension (e.g., Zabinksy 2003). However, what hinders its application to solving practical problems is that sampling exactly from a Boltzmann distribution is known to be extremely difficult. In attempts at resolving this difficulty, prior work has mostly focused on using Markov chain-based sampling techniques within the AAS framework to sample asymptotically from a Boltzmann distribution (cf. e.g., Romeijn and Smith 1994b, Zabinksy 2003). In this paper, we provide an alternative approach called Model-based Annealing Random Search (MARS). The underlying idea is to use a sequence of easy-to-sample distribution functions to approximate the target Boltzmann distributions and then use the sequence as surrogate

distributions to generate candidate points. The approximation technique involved in MARS inherits ideas from CE and MRAS, and is carried out by minimizing the Kullback-Leibler (KL) divergence between a family of parameterized distributions and the target Boltzmann distribution. However, we note that our approach does not require the quantile estimation of the distribution of the (unknown) objective function, a critical component used in the selection step of both CE and MRAS.

We also discuss a natural connection between MARS and the well-known stochastic approximation (SA) method (cf. e.g., Robbins and Monro 1951, Kushner and Yin 1997, Spall 2003). In particular, we show that, regardless of the type of decision variables of the original problem, MARS can be equivalently formulated into the form of a generalized stochastic approximation procedure on the parameter space (of the parameterized distribution family) for solving a sequence of *time-varying stochastic* optimization problems with differentiable structures. This viewpoint, which is new to this type of random search algorithms, allows us to study the asymptotic performance of MARS for a general class of global optimization problems, both continuous and discrete combinatorial, by using existing theory and analytical tools from SA.

The outline of the paper is as follows. In Section 2, we describe the MARS algorithm and establish its connection to SA. In Section 3, we present the global convergence property of MARS, followed by an asymptotic normality result in Section 4. Preliminary numerical results are reported in Section 5 and concluding remarks are given in Section 6. Due to space limitation, most of the proofs are omitted. A more comprehensive development of the approach and additional numerical results can be found in Hu and Hu (2010).

2 THE MARS ALGORITHM

We consider the global optimization problem

$$x^* \in \arg\max_{x \in \mathbb{X}} H(x), \tag{1}$$

where $H : \mathbb{X} \to \Re$ is a deterministic bounded objective function, $\mathbb{X} \subseteq \Re^n$ is a compact feasible region, which may either be continuous or discrete. We assume the existence of a unique global optimal solution x^* to (1); however, there could be multiple local optima.

The idealistic AAS algorithm iteratively approximates the global optimal solution x^* of (1) by assuming that solutions can be sampled exactly from the Boltzmann distribution

$$g_k(x) = \frac{e^{H(x)/T_k}}{\int_{\mathbb{X}} e^{H(x)/T_k} \nu(dx)} \tag{2}$$

at each iteration k, where T_k is an iteration-dependent temperature parameter and ν is the Lebesgue/discrete measure on $\mathbb{X}$. The idea is that as T_k decreases to a small constant $T^* \geq 0$, the sequence of $\{g_k\}$ will converge to a limiting distribution g^* that assigns most of its probability mass around x^*, so that near-optimal solutions will be sampled with higher probabilities as the search goes along. For a class of Lipschitz optimization problems, AAS is known to have several important theoretical properties, the most appealing of which is that its complexity increases at most linearly with the problem dimension (Romeijn and Smith 1994b, Zabinksy 2003). Unfortunately, the algorithm is not readily implementable to solving optimization problems, because the practical problem of sampling exactly from the Boltzmann distribution g_k is intractable in general. In the proposed MARS algorithm, we address this implementation difficulty of AAS by sampling points from a surrogate distribution that approximates g_k. The idea is to select a family of (easy-to-sample) parameterized distributions $\{f_\theta, \theta \in \Theta\}$ (Θ is the parameter space), and then project $\{g_k\}$ onto the parameterized family to obtain a sequence of sampling distributions. In particular, we borrow ideas from CE and MRAS, and carry out the projection at each iteration k of MARS by finding an optimal parameter θ_k that minimizes the

KL divergence between the family $\{f_\theta, \theta \in \Theta\}$ and g_k, i.e.,

$$\theta_k = \arg\min_{\theta\in\Theta} \mathscr{D}(g_k, f_\theta) := \arg\min_{\theta\in\Theta} E_{g_k}\Big[\ln \frac{g_k(X)}{f_\theta(X)}\Big], \tag{3}$$

where X denotes a random vector taking values in $\mathbb{X}$, and $E_g[\cdot]$ denotes the expectation taken with respect to the density/mass function g; also, throughout this paper, for a distribution parameterized by θ, we use $E_\theta[\cdot]$ to represent the expectation with respect to the underlying parameterized distribution. The primary reason for adopting the KL divergence is that for the natural exponential family (NEF) of distributions (cf. e.g., Morris 1982), the optimization problem (3) can be solved analytically in closed form for an arbitrary g_k, which makes the approach very convenient to implement efficiently.

Definition 1. *A parameterized family of density/mass functions* $\{f_\theta(\cdot),\ \theta \in \Theta \subseteq \Re^d\}$ *on* $\mathbb{X}$ *is said to belong to the natural exponential family (NEF) if there exist mappings* $\Gamma(\cdot) : \Re^n \to \Re^d$ *and* $K(\cdot) : \Re^d \to \Re$ *such that*

$$f_\theta(x) = \exp\left(\theta^T \Gamma(x) - K(\theta)\right), \quad \forall\, \theta \in \Theta,$$

where $K(\theta) = \ln \int_{\mathbb{X}} \exp\left(\theta^T \Gamma(x)\right) \nu(dx)$ *is called the log partition function, and* $\Theta = \{\theta \in \Re^d : |K(\theta)| < \infty\}$ *is called the natural parameter space.*

Let $int(\Theta)$ denote the interior of Θ. It is well-known (e.g., Morris 1982) that the function $K(\theta)$ is strictly convex on $int(\Theta)$ with gradient $\nabla K(\theta) = E_\theta[\Gamma(X)]$ and Hessian matrix $\mathrm{Cov}_\theta[\Gamma(X)]$. Therefore, the Jacobian of the mean parameter function $m(\theta) := E_\theta[\Gamma(X)]$ is strictly positive definite and invertible. From the inverse function theorem, it follows that $m(\theta)$ is also invertible on $int(\Theta)$.

In MARS, instead of directly using the sequence $\{g_k\}$ to minimize the KL-divergence in (3), we consider a general distribution sequence in the recursive form

$$\widetilde{g}_{k+1}(x) = \alpha_k g_{k+1}(x) + (1-\alpha_k) f_{\theta_k}(x) \ \text{ with } \alpha_k \in (0,1]\ \forall k = 0,1,\ldots, \tag{4}$$

where each $\widetilde{g}_{k+1}$ is a mixture of the Boltzmann distribution g_{k+1} parameterized by temperature T_{k+1} (cf. (2)) and the sampling distribution f_{θ_k} obtained at the kth iteration. Intuitively, such a mixture $\widetilde{g}_{k+1}$ retains the properties of the Boltzmann distribution g_{k+1}, while on the other hand, ensures that it does not stay too far apart from the sampling distribution f_{θ_k}.

When NEF is used to approximate the mixture distribution $\widetilde{g}_{k+1}$, the following lemma establishes a key link between the two successive mean parameter functions.

Lemma 1. *If* f_θ *belongs to the NEF and the new parameter* θ_{k+1} *obtained via minimizing* $\mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$ *satisfies* $\theta_{k+1} \in int(\Theta)$ *for all k, then*

$$m(\theta_{k+1}) - m(\theta_k) = -\alpha_k \nabla_\theta \mathscr{D}(g_{k+1}, f_\theta)|_{\theta=\theta_k} \quad \forall k = 0,1,2,\ldots. \tag{5}$$

Sketch of Proof: Since $\theta_{k+1} \in int(\Theta)$, it satisfies the first order necessary condition for optimality of the optimization problem $\min_{\theta\in\Theta} \mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$. Thus, setting the gradient $\nabla_\theta \mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$ to zero yields $m(\theta_{k+1}) = E_{\theta_{k+1}}[\Gamma(X)] = E_{\widetilde{g}_{k+1}}[\Gamma(X)]$. It follows from (4) that

$$m(\theta_{k+1}) = E_{\widetilde{g}_{k+1}}[\Gamma(X)] = \alpha_k E_{g_{k+1}}[\Gamma(X)] + (1-\alpha_k) m(\theta_k). \tag{6}$$

Finally, the result follows by rearranging the terms in (6) and then applying the dominated convergence theorem to switch the order of integral and derivative. □

Lemma 1 shows that by minimizing the KL divergence $\mathscr{D}(\widetilde{g}_{k+1}, f_\theta)$, the mean parameter function $m(\theta_{k+1})$ of the new sampling distribution $f_{\theta_{k+1}}$ can be viewed as an iterate generated by

a gradient descent algorithm for solving the *iteration-varying stochastic* minimization problem $\min_{\theta\in\Theta} \mathscr{D}(g_{k+1}, f_\theta) = \min_{\theta\in\Theta} E_{g_{k+1}}\left[\ln \frac{g_{k+1}(X)}{f_\theta(X)}\right]$ on the transformed parameter space Θ, whose solution, as k goes to infinity, is an optimal parameter $\theta^* \in int(\Theta)$ that provides the best possible approximation to the limiting Boltzmann distribution g^*. We remark that this observation is independent of the type of decision variables involved in the original optimization problem (1).

Note that in order to implement the above projection idea, we would still require the full information about the Boltzmann distribution g_{k+1}, which is generally unavailable unless the entire solution space $\mathbb{X}$ can be enumerated. Therefore, a rational approach in practice is to use only a finite number of samples generated at each iteration k to construct an empirical distribution $\bar{g}_{k+1}$, and then use $\bar{g}_{k+1}$ to approximate g_{k+1}. This results in the following implementable version of MARS:

Model-based Annealing Random Search (MARS)

Step 0: Choose an initial density/mass function $f_{\hat{\theta}_0}(x)$ on $\mathbb{X}$, $\hat{\theta}_0 \in int(\Theta)$. Specify an annealing schedule $\{T_k\}$, a step-size sequence $\{\alpha_k\}$, a sample size sequence $\{N_k\}$, and an exploration parameter sequence $\{\lambda_k\}$. Set iteration counter $k = 0$.

Step 1: Independently generate a population of N_k candidate solutions $\Lambda_k = \{X_1, \ldots, X_{N_k}\}$ as follows:

for $i = 1$ to N_k

generate a random number $u \sim U[0,1]$.

if $u < \lambda_k$, then sample a solution X_i from $f_{\hat{\theta}_0}$.

elseif $u \geq \lambda_k$, then generate a solution X_i according to $f_{\hat{\theta}_k}$.

endfor

Step 2: Compute the new parameter $\hat{\theta}_{k+1} = \arg\min_{\theta\in\Theta} \mathscr{D}(\widehat{g}_{k+1}, f_\theta)$, where $\widehat{g}_{k+1}$ is given in (7).

Step 3: If a stopping rule is satisfied, then terminate; otherwise set $k = k+1$ and go to Step 1.

In MARS, the initial density/mass function $f_{\hat{\theta}_0}$ can either be chosen based on prior knowledge of the underlying problem or be chosen in a way that any region in the solution space will have a positive probability of being sampled. In addition to the annealing temperature $\{T_k\}$ and the step-size sequence $\{\alpha_k\}$, the algorithm requires specifications of two parameter sequences $\{N_k\}$ and $\{\lambda_k\}$, where N_k specifies the number of candidate solutions to be generated at each iteration, and the exploration parameter λ_k allows the algorithm to explore the entire feasible region so that there is a positive probability for the algorithm to reach anywhere in $\mathbb{X}$ at each single iteration. At Step 2, the KL divergence is with respect to $\widehat{g}_{k+1}$, an estimate of $\widetilde{g}_{k+1}$ (cf. (4)) based on the sampled solutions in Λ_k, i.e.,

$$\widehat{g}_{k+1}(x) = \alpha_k \bar{g}_{k+1}(x) + (1-\alpha_k) f_{\hat{\theta}_k}(x), \quad x \in \Lambda_k. \tag{7}$$

Note that we have replaced the Boltzmann distribution g_{k+1} in (4) by an empirical distribution

$$\bar{g}_{k+1}(x) := \frac{e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\hat{\theta}_k}(x)}{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\hat{\theta}_k}(x)} \quad \forall x \in \Lambda_k, \tag{8}$$

where $\widehat{f}_{\hat{\theta}_k}(x) := (1-\lambda_k) f_{\hat{\theta}_k}(x) + \lambda_k f_{\hat{\theta}_0}(x)$ is the overall density/mass function that a candidate solution will be sampled at Step 1 of MARS. Intuitively, the division by $\widehat{f}_{\hat{\theta}_k}$ in $\bar{g}_{k+1}$ is used to compensate for solutions that are unlikely to be chosen, which makes $\bar{g}_{k+1}$ a good estimate of the Boltzmann distribution g_{k+1}.

Similar to Lemma 1, the following result shows the connection between the successive mean parameter vectors obtained in MARS.

Lemma 2. *If $\hat{\theta}_k \in int(\Theta)\ \forall k$, then the mean parameter function $m(\hat{\theta}_{k+1})$ of $f_{\hat{\theta}_{k+1}}$ satisfies*

$$m(\hat{\theta}_{k+1}) - m(\hat{\theta}_k) = -\alpha_k \Big(m(\hat{\theta}_k) - E_{\bar{g}_{k+1}}[\Gamma(X)] \Big) \quad \forall k = 0,1,2,\ldots. \tag{9}$$

Proof. Similar to the proof of Lemma 1. □

We conclude this section by relating MARS to stochastic gradient search. Note that (9) can be rewritten as follows:

$$\begin{aligned} m(\hat{\theta}_{k+1}) - m(\hat{\theta}_k) &= -\alpha_k \Big(m(\hat{\theta}_k) - E_{g_{k+1}}[\Gamma(X)] + E_{g_{k+1}}[\Gamma(X)] - E_{\bar{g}_{k+1}}[\Gamma(X)] \Big), \\ &= -\alpha_k \nabla_\theta \mathscr{D}(g_{k+1}, f_\theta)|_{\theta=\hat{\theta}_k} - \alpha_k \Big(E_{g_{k+1}}[\Gamma(X)] - E_{\bar{g}_{k+1}}[\Gamma(X)] \Big). \end{aligned} \tag{10}$$

This becomes a Robbins-Monro type stochastic approximation algorithm in terms of the true gradient and a noise term due to the approximation error between $\bar{g}_{k+1}$ and g_{k+1}. Thus, with the help of existing tools from stochastic gradient search and stochastic approximation, the asymptotic performance analysis of MARS essentially boils down to the issue of inspecting whether the Boltzmann distribution g_{k+1} can be closely approximated by its empirical estimate $\bar{g}_{k+1}$.

3 GLOBAL CONVERGENCE OF MARS

Since MARS is randomized, it induces a probability distribution over the set of all sampled solutions. We denote by $P(\cdot)$ and $E[\cdot]$ the probability and expectation taken with respect to this distribution. In the rest of the paper, probability one convergence is to be understood with respect to P. We also define $\mathscr{F}_k = \sigma\{\Lambda_0, \Lambda_1, \ldots, \Lambda_{k-1}\},\ k = 1,2,\ldots$ as the sequence of increasing σ-fields generated by the set of all sampled solutions up to iteration $k-1$. We use $\widehat{P}_{\hat{\theta}_k}(\cdot|\mathscr{F}_k)$ and $\widehat{E}_{\hat{\theta}_k}[\cdot|\mathscr{F}_k]$ to denote the conditional probability and expectation taken with respect to $\widehat{f}_{\hat{\theta}_k}$.

To present the main convergence result, we make the following assumptions, where Assumptions A1 and A2 are mild regularity conditions on the objective function, whereas A3−A5 are conditions on the input parameters.

Assumptions:

A1. *For any constant $\varepsilon < H(x^*)$, the set $\{x \in \mathbb{X} : H(x) \geq \varepsilon\}$ has a strictly positive Lebesgue or discrete measure.*

A2. *For any $\delta > 0$, $\sup_{x \in A_\delta} H(x) < H(x^*)$, where $A_\delta := \{x \in \mathbb{X} : \|x - x^*\| \geq \delta\}$.*

A3. *The mapping $\Gamma(x)$ given in Definition 1 is bounded on $\mathbb{X}$. Moreover, for any $\xi > 0$, there exists $\delta > 0$ such that $\|\Gamma(x) - \Gamma(x^*)\| \leq \xi$ whenever $\|x - x^*\| \leq \delta$.*

A4. *The step-size sequence $\{\alpha_k\}$ satisfies $\alpha_k > 0\ \forall k$, $\sum_{k=0}^{\infty} \alpha_k = \infty$, and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.*

A5. *(a) The annealing schedule $\{T_k\}$ satisfies $T_k > 0\ \forall k$ and $T_k \to T^* \geq 0$ as $k \to \infty$;*
(b) The exploration parameter sequence $\{\lambda_k\}$ satisfies $\lambda_k > 0\ \forall k$ and $\lambda_k \to \lambda^* \in [0,1)$ as $k \to \infty$;
(c) Moreover, $\frac{e^{2H^*/T_k}}{N_k \lambda_k} \to 0$ as $k \to \infty$, where $H^* = H(x^*)$.

We have the following convergence theorem for MARS.

Theorem 3. *If Assumptions A1 to A5 hold and $\hat{\theta}_k \in int(\Theta)\ \forall k$, then*

$$m(\hat{\theta}_k) \to E_{g^*}[\Gamma(X)] \ \ as \ \ k \to \infty \ \ w.p.1,$$

where the limit is taken component-wise and g^* is the limiting Boltzmann distribution parameterized by $T^* \geq 0$.

The result of Theorem 3 is much stronger than it appears to be. Its interpretation depends on the parameterized distribution family used in MARS and, in particular, on the specific form of the function $\Gamma(x)$. For example, in continuous optimization, if $T^* = 0$ and normal distributions are used as parameterized family, then Theorem 3 implies that the sequence of sampling distributions $\{f_{\hat{\theta}_k}\}$ in MARS will converge to a delta distribution with all mass concentrated at the global optimizer x^*, in the sense that $\lim_{k\to\infty} E_{\hat{\theta}_k}[X] = x^*$ and $\lim_{k\to\infty} \mathrm{Cov}_{\hat{\theta}_k}[X] = 0$ w.p.1. Another case of interest is when independent univariate density/mass functions are used and the parameterized family takes the form $f_\theta(x) = \prod_{i=1}^n \exp(x_i\vartheta_i - K(\vartheta_i))$, where x_i and ϑ_i are the respective ith components of $x = (x_1,\ldots,x_n)^T$ and the parameter vector $\theta = (\vartheta_1,\ldots,\vartheta_n)^T$, in which case, we have $\Gamma(x) = x$ and $m(\hat{\theta}_k) = E_{\hat{\theta}_k}[X]$. Thus, if $T^* = 0$, then the result of Theorem 3 reduces to $\lim_{k\to\infty} E_{\hat{\theta}_k}[X] = x^*$ w.p.1, i.e., the means of the sequence of sampling distributions converge to x^* w.p.1. As a third example, consider a discrete optimization problem with a feasible region $\mathbb{X}$ that contains m distinct values. To approach the problem, we can specify an m-by-1 probability vector Q, whose ith entry q_i indicates the probability that a solution will take the ith value $\mathbf{x}_i \in \mathbb{X}$. When parameterized by Q, the probability of sampling a solution x can be written as

$$f_\theta(x) = \prod_{i=1}^m q_i^{I\{x=\mathbf{x}_i\}} := e^{\theta^T\Gamma(x)},$$

where $\theta = [\ln q_1,\ldots,\ln q_m]^T$ and $\Gamma(x) = [I\{x=\mathbf{x}_1\},\ldots,I\{x=\mathbf{x}_m\}]^T$. Note that $\Gamma(x)$ satisfies Assumption A4. Thus, when $T^* = 0$, a straightforward interpretation of Theorem 3 yields

$$\lim_{k\to\infty} \sum_{x\in\mathbb{X}} \prod_{i=1}^m \left(q_i^k\right)^{I\{x=\mathbf{x}_i\}} I\{x=\mathbf{x}_j\} = I\{x^*=\mathbf{x}_j\} \quad \forall\, j \ \text{ w.p.1},$$

where q_i^k is the ith entry of the vector Q_k obtained at the kth iteration of MARS. This implies that $\lim_{k\to\infty} q_i^k = I\{x^*=\mathbf{x}_i\}$ w.p.1 $\forall\, i$. In other words, the sequence of probability vectors Q_k will converge to a limiting vector that assigns unit mass to the global optimum x^*.

Proof Sketch of Theorem 3: Since the function $\mathscr{D}(g_{k+1}, f_\theta)$ may change shape with k, our convergence proof is based on the analysis of a time-varying SA recursion given in Evans and Weber (1986). To proceed, we rewrite (9) in the form

$$\eta_{k+1} = \eta_k - \xi_k,$$

where $\eta_k := m(\hat{\theta}_k) - E_{g^*}[\Gamma(X)]$ and $\xi_k = \alpha_k\big(m(\hat{\theta}_k) - E_{\bar{g}_{k+1}}[\Gamma(X)]\big)$. To show the desired result, it is equivalent to show that $\eta_k \to 0$ as $k \to \infty$ w.p.1.

Let $M_k = \widehat{E}_{\hat{\theta}_k}[\xi_k|\mathscr{F}_k]$ and $Z_k = \xi_k - M_k$. We establish that the multivariate versions of conditions (i)-(iv) in Evans and Weber (1986) hold.

[i] First we show that for every $\varepsilon > 0$, $P(\|\eta_k\| > \varepsilon,\ \eta_k^T M_k < 0\ i.o.) = 0$. By the definition of M_k, it is easy to see that

$$M_k = \alpha_k\big(m(\hat{\theta}_k) - E_{g^*}[\Gamma(X)] + E_{g^*}[\Gamma(X)] - E_{g_{k+1}}[\Gamma(X)] + E_{g_{k+1}}[\Gamma(X)] - \widehat{E}_{\hat{\theta}_k}\left[E_{\bar{g}_{k+1}}[\Gamma(X)]\big|\mathscr{F}_k\right]\big). \quad (11)$$

Therefore,

$$\eta_k^T M_k = \alpha_k\Big(\|\eta_k\|^2 + \eta_k^T\left(E_{g^*}[\Gamma(X)] - E_{g_{k+1}}[\Gamma(X)]\right) + \eta_k^T\left(E_{g_{k+1}}[\Gamma(X)] - \widehat{E}_{\hat{\theta}_k}\left[E_{\bar{g}_{k+1}}[\Gamma(X)]\big|\mathscr{F}_k\right]\right)\Big).$$

Since η_k is bounded, by A1 and A2, it can be seen that the sequence of Boltzmann distributions $\{g_k\}$ converges to g^* in the sense that $\lim_{k\to\infty} E_{g_k}[\Gamma(X)] = E_{g^*}[\Gamma(X)]$. Moreover, by A3 and A5, it can be shown that the third term (i.e., the noise term) in the parenthesis above also vanishes to zero as

$k \to \infty$ w.p.1. Therefore, for almost every sample path generated by MARS, we must have $\eta_k^T M_k > 0$ whenever $\|\eta_k\| > \varepsilon$ for k sufficiently large, i.e., $P(\|\eta_k\| > \varepsilon,\ \eta_k^T M_k < 0\ i.o.) = 0$.

[ii] Since the mapping Γ is bounded on $\mathbb{X}$ by A3, both $m(\hat{\theta}_k)$ and $E_{\bar{g}_{k+1}}[\Gamma(X)]$ are bounded. Moreover, we have from Assumption A4 that $\alpha_k \to 0$ as $k \to \infty$. Therefore, $\|M_k\|(1+\|\eta_k\|)^{-1} \to 0$ as $k \to \infty$ w.p.1.

[iii] By the definition of Z_k, we have

$$\sum_{k=1}^{\infty} E[\|Z_k\|^2] = \sum_{k=1}^{\infty} \alpha_k^2 E\left[\left(\widehat{E}_{\hat{\theta}_k}\left[E_{\bar{g}_{k+1}}[\Gamma(X)]\middle|\mathscr{F}_k\right] - E_{\bar{g}_{k+1}}[\Gamma(X)]\right)^T \left(\widehat{E}_{\hat{\theta}_k}\left[E_{\bar{g}_{k+1}}[\Gamma(X)]\middle|\mathscr{F}_k\right] - E_{\bar{g}_{k+1}}[\Gamma(X)]\right)\right] < \infty,$$

since Γ is bounded and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$ by A4.

[iv] Finally, we show that $P\big(\liminf_{k\to\infty} \|\eta_k\| > 0,\ \sum_{k=1}^{\infty} \|M_k\| < \infty\big) = 0$. From (11), we have

$$\|M_k\| \geq \alpha_k\Big(\|\eta_k\| - \|E_{g^*}[\Gamma(X)] - E_{g_{k+1}}[\Gamma(X)]\| - \big\|E_{g_{k+1}}[\Gamma(X)] - \widehat{E}_{\hat{\theta}_k}\big[E_{\bar{g}_{k+1}}[\Gamma(X)]\big|\mathscr{F}_k\big]\big\|\Big).$$

Let $\Omega_1 = \{\liminf_{k\to\infty} \|\eta_k\| > 0\}$ and $\Omega_2 = \{\sum_{k=1}^{\infty} \|M_k\| < \infty\}$. For every sample point $\omega \in \Omega_1$, we can find a $\delta > 0$ such that $\liminf_{k\to\infty} \|\eta_k\| > \delta > 0$. This implies that there exists a $K_\delta(\omega)$ such that $\|\eta_k\| \geq \delta\ \forall k \geq K_\delta(\omega)$. In addition, let $\Omega_3 = \{\big\|E_{g_{k+1}}[\Gamma(X)] - \widehat{E}_{\hat{\theta}_k}\big[E_{\bar{g}_{k+1}}[\Gamma(x)]\big|\mathscr{F}_k\big]\big\| \to 0\}$. It can be shown that $P(\Omega_3) = 1$ and $E_{g_{k+1}}[\Gamma(X)] \to E_{g^*}[\Gamma(X)]$ as $k \to \infty$. Therefore, there exists a $\bar{K}_{\delta/2}(\omega)$ for every $\omega \in \Omega_3$ such that

$$\big\|E_{g^*}[\Gamma(X)] - E_{g_{k+1}}[\Gamma(X)]\big\| + \Big\|E_{g_{k+1}}[\Gamma(X)] - \widehat{E}_{\hat{\theta}_k}\big[E_{\bar{g}_{k+1}}[\Gamma(X)]\big|\mathscr{F}_k\big]\Big\| < \frac{\delta}{2}$$

for all $k \geq \bar{K}_{\delta/2}(\omega)$. Consequently, we have for every $\omega \in \Omega_1 \cap \Omega_3$, $\|M_k\| > \frac{\delta}{2}\alpha_k$ for all $k \geq K^*(\omega) := \max\{K_\delta(\omega), \bar{K}_{\delta/2}(\omega)\}$. Thus by A4,

$$\sum_{k=1}^{\infty} \|M_k\| \geq \sum_{k=K^*(\omega)}^{\infty} \|M_k\| \geq \frac{\delta}{2} \sum_{k=K^*(\omega)}^{\infty} \alpha_k = \infty \quad \forall\, \omega \in \Omega_1 \cap \Omega_3.$$

This implies $P(\Omega_1 \cap \Omega_2 \cap \Omega_3) = 0$. Thus, it follows that $P(\Omega_1 \cap \Omega_2) = P(\Omega_1 \cap \Omega_2 \cap \Omega_3) + P(\Omega_1 \cap \Omega_2 \cap \Omega_3^c) \leq P(\Omega_3^c) = 0$.

Finally, combining [i]−[iv] and directly applying the result of Evans and Weber (1986), we have $\eta_k \to 0$ as $k \to \infty$ w.p.1, which completes the proof of the theorem. □

4 ASYMPTOTIC CONVERGENCE RATE

To fix ideas, we consider a sample size sequence $N_k = O(k^\beta)$ and a step-size sequence of the form $\alpha_k = c/k^\alpha$ for some constants $\beta > 0$, $c > 0$, and $\alpha \in (\frac{1}{2}, 1)$. Note that such a choice of α_k satisfies A4. In addition, we require that $\{T_k\}$ and $\{\lambda_k\}$ satisfy the following strengthened version of Assumption A5.

Assumption B1. *For a given sample size sequence $N_k = O(k^\beta)$ and a step-size sequence $\alpha_k = O(k^{-\alpha})$, the sequence $\{T_k\}$ satisfies $T_k > T^* > 0\ \forall k$ and $\lim_{k\to\infty} k^{\frac{\alpha+\beta}{2}}\left(\frac{1}{T^*} - \frac{1}{T_k}\right) = 0$, and the sequence $\{\lambda_k\}$ satisfies $\lambda_k > 0\ \forall k$, $\lambda_k \to \lambda^* \in [0,1)$ as $k \to \infty$, and $\lambda_k = \Omega(k^{-\gamma})$ for some positive constant $\gamma < \frac{\beta}{2}$.*

It is easy to see that Theorem 3 still holds true with Assumption A5 replaced by B1. Thus, by the invertibility of $m(\cdot)$, the sequence of parameters $\{\hat{\theta}_k\}$ generated by MARS converges to a limiting parameter $\hat{\theta}^*$ w.p.1. Throughout this section, we should also assume that that $\hat{\theta}^* \in int(\Theta)$, i.e., the convergence of $\{\hat{\theta}_k\}$ occurs to a limiting point that lies in the interior of Θ. Since $m(\cdot)$ is continuously

differentiable on $int(\Theta)$, this assumption implies that the Jacobian of $m(\cdot)$ at $\hat{\theta}^*$ is strictly positive definite. Therefore, by inverse function theorem, there exists an open neighborhood of $m(\hat{\theta}^*)$ such that $m^{-1}(\cdot)$ is continuously differentiable on that neighborhood. This, together with the boundedness of Γ, further implies that the sequence of sampling distributions $\{f_{\hat{\theta}_k}\}$ converges point-wise to a limiting distribution $f_{\hat{\theta}^*}$ w.p.1.

We have the following asymptotic convergence rate result for MARS.

Theorem 4. *Let $\alpha_k = c/k^\alpha$ and $N_k = O(k^\beta)$ for constants $c > 0$, $\alpha \in (\frac{1}{2}, 1)$, and $\beta > \alpha$. Assume Assumptions A1–A3 and B1 hold, $\hat{\theta}_k \in int(\Theta)\ \forall k$, $\hat{\theta}^* \in int(\Theta)$, then*

$$k^{\frac{\alpha+\beta}{2}}\Big(m(\hat{\theta}_k) - E_{g^*}[\Gamma(X)]\Big) \xrightarrow{dist} \mathbf{N}\big(0,\Sigma\big) \quad as\ k \to \infty,$$

where $\Sigma = \Upsilon\widehat{Cov}_{\hat{\theta}^*}\Big[\big(\Gamma(X) - E_{g^*}[\Gamma(X)]\big)g^*(X)/\widehat{f}_{\hat{\theta}^*}(X)\Big]$ for some constant $\Upsilon > 0$, and $\widehat{Cov}_{\hat{\theta}^*}[\cdot]$ represents the covariance under $\widehat{f}_{\hat{\theta}^*}$.

Proof Sketch of Theorem 4: Given the specific forms of N_k and α_k, we can rewrite (9) in the form of a recursion in Fabian (1968):

$$\eta_{k+1} = (1 - ck^{-\alpha})\eta_k + k^{-(2\alpha+\beta)/2}R_k + k^{-(3\alpha+\beta)/2}W_k,$$

where $\eta_k = m(\hat{\theta}_k) - E_{g^*}[\Gamma(X)]$,

$$R_k = ck^{\beta/2}\Big(E_{\bar{g}_{k+1}}[\Gamma(X)] - \widehat{E}_{\hat{\theta}_k}\big[E_{\bar{g}_{k+1}}[\Gamma(X)]\big|\mathscr{F}_k\big]\Big),\ W_k = ck^{(\alpha+\beta)/2}\Big[\widehat{E}_{\hat{\theta}_k}\big[E_{\bar{g}_{k+1}}[\Gamma(X)]\big|\mathscr{F}_k\big] - E_{g^*}[\Gamma(X)]\Big].$$

Under conditions A1-A3 and B1, it can be verified that the term R_k satisfies the following two properties: (1) $\widehat{E}_{\hat{\theta}_k}[R_kR_k^T|\mathscr{F}_k] \to \Sigma$ as $k \to \infty$ w.p.1, where Σ is given in the statement of the theorem; (2) the sequence $\{R_k\}$ is uniformly square integrable in the sense that $\lim_{k\to\infty} E\big[I\{\|R_k\|^2 \geq rk^\alpha\}\|R_k\|^2\big] = 0\ \ \forall r > 0$. Moreover, the term W_k satisfies $k^{(\alpha+\beta)/2}W_k \to 0$ as $k \to \infty$ w.p.1. The desired result then follows from Theorem 2.2. in Fabian (1968). □

It is interesting to note that in contrast to general stochastic approximation algorithms, which have an optimal asymptotic rate of $O(1/\sqrt{k})$, Theorem 4 states that the asymptotic rate of convergence for MARS is at least $O(1/\sqrt{k})$ (i.e., when the values of α and β are chosen close to $1/2$). Moreover, this rate can be made arbitrarily fast by using a sample size sequence $\{N_k\}$ that increases sufficiently fast as $k \to \infty$. However, increasing sample sizes too fast may have a negative impact on the algorithm's practical performance, as the normality result is expressed in terms of the number of algorithm iterations, not the sample size. Therefore, there is a trade-off between the need for large values of β to increase the algorithm's (asymptotic) convergence speed and the desirability of using small values of β to reduce the per iteration computational cost.

5 NUMERICAL EXAMPLES

We illustrate the performance of MARS on multi-modal optimization problems and compare its performance with those of simulated annealing (SAN) and the Hide-and-Seek (HAS) algorithm (cf. e.g., Romeijn and Smith 1994b, Zabinksy 2003). The following four benchmark functions, taken from Hu, Fu, and Marcus (2007), are used in our experiment. The problem dimensions vary from 4 to 100. In particular, H_1 is low dimensional with only a few local optima; however, the maxima are separated by relatively flat regions and are far apart from each other. Functions H_2 has many wide-spread local optima, and the number of local maxima increases exponentially with the problem dimension. H_3 is a well-known badly-scaled problem, whereas H_4 is both highly multimodal and badly scaled.

(1) Shekel's function ($n = 4$, $0 \le x_i \le 10,\ i = 1,\ldots,n$)

$$H_1(x) = \sum_{j=1}^{5} \Big(\sum_{i=1}^{4} (x_i - A_{i,j})^2 + B_j \Big)^{-1} - 10.1532,$$

with $B = (0.1, 0.2, 0.2, 0.4, 0.4)^T$, $A_1 = A_3 = (4,1,8,6,3)$, and $A_2 = A_4 = (4,1,8,6,7)$, where A_i represents the ith row of A. The function has a global maxima $x^* = (4,4,4,4)^T$ and $H_1(x^*) = 0$.

(2) Trigonometric function ($n = 100$, $-10 \le x_i \le 10$, $i = 1\ldots,n$)

$$H_2(x) = -1 - \sum_{i=1}^{n} \Big[8\sin^2\big(7(x_i - 0.9)^2\big) + 6\sin^2\big(14(x_i - 0.9)^2\big) + (x_i - 0.9)^2 \Big],$$

where $x^* = (0.9,\ldots,0.9)^T$, $H_2(x^*) = -1$.

(3) Powell function ($n = 100,\ -10 \le x_i \le 10,\ i = 1\ldots,n$)

$$H_3(x) = -1 - \sum_{i=1}^{(n-2)/2} \Big[(x_{2i-1} + 10x_{2i})^2 + 5(x_{2i+1} - x_{2i+2})^2 + (x_{2i} - 2x_{2i+1})^4 + 10(x_{2i-1} - x_{2i+2})^4 \Big],$$

where $x^* = (0,\ldots,0)^T$ and $H_3(x^*) = -1$.

(4) Pinter's function ($n = 50$, $-10 \le x_i \le 10,\ i = 1,\ldots,n$)

$$\begin{aligned} H_4(x) \quad = \quad & -\sum_{i=1}^{n} i x_i^2 - \sum_{i=1}^{n} 20i \sin^2\big(x_{i-1}\sin x_i - x_i + \sin x_{i+1}\big) \\ & -\sum_{i=1}^{n} i \log_{10}\big(1 + i(x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2\big) - 1, \end{aligned}$$

where $x_0 = x_n$, $x_{n+1} = x_1$, $x^* = (0,\ldots,0)^T$, $H_4(x^*) = -1$.

In our implementation of MARS, we have used the independent multi-variate normal distributions as the parameterized distributions. Specifically, at the kth iteration of the algorithm, the parameterized sampling density takes the form

$$f_{\hat{\theta}_k}(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma_k^i)^2}} \exp\Big(-\frac{(x_i - \mu_k^i)^2}{2(\sigma_k^i)^2} \Big),$$

where the initial means are uniformly selected from the feasible region and initial variances $(\sigma_0^i)^2$ are set to 100 for all $i = 1,\ldots,n$. It is easy to verify that the new parameters are updated at Step 2 of MARS as

$$\mu_{k+1}^i = \alpha_k \frac{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\hat{\theta}_k}(x) x}{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\hat{\theta}_k}(x)} + (1 - \alpha_k)\mu_k^i$$

$$(\sigma_{k+1}^i)^2 = \alpha_k \frac{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\hat{\theta}_k}(x)(x - \mu_{k+1}^i)^2}{\sum_{x\in\Lambda_k} e^{\frac{H(x)}{T_{k+1}}} / \widehat{f}_{\hat{\theta}_k}(x)} + (1 - \alpha_k)\big((\sigma_k^i)^2 + (\mu_{k+1}^i - \mu_k^i)^2\big),$$

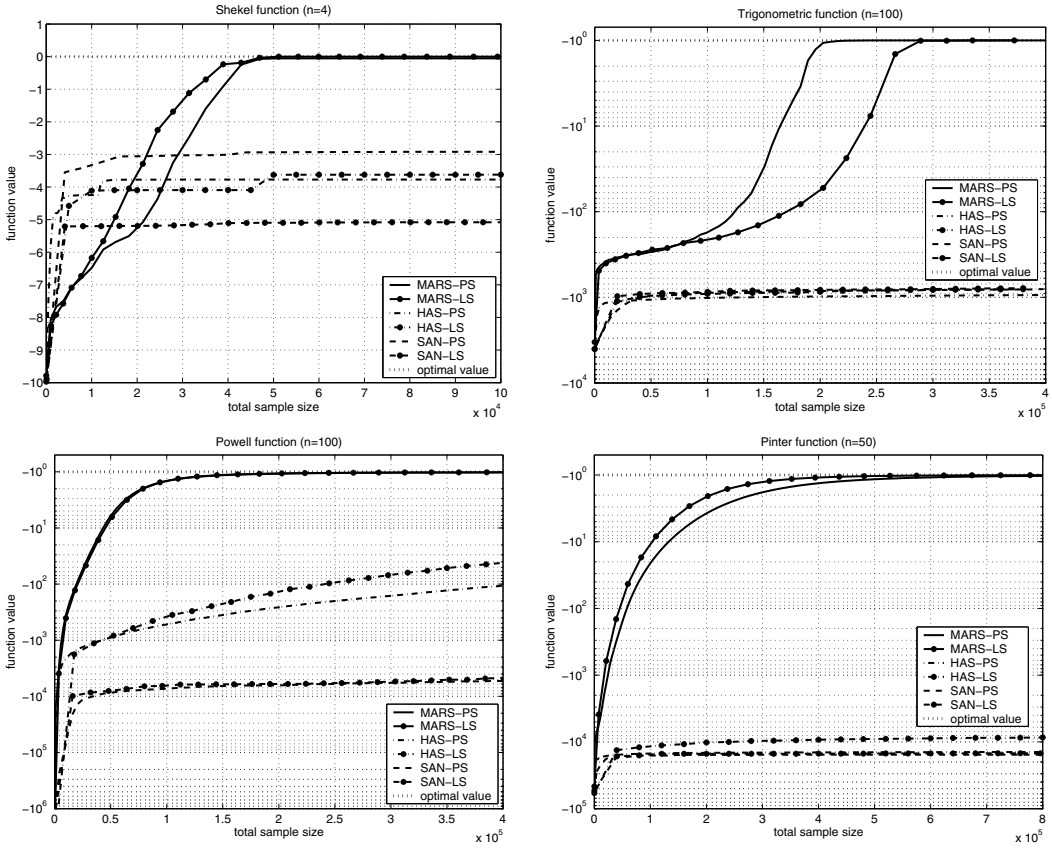

Figure 1: Averaged performance of MARS, HAS, and SAN on test functions H_1 to H_4.

for all $i = 1, \ldots, n$. In our implementation of SAN, we have used the following neighborhood structure: $\mathcal{N}(x) = \{y \in \mathbb{X} : \|x - y\|_\infty \leq 1\}$, which yields reasonable performance for all test problems. In HAS, the Markov chain sampler is implemented using a hyperspherical direction (e.g., Zabinksy 2003).

We consider two different annealing schedules in all three algorithms: (1) a polynomial schedule (PS): $T_k = 10^{-5} + |H(x_k^*)|/(1+k^{0.6})$; and (2) a logarithmic schedule (LS): $T_k = 10^{-5} + 0.1|H(x_k^*)|/\log(1+k)$, where x_k^* signifies the current best solution at the kth iteration of an algorithm, and the scaling factor $|H(x_k^*)|$ is introduced to counterbalance the effect of the magnitude of H in the term $e^{H(x)/T_k}$. Note that since H is bounded, both schedules PS and LS satisfy condition A5(a).

As in a typical stochastic approximation algorithm, we found empirically that the performance of MARS is primarily determined by the choice of the step-size sequence $\{\alpha_k\}$, but is insensitive to the choices of $\{N_k\}$ and $\{\lambda_k\}$. So a relatively conservative step-size $\alpha_k = 1/(k+100)^{0.501}$ is used in all four test cases, where the constant 100 is used to keep initial step sizes small in early iterations of the algorithm to prevent unstable behavior, whereas a slow decay rate 0.501 is used to produce non-negligible step sizes and prevent slow improvement in later iterations; see, e.g., Spall (2003) for a detailed discussion of step-size sequences of such a form. The other parameters in MARS are chosen as follows: $\lambda_k = 1/(1+k)^{0.5}$ and $N_k = \max\{10, \lfloor k^{0.502} \rfloor\}$, where $\lfloor a \rfloor$ is the largest integer no greater than a. Note that the above parameter settings satisfy the relevant conditions in Theorem 3 for convergence.

For each test case, we performed 50 independent replication runs of all three algorithms. The performances are shown in Figure 1, which plots the averaged function values at the best solutions found by the three comparison algorithms as a function of the number of function evaluations consumed thus far. Numerical results clearly indicate convergence of MARS with both annealing schedules

as well as its superior performance over both SAN and HAS. Since SAN combines local search, it shows a fast initial improvement, but the algorithm frequently stagnates at solutions that are far from optimal, especially in higher-dimensional cases. However, we note that the performance of both SAN and HAS may be improved by careful selections of neighborhood structures and adaptive annealing schedules tailored to specific problems.

6 CONCLUSIONS

In this paper, by combining ideas from AAS, CE, and MRAS, we have presented an algorithm called Model-based Annealing Random Search (MARS) for solving general global optimization problems with little structure. In addition, we have established a novel connection between the proposed algorithm and the well-known stochastic approximation method. This connection allows us to analyze the asymptotic performance of the algorithm for a general class of global optimization problems. Preliminary numerical results on high-dimensional multi-extremal benchmark problems show that MARS may yield high-quality solutions within a modest number of function evaluations.

ACKNOWLEDGMENTS

This work was supported in part by the Air Force Office of Scientific Research under Grant FA95501010340, and by the National Science Foundation under Grant DMI-0900332.

REFERENCES

Dorigo, M., and L. M. Gambardella. 1997. Ant colony system: A cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation* 1:53–66.

Evans, S. N., and N. C. Weber. 1986. On the almost sure convergence of a general stochastic approximation procedure. *Bulletin of the Australian Mathematical Society* 34:335–342.

Fabian, V. 1968. On asymptotic normality in stochastic approximation. *The Annals of Mathematical Statistics* 39:1327–1332.

Glover, F. W. 1990. Tabu search: a tutorial. *Interfaces* 20:74–94.

Goldberg, D. E. 1989. *Genetic algorithms in search, optimization, and machine learning*. Boston, MA: Academic Publishers.

Hu, J., M. C. Fu, and S. I. Marcus. 2007. A model reference adaptive search algorithm for global optimization. *Operations Research* 55:549–568.

Hu, J., and P. Hu. 2010. Annealing adaptive search, cross-entropy, and stochastic approximation in global optimization. working paper.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Kushner, H. J., and G. G. Yin. 1997. *Stochastic approximation algorithms and applications*. New York, NY: Springer-Verlag.

Larranaga, P., and J. A. Lozano. 2002. *Estimation of distribution algorithms: a new tool for evolutionary computation*. Boston, MA: Kluwer Academic Publishers.

Morris, C. N. 1982. Natural exponential families with quadratic variance functions. *Annals of Statistics* 10:65–80.

Robbins, H., and S. Monro. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 22:400–407.

Romeijn, H. E., and R. L. Smith. 1994a. Simulated annealing and adaptive search in global optimization. *Probability in the Engineering and Informational Sciences* 8:571–590.

Romeijn, H. E., and R. L. Smith. 1994b. Simulated annealing for constrained global optimization. *Journal of Global Optimization* 5:101–126.

Rubinstein, R. Y., and D. P. Kroese. 2004. *The cross-entropy method: A unified approach to combinatorial optimization, monte-carlo simulation, and machine learning*. New York, NY: Springer.

Shi, L., and S. Ólafsson. 2000. Nested partitions method for global optimization. *Operations Research* 48:390–407.

Spall, J. C. 2003. *Introduction to stochastic search and optimization*. Hoboken, NJ: John Wiley & Sons.

Zabinksy, Z. B. 2003. *Stochastic adaptive search for global optimization*. Norwell, MA: Kluwer Academic Publishers.

Zabinsky, Z. B., and R. L. Smith. 1992. Pure adaptive search in global optimization. *Mathematical Programming* 53:323–338.

AUTHOR BIOGRAPHIES

JIAQIAO HU is an Assistant Professor in the Department of Applied Mathematics and Statistics at the State University of New York, Stony Brook. He received a B.S. in automation from Shanghai Jiao Tong University, an M.S. in applied mathematics from the University of Maryland, Baltimore County, and a Ph.D. in electrical engineering from the University of Maryland, College Park. His research interests include Markov decision processes, applied probability, and simulation-based optimization. His e-mail address is <jqhu@ams.sunysb.edu>.

PING HU is a Ph.D. candidate in the Department of Applied Mathematics and Statistic at the State University of New York, Stony Brook. He received a B.S. degree in mathematics from Peking University, China in 2006. His research interests are in the areas of optimization and simulation. His e-mail address is <maycher0808@gmail.com>.