MATHEMATICS OF OPERATIONS RESEARCH
Vol. 31, No. 1, February 2006, pp. 199–216
ISSN 0364-765X | EISSN 1526-5471 | 06 | 3101 | 0199

DOI 10.1287/moor.1050.0177

Relative Frequencies of Generalized Simulated Annealing

Jan Hannig
Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877,
jan.hannig@colostate.edu

Edwin K. P. Chong
Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado 80523-1373,
echong@colostate.edu

Sanjeev R. Kulkarni
Department of Electrical Engineering, Princeton University, Princeton, New Jersey 08544,
kulkarni@princeton.edu

We consider a class of nonhomogeneous 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.

Key words: simulated annealing; randomized optimization; coupling; relative frequency
MSC2000 subject classification: Primary: 65C40, 65C05, 60F15, 90C15; secondary: 60J05, 60J10, 60J20, 60J22
OR/MS subject classification: Primary: probability, Markov processes; secondary: programming, stochastic
History: Received January 19, 2005; revised June 13, 2005.

1. Introduction. For at least the last 20 years, there has been an interest in stochastic search algorithms for global optimization based on nonhomogeneous Markov chains. The prime example is *simulated annealing*, first suggested for optimization by Kirkpatrick et al. [12] based on techniques of Metropolis et al. [14]. An early application to image processing was described by Geman and Geman [9]. 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 nonhomogeneous Markov chain. The nonhomogeneity 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 §3).

In a seminal paper, Hajek [11] 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 [17] proves essentially the same result but using different techniques. Around the same time, Connors and Kumar [3] 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 [8] and Tsallis and Stariolo [16] consider a continuous-space version of simulated annealing, and Del Moral and Miclo [5], Cot and Catoni [4], and Trouvé [15] 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 nonhomogeneous 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 [11]. 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 nonhomogeneous Markov chains arising in simulated annealing and other stochastic search algorithms.

Our focus on relative frequencies in nonhomogeneous Markov chains 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 paper of Hajek [11], 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 nonglobal minimizers but also the *rate* at which these relative frequencies vanish.

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 [13]; Wang et al. [19], [20]; Wang and Chong [18]; Chong et al. [2]). 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 [11]: 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 [11] probabilistic analysis to characterize rates in addition to convergence. Since Hajek's [11] 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 [1] who uses results from Freidlin and Wentzell [7]). This paper and our previous work along similar lines suggest that the same is not the case in a purely sample path setting.

The rest of this paper is organized as follows. We begin below with some notation and terminology we will need and a brief discussion of relative frequencies. In §2, we define our generalized simulated annealing process and state our main results. In §3, we discuss three examples to show how our generalized simulated annealing framework applies to specific stochastic search algorithms. For convenience and ease of presentation, the proofs of our results are relegated to §4. We end with some final remarks in §5.

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

$$\mathscr{F}_n(\boldsymbol{x} \in A) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \in A).$$

If A is the singleton $\{v\}$, we write $\mathscr{F}_n(\boldsymbol{x} = v)$. Similarly, we use the notation $\mathscr{F}_n(\boldsymbol{x} \neq v) = \mathscr{F}_n(\boldsymbol{x} \notin \{v\})$.

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: $\boldsymbol{X} = \{X_1, X_2, \ldots\}$, $\mathscr{F}_n(\boldsymbol{X} = x^*)$, 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 $\mathscr{F}_n(\boldsymbol{X} = x^*)$. If the sequence is *independent*, then convergence in probability is stronger than its relative frequency counterpart, as illustrated in the simple lemma below. The proof of this lemma is of interest because it employs a technique we will use repeatedly in the proof of our main results (we elaborate on this below).

Lemma 1.1. *Let $\boldsymbol{X} = \{X_1, X_2, \dots\}$ be an independent, discrete state-space, random sequence that converges to x^* in probability. Then, $\mathscr{F}_n(\boldsymbol{X} = x^*) \to 1$ a.s.*

Proof. Suppose that $\boldsymbol{X}$ converges to x^* in probability, i.e., $P(X_n \neq x^*) \to 0$. Fix $\varepsilon > 0$. Then, $P(X_n \neq x^*) \leq \varepsilon$ eventually (for sufficiently large n). Let $\{U_n\}$ be an i.i.d. sequence with uniform distribution on $[0, 1]$, independent of $\boldsymbol{X}$. Define the Bernoulli sequence $\boldsymbol{B} = \{B_n\}$ by

$$B_n = I(X_n \neq x^*) + I(X_n = x^*) I\left(U_n \leq \frac{\varepsilon - P(X_n \neq x^*)}{P(X_n = x^*)}\right)$$

(if $P(X_n = x^*) = 0$, then take the second term to be 0; this issue disappears when n is sufficiently large). It is clear that $\{B_n\}$ is an independent sequence, and $P(B_n = 1) = \varepsilon$ for sufficiently large n. Moreover, if $X_n \neq x^*$, then $B_n = 1$ (a.s.), which implies that $\mathscr{F}_n(\boldsymbol{X} \neq x^*) \leq \mathscr{F}_n(\boldsymbol{B} = 1)$. By the strong law of large numbers, $\mathscr{F}_n(\boldsymbol{B} = 1) \to \varepsilon$ a.s. Therefore $\limsup_{n\to\infty} \mathscr{F}_n(\boldsymbol{X} \neq x^*) \leq \varepsilon$ a.s. Because this argument holds for all $\varepsilon > 0$, we conclude that $\mathscr{F}_n(\boldsymbol{X} \neq x^*) \to 0$ a.s., as required. □

We provide the proof above not because we believe the result to be original but to illustrate a method of proof that we will use repeatedly in proving our main results: *coupling*. This method involves considering a sequence that is related to another sequence such that some property involving sample paths of both sequences holds a.s. In the above proof, we explicitly constructed the sequence $\boldsymbol{B}$ from $\boldsymbol{X}$—we say that the sequence $\boldsymbol{B}$ is *coupled with* $\boldsymbol{X}$. The properties of $\boldsymbol{B}$ of interest to us here are that $P(B_n = 1) = \varepsilon$ and if $X_n \neq x^*$, then $B_n = 1$ (a.s.). This coupling property of $\boldsymbol{B}$ allows us to bound the relative frequency $\mathscr{F}_n(\boldsymbol{X} \neq x^*)$ by $\mathscr{F}_n(\boldsymbol{B} = 1)$, a quantity that is much easier to characterize. We use such coupling arguments repeatedly in proving our main results (see §4).

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 $\mathscr{F}_n(\boldsymbol{X} = x^*) \to 1$ a.s., but the sequence does not converge to x^* in probability. To see this, consider the sequence $\boldsymbol{X} = \{X_1, X_2, \dots\}$ on the state-space $\{0, 1\}$, where $X_n = 1$ a.s. for all n except for those n of the form $n = 2^k$, $k = 1, 2, \dots$, in which case $X_n = 0$ a.s. It is clear that $P(X_n = 1)$ does not converge to 1, but $\mathscr{F}_n(\boldsymbol{X} = 1) \to 1$ a.s.

In our generalized simulated annealing framework, the sequences are nonhomogeneous 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*.

2. Generalized simulated annealing.

In this section, we present our generalized simulated annealing framework and our main results. The proofs of these results will be provided in §4, after we give three example applications of our framework in §3.

Consider a finite, directed, connected graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V}$ is a set of vertices and $\mathscr{E}$ a set of directed edges. Assume that each vertex $v \in \mathscr{V}$ is assigned a value $f(v)$. Our goal is to find the minimum of the function f, i.e., we wish to find $v_{\min} \in \mathscr{V}$ such that $f(v_{\min}) \leq f(v)$ for all $v \in \mathscr{V}$.

We assume that all values of $f(v)$, $v \in \mathscr{V}$ are distinct. We make this assumption to simplify the presentation. In particular, under this assumption, $v_{\min} = \arg\min_{v\in\mathscr{V}} f(v)$ is unique. However, all our results remain valid with appropriate adjustments if we remove this assumption. We elaborate on this in our final remarks (§5).

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

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

and, as usual,

$$P(X_n = u \mid X_{n-1} = u) = 1 - \sum_{v\neq u} P(X_n = v \mid X_{n-1} = u).$$

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 §3, 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 $u \in \mathscr{V}$ two neighborhoods: $\mathscr{N}_{\mathrm{out}}(u) = \{v \neq u\colon uv \in \mathscr{E}\}$ and $\mathscr{N}_{\mathrm{in}}(u) = \{v \neq u\colon vu \in \mathscr{E}\}$. With this notation, we see that because probabilities must be bounded above by 1 for all u,

$$\sum_{v \in \mathscr{N}_{\mathrm{out}}(u)\colon\ g_r(u,v)=0} g_c(u,v) \leq 1.$$

We now describe an assumption that links the function $f(v)$ with the transition probabilities of $\{X_n\}$. 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 $u_{i+1} \in \mathscr{N}_{\mathrm{out}}(u_i)$, $i = 1, \ldots, k-1$. For a path $p = \{u, u_2, u_3, \ldots, u_{k-1}, v\}$, we define its *height* by

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

(This definition is motivated by the notion of "height" in Hajek [11] for simulated annealing.) For any two vertices u and v, we then define

$$h(u,v) = \min\bigl\{h(p)\colon p \text{ is a path from } u \text{ to } v\bigr\}. \tag{1}$$

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. Our notion of weak reversibility reduces to that of Hajek's [11] similarly named property for the special case of simulated annealing. Height normalization plays a key role in convergence.

Definition 2.1. 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 2.2. We say that the generalized simulated annealing process is *height normalized* if for any vertex $v \neq v_{\min}$, $h(v, v_{\min}) - f(v) \leq 1$.

Note that height normalization implies that the graph is connected. We are ready to give our main convergence result, which essentially states that height normalization is necessary and sufficient for global convergence of the process (i.e., a.s. convergence to a global minimizer regardless of initial condition) in the sense of relative frequencies.

Theorem 2.1. *Consider a weakly reversible generalized simulated annealing process* $\boldsymbol{X} = \{X_1, X_2, \ldots\}$. *If the process is height normalized, then* $\mathscr{F}_n(\boldsymbol{X} = v_{\min}) \to 1$ *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* $P(\mathscr{F}_n(\boldsymbol{X} = v) \to 1) > 0$ (*which implies that* $P(\mathscr{F}_n(\boldsymbol{X} = v_{\min}) \to 1) < 1$).

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

Theorem 2.2. *Consider a weakly reversible, height-normalized generalized simulated annealing process* $\boldsymbol{X} = \{X_1, X_2, \ldots\}$. *Suppose* v *is a vertex such that*

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

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

Otherwise, if (2) *is not satisfied but*

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

then $\mathscr{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} n^{-(f(v)-f(v_{\min}))}$ *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, in which case either* $\mathscr{F}_n(\boldsymbol{X} = v) = 0$ *or* $\mathscr{F}_n(\boldsymbol{X} = v) \overset{O}{=} n^{-1}$ *a.s.*

Remark 2.1. Suppose in the definition of the transition probabilities, we replace $g_c(u,v)n^{-g_r(u,v)}$ by $g_c(u,v)a_n^{-g_r(u,v)}$. If $a_n/n \to 0$, then convergence to the global minimizer holds, but at a slower rate. On the other hand, if $a_n/n \to \infty$, then the same conclusion as the converse to Theorem 2.1 holds—the process may not converge to the global minimizer.

Recall that "$\overset{O}{=}$" is stronger than "$\overset{O}{\approx}$." Thus, for simplicity, we can summarize the essence of Theorem 2.2 as follows: If v is visited infinitely often a.s., then $\mathscr{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} n^{-(f(v)-f(v_{\min}))}$ a.s. regardless of the starting point. In §3, we will use this simplified version of Theorem 2.2 in applying our framework to specific examples. The only thing missing from this simplified statement of the theorem is that in certain cases, we have the stronger result with "$\overset{O}{=}$" instead of "$\overset{O}{\approx}$."

The proofs of Theorems 2.1 and 2.2 are relegated to §4. In the next section, we describe three examples of stochastic search processes for which we can apply our results to characterize convergence and rates of convergence of relative frequencies.

3. 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 [21], and the "stochastic comparison" algorithm of Gong et al. [10] 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 [11], 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 nonhomogeneous 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. It is tempting, at first glance, to treat the objective function $l(v)$ as the function $f(v)$. However, 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. Some modification to $l(v)$ is needed to link it to the underlying graph. The needed modification to $l(v)$ to obtain $f(v)$ should become apparent in our discussion of the three examples in this section.

3.1. 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 nonhomogeneous Markov chain $\{X_n\}$ with transition probability

$$P\big(X_n = v \mid X_{n-1} = u\big) = R(u,v)\exp\big(-[l(v)-l(u)]_+/T_n\big),$$

where $[x]_+ = \max\{x, 0\}$, and $R(u,v)$ is a transition probability such that

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

and $\sum_{v\in\mathcal{N}_{\text{out}}(u)} R(u,v) = 1$. 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 [9]. In Hajek [11], shows that $\{X_n\}$ converges in probability to the global minimizer if, and only if, $d \ge d^*$, 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 §2) 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

$$d^* = d\Big(\max_{v \neq v_{\min}} \{h(v, v_{\min}) - f(v)\}\Big).$$

We conclude that the process is height normalized if, and only if, $d \geq d^*$.

Combining the above with Theorems 2.1 and 2.2 gives the following convergence theorem for simulated annealing. Note that weak reversibility in simulated annealing (in the Hajek [11] sense) implies that the corresponding generalized simulated annealing process is weakly reversible also.

THEOREM 3.1. *For a weakly reversible simulated annealing process 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 3.1 exactly parallels that of Hajek's (the necessary and sufficient condition for convergence is identical to that of Hajek [11]). This shows that convergence in probability (Hajek's result) is *equivalent* to a.s. convergence of the relative frequency. The second part of Theorem 3.1 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 [11] 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 might converge to a local minimizer, using a coupling argument involving a generalized simulated annealing process that does not satisfy the conditions of Theorem 2.1. 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.

3.2. Stochastic ruler algorithm. Yan and Mukai [21] 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 that 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 nonhomogeneous Markov chain X_n satisfying

$$P(X_n = v \mid X_{n-1} = u) = R(u, v)(p(v, a, b))^{M_n},$$

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 [21] impose the additional restriction that the graph has "symmetric neighborhoods," i.e., $v \in \mathcal{N}_{\text{out}}(u)$ if, and only if, $u \in \mathcal{N}_{\text{out}}(v)$. We adopt this assumption in the remainder of this section.

Yan and Mukai [21] 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: 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 [21] 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 [21] can be weakened—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

$$M_n \sim \frac{\log n}{d}.$$

In this case, we see that for $v \neq u$,

$$P(X_n = v \mid X_{n-1} = u) = R(u, v)(p(v, a, b))^{M_n} \sim R(u, v) n^{(\log p(v,a,b))/d}.$$

The transition probabilities of this nonhomogeneous 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 symmetric neighborhood assumption is sufficient for the above particular choice of $f(v)$ to guarantee 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

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

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 [11]. 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 [21] 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

$$d^* = d\Bigl(\max_{v \neq v_{\min}} \{h(v, v_{\min}) - f(v)\}\Bigr),$$

and hence conclude that the process is height normalized if, and only if, $d \geq d^*$.

By applying Theorems 2.1 and 2.2 to the stochastic ruler algorithm, we obtain the following convergence result.

Theorem 3.2. *For the stochastic ruler algorithm with testing sequence $M_n \sim (\log n)/d$ applied to a symmetric neighborhood graph, $\mathcal{F}_n(X = v_{\min}) \to 1$ 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 $\mathcal{F}_n(X = v) \overset{O}{\approx} n^{-(\log p(v,a,b) - \log p(v_{\min},a,b))/d}$ a.s. regardless of the starting point.*

We end with a brief discussion of some technical assumptions used in the analysis of Yan and Mukai [21]. First, some notation. Let $\mu(a, b) = \min_s p(s, a, b)$, and let r be the "radius" of the graph, i.e., $r = \min_s \max_{s'} d(s, s')$, where $d(s, s')$ is the number of edges in the shortest path from s to s'. Yan and Mukai [21] use a particular choice of d in their convergence analysis: $d = (\log \sigma)/c$, where $c \leq 1/r$ and $\sigma \geq 1/\mu(a, b)$. It is straightforward to show that this choice of d, in fact, satisfies the condition $d \geq d^*$ of Theorem 3.2. Hence our analysis shows that the assumptions of Yan and Mukai [21] can be weakened.

We note one more side benefit of our approach, besides improving the conditions under which the algorithm provably converges and giving us the rate of convergence of the relative frequencies: our approach gives us for free a direct generalization of this algorithm to nonsymmetric neighborhood graphs that still imply a weakly reversible generalized simulated annealing process.

3.3. Stochastic comparison algorithm. Gong et al. [10] consider a setup that is similar to that of Yan and Mukai [21], except that their Markov chain $\{X_n\}$ satisfies for $v \neq u$,

$$P(X_n = v \mid X_{n-1} = u) = R(u, v)(P(H(v) < H(u)))^{M_n}.$$

So, unlike in Yan and Mukai [21], 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. [10] call their algorithm the *stochastic comparison* algorithm. Moreover, the graph in Gong et al. [10] 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). We adopt this assumption in our analysis.

Gong et al. [10] analyze the convergence of their stochastic comparison algorithm using tools that are much the same as those of Yan and Mukai [21]. 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. [10] 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 $v \neq u$,

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

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

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

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

$$f(v_{(j)}) = \min_{i \in \{1, \ldots, j-1\}} \bigl\{f(v_{(i)}) + g_r(v_{(i)}, v_{(j)})\bigr\} - g_r(v_{(j)}, v_{(1)}).$$

Note that $g_r(v_{(j)}, v_{(1)}) \leq g_r(v_{(j)}, v)$ for all v (it is easier to go from $v_{(j)}$ to $v_{(1)}$, the global minimizer, than to any other v). This implies that the path of lowest height from $v_{(j)}$ to $v_{(1)}$ is the single-edge path. Therefore, by definition, $h(v_{(j)}, v_{(1)}) = f(v_{(j)}) + g_r(v_{(j)}, v_{(1)})$ for all $j = 1, \ldots, n$. On the other hand, depending on F, the path of lowest height from $v_{(1)}$ to $v_{(j)}$ may involve multiple edges. However, by induction on j, we can show that $h(v_{(1)}, v_{(j)}) = f(v_{(j)}) + g_r(v_{(j)}, v_{(1)})$, which implies 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

$$h(u, v) \leq h(p) = \max\bigl\{h(u, v_{(1)}), h(v_{(1)}, v)\bigr\}.$$

On the other hand, consider a minimal-height path from u to v: $p' = \{u, w, \ldots, v\}$. The fact that $g_r(u, v_{(1)}) \leq g_r(u, w)$ implies that

$$h(u, v) = h(p') \geq h(u, v_{(1)}).$$

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

$$h(v_{(1)}, v) \leq h(q') = h(p') = h(u, v).$$

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 $v \neq v_{(1)}$, 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

$$d^* = d\Big(\max_{v \neq v_{\min}} \{h(v, v_{\min}) - f(v)\}\Big).$$

from which we conclude once again that the process is height normalized if, and only if, $d \geq d^*$. Hence, by applying Theorems 2.1 and 2.2, we obtain the following result.

Theorem 3.3. *For the stochastic comparison algorithm with testing sequence $M_n \sim (\log n)/d$ applied to a complete graph, $\mathcal{F}_n(\boldsymbol{X} = v_{\min}) \to 1$ 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 $\mathcal{F}_n(\boldsymbol{X} = v) \overset{O}{\approx} n^{-f(v)}$ a.s. regardless of the starting point.*

In their convergence analysis, Gong et al. [10] follow Yan and Mukai [21] in setting $d = (\log \sigma)/c$, where $0 < c < 1$ and $\sigma \geq 1/\mu$. Here, $\mu = \min_{u \neq v} P(H(v) < H(u))$. (It is silently assumed in Gong et al. [10] that $0 < \mu < 1$.) Clearly, in this case,

$$d = \frac{\log \sigma}{c} > -\log \mu \geq -\log F(l(v_{(2)}) - l(v_{(1)})) = d^*,$$

which shows that the choice of d in Gong et al. [10] satisfies the condition $d \geq d^*$ of Theorem 3.3.

It is interesting to note that even though the original graph of Gong et al. [10] is complete, the resulting generalized simulated annealing process is only weakly reversible. 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, as in the complete graph case, the global minimizer of l is still the global minimizer of f. However, the function f might no longer be a monotone transformation of l, in contrast to the case of a complete graph. This has implications on rates. Specifically, if the transformation from l to f is monotone, then a (nonglobal minimizing) point v that has smaller $l(v)$ value also has a slower rate of decay of its relative frequency. However, if the transformation from l to f is not monotone, this natural monotone relationship between $l(v)$ and the rate of decay of the relative frequency of v no longer holds. This shows that generalization of this algorithm to graphs that are not complete might be nontrivial.

We also remark that we did not need any assumptions on the distribution function F, rendering Assumption 3.1 of Gong et al. [10] unnecessary.

4. Proofs.

This section is devoted entirely to the proofs of our main results: Theorems 2.1 and 2.2. To facilitate our presentation, we first state and prove several technical lemmas.

Lemma 4.1. *Let $0 < \Delta \leq 1$ and let $A_1, A_2, \ldots$ be independent events such that $P(A_n) = p_n \sim dn^{-\Delta}$. Then, a.s.*

$$\frac{1}{n}\sum_{i=1}^{n} I_{A_i} \sim \begin{cases} \dfrac{d}{1-\Delta} n^{-\Delta} & \text{if } 0 < \Delta < 1; \\ d(\log n) n^{-1} & \text{if } \Delta = 1. \end{cases} \tag{5}$$

Proof. Set

$$c_n = \frac{1}{n}\sum_{i=1}^{n} p_i, \quad Y_i = I_{A_i} - p_i, \quad \text{and} \quad S_n = \sum_{i=1}^{n} \frac{Y_i}{ic_i}.$$

We first consider the case where $0 < \Delta < 1$. Define $p'_n = dn^{-\Delta}$. By the integral approximation

$$c'_n = \frac{1}{n}\sum_{i=1}^{n} p'_i \sim \frac{d}{1-\Delta} n^{-\Delta}.$$

In general, if $p_n \sim d\,n^{-\Delta}$, then for all $0 < \epsilon < 1$ small enough,

$$(1-\varepsilon)p'_n < p_n < (1+\varepsilon)p'_n \quad \text{eventually}$$

and

$$(1-\varepsilon)c'_n < c_n < (1+\varepsilon)c'_n \quad \text{eventually.}$$

Thus $c_n \sim d(1-\Delta)^{-1}n^{-\Delta}$.

We now show that $\{S_n\}$ converges a.s. To do this, we verify the conditions of a special case of Kolmogorov's *three-series theorem* (see Durrett [6, Chapter 1, Theorem 8.3, p. 63]).

Notice that $EY_i/c_i = 0$. There is a constant C such that

$$\operatorname{Var}\left(\frac{Y_i}{ic_i}\right) = \frac{p_i(1-p_i)}{i^2c_i^2} \le \frac{C}{i^{2-\Delta}} \quad \text{eventually.}$$

Thus $\sum_{i=1}^{\infty} \operatorname{Var}(Y_i/(ic_i)) < \infty$, and the almost sure convergence of $\{S_n\}$ follows. Finally, Kronecker's lemma implies that $(\sum_{i=1}^{n} Y_i)/(nc_n) \to 0$ a.s. and

$$\frac{1}{nc_n}\sum_{i=1}^{n} I_{A_i} = \frac{1}{nc_n}\sum_{i=1}^{n} Y_i + \frac{1}{nc_n}\sum_{i=1}^{n} p_i \to 1 \text{ a.s.},$$

verifying the first part of (5).

To complete the proof, we have to consider the case where $\Delta = 1$. The only changes in this case are $c'_n \sim d(\log n)n^{-1}$ and

$$\operatorname{Var}\left(\frac{Y_i}{ic_i}\right) \le \frac{C}{i(\log i)^2} \quad \text{eventually.}$$

As this remains summable, the rest of the proof is unchanged. □

Lemma 4.2. *Let $\Delta \ge 0$ and let $G_1, G_2, \ldots$ be independent geometric random variables with parameters $p_1, p_2, \ldots,$ respectively. If $p_n \sim dn^{-\Delta}$, then, a.s.,*

$$\sum_{i=1}^{n} G_i \sim \frac{n^{1+\Delta}}{d(1+\Delta)}. \tag{6}$$

Proof. The proof is almost identical to that of Lemma 4.1. Set

$$c_n = \sum_{i=1}^{n} \frac{1}{p_i}, \quad Y_i = G_i - \frac{1}{p_i}, \quad \text{and} \quad S_n = \sum_{i=1}^{n} \frac{Y_i}{c_i}.$$

Define $p'_n = dn^{-\Delta}$. By the integral approximation

$$c'_n = \sum_{i=1}^{n} \frac{1}{p'_i} \sim \frac{n^{1+\Delta}}{d(1+\Delta)}.$$

In general, if $p_n \sim dn^{-\Delta}$, then for all $0 < \epsilon < 1$ small enough,

$$(1-\varepsilon)p'_n < p_n < (1+\varepsilon)p'_n \quad \text{eventually}$$

and

$$\frac{c'_n}{1+\varepsilon} < c_n < \frac{c'_n}{1-\varepsilon} \quad \text{eventually.}$$

We now show that $\{S_n\}$ converges a.s. To do this, once again we verify the conditions of the special case of Kolmogorov's three-series theorem (again, see Durrett [6, Chapter 1, Theorem 8.3, p. 63]).

Notice that $EY_i/c_i = 0$. There is a constant C such that

$$\operatorname{Var}\left(\frac{Y_i}{c_i}\right) = \frac{(1-p_i)}{p_i^2 c_i^2} \le \frac{C}{i^2} \quad \text{eventually.}$$

Thus $\sum_{i=1}^{\infty} \operatorname{Var}(Y_i/c_i) < \infty$, and the almost sure convergence of $\{S_n\}$ follows. Finally, Kronecker's lemma implies that $(\sum_{i=1}^{n} Y_i)/c_n \to 0$ a.s. and

$$\frac{1}{c_n}\sum_{i=1}^{n} G_i = \frac{1}{c_n}\sum_{i=1}^{n} Y_i + \frac{1}{c_n}\sum_{i=1}^{n}\frac{1}{p_i} \to 1 \text{ a.s.},$$

verifying (6). □

Remark 4.1. Let $\Delta > 0$ and let $G_1, G_2, \ldots$ be independent geometric random variables with parameters $p_1, p_2, \ldots$, respectively. If $p_n = de^{-\Delta n}$, then again Kolmogorov's three-series theorem implies that for all $\varepsilon > 0$,

$$\sum_{i=1}^{n} G_i \le e^{(1+\varepsilon)\Delta n} \quad \text{eventually a.s.}$$

To obtain a similar lower bound notice that for all $\varepsilon > 0$,

$$P(G_n < e^{(1-\varepsilon)\Delta n}) = 1 - (1 - de^{-\Delta n})^{e^{(1-\varepsilon)\Delta n}}.$$

Now, for sufficiently large n,

$$\log(1-de^{-\Delta n})^{e^{(1-\varepsilon)\Delta n}} = e^{(1-\varepsilon)\Delta n}\log(1-de^{-\Delta n}) \ge -e^{(1-\varepsilon)\Delta n} 2de^{-\Delta n} = -2de^{-\varepsilon\Delta n}.$$

Hence, for sufficiently large n,

$$P(G_n < e^{(1-\varepsilon)\Delta n}) \le 1 - e^{-2de^{-\varepsilon\Delta n}} \le 2de^{-\varepsilon\Delta n},$$

which is summable. The Borel-Cantelli lemma then implies that

$$\sum_{i=1}^{n} G_i \ge e^{(1-\varepsilon)\Delta n} \quad \text{eventually a.s.}$$

Of course, if $\Delta = 0$, then we have $\sum_{i=1}^{n} G_i \sim n/d$ a.s. by the strong law of large numbers.

The following lemma is the main technical instrument used in our proofs. It addresses the situation of a graph with two vertices.

Lemma 4.3. *Let $\{X_n\}$ be a nonhomogeneous Markov chain with state-space $\{1, 2\}$ and transition probabilities satisfying*

$$P(X_n = 2 \mid X_{n-1} = 1) \sim d_1 n^{-\Delta_1} \quad \text{and} \quad P(X_n = 1 \mid X_{n-1} = 2) \sim d_2 n^{-\Delta_2}.$$

Assume that $d_1, d_2 > 0$ and $0 \le \Delta_2 < \Delta_1 \le 1$.

If $\Delta_1 < 1$, then $\mathcal{F}_n(\boldsymbol{X} = 2) \overset{o}{=} n^{-(\Delta_1-\Delta_2)}$ a.s.

If $\Delta_1 = 1$, then $\mathcal{F}_n(\boldsymbol{X} = 2) \overset{o}{\approx} n^{-(\Delta_1-\Delta_2)}$ a.s.

Proof. Let τ_k^{21} denote the time when the process $\{X_n\}$ transitions from state 2 to state 1 for the kth time, with the convention that $X_0 = 2$. Next, let τ_k^{12} denote the time of the kth transition from state 1 to state 2. By definition, $1 \le \tau_1^{21} < \tau_1^{12} < \tau_2^{21} < \tau_2^{12} < \cdots$. For example, if $\{X_n, n \ge 1\} = \{1, 1, 2, \ldots\}$, then $\tau_1^{21} = 1$ and $\tau_1^{12} = 3$. For notational convenience, set $\tau_0^{12} = 1$.

To simplify our notation, let

$$p_{12,\,n} = P(X_n = 2 \mid X_{n-1} = 1) \quad \text{and} \quad p_{21,\,n} = P(X_n = 1 \mid X_{n-1} = 2).$$

The assumptions on the asymptotic behavior of $p_{12,\,n}$ and $p_{21,\,n}$ imply that $p_{12,\,n} < p_{21,\,n}$ eventually. Because our argument relies only on asymptotic properties, we can assume for convenience and without loss of generality that $p_{12,\,n} < p_{21,\,n}$ for all n.

Our first task is to estimate the asymptotic behavior of τ_k^{21}. To this end, let $\{U_n\}$ be an i.i.d. sequence with uniform distribution on [0, 1], independent of $\{X_n\}$. Next, define a sequence of Bernoulli random variables $\{B_n\}$, coupled with $\{X_n\}$, as follows: $B_1 = 1$, and for $n \geq 2$,

$$B_n = I(X_{n-1} = 1, X_n = 2) + I(X_{n-1} = 2, X_n = 1) I(U_n \leq p_{12,\,n}/p_{21,\,n}). \tag{7}$$

Let V_k denote the kth time when $B_n = 1$. The sequence $\{B_n\}$ satisfies the following properties:

(i) $\{B_n\}$ is an independent sequence;
(ii) $P(B_n = 1) \sim d_1 n^{-\Delta_1}$; and
(iii) $V_k \leq \tau_k^{21} \leq V_{2k}$ a.s.

Property (i) follows easily from the Markov property of $\{X_n\}$. Property (ii) follows from the fact that $P(B_n = 1) = p_{12,\,n}$, which is also easy to verify.

We now show that property (iii) holds. To get $\tau_k^{21} \geq V_k$, it suffices to show that $\tau_{k-1}^{12} \geq V_k$ (recall that $\tau_k^{21} > \tau_{k-1}^{12}$ by definition). To see this, notice that if X_n makes a transition from 1 to 2 ($X_{n-1} = 1$ and $X_n = 2$), then $B_n = 1$. Hence, by time τ_{k-1}^{12}, the number of times that $B_n = 1$ had already occurred is at least k (recall that $B_1 = 1$). This implies that $\tau_{k-1}^{12} \geq V_k$, as desired.

To show that $\tau_k^{21} \leq V_{2k}$, notice that if $B_n = 1$, then X_n had to make a transition: either $X_{n-1} = 1$ and $X_n = 2$, or $X_{n-1} = 2$ and $X_n = 1$. Therefore, by time V_{2k}, the process $\{X_n\}$ had to undergo at least $2k$ transitions, and hence at least k transitions from state 2 to state 1. This implies that $\tau_k^{21} \leq V_{2k}$, as desired.

Having established bounds for τ_k^{21} based on V_k (lower bound) and V_{2k} (upper bound), we can characterize the asymptotic behavior of τ_k^{21} by characterizing the behavior of V_k.

We first consider the case where $0 < \Delta_1 < 1$. After inverting the result of Lemma 4.1, we get

$$V_k \sim \left(\frac{1-\Delta_1}{d_1} k\right)^{1/1-\Delta_1} \quad \text{a.s.} \tag{8}$$

Choose a nondecreasing sequence of random variables $\{k_n\}$ satisfying $\tau_{k_n}^{21} \leq n < \tau_{k_n+1}^{21}$. The lower bound $\tau_k^{21} \geq V_k$ and Lemma 4.1 imply that for all $\varepsilon > 0$,

$$k_n \leq \frac{d_1(1+\varepsilon)}{1-\Delta_1} n^{1-\Delta_1} \quad \text{eventually a.s.} \tag{9}$$

The upper bound $\tau_k^{21} \leq V_{2k}$ implies that for all $\varepsilon > 0$,

$$k_n \geq \frac{d_1(1-\varepsilon)}{2(1-\Delta_1)} n^{1-\Delta_1} \quad \text{eventually a.s.} \tag{10}$$

We now can estimate the relative frequency $\mathscr{F}_n(\boldsymbol{X} = 2)$. Notice that

$$\mathscr{F}_n(\boldsymbol{X} = 2) = \frac{\text{time spent in 2}}{n}.$$

Thus, all we need to do is to bound the time spent in 2, i.e., $\sum_{i=1}^k (\tau_i^{21} - \tau_{i-1}^{12})$ (recall $\tau_0^{12} = 1$), from above and from below.

To bound $\sum_{i=1}^k (\tau_i^{21} - \tau_{i-1}^{12})$ from above, first fix a small $0 < \varepsilon < 1$. For $k = 1, 2, \ldots,$ set

$$\tilde{p}_k = \frac{d_2}{1+\varepsilon} \left(\frac{(1-\varepsilon)d_1}{2(1-\Delta_1)}\right)^{\Delta_2/1-\Delta_1} k^{-\Delta_2/1-\Delta_1}.$$

Note that for $n = \tau_{k-1}^{12} + 1, \ldots, \tau_k^{21}$, we have $p_{12,\,n} < \tilde{p}_k < p_{21,\,n}$ for sufficiently large k (a.s.), by the upper bound $\tau_k^{21} \leq V_{2k}$ and (8). We now construct a sequence $\{\widetilde{G}_k\}$, coupled with $\{X_n\}$, such that

(i) $\{\widetilde{G}_k\}$ is an independent sequence;
(ii) $\widetilde{G}_k$ has geometric distribution with parameter $\tilde{p}_k$; and
(iii) $\tau_k^{21} - \tau_{k-1}^{12} \leq \widetilde{G}_k$ eventually a.s.

To define $\widetilde{G}_k$ (for a given sufficiently large k), let $\{U_n^k\}$ be an i.i.d. sequence with uniform distribution on [0, 1], independent of $\{X_n\}$ (and independent across k). Then, define the Bernoulli sequence $\{C_n^k\}$ as follows. For $n = \tau_{k-1}^{12} + 1, \ldots, \tau_k^{21}$, set

$$C_n^k = I(X_{n-1} = 2, X_n = 1) I\left(U_n \le \frac{\tilde{p}_k}{p_{21,\,n}}\right) + I(X_{n-1} = 1, X_n = 2) + I(X_{n-1} = 1, X_n = 1) I\left(U_n \le \frac{\tilde{p}_k - p_{12,\,n}}{1 - p_{12,\,n}}\right).$$

For all other n, set $C_n^k = I(U_n \le \tilde{p}_k)$. Note that $\{C_n^k\}$ is an independent sequence (by the strong Markov property) and $P(C_n^k = 1) = \tilde{p}_k$. Now, define $\widetilde{G}_k = \min\{n \ge \tau_{k-1}^{12} + 1 : C_n^k = 1\} - \tau_{k-1}^{12}$. By the strong Markov property, $\{\widetilde{G}_k\}$ is an independent sequence (property (i)), and $\widetilde{G}_k$ has geometric distribution with parameter $\tilde{p}_k$ (property (ii)). To show that property (iii) holds, note that for any $n = \tau_{k-1}^{12} + 1, \ldots, \tau_k^{21} - 1$, if $X_{n-1} = 2$ and $X_n = 2$, then $C_n^k = 0$ (by definition). This implies that $\widetilde{G}_k \ge \tau_k^{21} - \tau_{k-1}^{12}$, as desired.

Recall the sequence $\{k_n\}$ introduced above, satisfying $\tau_{k_n}^{21} \le n < \tau_{k_n+1}^{21}$. Using Lemma 4.2 and Equations (8), (9), and (10), we get

$$\mathscr{F}_n(\boldsymbol{X} = 2) \le \frac{\widetilde{C} + \sum_{i=1}^{k_n} \widetilde{G}_i}{n} \overset{o}{=} n^{-(\Delta_1 - \Delta_2)} \quad \text{a.s.,} \tag{11}$$

where $\widetilde{C}$ is a random constant.

To bound $\sum_{i=1}^{k} (\tau_i^{21} - \tau_{i-1}^{12})$ from below, define a sequence $\{\bar{p}_k\}$ (analogous to $\{\tilde{p}_k\}$ above) as follows:

$$\bar{p}_k = \frac{d_2}{1-\varepsilon} \left(\frac{(1+\varepsilon)d_1}{1-\Delta_1}\right)^{\Delta_2/1-\Delta_1} k^{-\Delta_2/1-\Delta_1}.$$

As before, the lower bound $V_k \le \tau_k^{21}$ and (8) imply that for $n = \tau_{k-1}^{12} + 1, \ldots, \tau_k^{21}$, we have $\bar{p}_k > p_{21,\,n}$ for sufficiently large k (a.s.). Therefore we can use a construction similar to the one before to define an independent sequence $\{\overline{G}_k\}$, coupled with $\{X_n\}$, such that $\tau_k^{21} - \tau_{k-1}^{12} \ge \overline{G}_k$ eventually (a.s.), and $\overline{G}_k$ has geometric distribution with parameter $\bar{p}_k$. Again, using Lemma 4.2 and Equations (8), (9), and (10), we get

$$\mathscr{F}_n(\boldsymbol{X} = 2) \ge \frac{\overline{C} + \sum_{i=1}^{k_n} \overline{G}_i}{n} \overset{o}{=} n^{-(\Delta_1 - \Delta_2)} \quad \text{a.s.,} \tag{12}$$

where $\overline{C}$ is another random constant.

Combining (11) and (12), we get the statement of the lemma for the case where $0 < \Delta_1 < 1$.

Consider now the case where $\Delta_1 = 1$. Because we now have an exponential growth, we can no longer invert the result of Lemma 4.1, nor does the upper bound $\tau_k^{21} \le V_{2k}$ yield a sharp enough estimate. However, in this case, we can use the lower bound $V_k \le \tau_k^{21}$ and the estimates provided in Remark 4.1 to get, for all $\varepsilon > 0$,

$$e^{k/((1+\varepsilon)d_1)} \le \tau_k^{21} \le e^{k/((1-\varepsilon)d_1)} \quad \text{eventually a.s.}$$

Therefore we can replace (9) and (10) by

$$(1-\varepsilon)d_1 \log n \le k_n \le (1+\varepsilon)d_1 \log n \quad \text{eventually a.s.,}$$

and define

$$\tilde{p}_k = e^{-\Delta_2(k/((1-\varepsilon)d_1))}, \quad \bar{p}_k = e^{-\Delta_2((k-1)/((1+\varepsilon)d_1))}.$$

The rest of the proof can be done in the same fashion as before, replacing the use of Lemma 4.2 with Remark 4.1. □

Remark 4.2. If $0 \le \Delta_2 \le 1 < \Delta_1$, then the Borel-Cantelli lemma implies that the process $\boldsymbol{X} = \{X_n\}$ will eventually move to state 1 and stay there forever. In this case, $\mathscr{F}_n(\boldsymbol{X} = 2) \overset{o}{=} n^{-1}$.

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. Denote $v_{\max} = \arg\max_{v \in \mathscr{V}} f(v)$. We first prove that $\mathscr{F}_n(\boldsymbol{X} = v_{\max}) \to 0$ with a power-law decay. Then, we remove vertex $v_{\max}$ from the graph, and reconnect neighbors of $v_{\max}$, resulting in a new process (a subsequence of the original process) on $\{v_{\max}\}^{\complement} = \mathscr{V} \setminus \{v_{\max}\}$. We then show that the new process satisfies the conditions of Theorem 2.1, and that the properties of the relative frequencies are unchanged by this procedure. The statement of the theorem will then follow by mathematical induction.

To show that $\mathscr{F}_n(\boldsymbol{X} = v_{\max}) \to 0$, we will 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 $\mathscr{F}_n(\boldsymbol{X} = v_{\max}) \le \mathscr{F}_n(\boldsymbol{Y} = 2)$, so that it suffices to analyze the convergence of $\mathscr{F}_n(\boldsymbol{Y} = 2)$.

To construct the process $\boldsymbol{Y}$, first define

$$u_r = \mathop{\arg\min}_{u \in \mathcal{N}_{\text{in}}(v_{\max})} g_r(u, v_{\max}), \quad u_c = \mathop{\arg\max}_{u \in \mathcal{N}_{\text{in}}(v_{\max})} g_c(u, v_{\max}).$$

and

$$v_r = \mathop{\arg\min}_{v \in \mathcal{N}_{\text{out}}(v_{\max})} g_r(v_{\max}, v), \quad v_c = \mathop{\arg\min}_{v \in \mathcal{N}_{\text{out}}(v_{\max})} g_c(v_{\max}, v). \tag{13}$$

By the assumption of the theorem, we have $0 \le g_r(v_{\max}, v_r) \le 1$. Because $f(v) \le f(v_{\max})$ for all the neighbors v of $v_{\max}$, we conclude by weak reversibility that $g_r(v_{\max}, v_r) < g_r(u_r, v_{\max})$.

For convenience, let $p^X_{v_{\max}, n} = P(X_n \neq v_{\max} \mid X_{n-1} = v_{\max})$ and $p^X_{v,n} = P(X_n = v_{\max} \mid X_{n-1} = v)$. Define the sequences $\{p^Y_{12,n}\}$ and $\{p^Y_{21,n}\}$ such that

$$p^Y_{12,n} \sim g_c(u_c, v_{\max}) n^{-g_r(u_r, v_{\max})} \quad \text{and} \quad p^Y_{21,n} \sim g_c(v_{\max}, v_c) n^{-g_r(v_{\max}, v_r)}.$$

Since the conditions above are only asymptotic, we can choose $\{p^Y_{12,n}\}$ and $\{p^Y_{21,n}\}$, so that the following is true not only asymptotically but for all n:

$$\begin{aligned} p^Y_{21,n} &\le p^X_{v_{\max},n}, \\ p^Y_{12,n} &\ge p^X_{v,n} \quad \text{for all } v \neq v_{\max}, \text{ and} \\ p^Y_{21,n} &\le 1 - \sum_{v \neq v_{\max}} p^X_{v,n}. \end{aligned}$$

Next, let $\{U_n\}$ be an i.i.d. sequence with uniform distribution on $[0, 1]$, independent of $\boldsymbol{X}$. Define the Bernoulli sequences $\{B_n\}$ and $\{C_n\}$ as follows:

$$B_n = I(X_{n-1} = v_{\max}) I(U_n \le p^Y_{12,n}) + \sum_{v \neq v_{\max}} \left[I(X_{n-1} = v, X_n = v_{\max}) + I(X_{n-1} = v, X_n \neq v_{\max}) I\left(U_n \le \frac{p^Y_{12,n} - p^X_{v,n}}{1 - p^X_{v,n}} \right) \right]$$

and

$$C_n = I(X_{n-1} = v_{\max}, X_n \neq v_{\max}) I(U_n \le p^Y_{21,n} / p^X_{v_{\max},n}) + \sum_{v \neq v_{\max}} I(X_{n-1} = v, X_n \neq v_{\max}) I\left(U_n \le \frac{p^Y_{21,n}}{1 - p^X_{v,n}} \right).$$

A simple calculation using the Markov property of $\{X_n\}$ shows that the sequence of vectors $\{(B_n, C_n)\}$ is independent. Moreover, it is easy to verify that $P(B_n = 1) = p^Y_{12,n}$ and $P(C_n = 1) = p^Y_{21,n}$.

We are now ready to define the Markov chain $\{Y_n\}$ (with state-space $\{1, 2\}$), using $\{B_n\}$ and $\{C_n\}$. First, set $Y_1 = 2$ if, and only if, $X_1 = v_{\max}$. If $Y_{n-1} = 1$, then set $Y_n = 2$ if, and only if, $B_n = 1$. Similarly, if $Y_{n-1} = 2$, then set $Y_n = 1$ if, and only if, $C_n = 1$. Hence, by construction, $P(Y_n = 2 \mid Y_{n-1} = 1) = p^Y_{12,n}$ and $P(Y_n = 1 \mid Y_{n-1} = 2) = p^Y_{21,n}$.

The process $\{Y_n\}$ constructed above has the property that if $X_n = v_{\max}$, then $Y_n = 2$, as desired. We show this by induction on n. For $n = 1$, the property holds by definition of Y_1. Assume the property holds for $n - 1$. We now show that if $X_n = v_{\max}$, then $Y_n = 2$. So suppose that $X_n = v_{\max}$. There are three cases to consider, depending on the values of X_{n-1} and Y_{n-1}; in each case, we show that $Y_n = 2$.

Case 1. $X_{n-1} \neq v_{\max}$ and $Y_{n-1} = 1$. In this case, we have $B_n = 1$, and hence $Y_n = 2$.

Case 2. $X_{n-1} = v_{\max}$ and $Y_{n-1} = 2$. Here, we have $C_n = 0$, and hence $Y_n = 2$.

Case 3. $X_{n-1} \neq v_{\max}$ and $n - 1_n = 2$. Here, again we have $C_n = 0$, and hence $Y_n = 2$. (The fourth case where $X_{n-1} = v_{\max}$ and $Y_{n-1} = 1$ is precluded based on the induction hypothesis.) So the desired property now follows by induction.

Having constructed $\boldsymbol{Y} = \{Y_n\}$ such that $\mathscr{F}_n(\boldsymbol{X} = v_{\max}) \le \mathscr{F}_n(\boldsymbol{Y} = 2)$, we are now ready to show that $\mathscr{F}_n(\boldsymbol{X} = v_{\max}) \to 0$. By construction, the transition probabilities of $\{Y_n\}$ satisfy

$$P(Y_n = 2 \mid Y_{n-1} = 1) \sim g_c(u_c, v_{\max}) n^{-g_r(u_r, v_{\max})}$$

and

$$P(Y_n = 1 \mid Y_{n-1} = 2) \sim g_c(v_{\max}, v_c) n^{-g_r(v_{\max}, v_r)}.$$

If $g_r(u_r, v_{\max}) \le 1$, then Lemma 4.3 implies that

$$\mathscr{F}_n(\boldsymbol{X} = v_{\max}) \le \mathscr{F}_n(\boldsymbol{Y} = 2) \overset{O}{\approx} n^{-(g_r(u_r, v_{\max}) - g_r(v_{\max}, v_r))} \quad \text{a.s.}$$

On the other hand, if $g_r(u_r, v_{\max}) > 1$, Remark 4.2 gives

$$\mathcal{F}_n(\boldsymbol{X} = v_{\max}) \le \mathcal{F}_n(\boldsymbol{Y} = 2) \overset{o}{=} n^{-1} \quad \text{a.s.}$$

In either case,

$$\mathcal{F}_n\big(\boldsymbol{X} \in \{v_{\max}\}^{\complement}\big) \to 1 \quad \text{a.s.} \tag{14}$$

Let us now define η_n as the nth time the process $\boldsymbol{X}$ is in $\{v_{\max}\}^{\complement}$. Clearly, $\eta_n \ge n$ and (14) implies that $\eta_n \sim n$. Set $X'_n = X_{\eta_n}$. The new process $\{X'_n\}$ is a Markov chain on the graph with $v_{\max}$ removed and with edges added between all the neighbors of $v_{\max}$. Note that because $\eta_n \sim n$, we have $\mathcal{F}_n(\boldsymbol{X}' = v) \sim \mathcal{F}_n(\boldsymbol{X} = v)$, for all $v \ne v_{\max}$.

We now turn our attention to the transition probabilities of this new Markov chain $\{X'_n\}$ (we use "primes" for the notation here, e.g., $g'_c(u, v)$ and $h'(u, v)$). If there is a vertex from u to v and either $u \notin \mathcal{N}_{\text{in}}(v_{\max})$ or $v \notin \mathcal{N}_{\text{out}}(v_{\max})$, then the transition probability from u to v remains unchanged, i.e., $g'_r(u, v) = g_r(u, v)$ and $g'_c(u, v) = g_c(u, v)$.

Now, consider $u \in \mathcal{N}_{\text{in}}(v_{\max})$ and $v \in \mathcal{N}_{\text{out}}(v_{\max})$, $u \ne v$, and assume first that there is no direct edge between u and v. Then, a.s.

$$\begin{aligned} P(X'_n = v \mid X'_{n-1} = u) &\sim g_c(u, v_{\max}) n^{-g_r(u, v_{\max})} \frac{g_c(v_{\max}, v) n^{-g_r(v_{\max}, v)}}{\sum_{w \in \mathcal{N}_{\text{out}}(v_{\max})} g_c(v_{\max}, w) n^{-g_r(v_{\max}, w)}} \\ &\sim g'_c(u, v) n^{-g'_r(u, v)}, \end{aligned} \tag{15}$$

where $g'_r(u, v) = g_r(u, v_{\max}) + g_r(v_{\max}, v) - g_r(v_{\max}, v_r)$ (v_r was defined in (13)), and the a.s. set on which (15) holds is the same as in (14). The value of $g'_c(u, v)$ is determined by (15), though the exact formula could be complicated. If there is already an existing link from u to v, then we have to add its associated transition probability with the one in (15), in which case

$$g'_r(u, v) = \min\big\{g_r(u, v), g_r(u, v_{\max}) + g_r(v_{\max}, v) - g_r(v_{\max}, v_r)\big\}. \tag{16}$$

We now prove that for any two vertices u and v that are not the vertex we removed, we have $h(u, v) = h'(u, v)$. Since the only change to the function g_r is in the neighborhood of $v_{\max}$, it is enough to consider $h'(u, v)$ for $u \in \mathcal{N}_{\text{in}}(v_{\max})$ and $v \in \mathcal{N}_{\text{out}}(v_{\max})$.

First, notice that Equation (16) implies that

$$g'_r(u, v_r) = g_r(u, v_{\max}).$$

Then, use the weak reversibility of the original process and the definition of v_r to conclude that the minimal-height path from v_r back to $v_{\max}$ must have height $f(v_{\max}) + g_r(v_{\max}, v_r)$. Suppose that this path is $\{v_r, \dots, \bar{u}_r, v_{\max}\}$, i.e., this path enters $v_{\max}$ from the vertex $\bar{u}_r$. By weak reversibility and the definition of v_r, we have $f(v_{\max}) + g_r(v_{\max}, v_r) = f(\bar{u}_r) + g_r(\bar{u}_r, v_{\max})$, and (16) gives

$$f(\bar{u}_r) + g'_r(\bar{u}_r, v) = f(v_{\max}) + g_r(v_{\max}, v).$$

Finally, we see that again weak reversibility and the definition of v_r implies that

$$f(u) + g(u, v_{\max}) \ge f(v_{\max}) + g_r(v_{\max}, v_r).$$

Thus we conclude that

$$h(\{u, v_{\max}, v\}) = h'(\{u, v_r, \dots, \bar{u}_r, v\}).$$

The fact that $h(u, v) = h'(u, v)$ now follows.

The first part of Theorem 2.1 now follows by repeating the above argument until there is only one vertex left.

To prove the second part of Theorem 2.1, suppose that the process is not height normalized. We can remove vertices from the graph until we get a graph with a maximal vertex $v_{\max}$ for which $1 < g_r(v_{\max}, v_r) < g_r(u_r, v_{\max})$. Then the Borel-Cantelli lemma implies that we get trapped at $v_{\max}$ with positive probability. The second part of the theorem then follows with v being this particular $v_{\max}$. □

We now turn to the proof of Theorem 2.2.

Proof of Theorem 2.2. Fix a vertex v. Recall the proof of Theorem 2.1. There, we used an iterative procedure to remove the maximal vertex at each iteration, consisting of two steps. First, we showed that the relative frequency of the maximal vertex goes to zero; then, we showed that the graph with the maximal vertex removed still satisfies our assumptions.

The reason for removing the maximal vertex and not another vertex is to facilitate the first step. However, once we know that Theorem 2.1 holds, we can remove any $u \neq v_{\min}$ using the same argument as in the second part of the proof of Theorem 2.1. In particular, we can remove vertices in such a way that the last two remaining vertices are $v_{\min}$ and v. The transition probabilities on the remaining graph will satisfy

$$P(X'_n = v \mid X'_{n-1} = v_{\min}) \sim d_1 n^{-\Delta_1} \quad \text{and} \quad P(X'_n = v_{\min} \mid X'_{n-1} = v) \sim d_2 n^{-\Delta_2},$$

where

$$\Delta_1 = \min\{h(p) - f(v_{\min})\colon p \text{ is a path from } v_{\min} \text{ to } v\}$$

and

$$\Delta_2 = \min\{h(p) - f(v)\colon p \text{ is a path from } v \text{ to } v_{\min}\}.$$

The statement of Theorem 2.2 now follows from Lemma 4.3 and Remark 4.2. □

5. Final remarks. We have shown that it is possible to use elementary arguments to analyze the convergence of relative frequencies in a class of generalized simulated annealing processes. Our analysis characterizes not only when the relative frequencies converge, but also at what rate. The class of processes that falls under our framework includes the classical simulated annealing algorithm, and two other stochastic search algorithms not previously connected closely to simulated annealing: the stochastic ruler algorithm and the stochastic comparison algorithm. In particular, our results provide necessary and sufficient conditions for convergence, and a characterization of the convergence rates, for the relative frequencies in these algorithms.

In our analysis, we have assumed, for convenience, that the objective function values on the given graph are distinct. However, as pointed out before, the nondistinct case can be handled in much the same way. Specifically, a careful look at the proof of Theorem 2.1 shows that if the values of $f(v)$, $v \in \mathcal{V}$ are not distinct, we get the convergence result in Theorem 2.1* below. We need the following generalization of the notion of height normalization.

Definition 2.2*. Denote the set of global minimizers by $\mathcal{V}_{\min} = \arg\min_{v\in\mathcal{V}} f(v)$. We say that the generalized simulated annealing process is height normalized if for any vertex $v \notin \mathcal{V}_{\min}$, there is a $v_{\min} \in \mathcal{V}_{\min}$ such that $h(v, v_{\min}) - f(v) \leq 1$.

Note that unlike Definition 2.2, this generalization to the notion of height normalization no longer guarantees a connected graph. However, even in an unconnected case, height normalization implies that each connected subgraph contains a global minimizer.

Theorem 2.1*. *Consider a weakly reversible generalized simulated annealing process* $\boldsymbol{X} = \{X_1, X_2, \ldots\}$. *If the process is height normalized, then* $\mathcal{F}_n(\boldsymbol{X} \in \mathcal{V}_{\min}) \to 1$ *a.s. regardless of the starting point.*

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

The main idea here is the fact that multiple global minimizers will only cause us to stop the pruning of the graph described in the proof of Theorem 2.1 sooner, and the theorem will then follow by coupling with a process that lumps all the global minimizers together.

Another technical difficulty arises when, at some iteration, the set of global maximizers $\mathcal{V}_{\max} = \arg\max_{v\in\mathcal{V}} f(v)$ has more then one element, as it might not be clear which point to prune out first. In this case, we can consider a coupling with a process that perturbs the values of f on $\mathcal{V}_{\max}$ in such a way that for $v_{\max} \in \mathcal{V}_{\max}$, the values of $f(v_{\max})$ are distinct, and for $v \notin \mathcal{V}_{\max}$, $f(v_{\max}) > f(v)$. We also adjust the values of g_r in such a way that the heights of all links remain the same. One can do this perturbation in such a way that it is harder for the new process to escape $\mathcal{V}_{\max}$ than it was for the original process. This coupling will establish that $\mathcal{F}_n(\boldsymbol{X} \in \mathcal{V}_{\max}) \to 0$, and we can then prune the elements of $\mathcal{V}_{\max}$ in the usual fashion.

The generalization of Theorem 2.2 to the nondistinct case is more complicated. This is caused by the fact that some of the global maximizers might be visited only finitely many times. To state the generalization, we need the following definition.

Definition 5.1. A collection of vertices $\mathcal{B}$ is called a basin if (1) there is $v_{\min} \in \mathcal{V}_{\min}$ such that for any $v \in \mathcal{V}$ satisfying $h(v_{\min}, v) \leq 1$, $v \in \mathcal{B}$ and (2) for any $u, v \in \mathcal{B}$, $h(u, v) \leq 1$.

Notice that every basin contains at least one global minimizer because $h(v_{\min}, v_{\min}) = 0$. Moreover, no subset or superset of $\mathcal{B}$ satisfies conditions (1) and (2) above. Thus the set of vertices $\mathcal{V}$ decomposes into one or more basins around the global minimizers plus vertices that are not included in any basin. The Borel-Cantelli lemma implies that we will eventually enter one of the basins $\mathcal{B}$ and never leave it a.s. The probability that a particular basin will "trap" the process depends on the starting point. If we remove all the vertices that are not part of $\mathcal{B}$, then the analysis of Theorem 2.2 will apply with minor modifications. Specifically, if $\mathcal{B}$ contains only one global minimizer, the proof of Theorem 2.2 applies as is. If $\mathcal{B}$ contains multiple global minimizers (denote them by $\mathcal{V}'_{\min} = \mathcal{V}_{\min} \cap \mathcal{B}$), we need to couple the process $\boldsymbol{X}$ with two different processes, described next.

For a fixed $v \in \mathcal{B}$, we will prune the graph using Theorem 2.1* until we are left only with v and $\mathcal{V}'_{\min}$. Then, we first couple our process $\boldsymbol{X}$ with a process $\boldsymbol{X}'$ living on a two-vertex graph with all of $\mathcal{V}'_{\min}$ collapsed into a single vertex. This can be done so that $\boldsymbol{X}'$ spends less time in v than the process $\boldsymbol{X}$ does, which leads to

$$\mathcal{F}_n(\boldsymbol{X} = v) \geq \mathcal{F}_n(\boldsymbol{X}' = v).$$

Second, we can consider a coupling with a process $\boldsymbol{X}''$ that increases all but one of the values of f on $\mathcal{V}'_{\min}$ in such a way that for $v_{\min} \in \mathcal{V}'_{\min}$, the value of $f(v_{\min})$ is unique and $f(v_{\min}) < f(v)$. We also adjust the values of g_r in such a way that the heights of all links remain the same. One can do this perturbation in such a way that it is easier for the new process to escape $\mathcal{V}_{\min}$ than it was for the original process. Thus process $\boldsymbol{X}''$ spends more time in v than the process $\boldsymbol{X}$ does, which leads to

$$\mathcal{F}_n(\boldsymbol{X} = v) \leq \mathcal{F}_n(\boldsymbol{X}'' = v).$$

Notice that $\boldsymbol{X}'$ and $\boldsymbol{X}''$ have distinct f values so we can apply Theorem 2.2. This leads to the following generalization of Theorem 2.2.

THEOREM 2.2*. *Consider a weakly reversible, height-normalized generalized simulated annealing process* $\boldsymbol{X} = \{X_1, X_2, \ldots\}$. *Then, a.s., there is a basin* $\mathcal{B}$ *in which the process is eventually trapped. This basin* $\mathcal{B}$ *satisfies the following.*

If $v \in \mathcal{B}$ and $\min_{v_{\min} \in \mathcal{V}'_{\min}} \{h(v_{\min}, v) - f(v_{\min})\} < 1$, then $\mathcal{F}_n(\boldsymbol{X} = v) \overset{o}{=} n^{-(f(v) - f(v_{\min}))}$.

If $v \in \mathcal{B}$ and $\min_{v_{\min} \in \mathcal{V}'_{\min}} \{h(v_{\min}, v) - f(v_{\min})\} = 1$, then $\mathcal{F}_n(\boldsymbol{X} = v) \overset{o}{\approx} n^{-(f(v) - f(v_{\min}))}$.

Finally, if $v \notin \mathcal{B}$, *then* v *is visited finitely often, in which case either* $\mathcal{F}_n(\boldsymbol{X} = v) = 0$ or $\mathcal{F}_n(\boldsymbol{X} = v) \overset{o}{=} n^{-1}$.

It is worth pointing out that if the conditions of Theorem 2.2* are satisfied and there are multiple basins, then we get a generalized simulated annealing process that is convergent but, in general, not ergodic. For example, consider the case of a connected graph with multiple basins. If the process is height normalized, then for each basin, there is a positive probability of being trapped in that basin.

Acknowledgments The work of Hannig was supported in part by an IBM Faculty Award and by the National Science Foundation Grant DMS-0504737. The work of Chong was supported in part by the National Science Foundation Grants 0098089-ECS, 0099137-ANI, and ANI-0207892. The work of Kulkarni was supported in part by Army Research Office Grant DAAD19-00-1-0466, Draper Laboratory IR&D 6002 Grant DL-H-546263, and National Science Foundation Grant CCR-0312413.

References

[1] Catoni, O. 1992. Rough large deviation estimates for simulated annealing: Application to exponential schedules. *Ann. Probability* **20**(3) 1109–1146.

[2] Chong, E. K. P., I.-J. Wang, S. R. Kulkarni. 1999. Noise conditions for prespecified convergence rates of stochastic approximation algorithms. *IEEE Trans. Inform. Theory* **45**(2) 810–814.

[3] Connors, D. P., P. R. Kumar. 1989. Simulated annealing type Markov chains and their order balance equations. *SIAM J. Control Optim.* **27**(6) 1440–1461.

[4] Cot, C., O. Catoni. 1998. Piecewise constant triangular cooling schedules for generalized simulated annealing algorithms. *Ann. Appl. Probability* **8**(2) 375–396.

[5] Del Moral, P., L. Miclo. 1999. On the convergence and applications of generalized simulated annealing. *SIAM J. Control Optim.* **37**(4) 1222–1250.

[6] Durrett, R. 1996. *Probability: Theory and Examples*, 2nd ed. Duxbury Press, Belmont, CA.

[7] Freidlin, M. I., A. D. Wentzell, 1984. Random perturbations of dynamical systems. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*], Vol. 260. Springer-Verlag, New York. [J. Szücs, trans.]

[8] Gelfand, S. B., S. K. Mitter. 1991. Simulated annealing type algorithms for multivariate optimization. *Algorithmica* **6**(3) 419–436.

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

[10] Gong, W.-B., Y.-C. Ho, W. Zhai. 2000. Stochastic comparison algorithm for discrete optimization with estimation. *SIAM J. Optim.* **10**(2) 384–404.

[11] Hajek, B. 1988. Cooling schedules for optimal annealing. *Math. Oper. Res.* **13**(2) 311–329.

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

[13] Kulkarni, S. R., C. S. Horn. 1996. An alternative proof for convergence of stochastic approximation algorithms. *IEEE Trans. Automatic Control* **41**(3) 419–424.

[14] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, H. Teller, E. Teller. 1953. Equation of State Calculations by Fast Computing Machines. *J. Chemical Phys.* **21**(6) 1087–1092.

[15] Trouvé, A. 1992. Convergence optimale pour les algorithmes de recuits généralisés. *C. R. Acad. Sci. Paris Sér. I Math.* **315**(11) 1197–1202.

[16] Tsallis, C., D. A. Stariolo. 1996. Generalized simulated annealing. *Physica A* **233**(1–2) 395–406.

[17] Tsitsiklis, J. N. 1989. Markov chains with rare transitions and simulated annealing. *Math. Oper. Res.* **14**(1) 70–90.

[18] Wang, I.-J., E. K. P. Chong. 1998. A deterministic analysis of stochastic approximation with randomized directions. *IEEE Trans. Automat. Control* **43**(12) 1745–1749.

[19] Wang, I.-J., E. K. P. Chong, S. R. Kulkarni. 1996. Equivalent necessary and sufficient conditions on noise sequences for stochastic approximation algorithms. *Adv. Appl. Probability* **28**(3) 784–801.

[20] Wang, I.-J., E. K. P. Chong, S. R. Kulkarni. 1997. Weighted averaging and stochastic approximation. *Math. Control Signals Systems* **10**(1) 41–60.

[21] Yan, D., H. Mukai. 1992. Stochastic discrete optimization. *SIAM J. Control Optim.* **30**(3) 594–612.