

# Adaptive Search with Stochastic Acceptance Probabilities for Global Optimization

Archis Ghatge<sup>a\*</sup> and Robert L. Smith<sup>b</sup>

<sup>a</sup>Industrial Engineering, University of Washington, Box 352650, Seattle, Washington, 98195, USA, archis@u.washington.edu.

<sup>b</sup>Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan, 48109, USA, rlsmith@umich.edu.

We present an extension of continuous domain Simulated Annealing. Our algorithm employs a globally reaching candidate generator, adaptive stochastic acceptance probabilities, and converges in probability to the optimal value. An application to simulation-optimization problems with asymptotically diminishing errors is presented. Numerical results on a noisy protein-folding problem are included.

Keywords: Global Optimization, Simulated Annealing, Markov Chain Monte Carlo, Simulation-Optimization

## 1. Introduction

We present a Markov chain method to solve the problem

$$\min f(x) \text{ s.t. } x \in S \subset R^d, \quad (1)$$

where  $f$  is a continuous function and  $S$  is a compact set. In particular, let  $R(.,.)$  be a Markov kernel called the *candidate kernel* on  $(S, \mathcal{B})$ , where  $\mathcal{B}$  denotes the Borel sigma field on  $S$ . We focus on Algorithm 1.1.

### Algorithm 1.1. Markov Chain Algorithm

1. Start with  $X_0 \in S$ ,  $k = 0$ .
2. Generate  $Y_{k+1}$  using candidate kernel  $R(X_k, .)$ .
3. Let  $0 \leq \mathcal{P}_k \leq 1$  be a random variable and set

$$X_{k+1} = \begin{cases} Y_{k+1} & \text{with probability } \mathcal{P}_k, \\ X_k & \text{with probability } 1 - \mathcal{P}_k. \end{cases}$$

4. Go to step 2.

The distribution of  $\mathcal{P}_k$  may depend on the entire sequence of iterates  $X_0, Y_1, X_1, \dots, Y_{k+1}$  thus

\*Corresponding author.

allowing for adaptive tuning of algorithmic parameters. We provide conditions under which the sequence  $\{f(X_k)\}$  converges in probability to the global minimum function value  $f_*$  of  $f$  over  $S$  (value convergence) in the second section and present the proof in the third. This result is applied to variants of (1) with noisy objective functions in the fourth section. Numerical experiments on a noisy protein folding problem are presented in the fifth.

Note that Algorithm 1.1 includes Simulated Annealing (SA) and several of its variants as special cases. SA was originally proposed as a stochastic search algorithm for discrete optimization [11] using a Metropolis acceptance filter parameterized by the so-called ‘temperature’. Its convergence properties for a logarithmic temperature schedule are well-known [8]. Recent generalizations of SA in the discrete domain include [9,10]. SA has also been extended to continuous problems [6,12,13,17]. Convergence analyses of continuous SA similar to this paper include [6,13]. An elegant proof for value convergence of SA in probability was presented in [6] where a globally reaching candidate kernel was used to allow for adaptive temperature schedules. This proof technique was adapted in [13] where locally reaching kernels were allowed but a deterministic temperature schedule was employed.

Convergence of discrete SA for noisy objective functions has been analyzed frequently in the literature (see [2,3,7,16,19] and references therein). Recent work on continuous domain stochastic search methods for noisy problems includes [1,5]. Continuity of the objective function was exploited in [5] to develop a novel estimation procedure. However, the technique was analyzed only for the special case of Pure Random Search (PRS) rendering it impractical. The convergence analysis in [1] was based on the restrictive assumptions that the objective function values have independent, heteroskedastic Normal distributions, the candidate point distribution does not depend on the current iterate, and tuning of algorithmic parameters is not adaptive.

Thus, to the best of our knowledge, Algorithm 1.1 is the first rigorously analyzed continuous domain Markov Chain Monte Carlo framework that *simultaneously* allows for generalized, stochastic acceptance probabilities, adaptive tuning of algorithmic parameters, and as we will show, noisy objective functions without restrictive assumptions on their distribution.

## 2. Sufficient Conditions for Convergence

We state five sufficient conditions for value convergence in probability. The first three are from [6].

**Condition 2.1.** *The feasible region  $S$  is a bounded and closed subset of  $R^d$ .  $f$  is continuous on  $S$ . Hence, there exists an  $x_* \in S$  where  $f$  achieves its minimum  $f_*$ . Assume moreover that for every  $\epsilon > 0$ , the set  $\{x \in S : |x - x_*| < \epsilon\}$  has positive Lebesgue measure.*

**Condition 2.2.** *For every open subset  $G$  in  $S$ , the candidate Markov kernel  $R(x, G)$  is continuous in  $x$ .*

**Condition 2.3.** *The candidate Markov kernel  $R$  is absolutely continuous in its second argument and its density is uniformly bounded away from 0. That is,  $R$  is of the form*

$$R(x, B) = \int_B r(x, y) dy \text{ with} \quad (2)$$

$$\gamma = \inf_{x, y \in S} r(x, y) > 0. \quad (3)$$

The next two conditions are on the acceptance probability filter  $\mathcal{P}_k$ . For each  $\beta > 0$ , we define a subset  $U_\beta$  of  $S \times S$  as  $U_\beta = \{(x, y) : x \in S, y \in S, f(y) \geq f(x) + \beta\}$ . Let  $\mathcal{U}_k(\beta)$  be the event that  $(X_k, Y_{k+1}) \in U_\beta$ .

**Condition 2.4.** *For any initial iterate  $X_0$ ,  $\epsilon > 0$ ,  $\beta > 0$ , and  $\delta > 0$ , there exists an integer  $N$  such that*

$$P[\mathcal{P}_k > \epsilon, \mathcal{U}_k(\beta)] < \delta \quad \forall k \geq N.$$

We also define a subset  $V$  of  $S \times S$  as  $V = \{(x, y) : x \in S, y \in S, f(y) \leq f(x)\}$ .  $\mathcal{V}_k$  is defined as the event that  $(X_k, Y_{k+1}) \in V$ .

**Condition 2.5.** *For any initial iterate  $X_0$ ,  $\epsilon > 0$ ,  $\beta > 0$ , and  $\delta > 0$ , there exists an integer  $N$  such that*

$$P[1 - \mathcal{P}_k > \epsilon, \mathcal{V}_k] < \delta \quad \forall k \geq N.$$

The first two conditions are regularity conditions. Condition 2.1 ensures in the presence of Condition 2.3 that neighborhoods of a global minimizer are visited infinitely often. Note that the second part of Condition 2.1 is satisfied when  $S$  is ‘locally full dimensional’ at  $x_*$ , meaning that there exists a  $\delta > 0$  such that the open ball of radius  $\delta$  centered at  $x_*$  is contained in  $S$ . Equation (3) in Condition 2.3 states that the density function of the Markov kernel is globally reaching, which helps ensure value convergence under weak conditions on the acceptance probability filter. The Hit-and-Run candidate generator proposed by one of us [20] is perhaps the most effective (see [12] and references therein) globally reaching candidate generator Markov chain in the literature. Hit-and-Run also satisfies Condition 2.2 [18].

The intuition behind the last two conditions is as follows. Optimization algorithms based on Markov chain candidate generators typically compute the difference between the function values at the current iterate and the candidate point. Intuitively, the probability of accepting a candidate point whose objective function value is ‘strictly worse’ than the current iterate should asymptotically vanish. This is characterized by the definition of set  $U_\beta$  and Condition 2.4. On the other hand, if the function value at the candidate point

is at least as good as the current iterate, there is probably no harm in accepting it. In fact, as we will show later, it suffices to let the probability of accepting points that are ‘no worse’ asymptotically approach 1. This is characterized by the definition of set  $V$  and Condition 2.5. In short, Condition 2.4 requires the acceptance probability filter to employ asymptotically diminishing exploration, while Condition 2.5 ensures asymptotically increasing exploitation.

As an example, we illustrate that the acceptance probability in SA satisfies these two conditions. In that case, given  $(X_k = x, Y_{k+1} = y)$ ,  $\mathcal{P}_k$  is a deterministic function  $p_k(x, y)$  given by

$$p_k(x, y) = \begin{cases} \exp[-\frac{f(y)-f(x)}{t_k}] & \text{if } f(y) > f(x), \\ 1 & \text{if } f(y) \leq f(x), \end{cases}$$

where  $\{t_k\}$  is a deterministic sequence of numbers that converges to zero. Condition 2.5 is trivially satisfied for all  $k$ . To see that Condition 2.4 is also satisfied fix  $\beta > 0$  and  $\epsilon > 0$ . Since  $t_k$  converges to 0, there exists an integer  $M$  such that  $t_k < \frac{\beta}{1/\epsilon}$  for all  $k \geq M$ . For all such values of  $k$ ,  $p_k(x, y) < \epsilon$  whenever  $(x, y) \in U_\beta$ . Similarly, it is easy to confirm that Conditions 2.4 and 2.5 are met in essentially all variants of continuous SA, for example, SA with an adaptive temperature schedule [?, ?], Barker acceptance criterion [4] with a deterministic temperature schedule, generalized acceptance probability proposed by Tsallis [21] with a deterministic temperature schedule, and Improving Hit-and-Run [22]. The main result in this paper is

**Theorem 2.6.** *For every initial iterate  $X_0 \in S$ , the sequence of function values  $\{f(X_k), k = 0, 1, 2, \dots\}$  generated by Algorithm 1.1 converges in probability to  $f_*$  whenever Conditions 2.1-2.5 are satisfied.*

### 3. Proof of Convergence

The proof technique is similar to Belisle [6]. For any  $\epsilon > 0$ , let the set  $S_\epsilon = \{x \in S : f(x) \leq f_* + \epsilon\}$ . For positive integers  $l, m, n$  we define the following four events.

$A = A(m, n)$  = the event that none of the states  $X_n, X_{n+1}, \dots, X_{n+m}$  is in  $S_\epsilon$ .

$B = B(l, m, n)$  = the event that at least one of the transitions  $X_{n+(k-1)} \rightarrow X_{n+k}$ ,  $k = 1, 2, \dots, m$  is a move from  $S_\epsilon$  to  $\{x \in S : f_* + \epsilon < f(x) < f_* + \epsilon + 1/l\}$ .

$C = C(l, m, n)$  = the event that at least one of the transitions  $X_{n+(k-1)} \rightarrow X_{n+k}$ ,  $k = 1, 2, \dots, m$  is a move from  $S_\epsilon$  to  $\{x \in S : f(x) \geq f_* + \epsilon + 1/l\}$ .

$D$  = the event that  $X_{n+m} \notin S_\epsilon$ .

To prove theorem 2.6, we need to prove that for every  $x_0 \in S$ ,  $\epsilon > 0$ , and  $\delta > 0$ , there exists an integer  $N$  such that

$$P[X_n \notin S_\epsilon | (X_0 = x_0)] < \delta \quad \forall n \geq N. \quad (4)$$

We begin by fixing  $x_0 \in S$ ,  $\epsilon > 0$  and  $\delta > 0$ . For simplicity of notation, henceforth, we may not write the conditioning on  $(X_0 = x_0)$ . It is to be understood implicitly. Since  $D \subset A \cup B \cup C$ ,

$$P[X_{n+m} \notin S_\epsilon] = P[D] \leq P[A] + P[B] + P[C].$$

Also, the symbol  $\phi$  will be used to denote the Lebesgue measure. The proof requires several Lemmas which are now presented one by one. Note that there exists an integer  $m_0$  such that  $(1 - \gamma\phi(S_\epsilon))^{m_0} < \delta/6$ . Fix this  $m_0$ . For integers  $n, m$  and  $r$ , we define  $G_r \equiv G(n, m, r)$  as the event that the candidate point  $Y_{n+r}$  is in  $S_\epsilon$ , for  $1 \leq r \leq m$ .

**Lemma 3.1.** *There exists an integer  $n_0$  such that for any integer  $r$ ,  $1 \leq r \leq m_0$ ,  $P[A(m_0, n), G_r] \leq \frac{\delta}{6m_0} \quad \forall n \geq n_0$ .*

*Proof.* Fix an integer  $r$  such that  $1 \leq r \leq m_0$ . Then  $P[A(m_0, n), G_r]$  is bounded above by

$$P[X_{n+r-1} \notin S_\epsilon, X_{n+r} \notin S_\epsilon, Y_{n+r} \in S_\epsilon] \\ = P[X_{n+r-1} \notin S_\epsilon, Y_{n+r} \in S_\epsilon, Y_{n+r} \text{ not accepted}].$$

Let  $H_{n+r}$  be the event that candidate  $Y_{n+r}$  is not accepted by the acceptance probability filter. Moreover, let  $I_{n+r}(\delta)$  be the event that  $1 - \mathcal{P}_{n+r-1} > \frac{\delta}{12m_0}$ . The above probability is then bounded above by

$$P[X_{n+r-1} \notin S_\epsilon, Y_{n+r} \in S_\epsilon, H_{n+r}, I_{n+r}(\delta)] \\ + P[X_{n+r-1} \notin S_\epsilon, Y_{n+r} \in S_\epsilon, H_{n+r}, I_{n+r}^c(\delta)] \\ \leq P[\mathcal{V}_{n+r-1}, I_{n+r}(\delta)] + \frac{\delta}{12m_0} \leq \frac{\delta}{12m_0} + \frac{\delta}{12m_0} \\ = \frac{\delta}{6m_0} \quad \forall n \geq n_r \text{ for some } n_r,$$

where the last step follows from Condition 2.5. Thus, there exists an integer  $n_0$  such that for any integer  $r$  between 1 and  $m_0$ ,  $P[A(m_0, n), G(n, m_0, r)] \leq \frac{\delta}{6m_0}$  for all  $n \geq n_0$ . This completes the proof.  $\square$

**Lemma 3.2.** *Let  $m_0$  and  $n_0$  be as above. Then  $P[A(m_0, n)] < \delta/3$  for all integers  $n \geq n_0$ .*

*Proof.* We define  $F \equiv F(n, m)$  as the event that none of the candidate points  $Y_{n+1}, \dots, Y_{n+m}$  is in  $S_\epsilon$ . We have  $A \subset (A \cap F) \cup (\bigcup_{r=1}^m (A \cap G_r))$ . Therefore,  $P[A(m, n)]$  is bounded above by

$$\begin{aligned} & P[A(m, n), F] + \sum_{r=1}^m P[A(m, n), G(m, n, r)] \\ & \leq P[X_n \notin S_\epsilon, F] + \sum_{r=1}^m P[A(m, n), G(m, n, r)] \\ & \leq (1 - \gamma\phi(S_\epsilon))^m + \sum_{r=1}^m P[A, G_r], \end{aligned}$$

where the last inequality follows since the probability that the candidate point will be in  $S_\epsilon$  is at least  $\gamma\phi(S_\epsilon)$  regardless of the past at each iteration. Then, from the definition of  $m_0$ ,  $P[A(m_0, n)]$  is bounded above by

$$\frac{\delta}{6} + \sum_{r=1}^{m_0} P[A(m_0, n), G(m_0, n, r)] < \delta/3 \quad (5)$$

for all  $n \geq n_0$  by Lemma 3.1.  $\square$

**Lemma 3.3.** *Let  $m_0$  be as in Lemma 3.2. There exists an integer  $l_0$  such that*

$$P[B(l_0, m_0, n)] < \delta/3 \quad \forall n \geq 0.$$

*Proof.* The proof of this lemma depends only on Conditions 2.2 and 2.3 on the Markov kernel and compactness of the feasible region  $S$ . In particular, it does not depend on the form of the acceptance probability. Since our conditions on the Markov kernel are the same as those of Belisle [6], and our feasible region  $S$  is compact, his proof carries over to our case.  $\square$

**Lemma 3.4.** *Let  $m_0$  and  $l_0$  be as in Lemmas 3.2 and 3.3. There exists an integer  $n_1$  such that*

$$P[C(l_0, m_0, n)] < \delta/3 \quad \forall n \geq n_1.$$

*Proof.* For  $k = 1, 2, \dots, m_0$ , let  $C_k(l_0, m_0, n)$  be the event that the transition  $X_{n+k-1}$  to  $X_{n+k}$  is a move from  $S_\epsilon$  to  $\{x \in S : f(x) \geq f_* + \epsilon + 1/l_0\}$ . Observe that  $C(l_0, m_0, n) = \bigcup_{k=1}^{m_0} C_k(l_0, m_0, n)$ . Therefore,  $P[C(l_0, m_0, n)] \leq \sum_{k=1}^{m_0} P[C_k(l_0, m_0, n)]$ . The probability  $P[C_k(l_0, m_0, n)]$  is equal to

$$\begin{aligned} & P[C_k(l_0, m_0, n), \mathcal{P}_{n+k-1} > \frac{\delta}{6m_0}] \\ & + P[C_k(l_0, m_0, n), \mathcal{P}_{n+k-1} \leq \frac{\delta}{6m_0}] \\ & \leq P[\mathcal{U}_{n+k-1}, \mathcal{P}_{n+k-1} > \frac{\delta}{6m_0}] + \frac{\delta}{6m_0} \\ & \leq \frac{\delta}{6m_0} + \frac{\delta}{6m_0} = \frac{\delta}{3m_0} \quad \forall n \geq n_1 \end{aligned}$$

for some integer  $n_1$  by Condition 2.4.  $\square$

Let  $n_2 = \max(n_0, n_1)$ . Lemmas 3.2, 3.3, and 3.4 imply that equation (4) holds with  $N = n_2 + m_0$ . The proof of Theorem 2.6 is now complete.

#### 4. Application to Stochastic Optimization

In this section, we apply Theorem 2.6 to the important special case of global optimization problems with noisy objective functions. Consider a specific version of problem (1) given by

$$\min_{x \in S} E[G(x)], \quad (6)$$

i.e.,  $f(x)$  in (1) is now the well-defined expected value of a stochastic performance measure  $G(x)$ . We assume that analytical computation of the expected value is hard, forcing us to estimate  $f(x)$  subject to random errors. We use  $\tilde{f}$  to denote function estimates. In particular, if the acceptance probability filter in Algorithm 1.1 used function values for solving problem (1), it must now use estimates  $\tilde{f}$  instead while solving the noisy problem (6). We use  $\epsilon_{k+1}(x)$  to denote the estimation error at point  $x \in S$  in iteration  $k+1$ , i.e.,  $\epsilon_{k+1}(x) = \tilde{f}_{k+1}(x) - f(x)$ .

**Condition 4.1.**  $|\epsilon_{k+1}(X_k)| \rightarrow 0$  in probability and  $|\epsilon_{k+1}(Y_{k+1})| \rightarrow 0$  in probability, i.e., for any  $\epsilon > 0$  and  $\delta > 0$ , there exists an integer  $N$  such that  $P[|\epsilon_{k+1}(X_k)| > \epsilon] < \delta$  for all  $k \geq N$ , and similarly for  $Y_{k+1}$ .

Note that when the estimates  $\tilde{f}$  are unbiased, it suffices for the sequence of error variances  $\sigma_{k+1}^2(X_k)$  and  $\sigma_{k+1}^2(Y_{k+1})$  to converge to zero as  $k \rightarrow \infty$  in order for Condition 4.1 to be satisfied owing to Chebyshev's inequality.

**Corollary 4.2.** *Suppose Conditions 2.1-2.5 and 4.1 hold. Then for every initial state  $X_0$ , the sequence of objective function value estimates  $(\tilde{f}_k(X_k), k = 0, 1, 2, \dots)$  converges in probability to  $f_*$ .*

*Proof.* Based on Theorem 2.6 and the fact that  $\tilde{f}_k(X_k) \rightarrow f(X_k)$  in probability.  $\square$

Corollary 4.2 can be used to prove convergence of several simulation-optimization algorithms. As an example, we consider SA with an adaptive temperature schedule  $T_k$  for problems with noisy function evaluations. Intuitively, the following condition means that the errors converge to zero faster than the temperature.

**Condition 4.3.** *The sequence of ratios  $\{\frac{|\epsilon_{k+1}(X_k)|}{T_k}\}$  converges to 0 in probability. Similarly, sequence of ratios  $\{\frac{|\epsilon_{k+1}(Y_{k+1})|}{T_k}\}$  converges to 0 in probability.*

The next condition is typical when the temperature schedule is adaptive [6].

**Condition 4.4.** *The sequence of temperatures  $\{T_k\}$  converges to 0 in probability.*

**Corollary 4.5.** *Suppose Conditions 2.1-2.3 hold for problem (6) to which we apply SA such that Conditions 4.3-4.4 are satisfied. Then the sequence of objective function value estimates  $(\tilde{f}_k(X_k), k = 0, 1, 2, \dots)$  converges in probability to  $f_*$ .*

*Proof.* It suffices to show that Conditions 2.4, 2.5 and 4.1 hold. Conditions 4.3 and 4.4 clearly imply Condition 4.1. For brevity, let  $\Delta_k$  denote  $f(Y_{k+1}) - f(X_k)$  and  $\Xi_k$  denote  $\epsilon_{k+1}(Y_{k+1}) - \epsilon_{k+1}(X_k)$ . Then note that for SA with noisy objective function evaluations, the acceptance probability  $\mathcal{P}_k$  is given by

$$\exp\left\{-\frac{[\Delta_k + \Xi_k]^+}{T_k}\right\}.$$

Proof for Condition 2.4.

Let  $\delta > 0$  and  $\epsilon > 0$  be given. First notice that for any three real numbers  $a, b$ , and  $c \geq a$ ,  $[a + b]^+ \leq [c + b]^+$ . Hence,  $\exp(-[a + b]^+) \geq \exp(-[c + b]^+)$ . We will use this with  $a = \frac{\beta}{T_k}$ ,  $b = \frac{\Xi_k}{T_k}$ , and  $c = \frac{\Delta_k}{T_k}$ . Observe that the event  $\mathcal{U}_k(\beta)$  implies that  $\Delta_k \geq \beta$ . Hence, it implies that  $\mathcal{P}_k \leq \exp(-\frac{[\beta + \Xi_k]^+}{T_k})$ . Thus,  $\mathcal{P}_k > \epsilon$  implies  $\exp(-\frac{[\beta + \Xi_k]^+}{T_k}) > \epsilon$ . Therefore, we can write

$$\begin{aligned} P[\mathcal{P}_k > \epsilon, \mathcal{U}_k(\beta)] &\leq P\left[\exp\left(-\frac{[\beta + \Xi_k]^+}{T_k}\right) > \epsilon\right] \\ &\leq P\left[\frac{\beta}{2T_k} < \log \frac{1}{\epsilon}, E_k(\beta/2)\right] + P[E_k^c(\beta/2)] \end{aligned}$$

where  $E_k(\beta/2)$  is the event that  $|\Xi_k| < \beta/2$ . The second term in the inequality can be made smaller than  $\delta/2$  for all  $k \geq n_0$  due to Condition 4.1. Therefore, the above inequality reduces to

$$\begin{aligned} P[\mathcal{P}_k > \epsilon, \mathcal{U}_k(\beta)] &< P\left[\frac{\beta}{2T_k} < \log \frac{1}{\epsilon}\right] + \frac{\delta}{2} \\ &= P\left[\frac{\beta}{2 \log \frac{1}{\epsilon}} < T_k\right] + \frac{\delta}{2} \quad \forall k \geq n_0. \end{aligned}$$

The first term can be made smaller than  $\delta/2$  for all  $k \geq n_1$  due to Condition 4.4. Therefore, we have  $P[\mathcal{P}_k > \epsilon, \mathcal{U}_k(\beta)] < \delta$  for all  $k \geq \max\{n_0, n_1\}$ .

Proof for Condition 2.5.

Let  $\delta > 0$  and  $\epsilon > 0$  be given. First notice that for any two real numbers  $a$  and  $b$ ,  $[a + b]^+ \leq [a]^+ + [b]^+$ . Therefore,  $\exp(-[a + b]^+) \geq \exp(-([a]^+ + [b]^+))$ . Using this with  $a = \frac{\Delta_k}{T_k}$  and  $b = \frac{\Xi_k}{T_k}$ , we get  $\mathcal{P}_k \geq \exp(-\frac{[\Delta_k]^+}{T_k}) \exp(-\frac{[\Xi_k]^+}{T_k})$ . Thus, the event  $1 - \mathcal{P}_k > \epsilon$  implies the event  $1 - \exp(-\frac{[\Delta_k]^+}{T_k}) \exp(-\frac{[\Xi_k]^+}{T_k}) > \epsilon$ . Observe that  $\mathcal{V}_k$  implies  $[\Delta_k]^+ = 0$ . Therefore, the probability  $P[1 - \mathcal{P}_k > \epsilon, \mathcal{V}_k]$  is bounded above by

$$\begin{aligned} P\left[1 - \exp\left(-\frac{[\Xi_k]^+}{T_k}\right) > \epsilon\right] &\leq P\left[\log \frac{1}{1 - \epsilon} < \frac{[\Xi_k]^+}{T_k}\right] \\ &< \delta \quad \forall k \geq n_0 \text{ by Condition 4.3.} \end{aligned}$$

This completes the proof.  $\square$

As an example, suppose  $T_k$  is a deterministic sequence of temperatures that converges to 0 as  $k \rightarrow \infty$  and estimates  $\tilde{f}$  are unbiased. Then

Chebyshev’s inequality implies that Condition 4.3 is satisfied when the ratios  $\sigma_{k+1}^2(X_k)/T_k^2$  and  $\sigma_{k+1}^2(Y_{k+1})/T_k^2$  converge to zero as  $k \rightarrow \infty$ .

### 5. Numerical Experiments: Protein Folding with Noise

A challenging test problem commonly used for testing global optimization software is the following energy minimization problem from the protein folding literature [14,15]. Let  $x_1, x_2, \dots, x_n$  be the 3-dimensional coordinates of  $n$  molecules. Our goal is to place these  $n$  molecules in the box  $[-1, 1]^3$  in  $R^3$  such that the potential energy of interaction  $\psi(\cdot)$  between these molecules is minimized. This can be formulated as follows.

$$\min_{x_1, x_2, \dots, x_n \in R^3} \psi(x_1, \dots, x_n) := \sum_{i < j}^n r(\|x_i - x_j\|_2)$$

$$x_i \in [-1, 1]^3, \quad i = 1, 2, \dots, n,$$

where  $\|\cdot\|_2$  denotes the Euclidean distance, and the function  $r(s) := s^{-12} - 2s^{-6}$  is the Lennard-Jones potential between two molecules. The number of local minima in the above objective function is exponential in  $n$ .

We introduce the additional complication whence the potential energy of interaction among  $n$  molecules at  $x = (x_1, x_2, \dots, x_n)$  is computed by a ‘stochastic black-box’, which receives  $x$  as input and produces a random number as output with distribution  $\Theta(x)$  whose expectation is  $\psi(x)$  and variance is  $\sigma^2(x)$ . Since our goal is to minimize the expected value function  $\psi(\cdot)$ , we use estimates  $\tilde{\psi}_k(x)$  of the expected value  $\psi(x)$  in iteration  $k$  of Algorithm 1.1. For simplicity, we generate these estimates by repeatedly querying the black-box. In particular, let  $\Theta^1(x), \Theta^2(x), \dots, \Theta^{N_k}(x)$  be  $N_k$  independent and identically distributed outputs produced by the stochastic black-box on receiving  $x$  as input and  $\tilde{\psi}_k(x)$  be their sample average. Note that  $\tilde{\psi}_k(x)$  is unbiased and converges almost surely to  $\psi(x)$  as  $k \rightarrow \infty$  by the Strong Law of Large Numbers. Moreover, the variance of the estimation error is  $\sigma^2(x)/N_k$ .

We employed the Hit-and-Run candidate generator from [20] and SA with deterministic temperature schedules so that the ratio of variance to

temperature square converges to zero with iterations. Numerical results for the 3 molecule case, which is a 9-dimensional problem with optimum objective function value  $-3$  [14], are presented in Table 1. The first column lists an abbreviation for the distribution  $\Theta(x)$  whose expectation is our objective function. For each  $\Theta(x)$ , the temperature schedule  $T_k$  and sample size  $N_k$  were parameterized by real numbers  $\alpha$  and  $\beta$  as  $T_k = k^{-\alpha}$  and  $N_k = k^\beta$ . It suffices to choose any  $\alpha > 0$  to satisfy Condition 4.4, and any  $\beta > 2\alpha$  so that the variance of our objective function estimate decreases faster than the square of the temperature. We used  $0.5 < \alpha < 1$  so that the corresponding  $\beta$  value is between 1 and 2 ensuring a moderate increase in sample size with iterations. For every combination of  $\Theta(x)$ ,  $T_k$ ,  $N_k$ , we performed 30 independent runs of 1000 iterations each with an independently generated random starting point in  $[-1, 1]^9$ . The fourth column lists the mean of the estimated objective function value reached at termination with standard deviation in the fifth column. Figure 1 shows progress of the mean of the estimated objective function value for two such combinations. *The quality of these results is evident since the Adaptive SA algorithm in [14] is not able to decrease the objective function value below -2.5 in 12000 iterations on the noise free version of our example.*

Table 1

Noisy protein folding results. N1 is Normal( $\psi(x), 0.1$ ), N2 is Normal( $\psi(x), 1$ ), U1 is Uniform( $\psi(x) - 1, \psi(x) + 1$ ), U2 is Uniform( $\psi(x) - 0.1, \psi(x) + 0.1$ ).

$\Theta(x)$	$\alpha$	$\beta$	value	std.
N1	0.80	1.65	-2.9944	0.0046
	0.70	1.50	-2.9869	0.0103
N2	0.80	1.70	-3.0068	0.1187
	0.65	1.35	-2.9866	0.0860
U1	0.60	1.30	-2.9802	0.0207
	0.75	1.70	-2.9896	0.0087
U2	0.75	1.60	-2.9874	0.0086
	0.85	1.80	-2.9951	0.0006

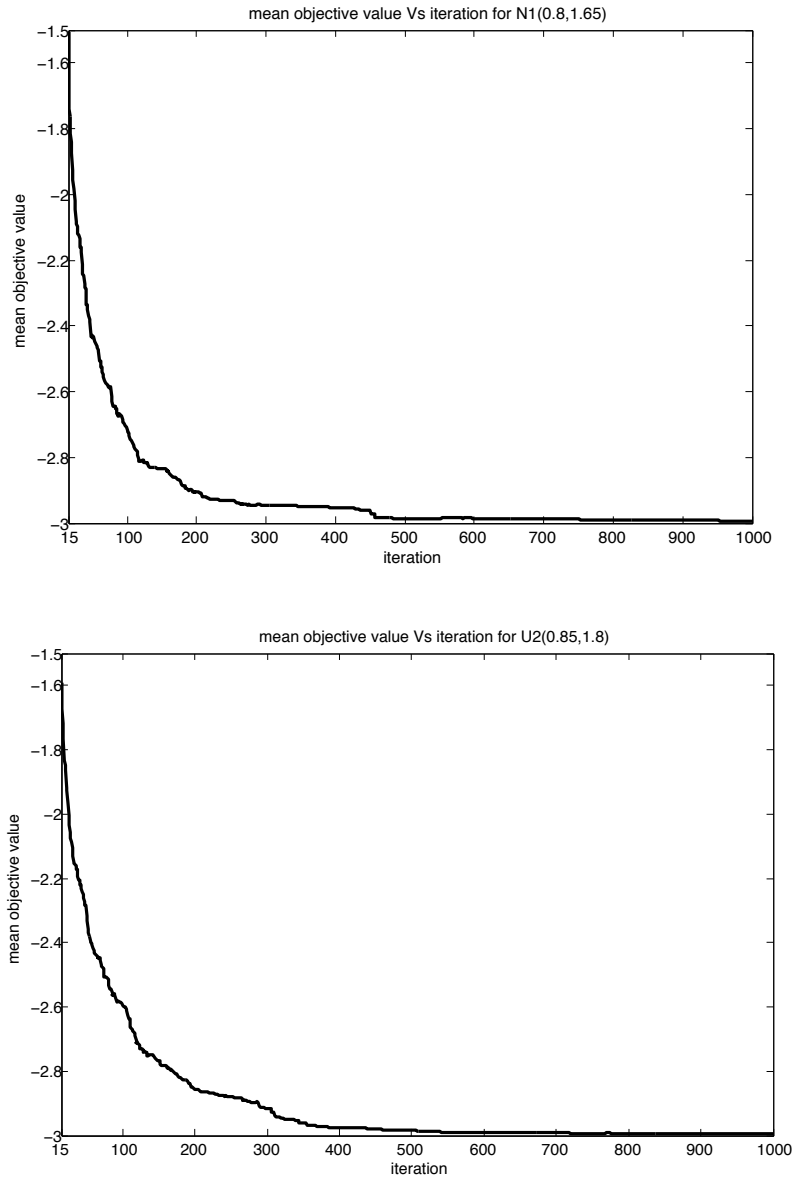


Figure 1. Evolution of the mean of the estimated objective function value with iterations for two example rows in Table 1. The first fourteen iterations are not shown as their objective function values were too big to fit in the plot. Notation such as  $N1(0.8, 1.65)$  in the title of a plot indicates that the distribution  $\Theta(x)$  is  $N1$  given in Table 1,  $\alpha = 0.8$  and  $\beta = 1.65$  respectively.

### Acknowledgements

Research supported in part by the NSF under grant DMI-0244291. We thank the referees and the Area Editor for their suggestions.

### REFERENCES

1. Alexander, D. L. J., Bulger, D. W., Calvin, J. M., Romeijn, H. E., and Sherriff, R. L. Approximate implementations of pure random search in the presence of noise, *Journal of Global Optimization*, 31 (2005) 601-612 .
2. Andradottir, S., Simulation Optimization: Integrating research and practice, *INFORMS Journal on Computing*, 14 (2002) 216-219.
3. Alrefaei, M. H., and Andradottir, S., A simulated annealing algorithm with constant temperature for discrete stochastic optimization, *Management Science*, 45 (1999) 748-764.
4. Barker, A. A., Monte Carlo calculations of the radial distribution functions for a proton-electron plasma, *Australian Journal of Physics*, 18 (1965) 119-133.
5. Baumert, S., and Smith, R. L., Pure random search for noisy objective functions, Technical Report TR01-03, Industrial and Operations Engineering, University of Michigan, Ann Arbor, (2001).
6. Belisle, C. J. P., Convergence theorems for a class of simulated annealing algorithms on  $R^d$ , *Journal of Applied Probability*, 29 (1992) 885 - 895.
7. Fu, M. C., Optimization for simulation: theory vs. practice, *INFORMS Journal on Computing*, 14 (2002) 192-215.
8. Hajek, B., Cooling schedules for optimal annealing, *Mathematics of Operations Research*, 13 (1988) 311-329.
9. Hanning, J., Chong, E. K. P., and Kulkarni, S., Relative frequencies of generalized simulated annealing, *Mathematics of Operations Research*, 31 (2006) 199-216.
10. Johnson, A. W., and Jacobson, S. H., A class of convergent generalized hill climbing algorithms, *Applied Mathematics and Computation*, 125 (2002) 359-373.
11. Kirkpatrick, S., Gelatt, C. D. Jr., Vecchi, M. P., Optimization by simulated annealing, *Science*, 220 (1983) 671 - 680.
12. Kalai A., and Vempala, S., Convex optimization by simulated annealing, *Mathematics of Operations Research*, 31 (2006), 253-266.
13. Locatelli, M., Simulated annealing algorithms for continuous global optimization: convergence conditions, *Journal of Optimization Theory and Applications*, 104 (2000), 121-133.
14. Mongeau, M., Karsenty, H., Rouze, V., and Hiriart-Urruty, J.-B., A comparison of public-domain software for black-box global optimization, <http://mip.ups-tlse.fr/publi/rapp99/99.50.html>
15. Neumaier, A., Molecular modeling of proteins and mathematical prediction of protein structure, *SIAM Review*, 39 (1997), 407-460.
16. Prudius, A. A., and Andradottir, S., Two Simulated Annealing algorithms for noisy objective functions, *Proceedings of the 2005 Winter Simulation Conference*, (2005) 797-801.
17. Romeijn, H. E., and Smith, R. L., Simulated annealing for constrained global optimization, *Journal of Global Optimization*, 5 (1994) 101 - 126.
18. Romeijn, H. E., Zabinsky, Z. B., Graesser, D. L., and Neogi, S., New reflection generator for simulated annealing in mixed-integer/continuous global optimization, *Journal of Optimization Theory and Applications*, 101 (1999) 403-427.
19. Spall, J. C., *Introduction to stochastic search and optimization: estimation, simulation and control*, Wiley, Hoboken, New Jersey, (2003).
20. Smith, R. L., Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions, *Operations Research*, 32 (1984) 1296-1308.
21. Tsallis, C., Possible generalization of Boltzmann-Gibbs statistics, *Journal of Statistical Physics*, 52 (1988) 479-487.
22. Zabinsky, Z. B., Smith, R. L., McDonald, J. F., Romeijn, H. E., and Kaufman, D. E., Improving Hit-and-Run for global optimization, *Journal of Global Optimization*, 3 (1993) 171-192.