Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Relative Frequencies of Non-homogeneous Markov Chains in Simulated Annealing and Related Algorithms

Jan Hannig, Edwin K. P. Chong, and Sanjeev R. Kulkarni

Abstract— **We consider a class of non-homogeneous Markov chains arising in simulated annealing and related stochastic search algorithms. Using only elementary first principles, we analyze the convergence and rate of convergence of the relative frequencies of visits to states in the Markov chain. We describe in detail three examples, including the standard simulated annealing algorithm, to show how our framework applies to specific stochastic search algorithms—these examples have not previously been recognized to be sufficiently similar to share common analytical grounds. Our analysis, though elementary, provides the strongest sample-path convergence results to date for simulated annealing type Markov chains. Our results serve to illustrate that by taking a purely sample-path view, surprisingly strong statements can be made using only relatively elementary tools.**

I. INTRODUCTION

For at least the last 20 years, there has been an interest in stochastic search algorithms for global optimization based on non-homogeneous Markov chains. The prime example is *simulated annealing*, first suggested for optimization by Kirkpatrick et al. [1] based on techniques of Metropolis et al. [2]. An early application to image processing was described by Geman and Geman [3]. The basic procedure in simulated annealing is to explore the search space by setting up a graph over the space and jumping from point (vertex) to point in this graph according to a non-homogeneous Markov chain. The non-homogeneity arises from the gradually decreasing probability of jumping from one point to a "worse" point in the course of the search (but such a jump also cannot be precluded, because of the need to "climb out" of "cups" around local minimizers). The speed at which this decrease in the transition probabilities occurs depends on a sequence called the "cooling schedule" (described in more detail in Section III).

In a seminal paper, Hajek [4] provides a detailed treatment of the behavior of the Markov chain associated with the simulated annealing algorithm. Specifically, he provides a necessary and sufficient condition on the cooling schedule for convergence in probability of the algorithm to the set of global minimizers. Tsitsiklis [5] proves essentially the same result, but using different techniques. Around the same time, Connors and Kumar [6] also study simulated annealing type Markov chains, providing yet a different view of such processes.

In the last 15 years, the literature on the analysis of simulated annealing has grown significantly. In particular, there have been several generalizations of simulated annealing. For example, Gelfand and Mitter [7] and Tsallis and Stariolo [8] consider a continuous-space version of simulated annealing, and Morai and Miclo [9], Cot and Catoni [10], and Trouve [11] consider an even further generalization of the Markov process in standard simulated annealing. The analysis of these generalizations of simulated annealing involve relatively sophisticated tools.

In this paper, we study a non-homogeneous Markov chain that is also a generalization of simulated annealing. Our generalization is different from those of the above papers—ours is much closer to the original simulating annealing framework of Hajek [4]. For convenience, in this paper we refer to our generalization simply as *generalized simulated annealing* (even though this same term is used also for other generalizations). The main reason for introducing our generalization is to facilitate the analysis of *relative frequencies* in non-homogeneous Markov chains arising in simulated annealing and other stochastic search algorithms.

Our focus on relative frequencies in our non-homogeneous Markov chain sharply differentiates our study from previous studies in the literature. At the same time, our approach offers several advantages. First, we use only elementary first principles—our tools consist essentially of applications of Kolmogorov's three-series theorem and coupling. In contrast, the seminal paper of Hajek [4], which was also based on first principles, requires relatively complex arguments. Second, our generalization, while simple, allows studying rather disparate search algorithms within a single unified framework. We illustrate this claim by considering two other search algorithms (besides standard simulated annealing)—these two other algorithms have not previously been recognized to be sufficiently akin to simulated annealing to have a common analytical "ancestry." Third, our approach provides what we believe to be the strongest *sample-path* characterizations of simulated annealing type Markov chains to date. We establish not only the convergence to zero of the relative frequencies of all non-global-minimizers, but also the *rate* at which these relative frequencies vanish.

The work of Hannig was supported in part by an IBM Faculty Award. The work of Chong was supported in part by the NSF under grants 0098089-ECS, 0099137-ANI, and ANI-0207892. The work of Kulkarni was supported in part by the ARO under grant DAAD19-00-1-0466, Draper Laboratory under IR&D 6002 grant DL-H-546263, and the NSF under grant CCR-0312413.

Jan Hannig is with the Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, USA. Jan.Hannig@ColoState.edu

Edwin K. P. Chong is with the Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523-1373, USA. EChong@Engr.ColoState.edu

Sanjeev R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. Kulkarni@EE.Princeton.edu.

0-7803-9568-9/05/$20.00 ©2005 IEEE

There is significant appeal in characterizing convergence and rates in purely sample-path terms. Our commitment to this program of study is evident in our previous work on sample-path analyses of various stochastic algorithms; see Kulkarni and Horn [12], Wang et al. [13], Wang et al. [14], Wang and Chong [15], and Chong et al. [16]. The typical conclusion we find is that although these purely sample-path analyses involve only elementary tools, the results are surprisingly strong—the results in this paper corroborate this conclusion. We contrast this with the probabilistic analysis of Hajek [4]: although his analysis provides the strongest possible condition for convergence based on first principles, rates of convergence do not fall out easily. In our analysis of relative frequencies, on the other hand, rate estimates follow relatively easily and naturally. From first principles it is extremely difficult to obtain the kind of "sharp" estimates needed in Hajek's probabilistic analysis to characterize rates in addition to convergence. Since Hajek's paper, there have certainly been results on convergence rates of *probabilities* in simulated annealing and its generalizations; however, more sophisticated machinery than Hajek's first-principles approach has to be brought to bear (e.g., see Catoni [17], who uses results from Freidlin and Wentzell [18]). This paper and our previous work along similar lines suggest that the same is not the case in a purely sample-path setting.

Some notation and terminology

We first introduce some notation used throughout this paper. For two positive sequences $\{a_n\}$ and $\{b_n\}$, we write:

- $a_n \sim b_n$ if $a_n/b_n \to 1$;
- $a_n \overset{O}{=} b_n$ if $\limsup a_n/b_n < \infty$, and $\limsup b_n/a_n < \infty$; and
- $a_n \overset{O}{\approx} b_n$ if $(\log a_n - \log b_n)/\log n \to 0$.

The difference between $a_n \overset{O}{=} b_n$ and $a_n \overset{O}{\approx} b_n$ is that while "$\overset{O}{=}$" implies that the two sequences are of the same "order," the weaker "$\overset{O}{\approx}$" allows their order to differ by a slowly varying function, e.g., a power of $\log n$.

Given a sequence $\boldsymbol{x} = \{x_n\} = \{x_1, x_2, \ldots\}$ and a set A, we define the notation

$$I(x_i \in A) = \begin{cases} 1 & \text{if } x_i \in A \\ 0 & \text{otherwise.} \end{cases}$$

The notation $I(x_i \in A)$ represents an "indicator" of the condition $x_i \in A$. We define the *relative frequency of visits to* A up to time n as $\mathcal{F}_n(\boldsymbol{x} \in A) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \in A)$. If A is the singleton $\{v\}$, we write $\mathcal{F}_n(\boldsymbol{x} = v)$. Similarly, we use the notation $\mathcal{F}_n(\boldsymbol{x} \neq v) = \mathcal{F}_n(\boldsymbol{x} \notin \{v\})$.

If $x_i \in A$ for an infinite number of i, then we say that A *is visited infinitely often*. Otherwise, we say that A *is visited finitely often*.

When considering random sequences, we use capital letters: $\boldsymbol{X} = \{X_1, X_2, \ldots\}$, $\mathcal{F}_n(\boldsymbol{X} = x^*)$, etc.

Relative frequencies of random sequences

Our results are stated in terms of convergence (a.s.) of relative frequencies. In general, a (discrete state-space) random sequence $\boldsymbol{X} = \{X_1, X_2, \ldots\}$ that converges *in probability* to x^* may or may not also have convergent relative frequencies of the form $\mathcal{F}_n(\boldsymbol{X} = x^*)$. If the sequence is *independent*, then convergence in probability is stronger than its relative frequency counterpart, as stated in this simple lemma.

Lemma 1: Let $\boldsymbol{X} = \{X_1, X_2, \ldots\}$ be an independent, discrete state-space, random sequence that converges to x^* in probability. Then, $\mathcal{F}_n(\boldsymbol{X} = x^*) \to 1$ a.s.

We should point out that in the independent case, convergence in probability is *strictly* stronger than its relative frequency counterpart, because there are instances where $\mathcal{F}_n(\boldsymbol{X} = x^*) \to 1$ a.s. but the sequence does not converge to x^* in probability. To see this, consider the sequence $\boldsymbol{X} = \{X_1, X_2, \ldots\}$ on the state-space $\{0, 1\}$, where $X_n = 1$ a.s. for all n except for those n of the form $n = 2^k$, $k = 1, 2, \ldots$, in which case $X_n = 0$ a.s. Thus, $P(X_n = 1)$ does not converge to 1, but $\mathcal{F}_n(\boldsymbol{X} = 1) \to 1$ a.s.

In our generalized simulated annealing framework, the sequences are non-homogeneous Markov chains. In these cases, it is not clear *a priori* whether convergence in probability is weaker or stronger than its relative frequency counterpart. It will turn out that in fact they are *equivalent*.

II. GENERALIZED SIMULATED ANNEALING

In this section we present our generalized simulated annealing framework and our main results. Consider a finite, oriented, connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ a set of directed edges. Assume that each vertex $v \in \mathcal{V}$ is assigned a value $f(v)$. Our goal is to find the minimum of the function f; i.e., we wish to find $v_{\min} \in \mathcal{V}$ such that $f(v_{\min}) \leq f(v)$ for all $v \in \mathcal{V}$.

We assume that all values of $f(v)$, $v \in \mathcal{V}$, are distinct. We make this assumption to simplify the presentation. In particular, under this assumption, $v_{\min} = \arg\min_{v \in \mathcal{V}} f(v)$ is unique. However, all our results remain valid with appropriate adjustments if we remove this assumption.

Now define a non-homogeneous Markov process $\{X_n\}$ on the graph $\mathcal{G}$, as follows. Associate with each edge $uv \in \mathcal{E}$, $u \neq v$, two values $g_r(u, v) \geq 0$ and $g_c(u, v) > 0$. The transition probabilities of $\{X_n\}$ satisfy, for $u \neq v$,

$$P(X_n = v | X_{n-1} = u) \begin{cases} \sim g_c(u,v) n^{-g_r(u,v)} & \text{if } uv \in \mathcal{E} \\ = 0 & \text{otherwise,} \end{cases}$$

and, as usual, $P(X_n = u | X_{n-1} = u) = 1 - \sum_{v \neq u} P(X_n = v | X_{n-1} = u)$. Thus, the asymptotic behavior of the transition probabilities is determined by the values of $g_r(u, v)$ and $g_c(u, v)$. We will call $\{X_n\}$ a *generalized simulated annealing* process. As we will see in Section III, generalized simulated annealing reduces not only to the familiar simulated annealing process, but also processes associated with other stochastic search algorithms.

For convenience, define for each vertex $u \in \mathcal{V}$ two neighborhoods: $\mathcal{N}_{\text{out}}(u) = \{v \neq u : uv \in \mathcal{E}\}$ and $\mathcal{N}_{\text{in}}(u) = \{v \neq u : vu \in \mathcal{E}\}$. With this notation, we see that because probabilities must be bounded above by 1, for all u, $\sum_{v \in \mathcal{N}_{\text{out}}(u):\ g_r(u,v)=0} g_c(u, v) \leq 1$.

We now describe the notion of *weak reversibility*, the main assumption that links the function $f(v)$ with the transition

probabilities of $\{X_n\}$ (the same term is used in Hajek [4] for simulated annealing—our definition reduces to that of Hajek's in that special case). As usual, we say that $p = \{u_1, u_2, u_3, \ldots, u_{k-1}, u_k\}$ is a *path* from u to v if $u_1 = u$, $u_k = v$, and $u_{i+1} \in \mathcal{N}_{\text{out}}(u_i)$, $i = 1, \ldots, k-1$. For a path $p = \{u, u_2, u_3, \ldots, u_{k-1}, v\}$ we define its *height* by

$$h(p) = \max\{f(u) + g_r(u, u_2), f(u_2) + g_r(u_2, u_3), \ldots, f(u_{k-1}) + g_r(u_{k-1}, v)\}.$$

(This definition is again motivated by the notion of "height" in Hajek [4] for simulated annealing.) For any two vertices u and v, we then define

$$h(u, v) = \min\{h(p) : \ p \text{ is a path from } u \text{ to } v\}. \quad (1)$$

Next, we introduce two definitions involving the notion of heights: *weak reversibility* and *height normalization*. These are needed in the statements of our main results. The notion of weak reversibility follows that of Hajek [4]. Height normalization plays a key role in convergence.

Definition 1: We say that the generalized simulated annealing process is *weakly reversible* if, for any two vertices u and v, $h(u, v) = h(v, u)$.

Definition 2: We say that the generalized simulated annealing process is *height-normalized* if, for any vertex $v \neq v_{\min}$, $h(v, v_{\min}) - f(v) \leq 1$.

We are ready to state our main convergence result.

Theorem 1: Consider a weakly reversible generalized simulated annealing process $\boldsymbol{X} = \{X_1, X_2, \ldots\}$. If the process is height-normalized, then $\mathcal{F}_n(\boldsymbol{X} = v_{\min}) \to 1$ a.s. regardless of the starting point.

On the other hand, suppose that the process is not height-normalized. Then, there is a vertex $v \neq v_{\min}$ such that $h(v, v_{\min}) - f(v) > 1$, and if $X_1 = v$, then $P(\mathcal{F}_n(\boldsymbol{X} = v) \to 1) > 0$ (which implies that $P(\mathcal{F}_n(\boldsymbol{X} = v_{\min}) \to 1) < 1$).

Our next result characterizes the *rate* of convergence in terms of relative frequencies.

Theorem 2: Consider a weakly reversible, height-normalized generalized simulated annealing process $\boldsymbol{X} = \{X_1, X_2, \ldots\}$. Suppose v is a vertex such that

$$h(v_{\min}, v) - f(v_{\min}) < 1. \quad (2)$$

Then, $\mathcal{F}_n(\boldsymbol{X} = v) \stackrel{O}{=} n^{-(f(v) - f(v_{\min}))}$ a.s. regardless of the starting point.

Otherwise, if (2) is not satisfied but

$$h(v_{\min}, v) - f(v_{\min}) = 1, \quad (3)$$

then $\mathcal{F}_n(\boldsymbol{X} = v) \stackrel{O}{\approx} n^{-(f(v) - f(v_{\min}))}$ a.s. regardless of the starting point.

Finally, if for some v neither (2) nor (3) is satisfied, then v is visited finitely often a.s. regardless of the starting point, whence either $\mathcal{F}_n(\boldsymbol{X} = v) = 0$ or $\mathcal{F}_n(\boldsymbol{X} = v) \stackrel{O}{=} n^{-1}$ a.s.

Recall that "$\stackrel{O}{=}$" is stronger than "$\stackrel{O}{\approx}$." Thus, for simplicity, we can summarize the essence of Theorem 2 as follows: If v is visited infinitely often a.s., then $\mathcal{F}_n(\boldsymbol{X} = v) \stackrel{O}{\approx} n^{-(f(v) - f(v_{\min}))}$ a.s. regardless of the starting point. In Section III, we will use this simplified version of Theorem 2 in applying our framework to specific examples.

Because of space restrictions, we are unable to provide our proofs, which are based on elementary *coupling* arguments (see [19] for details). The main technical instrument used in our proofs is the following lemma.

Lemma 2: Let $\{X_n\}$ be a non-homogeneous Markov chain with state-space $\{1, 2\}$ and transition probabilities satisfying $P(X_n = 2 | X_{n-1} = 1) \sim d_1 n^{-\Delta_1}$ and $P(X_n = 1 | X_{n-1} = 2) \sim d_2 n^{-\Delta_2}$. Assume that $d_1, d_2 > 0$ and $0 \leq \Delta_2 < \Delta_1 \leq 1$. If $\Delta_1 < 1$, then $\mathcal{F}_n(\boldsymbol{X} = 2) \stackrel{O}{=} n^{-(\Delta_1 - \Delta_2)}$ a.s. If $\Delta_1 = 1$, then $\mathcal{F}_n(\boldsymbol{X} = 2) \stackrel{O}{\approx} n^{-(\Delta_1 - \Delta_2)}$ a.s.

To prove Theorem 1, denote $v_{\max} = \arg\max_{v \in \mathcal{V}} f(v)$. We first prove that $\mathcal{F}_n(\boldsymbol{X} = v_{\max}) \to 0$ with a power-law decay. Then, we remove $v_{\max}$ from the graph, and reconnect neighbors of $v_{\max}$, resulting in a new process (a subsequence of the original process) on $\mathcal{V} \setminus \{v_{\max}\}$. We then show that the new process satisfies the conditions of Theorem 1, and that the properties of the relative frequencies are unchanged by this procedure. The statement of the theorem will then follow by induction.

To show that $\mathcal{F}_n(\boldsymbol{X} = v_{\max}) \to 0$, we construct a two-state Markov chain $\boldsymbol{Y} = \{Y_n\}$ with state-space $\{1, 2\}$, coupled with $\boldsymbol{X}$, such that if $X_n = v_{\max}$ then $Y_n = 2$. This coupling property of $\boldsymbol{Y}$ ensures that $\mathcal{F}_n(\boldsymbol{X} = v_{\max}) \leq \mathcal{F}_n(\boldsymbol{Y} = 2)$, so that it suffices to analyze the convergence of $\mathcal{F}_n(\boldsymbol{Y} = 2)$. This analysis uses Lemma 2, which also allows us to conclude the statement of Theorem 2.

III. APPLICATIONS

In this section we show that generalized simulated annealing provides a unifying framework to study various stochastic optimization algorithms. In particular, we show that the classical simulated annealing algorithm, the "stochastic ruler" algorithm of Yan and Mukai [20], and the "stochastic comparison" algorithm of Gong et al. [21] are all special cases of generalized simulated annealing. In doing so, our convergence results can be brought to bear in the analysis of these algorithms. We show that our analysis in fact yields stronger results than are available for these algorithms. For the case of simulated annealing, a necessary and sufficient condition for convergence in probability is already available in Hajek [4], though as far as we know, rates on the relative frequencies have not been previously obtained.

A stochastic optimization algorithm aims to minimize a function $l(v)$ defined on a discrete set $\mathcal{V}$ via a stochastic search process. The search process gives rise to a non-homogeneous Markov chain of the kind that we will show fits within our framework. When showing that a particular stochastic optimization algorithm is a special case of generalized simulated annealing, we first relate the functions $g_r(u, v)$ and $g_c(u, v)$ with the transition probabilities of the stochastic algorithm. We then show how to define a function $f(v)$ that makes the process weakly reversible. In general, we cannot use $l(v)$ directly in place of $f(v)$ because the function $f(v)$ contains information about the rate of convergence,

while $l(v)$ might not. The needed modification to $l(v)$ to obtain $f(v)$ should become apparent in our discussion of the three examples in this section.

A. Simulated Annealing

Consider the problem of minimizing $l(v)$ with $v \in \mathcal{V}$. In simulated annealing, we begin with a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and define a non-homogeneous Markov chain $\{X_n\}$ with transition probability $P(X_n = v | X_{n-1} = u) = R(u,v)\exp(-[l(v) - l(u)]_+/T_n)$, where $[x]_+ = \max(x, 0)$, and $R(u,v)$ is a transition probability such that

$$R(u,v) \begin{cases} > 0 & \text{if } v \in \mathcal{N}_{out}(u), \\ = 0 & \text{otherwise,} \end{cases}$$

and $\sum_{v \in \mathcal{N}_{out}(u)} R(u,v) = 1$. The sequence $\{T_n\}$ is a positive sequence called the *cooling schedule*. We focus our attention on cooling schedules of the form $T_n = d/\log n$, popularized by Geman and Geman [3]. In the seminal paper of Hajek [4], he shows that $\{X_n\}$ converges in probability to the global minimizer if and only if $d \geq d^*$, where d^* is a quantity Hajek calls the "depth of the second deepest cup," a parameter we define precisely below. Our goal here is to show that, based on our main results, the same condition as Hajek's (involving d^* above) is also necessary and sufficient for convergence in the relative-frequency sense. Moreover, we provide a characterization of the rate of convergence of the relative frequencies.

To begin, consider a cooling schedule satisfying $T_n \sim d/\log n$, with $d > 0$ fixed. Then, the simulated annealing algorithm above is readily seen to be an instance of generalized simulated annealing with

$$\begin{aligned} f(v) &= \frac{l(v)}{d}, \\ g_r(u,v) &= \frac{[l(v) - l(u)]_+}{d}, \\ g_c(u,v) &= R(u,v). \end{aligned}$$

It remains to see when the height-normalization condition holds. To this end, denote the set of all paths from u to v by $\mathcal{P}(u,v)$. Then define $d^* = \max_{v \neq v_{\min}} \min_{p \in \mathcal{P}(v, v_{\min})} \max_{u \in p} \{l(u) - l(v)\}$. Here, $u \in p$ means that the vertex u is part of the path p.

To understand the connection between d^* and height normalization, first observe that by the definitions of f, g_r, and g_c given above, the quantity $h(v, v_{\min})$ (defined by (1) in Section II) simplifies to $h(v, v_{\min}) = \frac{1}{d} \min_{p \in \mathcal{P}(v, v_{\min})} \max_{u \in p} l(u)$. From this, it is easy to see that d^* can be rewritten as $d^* = d\left(\max_{v \neq v_{\min}} \{h(v, v_{\min}) - f(v)\}\right)$. We conclude that the process is height-normalized if and only if $d \geq d^*$.

Combining the above with Theorems 1 and 2 gives the following convergence theorem for simulated annealing.

Theorem 3: For simulated annealing with cooling schedule $T_n \sim d/\log n$, $\mathcal{F}_n(\boldsymbol{X} = v_{\min}) \to 1$ a.s. regardless of the starting point if and only if $d \geq d^*$. Moreover, assuming $d \geq d^*$, if $v \neq v_{\min}$ is visited infinitely often, then $\mathcal{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} n^{-(l(v) - l(v_{\min}))/d}$ a.s. regardless of the starting point.

The first part of Theorem 3 exactly parallels that of Hajek's (the necessary and sufficient condition for convergence is identical to that of Hajek [4]). This shows that convergence in probability (Hajek's result) is *equivalent* to a.s. convergence of the relative frequency. The second part of Theorem 3 characterizes the rate of convergence in terms of relative frequencies. As noted before, we can sharpen the rate result to $\mathcal{F}_n(\boldsymbol{X} = v) \overset{O}{=} n^{-(l(v) - l(v_{\min}))/d}$ for those v such that $h(v_{\min}, v) - f(v_{\min}) < 1$. As far as we know, these results on relative frequencies for simulated annealing have not previously been available.

The convergence result in Hajek [4] goes beyond the case where $T_n \sim d/\log n$. In particular, he also shows that if $T_n \log n \to 0$ then simulated annealing might converge to a "local" rather than global minimizer, and if $T_n \log n \to \infty$ then the algorithm converges to the global minimizer. Our framework addresses the case of $T_n \log n \to 0$ as one can show that the algorithm does not converge, using a coupling argument involving a generalized simulated annealing process that does not satisfy the conditions of Theorem 1. On the other hand, we do not directly recover the case of $T_n \log n \to \infty$. However, in this case a coupling argument with a generalized simulated annealing process shows that the rate of convergence is slower then any power; i.e., for all v, we have $\mathcal{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} 1$, suggesting that a cooling schedule for which $T_n \log n \to \infty$ should not be used.

B. Stochastic Ruler Algorithm

Yan and Mukai [20] consider the problem of minimizing an objective function $l(v)$, $v \in \mathcal{V}$, that is assumed to be of the form $l(v) = EH(v)$, where $H(v)$ is random with finite variance. They assume we do not actually have access to $l(v)$; instead, we can only observe independent samples (realizations) of $H(v)$. They convert the problem to one of maximizing $p(v,a,b) = P(H(v) \leq \Theta(a,b))$, where $\Theta(a,b)$ is a random variable uniformly distributed on (a,b) (and independent of $H(v)$). They prove that for a small enough and b large enough, any u that maximizes $p(u,a,b)$ also minimizes $l(v)$. (We assume henceforth that a and b are chosen such that this conclusion holds.)

To find the maximizer of $p(u,a,b)$ they set up a non-homogeneous Markov chain X_n satisfying $P(X_n = v | X_{n-1} = u) = R(u,v)(p(v,a,b))^{M_n}$, where $M_n \to \infty$ is called the "testing sequence." (It is useful to think of the testing sequence as the reciprocal of a cooling schedule.) As in simulated annealing, the probabilities $R(u,v)$ satisfy $R(u,v) > 0$ if and only if $v \in \mathcal{N}_{\text{out}}(u)$. Yan and Mukai [20] impose the additional restriction that the process is "strongly" reversible, i.e., $v \in \mathcal{N}_{out}(u)$ if and only if $u \in \mathcal{N}_{out}(v)$.

Yan and Mukai [20] consider the specific testing sequence $M_n = \lfloor \log(n + n_0 + 1)/d \rfloor$, where $\lfloor x \rfloor$ is the integer part of x, and n_0 and $d > 0$ are fixed constants. They show how to implement the search algorithm using only samples of $H(\cdot)$: suppose that at the nth iteration the process is in state u,

and a random candidate next-state v is generated according to $R(u,v)$. Then, generate $\lfloor \log(n+n_0+1)/d \rfloor$ independent samples (realizations) of $H(v)$ and $\Theta(a,b)$, and transition to v if and only if $H(v) \leq \Theta(a,b)$ for all the samples. It is convenient to call the above algorithm the *stochastic ruler* algorithm, because the samples of $H(v)$ are compared to a "stochastic ruler" $\Theta(a,b)$.

The main convergence result in Yan and Mukai [20] is that with the above testing sequence, provided some technical assumptions hold (which we elaborate below), $\{X_n\}$ converges in probability to the global minimizer. Below, we show that the stochastic ruler algorithm falls within the framework of generalized simulated annealing, and hence our relative-frequency convergence results apply, including a characterization of the convergence rates of the relative frequencies. Moreover, as we will see below, the technical assumptions in Yan and Mukai [20] can be weakened considerably—we provide a necessary and sufficient condition for convergence.

In our analysis, we consider the slightly more general case where the testing sequence $\{M_n\}$ satisfies $M_n \sim (\log n)/d$. In this case, we see that for $v \neq u$, $P(X_n = v | X_{n-1} = u) = R(u,v)(p(v,a,b))^{M_n} \sim R(u,v)n^{(\log p(v,a,b))/d}$. The transition probabilities of this non-homogeneous Markov chain suggest the following specialization of generalized simulated annealing:

$$\begin{aligned} f(v) &= \frac{-\log p(v,a,b)}{d}, \\ g_r(u,v) &= \frac{-\log p(v,a,b)}{d}, \\ g_c(u,v) &= R(u,v). \end{aligned}$$

Notice that $f(v) \geq 0$, and maximizing $p(v,a,b)$ is equivalent to minimizing $f(v)$. Moreover, the above choice guarantees weak reversibility of the resulting process.

As before, denote the set of all paths from u to v by $\mathcal{P}(u,v)$. Then define

$$d^* = \max_{v \neq v_{\min}} \min_{p \in \mathcal{P}(v,v_{\min})} \max_{uu' \in p} \{\log p(v,a,b) - \log p(u,a,b) - \log p(u',a,b)\}. \tag{4}$$

Here, $uu' \in p$ means that the link uu' is part of the path p. This value of d^* is analogous to Hajek's notion of the "depth of the second deepest cup" for simulated annealing Hajek [4]. It will turn out that d^* characterizes a necessary and sufficient condition for convergence of the stochastic ruler algorithm (see below). Therefore, although Yan and Mukai [20] are careful to point out that their approach is "different from the technique of simulated annealing," the analysis of simulated annealing actually bears on the analysis of the stochastic ruler algorithm, through our generalized simulated annealing framework.

To see why d^* plays the same role here as in simulated annealing, note that by the above definitions of f and g_r, we can once again write d^* in the form $d^* = d\left(\max_{v \neq v_{\min}} \{h(v,v_{\min}) - f(v)\}\right)$. and hence conclude that the process is height-normalized if and only if $d \geq d^*$.

Applying Theorems 1 and 2 to the stochastic ruler algorithm, we obtain the following convergence result.

Theorem 4: For the stochastic ruler algorithm with testing sequence $M_n \sim (\log n)/d$, $\mathcal{F}_n(\boldsymbol{X} = v_{\min}) \to 1$ a.s. regardless of the starting point if and only if $d \geq d^*$. Moreover, assuming $d \geq d^*$, if $v \neq v_{\min}$ is visited infinitely often, then $\mathcal{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} n^{-(\log p(v,a,b) - \log p(v_{\min},a,b))/d}$ a.s. regardless of the starting point.

C. Stochastic Comparison Algorithm

Gong et al. [21] consider a set up that is similar to that of Yan and Mukai [20], except that their Markov chain $\{X_n\}$ satisfies, for $v \neq u$, $P(X_n = v | X_{n-1} = u) = R(u,v)(P(H(v) < H(u)))^{M_n}$. So, unlike in Yan and Mukai [20], the transition probability from u to v here involves comparing $H(u)$ with $H(v)$ (instead of with an independent "ruler"). For this reason, Gong et al. [21] call their algorithm the *stochastic comparison* algorithm. Moreover, the graph in Gong et al. [21] satisfies, for all $u \in \mathcal{V}$, $\mathcal{N}_{\text{out}}(u) = \{v \in \mathcal{V} : v \neq u\}$. In other words, they assume a complete graph—any two vertices are connected with an edge (in both directions).

Gong et al. [21] analyze the convergence of their stochastic comparison algorithm using tools that are much the same as those of Yan and Mukai [20]. Specifically, they first assume that $H(v) = l(v) + W$, where W has zero mean, finite variance, and a symmetric density that does not depend on v. Then, under certain technical assumptions, they show that $\{X_n\}$ converges in probability to the global minimizer. Below, we show that the stochastic comparison algorithm also falls within the framework of generalized simulated annealing. As was the case in our analysis of the stochastic ruler algorithm, the technical assumptions in Gong et al. [21] can be weakened considerably—we provide a necessary and sufficient condition for convergence of the stochastic comparison algorithm. Our analysis also reveals significant differences between the stochastic comparison algorithm and the stochastic ruler algorithm.

Once again, we consider the slightly more general case where $M_n \sim (\log n)/d$. To simplify the notation, let F be the distribution function of $W_1 - W_2$, where W_1 and W_2 are independent random variables with the same density as W defined above. Then, $P(H(v) < H(u)) = F(l(u) - l(v))$. In this case, we see that for $v \neq u$,

$$\begin{aligned} P(X_n = v | X_{n-1} = u) &= R(u,v)F(l(u) - l(v))^{M_n} \\ &\sim R(u,v)n^{(\log F(l(u)-l(v)))/d}. \end{aligned}$$

The transition probabilities of this non-homogeneous Markov chain suggest the following correspondence with generalized simulated annealing:

$$g_r(u,v) = \frac{-\log F(l(u) - l(v))}{d}, \quad g_c(u,v) = R(u,v).$$

The definition of f to satisfy weak reversibility involves a little more work. First, order the vertices in ascending order according to their values of the objective function l; denote the ordered vertices by $v_{(1)}, \ldots, v_{(N)}$. Note that $v_{(1)} =$

$v_{\min}$ is the global minimizer. Then set $f(v_{(1)}) = 0$ and $f(v_{(j)}) = \min_{i\in\{1,\ldots,j-1\}} \big(f(v_{(i)}) + g_r(v_{(i)}, v_{(j)})\big) - g_r(v_{(j)}, v_{(1)})$. Because $g_r(v_{(j)}, v_{(1)}) \le g_r(v_{(j)}, v)$ for all v, and the definitions of f and h (see (1)) imply that $h(v_{(1)}, v_{(j)}) = f(v_{(j)}) + g_r(v_{(j)}, v_{(1)})$, we conclude that $h(v_{(j)}, v_{(1)}) = h(v_{(1)}, v_{(j)})$ for all $j = 1, \ldots, n$.

To show that the resulting process is weakly reversible, consider two vertices u and v. Consider a path $p = \{u, v_{(1)}, \ldots, v\}$ where $\{v_{(1)}, \ldots, v\}$ is a "minimal-height" path from $v_{(1)}$ to v (i.e., a path whose height is equal to $h(v_{(1)}, v)$). We see that $h(u, v) \le h(p) = \max\big(h(u, v_{(1)}), h(v_{(1)}, v)\big)$. On the other hand, consider a minimal-height path from u to v: $p' = \{u, w, \ldots, v\}$. The fact that $g_r(u, v_{(1)}) \le g_r(u, w)$ implies that $h(u, v) = h(p') \ge h(u, v_{(1)})$. Now consider a minimal-height path from $v_{(1)}$ to u: $q = \{v_{(1)}, \ldots, u\}$. Combining q with p', we get a path $q' = \{v_{(1)}, \ldots, u, w, \ldots, v\}$. Thus $h(v_{(1)}, v) \le h(q') = h(p') = h(u, v)$. Combining the above, we get $h(u, v) = \max\big(h(v_{(1)}, u), h(v_{(1)}, v)\big)$ and weak reversibility follows by symmetry.

Finally, define $d^* = -\log F(l(v_{(2)}) - l(v_{(1)}))$. As before, d^* characterizes a necessary and sufficient condition for convergence of the stochastic comparison algorithm. To elaborate, first note that for any node v, the path that goes directly from v to $v_{(1)}$ is a minimal-height path from v to $v_{(1)}$. Next, among all $v \ne v_{(1)}$, the height of this minimal-height path to $v_{(1)}$ is maximized for $v = v_{(2)}$ (because $v_{(2)}$ is the node with the lowest probability to transition to $v_{(1)}$). Hence, just as in the two previous examples, our choice of f and g_r allows us to write $d^* = d\left(\max_{v\ne v_{\min}}\{h(v, v_{\min}) - f(v)\}\right)$. from which we conclude once again that the process is height-normalized if and only if $d \ge d^*$. Hence, by applying Theorems 1 and 2, we obtain the following result.

Theorem 5: For the stochastic comparison algorithm with testing sequence $M_n \sim (\log n)/d$, $\mathcal{F}_n(\boldsymbol{X} = v_{\min}) \to 1$ a.s. regardless of the starting point if and only if $d \ge d^*$. Moreover, assuming $d \ge d^*$, if $v \ne v_{\min}$ is visited infinitely often, then $\mathcal{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} n^{-f(v)}$ a.s. regardless of the starting point.

In our rate result above, we have not explicitly provided an expression for $f(v)$, because such an expression would be complicated to state, given the definition of $f(\cdot)$. However, it is easy to bound the rate in terms of $l(\cdot)$:

$$\mathcal{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} n^{-f(v)} \le n^{(\log F(l(v_{(1)}) - l(v)) - \log F(l(v) - l(v_{(1)})))/d}.$$

In their convergence analysis, Gong et al. follow Yan and Mukai in setting $d = (\log\sigma)/c$, where $0 < c < 1$ and $\sigma \ge 1/\mu$. Here, $\mu = \min_{u\ne v} P(H(v) < H(u))$. (It is silently assumed in Gong et al. [21] that $0 < \mu < 1$.) Clearly, in this case, $d = \frac{\log\sigma}{c} > -\log\mu \ge -\log F(l(v_{(2)}) - l(v_{(1)})) = d^*$, which shows that the choice of d in Gong et al. [21] satisfies the condition $d \ge d^*$ of Theorem 5.

It is interesting to note that even though the original graph of Gong et al. is complete, we get only weak reversibility of our generalized simulated annealing process. Moreover, the minimal-height path between any two vertices always goes through the global minimizer. Thus, we can generalize our result to graphs that are not complete but where every node is a neighbor of the global minimizer. In this case, the function $f(\cdot)$ might not be a monotone transformation of $l(\cdot)$, in contrast to the case of a complete graph. This shows that generalization of this algorithm to graphs that are not complete might be non-trivial.

REFERENCES

[1] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[2] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, H. Teller, and E. Teller, "Equation of State Calculations by Fast Computing Machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[3] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intelligence*, vol. 6, pp. 721 – 741, 1984.

[4] B. Hajek, "Cooling schedules for optimal annealing," *Math. Oper. Res.*, vol. 13, no. 2, pp. 311–329, 1988.

[5] J. N. Tsitsiklis, "Markov chains with rare transitions and simulated annealing," *Math. Oper. Res.*, vol. 14, no. 1, pp. 70–90, 1989.

[6] D. P. Connors and P. R. Kumar, "Simulated annealing type Markov chains and their order balance equations," *SIAM J. Control Optim.*, vol. 27, no. 6, pp. 1440–1461, 1989.

[7] S. B. Gelfand and S. K. Mitter, "Simulated annealing type algorithms for multivariate optimization," *Algorithmica*, vol. 6, no. 3, pp. 419–436, 1991.

[8] C. Tsallis and D. A. Stariolo, "Generalized simulated annealing," *Phys. A*, p. 395, 1996.

[9] P. Del Moral and L. Miclo, "On the convergence and applications of generalized simulated annealing," *SIAM J. Control Optim.*, vol. 37, no. 4, pp. 1222–1250 (electronic), 1999.

[10] C. Cot and O. Catoni, "Piecewise constant triangular cooling schedules for generalized simulated annealing algorithms," *Ann. Appl. Probab.*, vol. 8, no. 2, pp. 375–396, 1998.

[11] A. Trouvé, "Convergence optimale pour les algorithmes de recuits généralisés," *C. R. Acad. Sci. Paris Sér. I Math.*, vol. 315, no. 11, pp. 1197–1202, 1992.

[12] S. R. Kulkarni and C. S. Horn, "An alternative proof for convergence of stochastic approximation algorithms," *IEEE Trans. Automat. Control*, vol. 41, no. 3, pp. 419–424, 1996.

[13] I.-J. Wang, E. K. P. Chong, and S. R. Kulkarni, "Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms," *Adv. in Appl. Probab.*, vol. 28, no. 3, pp. 784–801, 1996.

[14] ——, "Weighted averaging and stochastic approximation," *Math. Control Signals Systems*, vol. 10, no. 1, pp. 41–60, 1997.

[15] I.-J. Wang and E. K. P. Chong, "A deterministic analysis of stochastic approximation with randomized directions," *IEEE Trans. Automat. Control*, vol. 43, no. 12, pp. 1745–1749, 1998.

[16] E. K. P. Chong, I.-J. Wang, and S. R. Kulkarni, "Noise conditions for prespecified convergence rates of stochastic approximation algorithms," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 810–814, 1999.

[17] O. Catoni, "Rough large deviation estimates for simulated annealing: application to exponential schedules," *Ann. Probab.*, vol. 20, no. 3, pp. 1109–1146, 1992.

[18] M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*, ser. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. New York: Springer-Verlag, 1984, vol. 260, translated from the Russian by Joseph Szücs.

[19] J. Hannig, E. K. P. Chong, and S. R. Kulkarni, "Relative frequencies of generalized simulated annealing," *Mathematics of Operations Research*, to appear.

[20] D. Yan and H. Mukai, "Stochastic discrete optimization," *SIAM J. Control Optim.*, vol. 30, no. 3, pp. 594–612, 1992.

[21] W.-B. Gong, Y.-C. Ho, and W. Zhai, "Stochastic comparison algorithm for discrete optimization with estimation," *SIAM J. Optim.*, vol. 10, no. 2, pp. 384–404 (electronic), 2000.