Sequential Monte Carlo Simulated Annealing

Enlu Zhou
Xi Chen

Department of Industrial & Enterprise Systems Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, U.S.A.

ABSTRACT

In this paper, we propose a population-based optimization algorithm, Sequential Monte Carlo Simulated Annealing (SMC-SA), for continuous global optimization. SMC-SA incorporates the sequential Monte Carlo method to track the converging sequence of Boltzmann distributions in simulated annealing. We prove an upper bound on the difference between the empirical distribution yielded by SMC-SA and the Boltzmann distribution, which gives guidance on the choice of the temperature cooling schedule and the number of samples used at each iteration. We also prove that SMC-SA is more preferable than the multi-start simulated annealing method when the sample size is sufficiently large.

I. INTRODUCTION

Simulated annealing (SA) is an attractive algorithm for optimization, due to its theoretical guarantee of convergence, good performance on many practical problems, and ease of implementation. It was first proposed in [16] by drawing an analogy between optimization and the physical process of annealing. The early study of simulated annealing focused on combinatorial optimization, and some fundamental theoretical work include [10], [11], [1], and [14]. Later, simulated annealing was extended to continuous global optimization and rigorous convergence results were proved under various conditions, such as [7], [2], [31], [19], [20], and [36]. Meanwhile, connections were exploited between simulated annealing and some other optimization algorithms, and many variations of simulated annealing were developed. The book [35] has a complete summary on simulated annealing for combinatorial optimization, and a recent survey paper [15] provides a good overview of the theoretical development of simulated annealing in both combinatorial and continuous optimization. The standard simulated annealing generates one candidate solution at each iteration, and the sequence of candidate solutions converge asymptotically to the optima in probability. To speed up the convergence, many

variations such as [33], [21], [4], [27], [34], [23], and [24], extend simulated annealing to population-based algorithms where a number of candidate solutions are generated at each iteration.

In this paper, we introduce a new population-based simulated annealing algorithm, Sequential Monte Carlo Simulated Annealing (SMC-SA), for continuous global optimization. It is well known that the Boltzmann distribution converges weakly to the uniform distribution concentrated on the set of global optima as the temperature decreases to zero [31]. Therefore, the motivation is to "track" closely this converging sequence of Boltzmann distributions. At each iteration, the standard simulated annealing essentially simulates a Markov chain whose stationary distribution is the Boltzmann distribution of the current temperature, and the current state becomes the initial state for a new chain at the next iteration. Hence, the temperature has to decrease slowly enough such that the chain does not vary too much from iteration to iteration, which ensures the overall convergence of simulated annealing. Motivated by this observation, our main idea is to provide a better initial state for the subsequent chain using a number of samples by drawing upon the principle of importance sampling. The resultant algorithm can be viewed as a sequential Monte Carlo method [8] used in tracking the sequence of Boltzmann distributions, and that is why the algorithm is named as SMC-SA. Sequential Monte Carlo (SMC) includes a broad class of statistical Monte Carlo methods engineered to track a sequence of distributions with minimum error in certain sense [9][3] .

Compared with the aforementioned population-based simulated annealing algorithms, SMC-SA differs in two main aspects: (i) SMC-SA has theoretical convergence results, which are lacking in most of them; (ii) The motivation of SMC-SA is to "track" the sequence of Boltzmann distributions as closely as possible. SMC-SA bears some similarity with the multi-particle version of simulated annealing, introduced in [23] and [24], which consists of N-particle exploration and N-particle selection steps with a meta-control of the temperature. The exploration step in their method can be viewed as a variation of the resampling step in SMC-SA, and the selection step is essentially the SA move step in SMC-SA. However, SMC-SA has an importance updating step which plays an important role, making it very different from the multi-particle version of simulated annealing. Although starting from a completely different motivation, the algorithm of SMC-SA falls into the broad framework under the name of "generation methods" (c.f. Algorithm 3.8 in [39], Chapter 5 in [38]). The convergence analysis of SMC-SA bears some similarity with that of the generation methods, but SMC-SA has its unique convergence properties due to its special structure.

The combination of the resampling and SA move steps in SMC-SA is also similar to that in the resample-move particle filter introduced in [12], which is developed for filtering (i.e. sequential state estimation). In the SMC community, [25] studied the annealed properties of Feynman-Kac-Metropolis

model, which can be interpreted as an *infinite-population* nonlinear simulated annealing random search and is only theoretical. [26] proposed an SMC sampler, and mentioned it can be used for global optimization but without further development. On an abstract level, SMC-SA can be viewed as an application of the SMC sampler with the target distributions being the Boltzmann distributions, in the same spirit as that the standard SA can be viewed as an application of the Metropolis algorithm with the Boltzmann distributions as target distributions as well.

As a benchmark, we compare SMC-SA to the multi-start simulated annealing method both analytically and numerically. Multi-start SA is probably the most naive population-based simulated annealing algorithm. It runs multiple simulated annealing algorithms independently with initial points drawn uniformly from the solution space. We find that SMC-SA is more preferable than multi-start SA when the sample size is sufficiently large (but the same for both algorithms). That can be roughly explained as a result of the interaction among the samples in SMC-SA as opposed to the independence between the samples in multi-start SA. To summarize, the main contribution of the paper includes

- A well-motivated global optimization algorithm SMC-SA with convergence results;
- Analytical and numerical comparison between SMC-SA and multi-start simulated annealing, which gives an indication for the general comparison between interactive and independent population-based algorithms.

The rest of the paper is organized as follows: Section II revisits simulated annealing and motivates the development of SMC-SA; Section III introduces SMC-SA with explanations of the rationale behind it; Section IV provides rigorous analysis on the convergence of SMC-SA and multi-start SA, and also a direct comparison of SMC-SA and multi-start SA; Section V presents the numerical results of SMC-SA compared with the standard SA, multi-start SA, and the cross-entropy method; Section VI concludes the paper.

II. Revisiting Simulated Annealing

We consider the maximization problem

$$\max_{x\in\mathcal{X}} H(x), \tag{1}$$

where the solution space $\mathcal{X}$ is a nonempty compact set in $\mathbb{R}^n$, and $H : \mathcal{X} \to \mathbb{R}$ is a continuous real-valued function. Under the above assumption, H is bounded on $\mathcal{X}$, i.e., $\exists H_l > -\infty,\ H_u < \infty$ s.t. $H_l \leq H(x) \leq H_u$, $\forall x \in \mathcal{X}$. We denote the optimal function value as H^*, i.e., there exists an x^* such that $H(x) \leq H^* \triangleq H(x^*)$, $\forall x \in \mathcal{X}$.

For the above maximization problem (1), the most common simulated annealing algorithm is as follows.

Algorithm 1: Standard Simulated Annealing

At the k^{th} iteration,

- Generate y_k from a symmetric proposal distribution with density $g_k(y|x_k)$.
- Calculate acceptance probability

$$\rho_k = \min\left\{\exp\left(\frac{H(y_k) - H(x_k)}{T_k}\right), 1\right\}.$$

- Accept/Reject

$$x_{k+1} = \begin{cases} y_k, & \text{w.p. } \rho_k; \\ x_k, & \text{w.p. } 1-\rho_k. \end{cases}$$

- Stopping: if a stopping criterion is satisfied, return x_{k+1}; otherwise, $k := k+1$ and continue.

The k^{th} iteration of the above Algorithm 1 is essentially one iteration of the Metropolis algorithm for drawing samples from the target distribution with density proportional to $\exp\{H(x)/T_k\}$. The Metropolis algorithm is one among the class of Markov Chain Monte Carlo methods [22][29], which draw samples by simulating an ergodic Markov chain whose stationary distribution is the target distribution. Starting from any initial state, the ergodic chain will go to stationarity after an infinite number of transitions, and thus at that time the samples are distributed exactly according to the target distribution. If the initial state happens to be in stationarity, then the chain stays in stationarity and the following states (samples) are always distributed according to the stationary distribution (target distribution).

From the interpretation of the Metropolis algorithm, theoretically at each fixed temperature we have to simulate the chain for an infinite number of transitions before a sample is truly drawn from the Boltzmann distribution at this temperature. Once the stationarity of the chain is achieved, we decrease the temperature, and then again have to simulate the new chain for an infinite number of transitions to achieve the stationary distribution which is the Boltzmann distribution at the new temperature. This type of SA is conceptually simple and easier to analyze, but is clearly impractical. In practice, the most commonly used SA iteratively decreases the temperature and draws one sample, as shown in Algorithm 1. This is equivalent to simulating each Markov chain for only one transition, and hence, the chain almost never achieves stationarity before the temperature changes. Obviously there could be some algorithms in between these two extremes, such as iteratively decreasing the temperature and drawing a few finite number of samples, which is equivalent to simulating each Markov chain for a few number of transitions before switching to the next chain. The two extreme cases described above are summarized as follows:

- Infinite-Transition SA (ITSA): It can be viewed as a sequence of Markov chains. Each Markov chain is of infinite length, and converges to the Boltzmann distribution at the current temperature.

The temperature is decreased in between subsequent Markov chains.

- Single-Transition SA (STSA): It can be viewed as a sequence of Markov chains. Each Markov chain has only one transition. The temperature is decreased in between subsequent Markov chains.

ITSA and STSA can be also viewed as "homogeneous SA" and "inhomogeneous SA" respectively as mentioned in [35], since ITSA can be viewed as a sequence of homogeneous Markov chains, and STSA as one single inhomogeneous Markov chain of infinite length, where the temperature is decreased in between subsequent transitions. For the algorithm to converge to the global optima in probability, STSA requires the temperature to decrease slowly enough, whereas there is no such requirement on ITSA [35]. That can be intuitively explained as a result that the Markov chain corresponding to each temperature almost never achieves stationarity in STSA. If the temperature decreases slowly enough, then the subsequent Markov chains do not differ too much, such that when the current state becomes the initial state for the next Markov chain, it is not too far away from the stationary distribution. On the other hand, in continuous global optimization, there exists some seemingly surprising result saying that SA converges regardless of how fast the temperature decreases under certain conditions [2]. Please note that it does not mean that the actual algorithm can converge arbitrarily fast: although the sequence of Boltzmann distributions converge in a fast rate, the sequence of distributions of the candidate solutions may converge in a much slower rate due to that fact that the difference between the two sequences becomes larger as iterations continue.

In view of the respective advantages of ITSA and STSA, we ask the question: *Can we follow the stationary distribution of each subsequent chain as closely as possible in one step?* Our main idea is to provide an initial state that is closer to the stationary distribution of the subsequent chain by drawing upon the principle of importance sampling. The resultant algorithm can be viewed as a Sequential Monte Carlo method used in tracking the converging sequence of Boltzmann distributions.

III. Sequential Monte Carlo Simulated Annealing

In this section, we propose the sequential Monte Carlo simulated annealing (SMC-SA) algorithm. The idea is to incorporate sequential Monte Carlo method to track the sequence of Boltzmann distributions in simulated annealing. It has three main steps: importance updating, resampling, and SA move. The importance updating step updates the empirical distribution from the previous iteration to a new one that is close to the target distribution of this iteration. More specifically, it takes the current Boltzmann distribution π_k as the target distribution, and the previous Boltzmann distribution π_{k-1} as the proposal distribution. Thus, given the previous samples are already distributed approximately according to π_{k-1} and the weights of these samples are updated in proportion to $\pi_k(\cdot)/\pi_{k-1}(\cdot)$, the new empirical distribution formed by these weighted samples will closely follow π_k. The resampling

step redistributes the samples such that they all have equal weights. The SA move step performs one iteration of simulated annealing on each sample to generate a new sample or candidate solution. This step essentially takes the current empirical distribution as the initial distribution, and simulates one transition of the Markov chain whose stationary distribution is the current Boltzmann distribution π_k. Hence, the resultant empirical distribution will be brought even closer to π_k. The resampling step together with the SA move step prevents sample degeneracy, or in other words, keeps the sample diversity and thus the exploration of the solution space. We explain the main steps in more detail in the following.

A. Importance Updating

The importance updating step is based on importance sampling [29], which essentially performs a change of measure. Thus, the expectation under one distribution can be estimated using the samples drawn from another distribution with appropriate weighting. Specifically, let f and g denote two probability density functions. For any integrable function ϕ, its integration with respect to f equals to

$$I_\phi = \int \phi(x) f(x) dx = \int \phi(x) \frac{f(x)}{g(x)} g(x) dx. \tag{2}$$

If we draw independent and identically distributed (i.i.d.) samples $\{x^i\}_{i=1}^N$ from g and set their weights $\{w^i\}_{i=1}^N$ according to

$$W^i = \frac{f(x^i)}{g(x^i)}, \quad w^i = \frac{W^i}{\sum_{j=1}^N W^j},$$

then in view of (2), an estimate of I_ϕ is

$$\hat{I}_\phi = \frac{1}{N} \sum_{i=1}^N W^i \phi(x^i), \quad x^i \overset{\text{iid}}{\sim} g,$$

and an approximation of f is

$$\hat{f}(x) = \sum_{i=1}^N w^i \delta_{x^i}(x), \tag{3}$$

where δ denotes the Dirac delta function, which satisfies

$$\int \phi(x) \delta_y(x) dx = \phi(y).$$

In other words, $\{x^i, w^i\}_{i=1}^N$ is a weighted sample from f, and $\hat{f}$ defined in (3) is an empirical distribution of f.

In simulated annealing, suppose we already have i.i.d. samples $\{x_{k-1}^i\}_{i=1}^{N_{k-1}}$ from the previous Boltzmann distribution π_{k-1}, based on the importance sampling described above we can obtain

weighted samples $\{x_{k-1}^i, w_k^i\}_{i=1}^{N_{k-1}}$ that are distributed according to the current Boltzmann distribution π_k. More specifically, the Boltzmann distribution at time k has the density function

$$\pi_k^d(x) = \frac{1}{Z_k} \exp\left\{\frac{H(x)}{T_k}\right\},$$

where $Z_k = \int \exp\{H(x)/T_k\}dx$ is the normalization constant, $H(x)$ is the objective function in (1), and T_k is often referred to as the *temperature* at time k. Noticing that

$$\frac{\pi_k^d(x)}{\pi_{k-1}^d(x)} = \frac{\exp\left\{H(x)\left(\frac{1}{T_k} - \frac{1}{T_{k-1}}\right)\right\}}{Z_k/Z_{k-1}}, \quad k = 2, \ldots,$$

an approximation of π_k is

$$\hat{\pi}_k(x) = \sum_{i=1}^{N_{k-1}} w_k^i \delta_{x_{k-1}^i}(x), \quad x_{k-1}^i \overset{\text{iid}}{\sim} \pi_{k-1},$$

where

$$w_k^i \propto \exp\left\{H(x_{k-1}^i)\left(\frac{1}{T_k} - \frac{1}{T_{k-1}}\right)\right\}, \quad \sum_{i=1}^{N_{k-1}} w_k^i = 1, \quad k = 2, \ldots.$$

Assuming that we do not have any prior knowledge about the optimal solution(s), we draw the initial samples from a uniform distribution over the solution space, i.e.,

$$\pi_0^d(x) \propto 1, \quad \forall x \in \mathcal{X}.$$

Since

$$\frac{\pi_1^d(x)}{\pi_0^d(x)} \propto \exp\left\{\frac{H(x)}{T_1}\right\},$$

the weights at the 1^{st} iteration should satisfy

$$w_1^i \propto \exp\left\{\frac{H(x_0^i)}{T_1}\right\}, \quad \sum_{i=1}^{N_0} w_1^i = 1.$$

In the following, we refer to N_k as the *sample size*, i.e., the number of samples or candidate solutions generated at the k^{th} iteration. The choice of $\{N_k\}$ is related with the choice of $\{T_k\}$, and that will be discussed in Section IV on the convergence results.

B. Resampling

The importance updating step yields $\hat{\pi}_k = \sum_{i=1}^{N_{k-1}} w_k^i \delta_{x_{k-1}^i}$, an approximation of π_k. However, the weighted samples $\{x_{k-1}^i, w_k^i\}_{i=1}^{N_{k-1}}$ will suffer from the problem of degeneracy. That is, after a few iterations, only few samples have dominating weights and most others have weights close to zero. These negligible samples waste future computation effort, since they do not contribute much to the updating of the empirical distribution. Therefore, the resampling step is needed to sample from the weighted samples $\{x_{k-1}^i, w_k^i\}_{i=1}^{N_{k-1}}$ in order to generate N_k i.i.d. new samples $\{\tilde{x}_k^i\}_{i=1}^{N_k}$, which are still approximately distributed according to π_k. In SMC-SA, we use sampling with replacement scheme for

the resampling step. There are several other resampling schemes mainly for the purpose of variance reduction, such as stratified resampling, residual resampling [18], and multinomial resampling [13], and their effects on the algorithm performance will be studied in the future.

The purpose of resampling can be explained from different perspectives. From sampling perspective, the resampling step together with the SA move step help to overcome sample degeneracy. After resampling, samples with large weights would have multiple copies, and these identical copies lead to different samples because of the SA move step next. Hence, resampling keeps the diversity of samples and ensure that every sample is useful. From the optimization perspective, resampling brings more exploration to the neighborhood of good solutions. It is similar to the selection step in genetic algorithms, where the elite parents would have more offsprings.

C. SA Move

At iteration k, the SA move step is in fact one step of the Metropolis algorithm with the target distribution being the Boltzmann distribution π_k. As $\{\tilde{x}_k^i\}_{i=1}^{N_k}$ are the initial states of the Markov chain and are distributed "closely" according to π_k, new samples $\{x_k^i\}_{i=1}^{N_k}$ generated from $\{\tilde{x}_k^i\}_{i=1}^{N_k}$ by the SA move step are even "closer" to the stationary distribution π_k. The SA move step is essentially the same as the SA algorithm, and is described below for clarity of notations.

Algorithm 2: SA Move at iteration k in SMC-SA

- Choose a symmetric proposal distribution with density $g_k(\cdot|x)$, such as a normal distribution with mean x.
- Generate $y_k^i \sim g_k(y|\tilde{x}_k^i), i = 1, \ldots, N_k$.
- Calculate acceptance probability

$$\rho_k^i = \min\left\{\exp\left(\frac{H(y_k^i) - H(\tilde{x}_k^i)}{T_k}\right), 1\right\}.$$

- Accept/Reject

$$x_k^i = \begin{cases} y_k^i, & \text{w.p. } \rho_k^i; \\ \tilde{x}_k^i, & \text{w.p. } 1 - \rho_k^i. \end{cases}$$

In summary, for the maximization problem (1), our proposed algorithm is as follows.

Algorithm 3: Sequential Monte Carlo Simulated Annealing (SMC-SA)

- Input: sample sizes $\{N_k\}$, cooling schedule for $\{T_k\}$.
- Initialization: generate $x_0^i \overset{\text{iid}}{\sim} \text{Unif}(\mathcal{X})$, $i = 1, 2, \ldots, N_0$. Set $k = 1$.
- At iteration k:
  - Importance Updating: compute normalized weights according to $w_k^i \propto \exp\left\{\frac{H(x_0^i)}{T_1}\right\}$ if $k = 1$, and $w_k^i \propto \exp\left\{H(x_{k-1}^i)\left(\frac{1}{T_k} - \frac{1}{T_{k-1}}\right)\right\}$ if $k > 1$.

– Resampling: draw i.i.d. samples $\{\tilde{x}_k^i\}_{i=1}^{N_k}$ from $\{x_{k-1}^i, w_k^i\}_{i=1}^{N_{k-1}}$.
– SA Move: generate x_k^i from $\tilde{x}_k^i$ for each i, $i = 1, \ldots, N_k$, according to Algorithm 2.
– Stopping: if a stopping criterion is satisfied, return $\max_i H(x_k^i)$; otherwise, $k := k + 1$ and continue.

IV. Convergence Analysis

A. Error Bounds of SMC-SA

It has been shown in [31] that under our assumptions on $\mathcal{X}$ and H, the Boltzmann distribution converges weakly to the uniform distribution on the set of optimal solutions as the temperature decreases to zero. In particular, if there is only one unique optimum solution, it converges weakly to a degenerate distribution concentrated on that solution. It has been stated formally as follows.

Proposition 1 (Proposition 3.1 in [31])**:** For all $\xi > 0$,

$$\lim_{T_k \to 0} \pi_k(\mathcal{X}_\xi) = 1,$$

where $\mathcal{X}_\xi = \{x \in \mathcal{X} : H(x) > H^* - \xi\}$.

Therefore, it is sufficient for us to show that SMC-SA tracks the sequence of Boltzmann distributions such that SMC-SA also converges to the optimal solution(s). More specifically, denoting the distribution yielded by SMC-SA as μ_k, we want to find out how the "difference" between μ_k and π_k evolves over time as a function of the temperature sequence $\{T_j\}_{j=1}^k$ and the sample size sequence $\{N_j\}_{j=0}^k$. To proceed to our formal analysis, we introduce the following notations and definitions. Let $\mathcal{F}$ denote the σ-field on $\mathcal{X}$, $\mathcal{B}(\mathcal{X})$ denote the set of measurable and bounded functions $\phi : \mathcal{X} \to \mathbb{R}$, and $\mathcal{B}^+(\mathcal{X})$ denote the set of measurable and bounded functions $\phi : \mathcal{X} \to \mathbb{R}^+$. We use $\mathcal{F}_k = \sigma\left(\{x_0^i\}_{i=1}^{N_0}, \{\tilde{x}_1^i, x_1^i\}_{i=1}^{N_1}, \ldots, \{\tilde{x}_k^i, x_k^i\}_{i=1}^{N_k}\right)$ to denote the sequence of increasing sigma-fields generated by all the samples up to the k^{th} iteration. For a measure ν defined on $\mathcal{F}$, we often use the following representation:

$$\langle \nu, \phi \rangle = \int \phi(x)\nu(dx), \quad \forall \phi \in \mathcal{B}(\mathcal{X}).$$

Definition 1: For any $\phi \in \mathcal{B}(\mathcal{X})$, its *supremum norm* is defined as

$$\|\phi\| = \sup_{x \in \mathcal{X}} |\phi(x)|.$$

Definition 2: Consider two probability measures ν_1 and ν_2 on a measurable space $(\mathcal{X}, \mathcal{F})$, then the *total variation distance* between ν_1 and ν_2 is defined as

$$\|\nu_1 - \nu_2\|_{TV} = \sup_{A \in \mathcal{F}} \|\nu_1(A) - \nu_2(A)\|.$$

We summarize the notations for all the probability distributions involved at the k^{th} iteration of SMC-SA as follows:

$$\begin{aligned}
\pi_k^d &= \frac{\exp(H(x)/T_k)}{\int \exp(H(x)/T_k)dx} \\
\tilde{\mu}_k &= \sum_{i=1}^{N_{k-1}} w_k^i \delta_{x_{k-1}^i} \\
\tilde{\mu}_k^{N_k} &= \frac{1}{N_k}\sum_{i=1}^{N_k} \delta_{\tilde{x}_k^i} \\
\mu_k &= \frac{1}{N_k}\sum_{i=1}^{N_k} \delta_{x_k^i},
\end{aligned}$$

where δ_x denotes the Dirac mass in x; π_k^d is the density function of the Boltzmann distribution π_k; $\tilde{\mu}_k$ is the distribution after the importance sampling step; $\tilde{\mu}_k^{N_k}$ is the distribution after the resampling step, i.e., an empirical distribution of $\tilde{\mu}_k$ with N_k i.i.d. samples; μ_k is the distribution after the SA move step, i.e., the output distribution of SMC-SA at iteration k. Using the above notations and denoting

$$\Psi_k \triangleq \frac{\pi_k^d}{\pi_{k-1}^d},$$

the relationship between the distributions according to the timeline of SMC-SA can be shown as:

$$\mu_{k-1} \quad \longrightarrow \quad \tilde{\mu}_k = \tfrac{\mu_{k-1}\Psi_k}{\langle \mu_{k-1}, \Psi_k \rangle} \quad \longrightarrow \quad \tilde{\mu}_k^{N_k} \quad \longrightarrow \quad \mu_k = \tilde{\mu}_k^{N_k} P_k$$

importance updating — resampling — SA Move

Here P_k denotes the transition kernel of the Markov chain associated with the SA move step, and it can be written as

$$P_k(x, dy) = \min\left\{1, \frac{\pi_k^d(y)}{\pi_k^d(x)}\right\} g_k(y|x)dy + (1 - r(x))\delta_x(dy), \tag{4}$$

where $r(x) = \int_{\mathcal{X}} \min\left\{1, \pi_k^d(y)/\pi_k^d(x)\right\} g_k(y|x)dy$.

The idea of the analysis is intuitive: at iteration k, the target distribution is π_k, which is the stationary distribution of the Markov chain associated with the SA move step; the importance sampling step brings the initial distribution of the chain close to π_k but not exactly π_k due to the sampling error. The hope is that the SA move step, corresponding to one transition of the chain, will bring the distribution even closer to π_k and thus combat the approximation error introduced by sampling. That can be achieved if the chain satisfies some ergodicity property that depends on the following assumption.

Assumption 1: The proposal density in the SA move step satisfies $g_k(y|x) \geq \varepsilon_k > 0$, $\forall x, y \in \mathcal{X}$.

Assumption 1 ensures that it is possible for the SA move step to visit any subset that has a positive Lebesgue measure in the solution space. Since the objective function H is continuous, the ξ-optimal

solution set $\{x \in \mathcal{X} : H(x) \geq H^* - \xi\}$ for any constant $\xi > 0$ has a positive Lebesgue measure, and thus always has a positive probability to be sampled from.

To consider the effect of the SA move step, we first show that the Markov chain associated with each SA move step is uniformly ergodic based on the following theorem.

Theorem 1 (Theorem 8 in [30])**:** Consider a Markov chain with transition kernel $P(x, dy)$ for $x, y \in \mathcal{X}$ and stationary probability distribution $\pi(\cdot)$. The entire space $\mathcal{X}$ is *small* if there exists a positive integer n_0, a constant $\epsilon \in (0, 1)$, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that the following *minorisation* condition holds:

$$P^{n_0}(x, A) \geq \epsilon \nu(A), \quad \forall x \in \mathcal{X}, \forall A \in \mathcal{F}. \tag{5}$$

Then the chain is *uniformly ergodic*, and in fact

$$\|P^n(x) - \pi\|_{TV} \leq (1 - \epsilon)^{\lfloor n/n_0 \rfloor}, \quad \forall x \in \mathcal{X},$$

where $\lfloor r \rfloor$ is the greatest integer not exceeding r.

Corollary 1.1: Under Assumption 1, the Markov chain corresponding to the SA move step at each iteration k is uniformly ergodic, and in particular, there exists $\epsilon_k \in (0, 1)$ such that

$$\|P_k^n(x) - \pi_k\|_{TV} \leq (1 - \epsilon_k)^n, \quad \epsilon_k = \varepsilon_k \exp\left\{\frac{H_l - H_u}{T_k}\right\} \lambda(\mathcal{X}), \tag{6}$$

where P_k is the transition kernel of the chain as defined in (4), π_k is the Boltzmann distribution at iteration k, and ε_k is the lower bound of the proposal distribution g_k as defined in Assumption 1, H_l and H_u are the lower and upper bounds of the objective function $H(x)$, and $\lambda(\mathcal{X})$ is the Lebesgue measure of $\mathcal{X}$.

Proof of Corollary 1.1: The SA move step is essentially one iteration of the Metropolis algorithm, which simulates one transition of a Markov chain with the stationary distribution π_k and the transition kernel $P_k(x, dy)$ as defined in (4). According to Assumption 1, the proposal density $g_k(y|x) \geq \varepsilon_k$, $\forall x, y \in \mathcal{X}$. Since $H_l \leq H(x) \leq H_u$, $\forall x \in \mathcal{X}$, we then have

$$P_k(x, dy) \geq \min\left\{1, \frac{\pi_k^d(y)}{\pi_k^d(x)}\right\} g_k(y|x) dy \geq \varepsilon_k \exp\left\{\frac{H_l - H_u}{T_k}\right\} dy,$$

which is a positive measure independent of x. Hence,

$$P_k(x, A) \geq \varepsilon_k \exp\left\{\frac{H_l - H_u}{T_k}\right\} \lambda(\mathcal{X}) \nu(A), \quad \forall x \in \mathcal{X}, \forall A \in \mathcal{F}.$$

where $\lambda(\cdot)$ is the Lebesgue measure, and $\nu(\cdot) \triangleq \lambda(\cdot)/\lambda(\mathcal{X})$ defines a probability measure on $\mathcal{X}$. This means that the minorisation condition (5) is satisfied with $n_0 = 1$, $\epsilon_k = \varepsilon_k \exp\left\{\frac{H_l - H_u}{T_k}\right\} \lambda(\mathcal{X})$, and the probability measure $\nu(\cdot)$. It can be easily verified that $\epsilon_k \in (0, 1)$. Therefore, (6) holds according to Theorem 1. ■

In words, the uniform ergodicity means that the distribution of the chain converges to the stationary distribution exponentially fast in a rate that is the same for every initial state $x \in \mathcal{X}$. The following corollary generalizes Theorem 1 to the case when the chain starts from an initial distribution.

Corollary 1.2: Consider a Markov chain with initial distribution μ, transition kernel P, and stationary probability distribution π. Suppose $|\langle \mu - \pi, \phi \rangle| \leq c\|\phi\|$ for all $\phi \in \mathcal{B}(\mathcal{X})$, where c is a positive constant. If the chain is uniformly ergodic with $\|P^n(x) - \pi\|_{TV} \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor}$ for all $x \in \mathcal{X}$, then

$$|\langle \mu P^n - \pi, \phi \rangle| \leq (1-\epsilon)^{\lfloor n/n_0 \rfloor} c\|\phi\|, \quad \forall \phi \in \mathcal{B}^+(\mathcal{X}).$$

Proof of Corollary 1.2: Since $\pi = \pi P^n$ and $\langle \mu - \pi, \pi\phi \rangle = 0$, we have

$$\begin{aligned} |\langle \mu P^n - \pi, \phi \rangle| &= |\langle \mu - \pi, P^n \phi \rangle| \\ &= |\langle \mu - \pi, (P^n - \pi)\phi \rangle| \\ &\leq c\|\Phi\|, \end{aligned}$$

where

$$\Phi(x) \triangleq \langle P^n(x) - \pi, \phi \rangle.$$

For all $\phi \in \mathcal{B}^+(\mathcal{X})$,

$$\begin{aligned} |\Phi(x)| &= \left| \int \phi(y) P^n(x, dy) - \int \phi(y)\pi(dy) \right| \\ &\leq \|\phi\| \|P^n(x) - \pi\|_{TV} \\ &\leq \|\phi\|(1-\epsilon)^{\lfloor n/n_0 \rfloor}. \end{aligned}$$

Since the above inequality holds for every $x \in \mathcal{X}$, we have

$$|\langle \mu P^n - \pi, \phi \rangle| \leq c\|\phi\|(1-\epsilon)^{\lfloor n/n_0 \rfloor}.$$

■

The following Lemma considers the approximation error introduced by resampling.

Lemma 1: Suppose conditionally with respect to $\mathcal{F}$, the random variables $(x^1, \ldots, x^N)$ are i.i.d. with the (conditional) probability distribution ν. Denoting $\nu^N \triangleq \frac{1}{N}\sum_{i=1}^N \delta_{x^i}$, it holds that

$$E\left[|\langle \nu - \nu^N, \phi \rangle| | \mathcal{F}\right] \leq \frac{\|\phi\|}{\sqrt{N}}, \quad \forall \phi \in \mathcal{B}(\mathcal{X}).$$

Proof of Lemma 1: For all $\phi \in \mathcal{B}(\mathcal{X})$, we have

$$\begin{aligned}
& E\left[\left|\langle \nu - \nu^N, \phi\rangle\right| |\mathcal{F}\right]^2 \\
\leq\ & E\left[\left|\int \phi d\nu - \frac{1}{N}\sum_{i=1}^N \phi(x^i)\right|^2 \Big|\mathcal{F}\right] \\
=\ & E\left[\left|\frac{1}{N}\sum_{i=1}^N \left(\int \phi d\nu - \phi(x^i)\right)\right|^2 \Big|\mathcal{F}\right] \\
=\ & \frac{1}{N^2}\sum_{i=1}^N E\left[\left(\int \phi d\nu - \phi(x^i)\right)^2 \Big|\mathcal{F}\right] \\
\leq\ & \frac{1}{N^2}\sum_{i=1}^N E\left[\phi(x^i)^2|\mathcal{F}\right] \leq \frac{\|\phi\|^2}{N}.
\end{aligned}$$

■

The following Lemma essentially considers the propagation of the distance between two distributions in importance updating.

Lemma 2: Suppose $|\langle \mu - \nu, \varphi\rangle| \leq c\|\varphi\|$ for all $\varphi \in \mathcal{B}(\mathcal{X})$, where c is a positive constant, and

$$\mu' = \frac{\mu\Psi}{\langle \mu, \Psi\rangle},$$

where $\Psi = \nu'_d/\nu_d$, ν'_d and ν_d are the densities of probability measures ν' and ν with respect to the Lesbesgue measure, respectively. Then

$$\left|\langle \mu' - \nu', \phi\rangle\right| \leq c\|\Psi\|\|\phi\|, \quad \forall \phi \in \mathcal{B}^+(\mathcal{X}).$$

Proof of Lemma 2: Since $\langle \nu, \Psi\phi\rangle = \langle \nu', \phi\rangle$ and $\langle \nu, \Psi\rangle = 1$, we can write

$$\begin{aligned}
|\langle \mu' - \nu', \phi\rangle| &= \left|\frac{\langle \mu, \Psi\phi\rangle}{\langle \mu, \Psi\rangle} - \frac{\langle \nu, \Psi\phi\rangle}{\langle \nu, \Psi\rangle}\right| \\
&= \left|\frac{\langle \mu, \Psi\phi\rangle}{\langle \mu, \Psi\rangle} - \frac{\langle \mu, \Psi\phi\rangle}{\langle \nu, \Psi\rangle} + \frac{\langle \mu, \Psi\phi\rangle}{\langle \nu, \Psi\rangle} - \frac{\langle \nu, \Psi\phi\rangle}{\langle \nu, \Psi\rangle}\right| \\
&= \left|\frac{\langle \mu, \Psi\phi\rangle\langle \nu - \mu, \Psi\rangle}{\langle \mu, \Psi\rangle} + \langle \mu - \nu, \Psi\phi\rangle\right| \\
&= \left|\left\langle \mu - \nu, \Psi\left(\phi - \frac{\langle \mu, \Psi\phi\rangle}{\langle \mu, \Psi\rangle}\right)\right\rangle\right| \\
&\leq c\left\|\Psi\left(\phi - E_{\mu'}[\phi]\right)\right\| \\
&\leq c\|\Psi\|\|\phi\|,
\end{aligned}$$

where the last inequality is because $\phi - E_{\mu'}[\phi] \in [-\|\phi\|, \|\phi\|]$ for all $\phi \in \mathcal{B}^+(\mathcal{X})$. ■

Finally, based on the above results, we show the evolution of the distance between μ_k and π_k over time in terms of the sample size and the temperature in the following theorem.

Theorem 2: Without loss of generality, we assume that $H(x) > 0$ for all $x \in \mathcal{X}$. Suppose the initial distribution is ν with the density ν^d with respect to the Lebesgue measure, then under Assumption 1,

$$E\left[|\langle \mu_k - \pi_k, \phi\rangle| \,|\, \mathcal{F}_{k-1}\right] \leq c_k \|\phi\|, \quad \forall \phi \in \mathcal{B}^+(\mathcal{X}),$$

where c_k satisfies the recursive equation

$$\begin{aligned} c_0 &= \frac{\|\pi_0^d/\nu^d\|^2}{N_0}, \\ c_k &= (1-\epsilon_k)\left(\frac{1}{\sqrt{N_k}} + \exp\left(H^*\Delta_k\right) c_{k-1}\right), \quad k = 1, 2, \ldots, \end{aligned} \tag{7}$$

where $\Delta_k = \left|\frac{1}{T_k} - \frac{1}{T_{k-1}}\right|$, ϵ_k is the constant defined in (6).

Proof of Theorem 2: First, consider the initialization step.

$$\langle \mu_0, \phi\rangle = \frac{\sum_{i=1}^{N_0} w(x_i)\phi(x_i)}{\sum_{i=1}^{N_0} w(x_i)}, \quad x_i \overset{\text{iid}}{\sim} \nu,$$

where $w(x_i) = \pi_0^d(x_i)/\nu^d(x_i)$. Using the Taylor expansion, it has been shown that

$$E\left[\langle \mu_0, \phi\rangle\right] = E_{\pi_0}[\phi] + \frac{E_{\pi_0}[\phi]\mathrm{Var}_\nu(w) - \mathrm{Cov}_\nu(w, w\phi)}{N_0} + O(N_0^{-2}).$$

Hence,

$$\begin{aligned} E\left[\langle \mu_0 - \pi_0, \phi\rangle\right] &= \frac{\left|\langle \pi_0, \phi\rangle\langle \nu, w^2\rangle - \langle \nu, w^2\phi\rangle\right|}{N_0} \\ &\leq \frac{\langle \nu, w^2(E_{\pi_0}[\phi] - \phi)\rangle}{N_0} \\ &\leq \frac{\|w\|^2\|\phi\|}{N} \\ &\triangleq c_0\|\phi\|, \end{aligned}$$

where the first equality is because $\langle \pi_0, \phi\rangle\langle \nu, w\rangle^2 - \langle \nu, w\rangle\langle \nu, w\phi\rangle = 0$ by plugging in $w = \pi_0^d/\nu^d$, and the last inequality is because $(\phi - E_{\pi_0}[\phi]) \in [-\|\phi\|, \|\phi\|]$ for all $\phi \in \mathcal{B}^+(\mathcal{X})$.

Next, we consider the k^{th} ($k \geq 1$) iteration. Recalling the timeline of SMC-SA, we have for all $\phi \in \mathcal{B}^+(\mathcal{X})$,

$$\begin{aligned} & E[|\langle \tilde{\mu}_k^{N_k} - \pi_k, \phi\rangle| \;|\mathcal{F}_{k-1}] \\ \leq\; & E[|\langle \tilde{\mu}_k^{N_k} - \tilde{\mu}_k, \phi\rangle| \;|\mathcal{F}_{k-1}] + E\left[|\langle \tilde{\mu}_k - \pi_k, \phi\rangle| \;|\mathcal{F}_{k-1}\right] \\ =\; & E[|\langle \tilde{\mu}_k^{N_k} - \tilde{\mu}_k, \phi\rangle| \;|\mathcal{F}_{k-1}] + E\left[\left|\frac{\langle \mu_{k-1}, \Psi_k\phi\rangle}{\langle \mu_{k-1}, \Psi_k\rangle} - \frac{\langle \pi_{k-1}, \Psi_k\phi\rangle}{\langle \pi_{k-1}, \Psi_k\rangle}\right| \;\middle|\mathcal{F}_{k-1}\right] \\ \leq\; & \left(\frac{1}{\sqrt{N_k}} + c_{k-1}\|\Psi_k\|\right)\|\phi\|, \end{aligned}$$

where the last inequality is a direct application of Lemma 1 and Lemma 2. Using Corollary 1.2 and Corollary 1.1, we further have

$$\begin{aligned} E\left[|\langle \mu_k - \pi_k, \phi\rangle| \,|\mathcal{F}_{k-1}\right] &= E[|\langle \tilde{\mu}_k^{N_k} P_k - \pi_k, \phi\rangle| \,|\mathcal{F}_{k-1}] \\ &\leq (1-\epsilon_k)\left(\frac{1}{\sqrt{N_k}} + c_{k-1}\|\Psi_k\|\right)\|\phi\|. \end{aligned}$$

To proceed, we need to find an upper bound on $\|\Psi_k\|$. Without loss of generality, we can assume $H(x) > 0$ for all $x \in \mathcal{X}$. Since $H(x)$ is lower bounded by H_l, we can always let $\tilde{H}(x) = H(x) + H_l > 0$ be the objective function, which has the same optimal solutions. Hence, we have

$$\begin{aligned} \left\|\pi_k^d / \pi_{k-1}^d\right\| &= \sup_{x\in\mathcal{X}} \left| \frac{\exp\left(\frac{H(x)}{T_k}\right) / \int_{\mathcal{X}} \exp\left(\frac{H(x)}{T_k}\right) dx}{\exp\left(\frac{H(x)}{T_{k-1}}\right) / \int_{\mathcal{X}} \exp\left(\frac{H(x)}{T_{k-1}}\right) dx} \right| \\ &\leq \exp\left\{ H^* \left| \frac{1}{T_k} - \frac{1}{T_{k-1}} \right| \right\}, \end{aligned}$$

Therefore,

$$\begin{aligned} E\left[|\langle \mu_k - \pi_k, \phi\rangle| \,|\mathcal{F}_{k-1}\right] &\leq (1-\epsilon_k)\left(\frac{1}{\sqrt{N_k}} + c_{k-1}\exp\left\{ H^* \left| \frac{1}{T_k} - \frac{1}{T_{k-1}} \right| \right\}\right)\|\phi\| \\ &\triangleq c_k \|\phi\|. \end{aligned}$$

■

Remark 1: There are a few important conclusions drawn from Theorem 2.

- We can ensure $E\left[|\langle \mu_k - \pi_k, \phi\rangle|\right]$ monotonically decreasing with respect to time by appropriately choosing the sample size sequence $\{N_k\}$ and the temperature cooling schedule $\{T_k\}$: noticing that $\exp\left(H^*\Delta_k\right) \downarrow 1$ as Δ_k goes to zero, we can choose sufficiently small Δ_k and sufficient large N_k to ensure that $c_k < c_{k-1}$. Corollary 2.1 below gives a special instance that ensures $c_k \to 0$ as $k \to 0$.
- If we want to decrease the temperature faster (i.e., Δ_k is larger), we have to increase the sample size N_k at each iteration in order to remain the same precision; and vice versa. This reveals the fundamental trade-off between temperature change rate and sample size. If the temperature decreases faster, then it is faster for the sequence of Boltzmann distributions to converge to the uniform distribution on the global optima, and hence we need less iterations to achieve a given precision; however, it takes more sampling effort to track the Boltzmann distribution at each iteration.
- The temperature $\{T_k\}$ does not have to be monotonically decreasing, as long as $T_k \to 0$ and the absolute change $\Delta_k = \left|\frac{1}{T_k} - \frac{1}{T_{k-1}}\right|$ is monotonically decreasing to 0 as $k \to 0$. This allows more flexible temperature cooling schedule.

Corollary 2.1: If $T_k = T_0/\log(k+1)$, $\varepsilon_k\lambda(\mathcal{X}) = \varepsilon < 1$, where $\varepsilon > (1/2)^{1-\frac{H_u-H_l}{T_0}}$ and $\frac{H_u-H_l}{T_0} < 1$, and $\{N_k\}$ increases sufficiently fast as k increases, then $\{c_k\} \to 0$ as $k \to \infty$.

Proof of Corollary 2.1: From (7) and (6), we have

$$\begin{aligned} c_k &= \left(1-\varepsilon\left(\frac{1}{k+1}\right)^{\Delta H/T_0}\right)\left(\frac{1}{\sqrt{N_k}}+\left(\frac{k+1}{k}\right)^{H^*/T_0}c_{k-1}\right) \\ &\le \left(1-\varepsilon\left(\frac{1}{k+1}\right)^{\Delta H/T_0}\right)\frac{1}{\sqrt{N_k}}+\left(\left(1+\frac{1}{k}\right)^{\Delta H/T_0}-\varepsilon\left(\frac{1}{k}\right)^{\Delta H/T_0}\right)c_{k-1}. \end{aligned}$$

where $\Delta H = H_u - H_l \ge H^*$. In the following we analyze the coefficient in front of c_{k-1}. To simplify notations, let $a \triangleq \Delta H/T_0$, so $0 < a < 1$. Since $0 < 1/k \le 1$, we consider the function

$$f(x) = (1+x)^a - \varepsilon x^a, \quad x \in (0,1].$$

It is easy to show under the condition $\varepsilon > (1/2)^{1-a}$ that $f'(x) < 0$, $f''(x) > 0$, and $\lim_{x\to 0^+} f'(x) = -\infty$. Hence, $f(x)$ is strictly decreasing and strictly convex on $(0,1]$ with $f(0) = 1$ and $f(1) = 1-\varepsilon > 0$. Consider another function

$$g(x) = (1-x)^b, \quad x \in (0,1],$$

where b is a constant in $(0,1)$. It is easy to verify that $g'(x) < 0$ and $g''(x) < 0$. Hence, $g(x)$ is strictly decreasing and strictly concave on $(0,1]$ with $g(0) = 1$ and $g(1) = 0$. From the above characterizations of $f(x)$ and $g(x)$ and the observation $g'(0) = -b > \lim_{x\to 0^+} f'(x)$, we know that the two functions must intersect at some point $\tilde{x} \in (0,1)$ and

$$f(x) < g(x), \quad \forall x \in (0, \tilde{x}).$$

This means there exists $K > 1/\tilde{x}$ such that

$$\left(1+\frac{1}{k}\right)^a - \varepsilon\left(\frac{1}{k}\right)^a < \left(1-\frac{1}{k}\right)^b, \quad \forall k \ge K.$$

Therefore, when N_k is sufficiently large such that it satisfies

$$\left(1-\varepsilon\left(\frac{1}{k+1}\right)^{\Delta H/T_0}\right)\frac{1}{\sqrt{N_k}} \le \left\{\left(1-\frac{1}{k}\right)^b - \left(1+\frac{1}{k}\right)^a + \varepsilon\left(\frac{1}{k}\right)^a\right\}c_{k-1},$$

we have

$$\begin{aligned} c_k &\le c_K \prod_{i=K}^{k}\left(1-\frac{1}{i}\right)^b \\ &= \frac{c_K(K-1)^b}{k^b} \to 0, \quad \text{as } k \to \infty. \end{aligned}$$

Since c_k is nonnegative, $\{c_k\} \to 0$ as $k \to \infty$.

■

B. Comparison with Multi-start Simulated Annealing

Multi-start simulated annealing, as the name suggests, is to run independent simulated annealing algorithms from multiple initial starting points in the solution space. If we have no prior information, these initial points can be drawn uniformly from the solution space. In multi-start SA, the multiple runs are independent, or in other words, the samples do not interact with each other. In SMC-SA, the samples interact with each other at every iteration through the importance updating and resampling steps, which guide the new samples to concentrate on the more promising area found so far, making the search more efficiently. On the other hand, the interaction among samples sometimes could be misleading; for instance, if the sample size is too small and one sample stands out, then all the samples may be guided to concentrate near this sample, making it harder to escape from the local optimum near this sample. As we will see shortly, SMC-SA has a faster convergence rate than multi-start SA, if the sample size is sufficiently large (the same for both algorithms). In the following, we will analytically compare multi-start SA and SMC-SA, and also derive a bound for multi-start SA in a similar approach as for SMC-SA.

Let ν denote the initial distribution for drawing the starting points. Let η_k^N denote the empirical distribution generated at the k^{th} iteration. The timeline of multi-start SA can be represented as:

$$\nu \quad \longrightarrow \quad \eta_0^N \quad \longrightarrow \quad \eta_1^N = \eta_0^N P_1 \quad \longrightarrow \quad \ldots \quad \longrightarrow \quad \eta_k^N = \eta_{k-1}^N P_k$$

sampling SA move SA move SA move

Here P_k denotes the transition kernel of the Markov chain corresponding to the SA move, expressed in (4). The following Lemma shows that the importance updating step in SMC-SA helps to reduce the error when the initial error is sufficiently small.

Lemma 3: Given a probability measure ζ that satisfies $|\langle \zeta - \pi_{k-1}, \phi \rangle| = |\langle \mu_{k-1} - \pi_{k-1}, \phi \rangle| \leq c_{k-1}\|\phi\|$, if c_{k-1} is sufficiently small, then

$$|\langle \tilde{\mu}_k - \pi_k, \phi \rangle| < |\langle \zeta - \pi_k, \phi \rangle|. \tag{8}$$

Proof of Lemma 3: Recalling that

$$\langle \tilde{\mu}_k, \phi \rangle = \frac{\langle \mu_{k-1} - \Psi_k \phi \rangle}{\langle \mu_{k-1}, \Psi_k \rangle},$$

after some simplification we have

$$|\langle \tilde{\mu}_k - \pi_k, \phi \rangle|^2 - |\langle \zeta - \pi_k, \phi \rangle|^2 = AB,$$

where

$$\begin{aligned} A &= \langle \mu_{k-1}, \Psi_k \phi \rangle - \langle \zeta, \phi \rangle \langle \mu_{k-1}, \Psi_k \rangle, \\ B &= \langle \mu_{k-1}, \Psi_k \phi \rangle + \langle \zeta, \phi \rangle \langle \mu_{k-1}, \Psi_k \rangle - 2\langle \pi_k, \phi \rangle \langle \mu_{k-1}, \Psi_k \rangle. \end{aligned}$$

Using $|\langle \mu_{k-1} - \pi_{k-1}, \phi \rangle| = |\langle \zeta - \pi_{k-1}, \phi \rangle| \leq c_{k-1}\|\phi\|$, we can show

$$\begin{aligned} |A - C| &\leq c_{k-1}\|\phi\|(2\|\Psi_k\| + 1), \\ |B + C| &\leq c_{k-1}\|\phi\|(4\|\Psi_k\| + 1), \end{aligned}$$

where $C = \langle \pi_k, \phi \rangle - \langle \pi_{k-1}, \phi \rangle - c_{k-1}^2\|\phi\|\|\Psi_k\|$. Since ϕ and Ψ_k are bounded, when c_{k-1} is sufficiently small such that it satisfies

$$c_{k-1}^2\|\phi\|\|\Psi_k\| + c_{k-1}\|\phi\|(4\|\Psi_k\| + 1) \leq |\langle \pi_k, \phi \rangle - \langle \pi_{k-1}, \phi \rangle|,$$

it is easy to verify that $AB < 0$. Hence, (8) is proved. ■

Remark 2: Lemma 3 shows that if $\zeta = \mu_{k-1}$, then after the importance updating step, the resultant $\tilde{\mu}_k$ is closer to π_k than μ_{k-1}. It verifies our earlier argument that in SMC-SA the importance updating step gives a head start for the current iteration. Lemma 3 also shows that if $\zeta = \eta_{k-1}^N$, the distribution at the $(k-1)^{th}$ iteration in multi-start SA, has the same sufficiently small error bound as μ_{k-1}, then SMC-SA after the importance updating step yields a smaller error than multi-start SA. This is the basis for the following theorem that directly compares one iteration of the two algorithms.

Theorem 3: Given $N_k = N$ and $|\langle \mu_{k-1} - \pi_{k-1}, \phi \rangle| = |\langle \eta_{k-1}^N - \pi_{k-1}, \phi \rangle| \leq c_{k-1}\|\phi\|$, if N_k is sufficiently large and c_{k-1} is sufficiently small, then

$$|\langle \mu_k - \pi_k, \phi \rangle| \leq |\langle \eta_k^N - \pi_k, \phi \rangle|, \quad w.p.1. \tag{9}$$

Proof of Theorem 3:

$$\begin{aligned} |\langle \mu_k - \pi_k, \phi \rangle| &= |\langle \tilde{\mu}_k^{N_k} P_k - \pi_k, \phi \rangle| \\ &\leq |\langle (\tilde{\mu}_k^{N_k} - \tilde{\mu}_k) P_k, \phi \rangle| + |\langle \tilde{\mu}_k P_k - \pi_k, \phi \rangle|. \end{aligned}$$

The first term $|\langle (\tilde{\mu}_k^{N_k} - \tilde{\mu}_k) P_k, \phi \rangle| \to 0$ w.p.1 as $N_k \to \infty$ by the law of large numbers, so it remains to show that the second term is less than $|\langle \eta_k^N - \pi_k, \phi \rangle|$ when c_{k-1} is sufficiently small. The second term can be rewritten as

$$\begin{aligned} |\langle \tilde{\mu}_k P_k - \pi_k, \phi \rangle| &= |\langle \tilde{\mu}_k - \pi_k, P_k\phi \rangle| \\ &\leq |\langle \eta_{k-1}^N - \pi_k, P_k\phi \rangle| \\ &= |\langle \eta_{k-1}^N P_k - \pi_k, \phi \rangle| \\ &= |\langle \eta_k^N - \pi_k, \phi \rangle|. \end{aligned}$$

where the inequality is a direct result of Lemma 3 when c_{k-1} is sufficiently small. ■

Remark 3: Theorem 3 shows that if the two algorithms have the same sufficiently small error bound at iteration $k-1$, then SMC-SA will yield a smaller error than the multi-start SA using the same

sufficiently large number of samples. Using that result repeatedly for every iteration, we can conclude that *given the same sample size and temperature cooling schedule, SMC-SA converges faster than multi-start SA when the sample size is sufficiently large.* This can be also explained intuitively. As the sample size increases, the interaction among samples tends to be more useful rather than misleading, and hence, it helps to guide the samples to become more concentrated around the global optima.

We also derive an explicit error bound for multi-start SA in the following theorem.

Theorem 4: Without loss of generality, we assume that $H(x) > 0$ for all $x \in \mathcal{X}$. Suppose the initial distribution is ν with the density ν^d with respect to the Lebesgue measure, then under Assumption 1,

$$E\left[\left|\left\langle \eta_k^N - \pi_k, \phi \right\rangle\right| | \mathcal{F}_{k-1}\right] \leq d_k \|\phi\|, \quad \forall \phi \in \mathcal{B}^+(\mathcal{X}),$$

where d_k satisfies the recursive equation

$$\begin{aligned} d_0 &= \frac{1}{\sqrt{N}} + \left\| 1 - \frac{\pi_0^d}{\nu^d} \right\|, \\ d_k &= (1-\epsilon_k)\left(d_{k-1} + \exp\left(H^* \Delta_k\right) - 1\right), \quad k = 1, 2, \ldots, \end{aligned} \tag{10}$$

where $\Delta_k = \left| \frac{1}{T_k} - \frac{1}{T_{k-1}} \right|$.

Proof of Theorem 4: First, consider the initial sampling step,

$$\begin{aligned} E\left[\left|\langle \eta_0^N - \pi_0, \phi \rangle\right|\right] &\leq E\left[\left|\langle \eta_0^N - \nu, \phi \rangle\right|\right] + \left|\langle \nu - \pi_0, \phi \rangle\right| \\ &\leq \frac{\|\phi\|}{\sqrt{N}} + \left| \langle \nu, \phi \rangle - \left\langle \nu, \frac{\pi_0^d}{\nu^d} \phi \right\rangle \right| \\ &\leq \left(\frac{1}{\sqrt{N}} + \left\| 1 - \frac{\pi_0^d}{\nu^d} \right\| \right) \|\phi\| \\ &\triangleq d_0 \|\phi\|, \end{aligned}$$

where the second inequality is due to Lemma 1.

Next, consider the SA move step at the k^{th} ($k \geq 1$) iteration.

$$E\left[\left|\left\langle \eta_{k-1}^N - \pi_k, \phi \right\rangle\right|\right] \leq E\left[\left|\left\langle \eta_{k-1}^N - \pi_{k-1}, \phi \right\rangle\right|\right] + \left|\langle \pi_{k-1} - \pi_k, \phi \rangle\right|,$$

where the second term on the righthand side can be further expressed as

$$\begin{aligned} \left|\langle \pi_{k-1} - \pi_k, \phi \rangle\right| &= \left|\langle \pi_{k-1}, \phi \rangle - \langle \pi_{k-1}, \Psi_k \phi \rangle\right| \\ &\leq \|1 - \Psi_k\| \, \|\phi\| \\ &\leq \left(\exp\left\{ H^* \left| \frac{1}{T_k} - \frac{1}{T_{k-1}} \right| \right\} - 1 \right) \|\phi\|, \end{aligned}$$

where the last inequality is derived in the same way as in the proof of Theorem 2. Therefore, by applying Corollary 1.2 and Corollary 1.1, we have

$$\begin{aligned} E\left[\left|\langle \eta_k^N - \pi_k, \phi\rangle\right|\right] &= E\left[\left|\langle \eta_{k-1}^N P_k - \pi_k, \phi\rangle\right|\right] \\ &\leq (1-\epsilon_k)\left(d_{k-1} + \exp\left\{H^*\left|\frac{1}{T_k} - \frac{1}{T_{k-1}}\right|\right\} - 1\right)\|\phi\| \\ &\triangleq d_k\|\phi\|. \end{aligned}$$

■

V. Numerical Experiments

In this section, we present results of numerical experiments to illustrate the effectiveness of the proposed SMC-SA algorithm. We test SMC-SA on six well-known unconstrained and continuous optimization problems: Dejong's 5th function (H_a), 20-dimensional Powel singular function (H_b), 20-dimensional Rosenbrock function (H_c), 20-dimensional Griewank function (H_d), 10-dimensional Trigonometric function (H_e), and 10-dimensional Pintér function (H_f). Their explicit expressions are listed in the Appendix.

As a comparison of the proposed SMC-SA method, we also solved the test problems using the standard SA algorithm, multi-start SA, and the cross-entropy (CE) method [32], and compared their average performance based on 100 independent runs.

For SMC-SA, standard SA, and multi-start SA, we use the logarithm cooling schedule $T_k = |H^*(x_{k-1})|/\log(k+1)$, where $H^*(x_{k-1})$ is the optimal sample function value at the $(k-1)^{th}$ iteration. The reason for using $|H^*(x_{k-1})|$ is because the weights w_k^i are calculated in proportion to the exponential function $\exp\left\{H(x_{k-1}^i)\left(\frac{1}{T_k} - \frac{1}{T_{k-1}}\right)\right\}$, which may get exploded if the argument of the exponential function is large, and may become identical values if the argument is in the flat tail of the exponential function. By using $|H^*(x_{k-1})|$ in the temperature, the weights will not depend too much on the value of $H(x_{k-1}^i)$. In these four methods, the initial candidate solutions are all chosen randomly according to the uniform distribution on $[-50, 50]^n$. For SMC-SA, the proposal distribution in the SA move step is the normal distribution with standard deviation $\alpha\beta^k$ at iteration k, where $\alpha = 10$, $\beta = 0.995$ for objective functions H_a and H_b, and $\beta = 0.998$ for H_c, H_d, H_e, and H_f. Although in the theoretical analysis in section IV, a sufficient condition for the convergence of the algorithm requires an increasing sample size, we find in practice that a fixed sample size may also provide good solutions with less computation effort. Fixing the sample size also makes SMC-SA comparable with multi-start SA, where the sample size does not change with iteration. The sample size is set to be $N = 200$ for H_a, H_b, H_d and H_f, and $N = 1000$ for the high-dimensional problems

H_c and H_e. For the standard SA and multi-start SA, the parameter settings are the same as in SMC-SA for each problem, i.e, the same temperature, proposal distributions, and the same sample size N in multi-start SA. For the CE method, we use the normal distributions as the parameterized family; the initial mean μ_0 is chosen randomly according to the uniform distribution on $[-50, 50]^n$, and the initial covariance matrix is set to be $\Sigma_0 = 500I_{n\times n}$; the quantile parameter ρ is set to be 0.01; the sample size N is 500 for H_a and H_b, and 5000 for $H_c - H_f$; the parameters are updated according to a smoothing scheme [6] with the smoothing coefficient set to be $\nu = 0.2$, which is found to work best by trial and error in our experiments.

Table I shows the average performance based on 100 independent runs, where H^* is the true optimal value of $H(\cdot)$, $\bar{H}^*$ is the average of the computed optimal values of the 100 runs, std_err is the standard error of the computed optimal values of the 100 runs, and M_ε is the number of ε-optimal solutions out of 100 runs. In this numerical experiment, we consider $\varepsilon = 10^{-5}$ for problem H_a, H_d, H_e, and H_f, and $\varepsilon = 0.01$ for problems H_b and H_c. Fig. 1 shows the average computed optimal value $\bar{H}^*$ versus the total number of samples for these four methods.

		SMC-SA		multi-start SA		standard SA		CE ($\nu = 0.2$)	
	H^*	$\bar{H}^*(std\_err)$	M_ε	$\bar{H}^*(std\_err)$	M_ε	$\bar{H}^*(std\_err)$	M_ε	$\bar{H}^*(std\_err)$	M_ε
H_a	-0.998	-0.998(1.34E-7)	100	-1.0024(0.0014)	19	-3.999(0.2117)	4	-1.544(0.0695)	51
H_b	-0.01	-0.0164(4.95E-4)	81	-20.46(4.26)	0	-89.63(1.277)	0	-113.3(66.39)	69
H_c	-1	-5.673(0.249)	5	-6.623(0.313)	4	-378.8(5.478)	0	-18.35(0.0113)	0
H_d	0	-1.80E-7(2.81E-9)	100	-2.44E-7(4.25E-9)	100	-0.274(0.0029)	0	-7.44E-12(3.09E-13)	100
H_e	-1	-1.225(0.0275)	56	-1.407(0.035)	2	-41.61(3.25)	0	-1(0.0E00)	100
H_f	-10^{-15}	-1.616E-15(1.86E-17)	100	-6.13E-6(4.87E-6)	98	-1.00E+3(85.91)	0	-0.1777(0.0037)	0

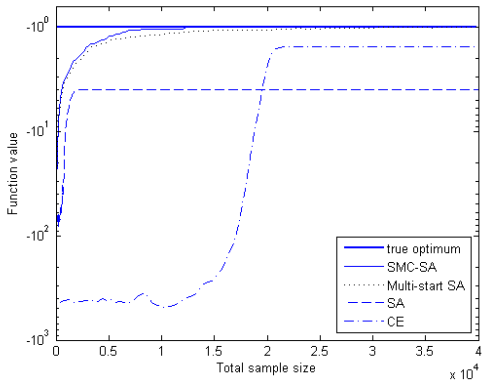
TABLE I

AVERAGE PERFORMANCE OF SMC-SA, MULTI-START SA, STANDARD SA AND CE ON BENCHMARK PROBLEMS

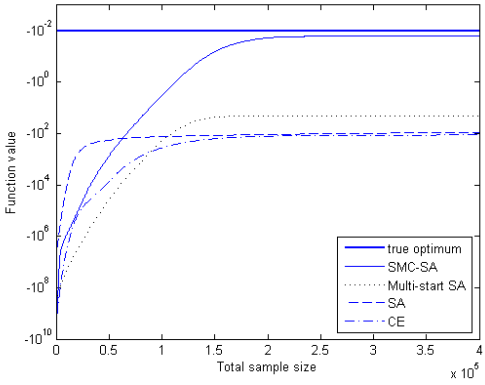
From the results, we may see that for all of these six test problems, SMC-SA outperforms the standard SA. SMC-SA provides much more accurate solutions with smaller standard error, and it also converges faster than standard SA on problems H_a, $H_d - H_f$. SMC-SA performs better than multi-start SA on problems H_a, H_b, H_e and H_f in both accuracy and convergence rate, and performs slightly better than multi-start SA on problems H_c and H_d. In all the problems except H_d and H_e, SMC-SA performs better than CE in accuracy. SMC-SA converges faster than CE on the first two problems; on the last four problems, it converges faster than CE at the very beginning and then slower.

In summary, SMC-SA is a great improvement of the standard SA on all the test problems; SMC-SA works better than multi-start SA and CE on badly-scaled problems and problems with a small number

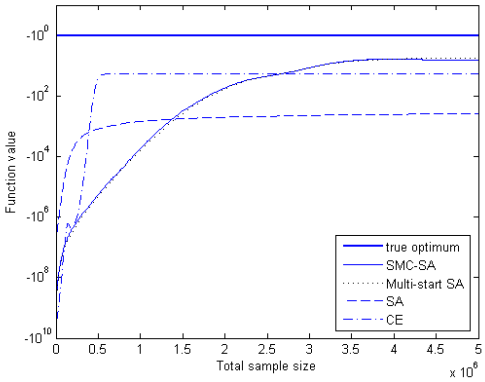
(a) 2-D Dejong's 5th function



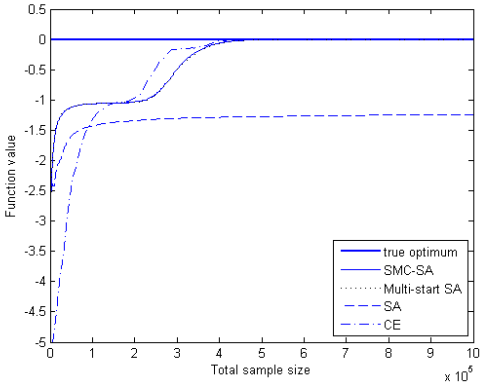
(b) 20-D Powel function



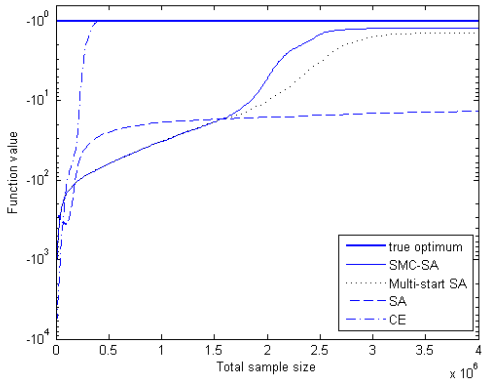
(c) 20-D Rosenbrock function



(d) 20-D Griewank function



(e) 10-D Trigonometric function



(f) 10-D Pintér's function

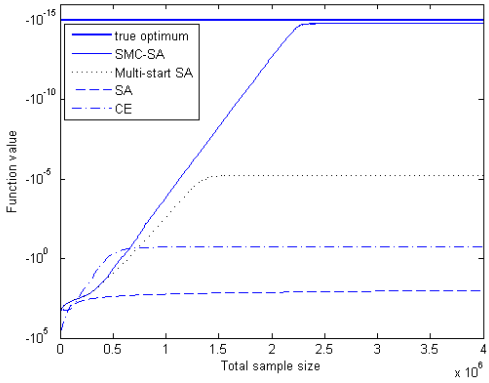


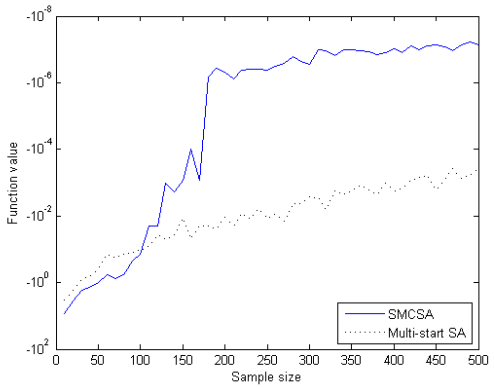
Fig. 1. Average Performance of SMC-SA, multi-start SA, standard SA and CE

of local optima; the CE method works better on well-scaled problems with a large number of local optima.

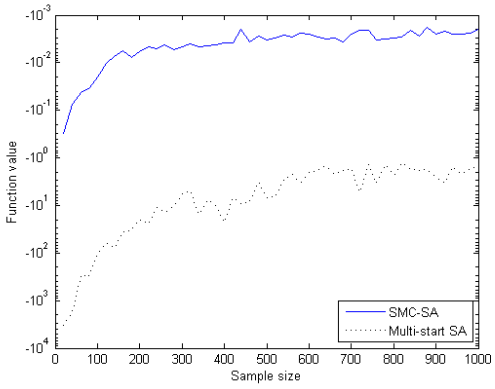
A. *Comparison of SMC-SA and Multi-Start SA with Different Sample Sizes*

In this section, we numerically compare the performance of SMC-SA and multi-start SA versus different sample sizes. Experiments are carried out on three different kinds of objective functions, low-dimensional functions with a limited number of local optima (e.g, Dejong's 5th function), high-dimensional badly-scaled functions (e.g., Powel singular function), and high-dimensional functions with a large number of local optima (e.g., Pintér's function). We vary the sample size N, and use the same total number of iterations and other parameters as in the previous section for all three functions. We compare the average computed optimal value $\bar{H}^*$ of 100 independent runs. The comparison results are shown in Fig. 2.

(a) Dejong's 5th function



(b) Powel function



(c) Pintér's function

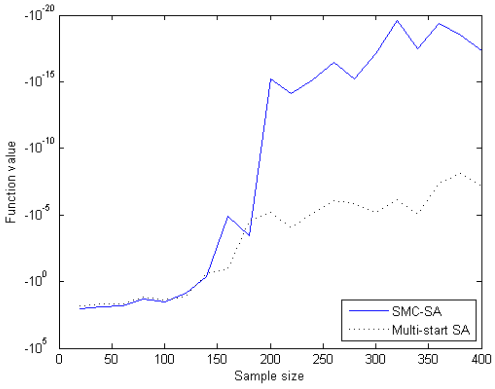


Fig. 2. Average Performance of SMC-SA and Multi-start SA vs Sample Size

For Dejong's 5th function and Pintér's function, Fig. 2 (a) and (c) show that SMC-SA outperforms multi-start SA when the sample size is large, and multi-start SA performs better when the sample size is small. This observation is consistent with the conclusion in Section IV-B that SMC-SA is more preferable than multi-start SA with sufficiently large sample size. For Powel singular function, SMC-SA always performs better than multi-start SA with different sample sizes. Intuitively, with

small sample size, the interaction among the samples is more likely to be misleading such that the samples may easily get concentrated around certain local optimum without exploring more area. This phenomenon is more severe in the problem with multiple local optima, such as Dejong's 5th and Pintér's function. In problems with few local optima, such as Powel function, the interaction among samples in SMC-SA is useful rather than misleading even with a small sample size.

VI. Conclusion and Future Research

In this paper, we proposed the sequential Monte Carlo simulated annealing (SMC-SA) algorithm for continuous global optimization. The main idea is to track the converging sequence of Boltzmann distributions using a population of samples via sequential Monte Carlo method, such that the empirical distributions yielded by SMC-SA also converge weakly to the uniform distribution concentrated on the set of global optima as the temperature decreases to zero.

We proved an upper bound on the difference between the Boltzmann distribution and the empirical distribution yielded by SMC-SA. The bound guides the choice of the sample size and the cooling schedule to ensure the error strictly decreases and eventually converges to zero. It shows a trade-off between the sample size and the rate of temperature decrease: by generating more samples at each iteration we can reduce the number of iterations to reach the same accuracy of the solution. Moreover, we also proved a bound for multi-start simulated annealing using a similar approach, and analytically compared multi-start SA with SMC-SA. The result shows that SMC-SA is more preferable than the multi-start SA when the sample size is sufficiently large.

We carried out numerical experiments on several benchmark problems. The numerical results show that SMC-SA is a great improvement of the standard SA on all the test problems; SMC-SA outperforms multi-start SA and CE on badly-scaled problems and problems with a small number of local optima; the CE method works better on well-scaled problems with a large number of local optima. We also compared the performance of SMC-SA and multi-start SA as the sample size varies, and the results verified our analytical results.

In our numerical experiments, we found that the Boltzmann distribution may not be very desirable due to its exponential form: the weights of different samples tend to be identical if they fall on the flat tail of the exponential function, and the weights may explode as T_k goes to zero (if $H(x) > 0$). That suggests a possibility for better choices of the target distributions. We should note that the idea of SMC-SA can be easily generalized to other target distributions, as long as they converge weakly to the degenerate distribution concentrated on one or more global optima and satisfy some regularity conditions.

VII. Acknowledgements

This work was supported by the National Science Foundation under Grant ECCS-0901543 and CMMI-1130273. We would like to thank the associate editor and the two anonymous reviewers for their careful reading of the paper and very constructive comments that led to a substantially improved paper. The first author would also like to thank Peter Glynn for an inspiring discussion.

The algorithm of SMC-SA and part of the numerical results have been presented at the 2010 Winter Simulation Conference [40].

VIII. Appendix: Test Problems

The six benchmark problems are originally presented in [5], [37], [17], [28] as minimization problems. Since SMC-SA is presented in maximization form, we take the negative value of the objective function and convert them to maximization problems.

(a) Dejong's 5th function (n=2)

$$H_a(x) = -\left[0.002 + \sum_{j=1}^{25} \frac{1}{j + \sum_{i=1}^{2}(x_i - a_{ji})^6}\right]^{-1}$$

where $a_{j1} = (-32, -16, 0, 16, 32, -32, -16, 0, 16, 32, -32, -16, 0, 16, 32, -32, -16, 0, 16, 32, -32, -16, 0, 16, 32)$ and $a_{j2} = (-32, -32, -32, -32, -32, -16, -16, -16, -16, -16, 0, 0, 0, 0, 0, 16, 16, 16, 16, 16, 32, 32, 32, 32, 32)$. The global maximum is at $x^* = (-32, -32)^T$, and $H_a^* \approx -0.998$.

(b) Powel singular function (n=20)

$$H_b(x) = -\sum_{i=2}^{n-2}\left[(x_{i-1} + 10x_i)^2 + 5(x_{i+1} - x_{i+2})^2 + (x_i - 2x_{i+1})^4 + 10(x_{i-1} - x_{i+2})^4\right] - 0.01$$

where $x^* = (0, \cdots, 0)^T$, $H_b^* = -0.01$.

(c) Rosenbrock function (n=20)

$$H_c(x) = -\sum_{i=1}^{n-1}\left[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2\right] - 1$$

where $x^* = (1, \cdots, 1)^T$, $H_c^* = -1$.

(d) Griewank function (n=20)

$$H_d(x) = -\left[\frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1\right]$$

where $x^* = (0, \cdots, 0)^T$, $H_d^* = 0$.

(e) Trigonometric function (n=10)

$$H_e(x) = -1 - \sum_{i=1}^{n}\left[8\sin^2(7(x_i - 0.9)^2) + 6\sin^2(14(x_i - 0.9)^2) + (x_i - 0.9)^2\right]$$

where $x^* = (0.9, \cdots, 0.9)^T$, $H_e^* = -1$.

(f) Pintér's function (n=10)

$$\begin{aligned} H_f(x) &= -\left[\sum_{i=1}^{n} i x_i^2 + \sum_{i=1}^{n} 20i \sin^2(x_{i-1} \sin x_i - x_i + \sin x_{i+1}) \right. \\ &\quad \left. + \sum_{i=1}^{n} i \log_{10}(1 + i(x_{i-1}^2 - 2x_i + 3x_{i+1} - \cos x_i + 1)^2)\right] - 10^{-15} \end{aligned}$$

where $x^* = (0, \cdots, 0)^T$, $H_f^* = -10^{-15}$.

REFERENCES

[1] S. Anily and A. Federgruen. Simulated annealing methods with general acceptance probabilities. *Journal of Applied Probability*, 24(3):657–667, 1987.

[2] C. J. P. Belisle. Convergence theorems for a class of simulated annealing algorithms on R^d. *Journal of Applied Probability*, 29:885–895, 1992.

[3] O. Cappé, S. J. Godsill, and E. Moulines. An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, 95(5):899–924, 2007.

[4] K. W. Chu, Y. Deng, and J. Reinitzy. Parallel simulated annealing by mixing of states. *Journal of Computational Physics*, 148:646–662, 1999.

[5] A. Corana, M. Marchesi, C. Martini, and S. Ridella. Minimizing multimodal functions of continuous variables with the simulated annealing algorithm. *ACM Transactions on Mathematical Software*, 13(3):262–208, 1987.

[6] P. T. DeBoer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67, 2005.

[7] A. Dekkers and E. Aarts. Global optimization and simulated annealing. *Mathematical Programming*, 50(3):367–393, 1991.

[8] A. Doucet, J. F. G. deFreitas, and N. J. Gordon, editors. *Sequential Monte Carlo Methods In Practice*. Springer, New York, 2001.

[9] A. Doucet, S. Godsill, and C. Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, 10(3):197 – 208, 2000.

[10] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE *Transactions of Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[11] B. Gidas. Nonstationary Markov chains and convergence of the annealing algorithm. *Journal of Statistical Physics*, 39(1-2):73–131, 1985.

[12] W. Gilks and C. Berzuini. Following a moving target - Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society*, 63(1):127–146, 2001.

[13] N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993.

[14] B. Hajek. Cooling schedules for optimal annealing. *Mathematics of Operations Research*, 13(2):311–329, 1988.

[15] D. Henderson, S. H. Jacobson, and A. W. Johnson. *Handbook of Metaheuristics*, volume 57 of *International Series in Operations Research & Management Science*, chapter The Theory and Practice of Simulated Annealing, pages 287–319. Springer, 2003.

[16] S. Kirkpatrick, C. D. Gelatt, and Jr. M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[17] D. P. Kroese, S. Porotsky, and R. Y. Rubinstein. The cross-entropy method for continuous multiextremal optimization. *Methodology and Computing in Applied Probability*, 8(3):383–407, 2006.

[18] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

[19] M. Locatelli. Convergence properties of simulated annealing for continuous global optimization. *Journal of Applied Probability*, 33:1127–1140, 1996.

[20] M. Locatelli. Simulated annealing algorithms for continuous global optimization: Convergence conditions. *Journal of Optimization Theory and Applications*, 104(1):121–133, 2000.

[21] S. W. Mahfoud and D. E. Goldberg. Parallel recombinative simulated annealing: a genetic algorithm. *Parallel Computing*, 21:1–28, 1995.

[22] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[23] O. Molvalioglu, Z. B. Zabinsky, and W. Kohn. *Models and Algorithms for Global Optimization*, volume 4 of *Springer Optimization and Its Applications*, chapter Multi-particle Simulated Annealing, pages 215–222. Springer, 2007.

[24] O. Molvalioglu, Z. B. Zabinsky, and W. Kohn. The interacting-particle algorithm with dynamic heating and cooling. *Journal of Global Optimization*, 43(2-3):329–356, 2009.

[25] P. Del Moral. F*eynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, New York, 2004.

[26] P. Del Moral, A. Doucet, and T. France. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B*, 68(3):411436, 2006.

[27] E. Onbaşoğlu and L. Özdamar. Parallel simulated annealing algorithms in global optimization. *Journal of Global Optimization*, 19(1):27–50, 2001.

[28] J. D. Pintér. *Global Optimization in Action*. Kluwer Academic Publishers, Dordrecht, The Neitherlands, 1996.

[29] C. P. Robert and G. Casella. M*onte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York, 2004.

[30] G. Roberts and J. S. Rosenthal. General state spact Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.

[31] H. E. Romeijn and R. L. Smith. Simulated annealing for constrained global optimization. *Journal of Global Optimization*, 5(2):101–126, 1994.

[32] R. Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2):127–190, 1999.

[33] G. Ruppeiner, J. M. Pedersen, and P. Salamon. Ensemble approach to simulated annealing. *Journal de Physics I*, 1:455–470, 1991.

[34] P. van Hentenryck and Y. Vergados. Population-based simulated annealing for traveling tournaments. In *Proceedings of the 22nd national conference on artificial intelligence*, volume 1, pages 267–272, 2007.

[35] P. J. M. van Laarhoven and E. H. L. Aarts. *Simulated Annealing: Theory and Applications*. Springer, 1987.

[36] R. L. Yang. Convergence of the simulated annealing algorithm for continuous global optimization. *Journal of Optimization Theory and Applications*, 104(3):691–716, 2000.

[37] X. Yao and Y. Liu. Fast evolutionary programming. In *Proceedings of the Fifth Annual Conference on Evolutionary Programming*, pages 451–460, Cambridge, MA, 1996. MIT Press.

[38] A. A. Zhigljavsky. *Theory of Global Random Search*. Kluwer Academic Publishers, 1991.

[39] A. Zhiljavsky and A. Zilinskas. *Stochastic Global Optimization*, volume 9 of *Springer Optimization and Its Applications*. Springer, 2008.

[40] E. Zhou and X. Chen. A new population-based simulated annealing algorithm. In *Proceedings of the 2010 Winter Simulation Conference*, pages 1211–1222, 2010.