SMART SIMULATED ANNEALING

by

George Winfield Heine, III

B. A., Reed College, 1971

M. S., University of Colorado, 1989

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado at Denver

in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Applied Mathematics

1994
This thesis for the Doctor of Philosophy

degree by

George Winfield Heine, III

has been approved for the

Department of

Mathematics

by

Bennett L. Fox

James Koehler

Burton Simon

Richard Davis

Thomas Russell

Richard Holley

Date April 27, 1994
Heine, III, George Winfield (Ph.D., Applied Mathematics) 
Smart Simulated Annealing 
Thesis directed by Professor Bennett L. Fox 

A naive implementation of simulated annealing is inefficient; at low temperatures it spends almost all its time proposing moves and then rejecting them. We present an algorithm, due to Fox, which reduces time at self-loops while preserving convergence properties. We show that the computer time spent in each distinct state by Fox's algorithm is bounded (a.s.) by a constant which can be determined in advance.

This result suggests a refinement of Fox's algorithm: when computer time at a given state exceeds the predetermined constant, exit to another state. A technical lemma needed to prove the refinement works has surprising corollaries. For example, we use the lemma to analyze simulated annealing when the objective function is not precisely known and must be estimated.

Another way of speeding up simulated annealing keeps temperature constant while in a state, updating when the state changes. Thus sojourn time in a state is easily simulated by sampling from a geometric distribution. Using methods of Connors and Kumar, we show that in this scenario the proportion of time spent in an optimal state converges to one almost surely.

It might be thought that the various convergence results in simulated annealing are an artifact, since at low temperatures the algorithm spends large blocks of time at optimal states. However, we look at the embedded chain of pairwise distinct states and find that it too exhibits convergence, though in general to a slightly different set of states.
This abstract accurately represents the content of the candidate's thesis. I recommend its publication.

Signed\[signature\]

Bennett L. Fox
ACKNOWLEDGEMENTS

Many people have contributed, directly or indirectly, to the completion of this thesis. Many U.C.D. faculty members have given me inspiration, technical advice, and moral support; I would like to especially thank Bill Briggs, Stan Payne, and Rich Lundgren, as well as the members of my thesis committee. Of course, the greatest thanks are due my advisor, Ben Fox. This project would not have been possible without his knowledge and insight, as well as his constant support.

Special thanks goes to Arlan Ramsay at C.U.-Boulder, both for having suggested I investigate the U.C.D. graduate program in the first place, and, together with his son, Bruce, for having written the software program which produced this thesis.

Thanks also to management and co-workers at the Bureau of Land Management, who have cheerfully put up with so many irregularities in my schedule.

I also gratefully acknowledge the unwavering faith in my ability to succeed given by friends and family over the years, especially my parents, George and Mary Heine, my sister Nancy and her family, and my in-laws, Charles and Audrey Barnett. Most of all, I am indebted to my wife, Janet Barnett, whose unswerving love has been put to the test in many difficult moments. It is to her that I dedicate this thesis.
# CONTENTS

## CHAPTER

1 INTRODUCTION ................................................. 1

1.1 The Simulated Annealing Algorithm .................. 2

1.2 Previous Work ........................................... 4

2 CONVERGENCE OF A LOOP-SKIPPING ALGORITHM ...... 7

2.1 Summary of Results .................................... 9

2.2 Proofs of Theorems .................................... 12

2.2.1 Proof of Theorem 2.1 ............................... 15

2.2.2 Proof of Theorem 2.2 ............................... 16

2.2.3 Proof of Theorem 2.3 ............................... 17

2.3 Efficiency of Fox’s Algorithm ....................... 20

3 SIMULATED ANNEALING WITH OVERRIDES ............ 25

3.1 Simple Examples ..................................... 27

3.1.1 Random Restart .................................. 27

3.1.2 Greedy Overrides ................................ 27

3.1.3 Perverse Overrides ................................ 28

3.2 Dynamic Neighborhoods ............................... 28

3.3 Overrides: The Main Result ............................ 30

4 SIMULATED ANNEALING WITH A NOISY OBJECTIVE

FUNCTION .................................................... 34

4.1 Formal Setup .......................................... 35
CHAPTER 1

INTRODUCTION

Simulated annealing is attracting increasing attention as a method of combinatorial optimization. As knowledge of the algorithm spreads, researchers find ever more diverse uses, reporting applications such as refinement of protein structures [35], redrawing of legislative districts [5], in-core nuclear fuel management [33], and contaminated groundwater remediation [34].

In certain senses and under certain conditions, made explicit below, the Markov chain generated by simulated annealing is convergent. However, the algorithm has been criticized for being slow and expensive. Perhaps this is partly because most theoretical studies have given little attention to practical ways to implement the algorithm.

The results in this thesis were originally motivated by attempts to find an efficient algorithm which maintains the desirable convergence properties. In the process, we have discovered additional theoretical results, interesting in their own right.

To probabilists, simulated annealing presents an interesting special case of a time-inhomogeneous Markov chain. The properties of stationary Markov chains have long been well-understood, but theory for the general nonstationary case still appears far off. Nonstationary Markov chains are better understood today than a few years ago, partly because of research in simulated
annealing. We hope that the results outlined in this thesis will contribute to the growth of knowledge in this important area.

1.1 The Simulated Annealing Algorithm

The goal of simulated annealing is to minimize an objective function $U$ on a set of states $S$. We take $S$ to be finite unless the contrary is explicitly stated. The algorithm generates a sequence of moves between successive neighboring states. At each step of the algorithm, a candidate move is proposed. If the move is to a state with lower or equal objective-function value, it is accepted; if the move is to a state with a higher objective-function value, it is accepted with a small probability.

We now formalize these notions. By problem we mean a triple $(S, U, R)$, where $S$ is a (finite) set of states, $U$ is a real-valued function on $S$, and $R$ is the $|S| \times |S|$ stochastic matrix of tentative-move probabilities. We assume without loss of generality that the minimum of $U$ over $S$ is 0. Sometimes we use energy level as shorthand for objective-function value.

Simulated annealing is a (generally nonstationary) Markov chain $\{X_t : t = 1, 2, \ldots\}$ on a set of states $S$. The transition probabilities are given by

$$\Pr(X_{t+1} = y | X_t = x) = R(x, y) e(t)^{(U(y) - U(x))^{\frac{1}{t}}}$$

for $x \neq y$, and

$$\Pr(X_{t+1} = x | X_t = x) = 1 - \sum_{y \neq x} \Pr(X_{t+1} = y | X_t = x),$$

where $e(\cdot)$ is implied by the cooling schedule, described below.

Denote by $S_0$ the set of states on which $U$ is minimized. The neighborhood $N(x)$ of a state $x$ consists of those states $y$ for which $R(x, y) > 0$. A
**path** is a sequence of states $x_0, \ldots, x_p$ such that $x_i \in N(x_{i-1})$ for $1 \leq i \leq p$.

As is customary, we assume that $R$ is irreducible, so that any two states are connected by a path. Following Chiang and Chow [7], we say that state $y$ is **reachable from** $x$ at **height** $h$ if there exists a path $x = x_0, x_1, \ldots, x_p = y$ such that $\max_{0 \leq i \leq p} U(x_i) \leq h$. A state $x$ is a **local minimum** if no state $y$ with $U(y) < U(x)$ can be reached from $x$ at height $U(x)$. For each pair $(x, y)$ of distinct states, let $h(x, y)$ be the smallest number $h$ such that $y$ is reachable from $x$ at height $h$. The **depth** $d(x)$ of a state is defined as

$$d(x) = \begin{cases} 
\min h(x, y), U(y) < U(x) & \text{if } x \notin S_0; \\
\min h(x, y), y \in S_0 & \text{if } x \in S_0.
\end{cases}$$

The problem $(S, U, R)$ has the property of **neighborhood symmetry** ($NS$) if $R(x, y) > 0$ whenever $R(y, x) > 0$. In most cases this property is unnecessarily strong and can be replaced with **weak reversibility** ($WR$), which requires that if a state $y$ is reachable from state $x$ at height $h$, then $x$ is also reachable from $y$ at height $h$. Assuming $WR$, as we do throughout, rather than $NS$ has important practical implications: Fox [14] shows how to integrate simulated annealing and tabu penalties by redefining $S$, $U$, and $R$; with additional structure, given in [14], the redefined problem satisfies $WR$, but it is generally not possible to satisfy $NS$.

The **cooling schedule** is a sequence $(T_k), k = 1, 2, \ldots$ of positive numbers, usually converging to zero, such that the probability of accepting an uphill move of height one at step $k$ is

$$\epsilon(k) = \exp(-1/T_k).$$

We find it more convenient to deal directly with the sequence $(\epsilon(k))$; by abuse
of terminology, we also call this sequence the cooling schedule; this causes no ambiguity, because the $\epsilon(k)$ and $T_k$ determine each other.

In most of what follows, we restrict our attention to schedules of the form

$$\epsilon(k) = k^{-1/c},$$

for $c > 0$, or more generally, to schedules satisfying

$$c_1(k + a)^{-1/c} \leq \epsilon(k) \leq c_2(k + a)^{-1/c}$$

for positive constants $a$, $c$, $c_1$ and $c_2$. These are the schedules that have attracted the most attention historically, at least among probabilists. See [6] and [21] for discussions of other kinds of schedules.

In the work that follows, we are often concerned with the dynamics of exit from a single state $x$; the most significant quantity in this analysis is how high the chain must climb to leave $x$. Denote this quantity by

$$\delta(x) = \min_{y \in N(x)} (U(y) - U(x))^+;$$

note that $\delta(x) > 0$ only if every neighbor of $x$ has an objective-function value strictly larger than $U(x)$. We call such a state an isolated local minimum.

1.2 Previous Work

A central problem of simulated annealing is how slowly the cooling schedule must go to zero in order to guarantee convergence to $S_0$. Geman and Geman [20] showed that, for problems satisfying $NS$, a constant $c$ exists such that

$$\sum_{t=1}^{\infty} \epsilon(t)^c = \infty \quad \Rightarrow \quad \lim_{t \to \infty} \Pr(X_t \in S_0) \to 1.$$
Hajek [25] identifies the constant as

\[ d^* = \max_{x \notin S_0} d(x), \]

where \( d(\cdot) \) is the depth function defined in Section 1.1. He requires \((S, U, R)\) to satisfy \( WR \) but not necessarily \( NS \). Hajek shows that his conditions are necessary as well as sufficient. Hwang and Sheu [30] prove essentially the same theorem by an application of the “cycles” of Freidlin and Wentzell [18]. Chiang and Chow [7],[8] repeat the result in continuous time, and (with certain restrictions) give the asymptotic probability distribution for large \( t \). Tsitsiklis [37] obtains similar results by working with small perturbations of probability matrices. Connors and Kumar [11], [12] show that, under similar conditions, the \( \lim \sup \) of the Cesàro average of \( \Pr(X_k \in S_0) \) is one; we return to this point in Section 5.1.

All of these results assume a finite state space. Although there is a large body of literature about simulated annealing on continuous state spaces, we are not aware of any exact results relating the speed of cooling to the existence of convergence. Bélysle [2] shows that convergence occurs on finite or infinite state spaces provided \( S_0 \) is reachable in a single step from anywhere in the state space. It is easy to modify his proofs to show that a slightly weaker condition suffices: namely, from anywhere in \( S \) there is a path with no uphill moves leading to \( S_0 \). Bélysle’s condition is strong enough that his result holds for any cooling schedule that approaches zero, or even a random schedule that approaches zero in probability.

The rest of this thesis focuses on efficiently implementing simulated
annealing without losing the convergence properties just discussed. In Chapter 2, we describe Fox’s [14] algorithm QUICKER. Using it eliminates explicitly rejected moves while leaving the sample path stochastically unchanged. We show that the computer time to execute QUICKER is asymptotically bounded (a.s.) by a constant that can be prescribed in advance. This complements Fox’s analysis of quadratic-mean convergence. In Chapter 3, we use these results to propose an improved algorithm QUICKER-j. A technical lemma needed to prove the convergence of QUICKER-j turns out to be useful in an important and little-studied problem: simulated annealing when the objective function is not precisely known; we analyze this situation in Chapter 4. In Chapter 5, we analyze an algorithm, equivalent to QUICKER-1, which amounts to using a random cooling schedule. It exhibits a different sort of convergence: the proportion of observed optimal states tends almost surely to one. In Chapter 6, we consider the embedded chain of pairwise distinct states (i.e., the chain \( X_k \), pruned of its self-loops) and show that converges to a set \( S_1 \). We relate \( S_0 \) and \( S_1 \). Chapter 7 indicates directions for future research.
CHAPTER 2

CONVERGENCE OF A LOOP-SKIPPING ALGORITHM

Implementing simulated annealing in a naive but straightforward fashion, the processing at step $t$ looks like:

**Algorithm SLOW($x, t$)**

1. Select $y$ from $N(x)$
2. If $U(y) \leq U(x)$ then
   - Set $X_{t+1} \leftarrow y$
3. Else
   - Generate a uniform random variate $V$
   - If $V < e(t)^{[U(y)-U(x)]}$, then
     - Set $X_{t+1} \leftarrow y$
   - Else
     - Set $X_{t+1} \leftarrow x$

Implemented on a computer, Algorithm SLOW is inefficient; the computer spends almost all of its time proposing moves and then rejecting them. Moreover, the longer the simulation runs, the worse the inefficiency becomes. Fox (preprint of [14]) notes that the algorithm is a “victim of its own success”: when the probability of accepting an uphill move gets small, the algorithm stays, with high probability, at local minima for very long periods of time. Algorithm SLOW, as well as its modifications discussed in this thesis, may also consume a lot of time in oscillating between adjacent states. Fox [14] inhibits this, while preserving convergence in probability, by integrating tabu penalties.
We discuss this further in Section 7.2.

Greene and Supowit [24] eliminate rejections by forcing the algorithm to change states at every step. The new state is chosen from the neighbors of the current state via a distribution conditioned on an accepted move. However, their method fails to update temperatures; thus, it destroys the convergence properties described in the previous chapter.

Fox [14, 15] proposes an algorithm QUICKER which greatly reduces the computer time in self-loops, yet (because simulated time is unaffected) updates temperatures in the right way to preserve convergence to the optimal states. Suppose that the algorithm enters state $x$ at time $k$. The point of QUICKER is to generate

$$L = \min\{l \geq k : X_l \neq X_k\},$$

the transition number on which a move to a different state next occurs. Let $\alpha(x, k)$ be the probability that the algorithm exits state $x$ at time $k + 1$; that is,

$$\alpha(x, k) = \Pr(X_{k+1} \neq x \mid X_k = x).$$

We assume $\alpha(x, k)$ is nonincreasing in $k$ for any fixed $x$. QUICKER is based on the observation that, to a first approximation, $L$ has the same distribution as a geometric variate with parameter $\alpha(x, k)$ (i.e., the trial number of the first success in iid Bernoulli trials with success probability $\alpha(x, k)$). The algorithm follows.

**Algorithm QUICKER** $(x, t)$
Set $k \leftarrow t$
Until exit, repeat
Generate a geometric r.v. $G$ with parameter $\alpha(x, k)$
Set $L \leftarrow k + G - 1$
Generate a uniform r.v. $V$
If $V < \alpha(x, L)/\alpha(x, k)$ then
  Execute NEXT($x, L$)
  Exit
Else
  Set $k \leftarrow L$
End

Above, “Exit” means go to the next state. The subroutine NEXT($x, k$) selects a state from $N(x)$, conditioned on a move away from $x$ at time $k$.

Fox [15] shows that the number of repetitions of the “Until...repeat” loop required by each invocation of QUICKER converges in quadratic mean to 1, provided $c(t) = t^{-1/c}$ and $c$ is strictly larger than a constant $\bar{d}$, to be introduced shortly. He shows by example that in general no convergence occurs when $c = \bar{d}$. We give an almost-sure counterpart, with interesting consequences.

Fox also shows that at isolated local minima QUICKER reduces computer time by an asymptotically infinite factor. We sharpen this observation by giving upper and lower bounds for this time reduction.

2.1 Summary of Results

As in the previous section, $\alpha(x, k)$ denotes the probability of accepting a move out of state $x$ at time $k$. We often suppress dependence on $x$, writing simply $\alpha_k$. 
On any given sample path \( \omega \), let

\[
Q(x, k, \omega) = \begin{cases} 
\text{the number of geometric variates (number of inner} \\
\text{loop iterations) generated by QUICKER if state } x \\
\text{is entered at time } k; \\
0, \text{ otherwise.}
\end{cases}
\tag{2.1}
\]

Again, we often suppress dependence on \( \omega \) or \( x \) and write \( Q(x, k) \) or just \( Q_k \).

Recall the function \( \delta(x) \) defined on page 4. Let

\[
\bar{d} = \max_{x \in S} \delta(x) = \max_{x \in S} \min_{y \in N(x)} (U(y) - U(x))^+. 
\]

There is in general no simple relationship between \( \bar{d} \) and Hajek’s constant \( d^* \); they are maxima of different functions over different sets. However, the case of greatest interest is probably that of a problem with a local, nonglobal minimum \( x \) whose depth \( d(x) \) is at least as great as the depth of all global optimizers.

In this case, \( \bar{d} \leq d^* \), and a necessary condition for \( \Pr(X_k \in S_0) \to 1 \) is that \( \sum_{k=1}^{\infty} \epsilon(k)^2 = \infty \). When \( \epsilon(k) = k^{-1/c} \), this is equivalent to \( \bar{d}/c \leq 1 \).

In contrast to \( d^* \), the constant \( \bar{d} \) depends for its value only on the immediate neighborhood of each local minimum. This reveals something of the flavor of QUICKER; its improvement over SLOW is greatest at those local minima \( x \) where \( \delta(x) \) is large.

As discussed above, \( \bar{d} \leq d^* \) in the cases of greatest interest. There \( \delta(x)/c \leq 1 \) is a necessary condition for \( \Pr\{X_k \in S_0\} \to 1 \). We shall see shortly that it is undesirable to have \( \delta(x)/c = 1 \). On the other hand, when \( \delta(x)/c < 1 \), QUICKER acquires some attractive properties.

Fox’s result in [15] is that, for \( c > \delta(x) \),

\[
\lim_{k \to \infty} \mathbb{E}\left( (Q(x, k) - 1)^2 \mid Q(x, k) > 0 \right) = 0.
\]
Actually, it is routine to show that all moments converge; that is, the above expression holds with any positive integer exponent.

The next two theorems constitute an almost-sure counterpart to Fox's result.

**Theorem 2.1.** When a state \( x \) is not an isolated local minimum \((\delta(x) = 0)\), then with probability one, \( Q(x, k) \leq 1 \) for all but finitely many \( k \).

**Theorem 2.2.** Let \( j \) be a positive integer. If \( x \) is an isolated local minimum, and \( \delta(x)/c < (j - 1)/j \), then with probability one, \( Q(x, k) \leq j \) for all but finitely many \( k \).

We do not have a converse to Theorem 2.2, but the following may provide some insight. Although the conditions of Theorem 2.3 would never be realized in a lone sample path, they might apply, for example, to massively parallel runs of QUICKER.

**Theorem 2.3** Assume that \( Q(x, k) > 0 \) for all \( k \), and that \( Q(x, k) \) and \( Q(x, l) \) are independent when \( k \neq l \). If \( \delta(x)/c \geq (j - 1)/j \), then \( Q(x, k) > j \) infinitely often on almost all sample paths.

The quantities \( Q(x, k) \) are certainly not independent, but Theorem 2.3 may indicate something about the worst-case behavior of QUICKER.

The following example (see Figure 2.1) shows what can go wrong in the critical case \( \delta(x)/c = 1 \). Let \( S = \{x, y, z\} \) with \( U(x) = 1, U(y) = 2, U(z) = 0 \), \( N(x) = N(z) = \{y\} \) and \( N(y) = \{x, z\} \), and take \( \epsilon(k) = k^{-1} \). A routine calculation shows that the expected time to exit from state \( x \), given \( X_k = x \), is infinite for all \( k \). Since the sum of a finite number of geometric variates with positive parameters is finite a.s., it follows that \( \Pr(Q(x, k) > j) > 0 \) for all
positive integers \( j \) and \( k \).

**Adaptive Cooling.** Consider the following modification: upon entry to a state at step \( k \), keep the temperature constant at \( \epsilon(k) \) until a move out of the state is finally accepted at time \( m \); then change the temperature to \( \epsilon(m) \). In effect, we modify the cooling schedule \( (\epsilon(t)) \) in a certain random way. The sojourn time in state \( x \) beginning at time \( k \) has a geometric distribution with parameter \( \alpha(x, k) \).

Suppose \((S, U, R)\) contains no isolated local minima, and \( \epsilon(t) = t^{-1/c} \), with \( c \geq d^* \). Theorem 2.1 makes it plausible that \( \Pr(X_k \in S_0) \to 1 \) when \( \epsilon(t) \) is replaced by the random schedule described above. We show in Chapter 3 that this is in fact the case. We give further convergence results for such schedules in Chapter 5.

### 2.2 Proofs of Theorems

All three proofs follow the same pattern: we write down an exact expression for \( \Pr(Q_k > j \mid Q_k > 0) \), then bound it by suitable inequalities to show that the sum over \( k \) either converges or diverges, then invoke the Borel-Cantelli lemmas to assert that the event \( Q_k > j \) does or does not occur
infinitely often. We begin with some relationships that will be useful in all the proofs.

The acceptance probability at a state $x$ is of the form

$$\alpha(x, k) = A_0 + \sum_{i=1}^{n} A_i (k + a)^{-d_i/c},$$

(2.2)

where $n \geq 0$ and we assume without loss of generality that $0 < d_1 < \ldots < d_n$ and all the $A_i$ except possibly $A_0$ are strictly positive. The case $n = 0$ occurs when and only when $x$ has no uphill neighbors; i.e., it is a local maximum. The case $A_0 = 0$ occurs when and only when all of the neighbors of $x$ are strictly uphill—that is, recalling the definition on page 4, $x$ is an isolated local minimum. In that case, $d_1 = \delta(x)$.

The inner loop of QUICKER will be executed more than once if and only if the first “V-test” fails. Conditioned on the first geometric variate being $l$, the probability of this is

$$1 - (\alpha_{k+l-1}/\alpha_k) = \alpha_k^{-1} (\alpha_k - \alpha_{k+l-1}).$$

Thus

$$\Pr(Q_k > 1) \leq \Pr(Q_k > 1 \mid Q_k > 0)$$

$$= \sum_{i=1}^{\infty} \alpha_k^{-1} \cdot (\alpha_k - \alpha_{k+l-1}) \cdot (1 - \alpha_k)^{l-1} \cdot \alpha_k$$

$$= \sum_{i=0}^{\infty} (\alpha_k - \alpha_{k+l})(1 - \alpha_k)^l$$

$$= \sum_{i=1}^{\infty} (\alpha_k - \alpha_{k+l})(1 - \alpha_k)^l.$$  

(2.3)

Likewise, if the first geometric variate is $l$ and the first “V-test” fails, the remaining number of geometric variates is $Q_{k+l-1}$. Thus we obtain the recursion

$$\Pr(Q_k > j + 1) \leq \Pr(Q_k > j + 1 \mid Q_k > 0)$$

13
\[
\begin{align*}
&= \sum_{i=1}^{\infty} (\alpha_k - \alpha_{k+i-1}) (1 - \alpha_k)^{l-1} \Pr(Q_{k+l-1} > j) \\
&= \sum_{i=0}^{\infty} (\alpha_k - \alpha_{k+i}) (1 - \alpha_k)^l \Pr(Q_{k+l} > j) \\
&= \sum_{i=1}^{\infty} (\alpha_k - \alpha_{k+i}) (1 - \alpha_k)^l \Pr(Q_{k+l} > j). \quad (2.4)
\end{align*}
\]

In the case \( j = 0 \), (2.4) reduces to (2.3).

Now we obtain a bound for the factor \( \alpha_k - \alpha_{k+l} \), which is the (possibly null) sum of differences of the form

\[
A_i k^{-d_i/c} - A_i (k + l)^{-d_i/c}.
\]

The exponent \(-d_i/c\) is strictly negative. We expand \( A_i (k + l)^{-d_i/c} \) in a Taylor series around \( A_i k^{-d_i/c} \) and note that the series alternates in sign, implying

\[
A_i (k + l)^{-d_i/c} > A_i k^{-d_i/c} - (d_i/c) l k^{-(d_i/c)-1}, \quad (2.5)
\]

or equivalently,

\[
A_i k^{-d_i/c} - A_i (k + l)^{-d_i/c} < (d_i/c) l k^{-(d_i/c)-1}. \quad (2.6)
\]

Therefore,

\[
\alpha_k - \alpha_{k+l} = (A_0 - A_0) + \sum_{i=1}^{n} (A_i k^{-d_i/c} - A_i (k + l)^{-d_i/c}) \leq \sum_{i=1}^{n} A_i (d_i/c) l k^{-(d_i/c)-1}. \quad (2.7)
\]

Equality occurs in (2.7) in the case \( n = 0 \) (a local maximum), since then both sides vanish.
By keeping the second order term in the Taylor series, we get the lower bounds corresponding to (2.6) and (2.7):

\[ A_i k^{-d_i/c} - A_i (k + l)^{-d_i/c} > (d_i/c) \cdot l \cdot k^{-(d_i/c) - 1} - \frac{1}{2} (d_i/c)(1 + d_i/c) \cdot l^2 \cdot k^{-(d_i/c) - 2}, \]

and

\[ \alpha_k - \alpha_{k+l} > \sum_{i=1}^{n} A_i (d_i/c) \cdot l \cdot k^{-(d_i/c) - 1} - \frac{1}{2} \sum_{i=1}^{n} A_i (d_i/c)(1 + d_i/c) \cdot l^2 \cdot k^{-(d_i/c) - 2}, \quad (2.8) \]

Finally, the expressions for the first and second moments of a geometric distribution imply

\[ \sum_{i=1}^{\infty} l \cdot (1 - \alpha_k)^i = \alpha_k^{-2} (1 - \alpha_k) \leq \alpha_k^{-2}, \quad (2.9) \]

and

\[ \frac{1}{2} \sum_{i=1}^{\infty} l^2 (1 - \alpha_k)^{i-1} = \frac{1}{2} \alpha_k^{-3} (2 - \alpha_k)(1 - \alpha_k) < \alpha_k^{-3}. \quad (2.10) \]

2.2.1 Proof of Theorem 2.1 Fix a state \( x \). If \( n = 0 \) (i.e., \( x \) has no uphill neighbors) then (2.3) and the remark following (2.7) imply that \( \Pr(Q_k > 1) = 0 \) for all \( k \), and the theorem is trivial. Assume for the rest of the proof that \( n > 0 \). We have from (2.2) that

\[ \alpha(x, k) \geq A_0; \]

By hypothesis, \( x \) is not an isolated local minimum. This means that \( A_0 > 0 \) and therefore,

\[ \alpha_k^{-2} \leq A_0^{-2}. \quad (2.11) \]
From (2.3) and (2.7),
\[
\Pr(Q_k > 1) = \sum_{l=1}^{\infty} \left( \alpha_k - \alpha_{k+l} \right) (1 - \alpha_k)^l \\
\leq \sum_{i=1}^{n} A_i (d_i/c) k^{- (d_i/c) - 1} \sum_{l=1}^{\infty} l (1 - \alpha_k)^l.
\]
Applying (2.9) and (2.11),
\[
\Pr(Q_k > 1) \leq \sum_{i=1}^{n} A_i (d_i/c) k^{- (d_i/c) - 1} \alpha_k^{-2} \\
\leq \sum_{i=1}^{n} A_i A_0^{-2} (d_i/c) k^{- (d_i/c) - 1},
\]
and summing over positive \(k\),
\[
\sum_k \Pr(Q_k > 1) \leq \sum_{i=1}^{n} A_i A_0^{-2} (d_i/c) \left( \sum_k k^{- (d_i/c) - 1} \right) < \infty. \tag {2.12}
\]
From the Borel-Cantelli lemmas, (2.12) is sufficient to prove that, almost surely, \(Q_k > 1\) only finitely often. \(\square\)

2.2.2 Proof of Theorem 2.2 By hypothesis, \(x\) is an isolated local minimum, so \(A_0 = 0\) and \(\alpha(x, k)\) is of the form
\[
\alpha_k = \sum_{i=1}^{n} A_i k^{-d_i/c},
\]
where \(n > 0\). As before, assume that \(0 < d_1 < \ldots < d_n < 1\) and all the \(A_i\) are strictly positive. Note that \(d_1 = \delta(x)\); for the rest of this proof we suppress the reference to \(x\) and write just \(\delta\). There is a positive constant \(C_3\) such that for all \(k\)
\[
A_1 k^{-\delta/c} \leq \alpha_k \leq C_3 k^{-\delta/c},
\]
or equivalently,
\[
C_3^{-1} k^{\delta/c} \leq \alpha_k^{-1} \leq A_1^{-1} k^{\delta/c}. \tag {2.13}
\]
We now prove by induction the following statement for nonnegative integers \(j\):

16
(Pj) Assume \( \delta/c < 1 \). Then there exists a constant \( D_j \), not dependent on \( k \), such that \( \Pr(Q_k > j) \leq D_j \cdot k^{j(\delta/c) - j} \).

By the Borel-Cantelli lemmas, proving statement \( P_j \) for positive \( j \) is sufficient to prove Theorem 2.2, since if \( j > 0 \) and \( \delta/c < (j-1)/j \), then \( \delta/c < 1 \); also, the quantity \( j(\delta/c) - j \) is strictly less than \(-1\), so that \( \sum_k \Pr(Q_k > j) \) is bounded by the convergent series \( \sum_k D_j k^{j(\delta/c) - j} \).

Statement \( P_0 \) is trivial. Assume statement \( P_j \) and \( \delta/c < 1 \). Then \( j(\delta/c) - j < 0 \), so

\[
(k + l)^{j(\delta/c) - j} < k^{j(\delta/c) - j}
\]  

(2.14)

Writing (2.4) and using (2.7), the induction hypothesis, and (2.14):

\[
\Pr(Q_k > j + 1) = \sum_{i=1}^{\infty} \Pr(Q_{k+i} > j)(\alpha_k - \alpha_{k+i})(1 - \alpha_k)^i
\]

\[
\leq \sum_{i=1}^{\infty} A_i D_j (d_i/c) k^{-j(\delta/c) - j} k^{j(\delta/c) - j} \sum_{i=1}^{\infty} i(1 - \alpha_k)^i.
\]

It now follows from (2.9) and the right inequality of (2.13) that

\[
\Pr(Q_k > j + 1) \leq \left( \sum_{i=1}^{\infty} A_i D_j d_i/c \right) k^{-j(\delta/c) - j} \cdot \alpha_k^{-2}
\]

\[
\leq \left( \sum_{i=1}^{\infty} A_i\alpha_k^{-2} D_j d_i/c \right) k^{-j(\delta/c) - j} \cdot \alpha_k^{-2},
\]

this proves statement \( P_{j+1} \) and finishes the proof of Theorem 2.2. \( \Box \)

2.2.3 Proof of Theorem 2.3 By hypothesis, \( Q_k > 0 \) for all \( k \); therefore,

\[
\Pr(Q_k > j) = \Pr(Q_k > j \mid Q_k > 0), \quad \text{a.s.} \tag{2.15}
\]

In the remainder of the proof, we use the the two sides of (2.15) interchangeably.

Since \( \alpha_k \) is strictly less than one (at least for large \( k \)), and is nonincreasing in \( k \), statement (2.13) implies that there is a constant \( C_4 \) such that

\[
\alpha_k^{-2}(1 - \alpha_k) \geq C_4 k^{\delta/c}.
\]  

(2.16)

17
We prove the following statement by induction for nonnegative integers \( j \):

\((\mathcal{P}'_j)\) Assume \( \delta/c < 1 \). Then there exist positive constants \( D'_j \) and \( D''_j \), not dependent on \( k \), such that

\[
\Pr(Q_k > j) \geq D'_j \cdot k^{j(\delta/c) - j} - D''_j \cdot k^{(j+1)(\delta/c) - (j+1)}.
\]

Theorem 2.3 will then follow from the Borel-Cantelli lemmas and the independence of the \( Q_k \).

Statement \( \mathcal{P}'_0 \) is trivial. Assume \( \mathcal{P}'_j \). Then from (2.4), (2.8), and the induction hypothesis,

\[
\begin{align*}
\Pr(Q_k > j + 1) &= \sum_{l=1}^{\infty} (\alpha_k - \alpha_{k+l})(1 - \alpha_k)^l \Pr(Q_{k+l} > j) \\
&\geq \sum_{l=1}^{\infty} (1 - \alpha_k)^l \\
&\quad \times \left( \sum_{i=1}^{n} A_i(d_i/c)k^{-(d_i/c)-1}l - \sum_{i=1}^{n} A_i(d_i/c)(1 + d_i/c)k^{-(d_i/c)-2}(l^2/2) \right) \\
&\quad \times \left( D'_j(k + l)^{j(\delta/c) - j} - D''_j(k + l)^{(j+1)(\delta/c) - (j+1)} \right). \tag{2.17}
\end{align*}
\]

Now

\[
(k + l)^{(j+1)(\delta/c) - (j+1)} \leq k^{(j+1)(\delta/c) - (j+1)}, \tag{2.18}
\]

while a Taylor-series argument similar to that used to derive (2.5) shows that

\[
(k + l)^{j(\delta/c) - j} \geq k^{j(\delta/c) - j} - jl(1 - \delta/c)k^{j(\delta/c) - j-1}. \tag{2.19}
\]

Applying (2.18) and (2.19) to the last factor of (2.17), expanding the product, and dropping some positive terms,

\[
\begin{align*}
\Pr(Q_k > j + 1) &\geq \sum_{l=1}^{\infty} (1 - \alpha_k)^l
\end{align*}
\]

18
\[
\begin{align*}
&\times \left( A_1 \left( \frac{\delta}{c} \right)^{k^{-\delta/c-1}l - \sum_{i=1}^{n} A_i \left( \frac{d_i}{c} \right) \left( 1 + \frac{d_i}{c} \right) k^{-\left(d_i/c\right)-2} \left( \frac{l^2}{2} \right) \right) \right) \\
&\times \left( D'_j k^{j(\delta/c)-j} - D'_j \left( 1 - \frac{\delta}{c} \right) k^{j(\delta/c)-j-1} l - D''_j k^{j+1)(\delta/c)-(j+1)} \right) \\
&\geq A_1 D'_j \left( \frac{\delta}{c} \right) k^{-\left(\delta/c\right)-1+j(\delta/c)-j} \left( \sum_{l=1}^{\infty} l(1 - \alpha_k)^l \right) \\
&- A_1 D'_j \left( \frac{\delta}{c} \right) j(1 - \delta/c) k^{-\left(\delta/c\right)-1+j(\delta/c)-j-1} \left( \sum_{l=1}^{\infty} l^2(1 - \alpha_k)^l \right) \\
&- A_1 D''_j \left( \frac{\delta}{c} \right) k^{-\left(\delta/c\right)-1+j+1)(\delta/c)-(j+1)} \left( \sum_{l=1}^{\infty} l(1 - \alpha_k)^l \right) \\
&- \sum_{i=1}^{n} A_i D'_j \left( d_i/c \right) \left( 1 + d_i/c \right) k^{-d_i/c-2+j(\delta/c)-j} \left( \sum_{l=1}^{\infty} \frac{l^2}{2} \left( 1 - \alpha_k \right)^l \right).
\end{align*}
\]

Finally, we use (2.9) and (2.10), and then (2.13) and (2.16) to conclude that

\[
\Pr(Q_k > j + 1) \geq A_1 D'_j \left( \frac{\delta}{c} \right) k^{j+1)(\delta/c)-(j+1)}(\alpha_k^{-2}) \\
- 2A_1 D'_j \left( \frac{\delta}{c} \right) j(1 - \delta/c) k^{(j+1)(\delta/c)-(j+2)}(\alpha_k^{-3}) \\
- A_1 D''_j \left( \frac{\delta}{c} \right) ck^{j(\delta/c)-(j+2)}(\alpha_k^{-2}) \\
- \sum_{i=1}^{n} A_i D'_j \left( d_i/c \right) \left( 1 + d_i/c \right) k^{(j+1)(d_i/c)-(j+2)}(\alpha_k^{-3}) \\
\geq D'_j k^{j+1)(\delta/c)-(j+1)} - D''_j k^{j+2)(\delta/c)-(j+2)},
\]

where

\[
D'_j = D'_j A_1 C_3^{-2}(\delta/c)
\]

and

\[
D''_j = D'_j A_1^{-3} \left[ 2A_1(\delta/c)(1 - \delta/c)j + \sum_{i=1}^{n} A_i(d_i/c)(1 + d_i/c) \right] + D''_j A_1^{-1}\delta/c.
\]

This finishes the induction proof of \( P'_{j+1} \) and the proof of Theorem 2.3. \( \square \)
2.3 Efficiency of Fox’s Algorithm

We compare the expected efficiency, work per unit cost, of algorithms SLOW and QUICKER. Say that computing the states $X_1, \ldots, X_k$ represents $k$ units of work, and that each iteration of SLOW represents one unit of cost, so that the efficiency of SLOW is 1. To find the relative efficiency of QUICKER we find its cost for computing $X_1, \ldots, X_k$.

We can break each invocation of QUICKER into three phases:

(i) Computation of the acceptance probability.

(ii) Iteration of the inner loop.

(iii) The call to the algorithm NEXT.

Steps (i) and (iii) involve scanning the neighborhood of state $x$ and thus their cost depends on the size of $N(x)$. It is reasonable to assume that this cost is bounded above and below by constant multiples $m$ and $M$, respectively, of the cost of one iteration of SLOW. The main work in the inner loop is the generation of two random numbers, (and, possibly, another neighborhood scan), so the cost of each iteration of the inner loop is bounded above and below by constants $m'$ and $M'$.

In QUICKER, the clock is advanced only during iterations of the inner loop. Thus

$$H(k) \cdot m + G(k) \cdot m' \leq \text{cost} (X_1, \ldots, X_k) \leq H(k) \cdot M + G(k) \cdot M',$$

where $H(k)$ is the number of iterations of QUICKER up to time $k$ and $G(k)$ is the number of iterations of the inner loop up to time $k$.

The import of the theorems in this chapter is that for all large $k$, $G(k)/H(k)$ is between 1 and $j$, for some $j$ depending only on the cooling
schedule. We use this to get

\[(m + m') \cdot H(k) \leq \text{cost } (X_1, \ldots, X_k) \leq (M + jM') \cdot H(k), \quad (2.20)\]

with probability arbitrarily close to one for large values of \(k\).

We now estimate \(E(H(k))\). For this purpose, consider the sequence of random variables \(\beta(k), k = 0, 1, \ldots\), defined by

\[\beta(k) = \alpha(X_{k-1}, k) = \Pr(X_k \neq X_{k-1}).\]

Suppose we know \(H(k)\) for some \(k\). If the algorithm moves to a different state at step \(k + 1\), then \(H(k + 1) = H(k) + 1\). Otherwise, \(H(k + 1) = H(k)\). Thus we have the recursive relationship

\[E(H(k + 1)|H(k), \beta(k + 1)) = (1 + H(k))\beta(k + 1) + H(k)(1 - \beta(k + 1))\]
\[= H(k) + \beta(k + 1). \quad (2.21)\]

Taking expectations in (2.21) and using the initial condition \(H(0) = 0\), we conclude by induction that

\[E(H(k)) = \sum_{j=1}^{k} E(\beta(k)). \quad (2.22)\]

Next, we estimate the quantities \(\beta(1), \ldots, \beta(k)\). Trivially \(\beta(j) \leq 1\) for all \(j\). Also, if \(\bar{d} > 0\), and \(X_j\) is an isolated local minimum, then

\[D \cdot j^{-\bar{d}/c} \leq \beta(j)\]

for a constant \(D\). If \(X_j\) is not an isolated local minimum, then \(\beta(j)\) is bounded away from zero. Thus in any case we have

\[D' \cdot j^{-\bar{d}/c} \leq \beta(j) \leq 1,\]

21
for some constant $D'$. Taking sums, we have

$$k \geq \sum_{j=1}^{k} \beta(j) \geq \sum_{j=1}^{k} D' \cdot j^{-\bar{d}/c} \geq C_0 \cdot (k^{1-\bar{d}/c})$$  \hspace{1cm} (2.23)

for another constant $C_0$; applying (2.22),

$$C_0 \cdot k^{1-(\bar{d}/c)} \leq \mathbb{E}(H(k)) \leq k.$$  \hspace{1cm} (2.24)

Recall that our measure for efficiency is work per unit cost, or

$$k/(\text{cost } (X_1, \ldots, X_k)).$$

Applying (2.24) to (2.20), taking multiplicative inverses, and then multiplying through by $k$, this means that there exist constants $C_1$ and $C_2$ such that

$$C_1 \leq \mathbb{E}(\text{efficiency of QUICKER at time } k) \leq C_2 k^{\bar{d}/c}.$$  \hspace{1cm} (2.25)

The efficiency of QUICKER relative to SLOW increases as the parameter $c$ in the exponent of the cooling schedule gets smaller. This should not be interpreted as an argument in favor of faster schedules. It reflects the fact that SLOW is becoming less efficient; as $c$ gets close to $\bar{d}$, the algorithm spends more and more of its time in self loops at local minima.

Efficiency is highest when many of the states $X_1, \ldots, X_k$ are isolated local minima. When there are no isolated local minima, the lower bound in (2.7), and thus the upper bound in (2.25), can be replaced by a constant. In this case, the relative efficiency of QUICKER is bounded both above and below by a constant, and there is no order-of-magnitude gain in speed from its use. On the other hand, if all local minima are isolated, then (2.7) is sharp. Heuristically, QUICKER tends to be an efficient algorithm to the degree that direct transitions between local minima are rare or difficult.
Some examples from the literature show that isolated or near-isolated local minima can actually occur. In the “maximum matching problem” of [25], all local minima are isolated. In the first formulation of the graph coloring problem in [31], direct transitions between local minima are not impossible, but would be extremely rare. Our results indicate that QUICKER would attain its maximum efficiency in such situations and would be worth trying there.

The Distribution of $H(k)$ Another approach to computing the expected value of $H(k)$ may give additional insight. Consider the following modification: fix a state $x_0$ and start the algorithm there. Run QUICKER as before, but every time there is a move to a state other than $x_0$, return immediately to $x_0$ before continuing. Let $\tilde{H}(k)$ denote the number of inner-loop iterations up to time $k$ in the modified process. Clearly, $H(\cdot)$ is an inhomogeneous discrete-time counting process. Letting $p_i(k)$ denote the probability that $\tilde{H}(k) = i$, the $k$-step transition vector $p(k)$ satisfies the forward equations

\[
p_0(k + 1) = (1 - \alpha(x_0, k + 1))p_0(k);
\]
\[
p_i(k + 1) = (1 - \alpha(x_0, k + 1))p_i(k) + \alpha(x_0, k + 1)p_{i-1}(k), \quad i = 1, 2, \ldots
\]

The generating function $\zeta(z, k) \overset{\text{def}}{=} \sum_{i=0}^{\infty} z^i p_i(k)$ then satisfies the first-order difference equation

\[
\zeta(z, k + 1) - \zeta(z, k) = -\alpha(x_0, k + 1)\zeta(z, k) + z\alpha(x_0, k + 1)\zeta(z, k),
\]

with initial condition $\zeta(z, 0) = 1$. It is easily verified by induction that the solution is

\[
\zeta(z, k) = \prod_{j=1}^{k} (1 + (z-1)\alpha(x_0, j)); \quad (2.26)
\]
taking the first derivative at $z = 1$ shows that

$$E(\tilde{H}(k)) = \sum_{j=1}^{k} \alpha(x_0, j),$$

(2.27)

exactly. Moreover, for $z$ near one, the right side of (2.26) is arbitrarily close to

$$\exp \left( (z - 1) \sum_{j=1}^{k} \alpha_j \right),$$

which is the probability generating function of a Poisson distribution with the same mean.

Observing that $\tilde{H}(k) \leq H(k)$ stochastically, we obtain (2.24) as before.
CHAPTER 3

SIMULATED ANNEALING WITH OVERRIDES

The results of Chapter 2 show that there is an integer $j$, determinable in advance, which bounds the number of inner-loop iterations of QUICKER($x$) for all $x$ after a random but finite number of moves. This suggests that we modify QUICKER to always force an exit after $j$ iterations. Formally:

**Algorithm QUICKER-$j$($j, x, t$)**

Set $k \leftarrow t$

Set $i \leftarrow 0$

Until exit, repeat

Set $i \leftarrow i + 1$

Generate a geometric r.v. $G$ with parameter $\alpha(x, k)$

Set $L \leftarrow k + G - 1$

Generate a uniform r.v. $V$

If $V < \alpha(x, L)/\alpha(x, k)$ or $i > j$

Execute NEXT($x, L$)

Exit

Else

Set $k \leftarrow L$

End

The boxes above outline the additional steps needed to convert QUICKER into QUICKER-$j$.

With QUICKER-$j$, more moves are accepted in fixed computer time than with QUICKER. We cannot say that QUICKER-$j$ shortens the time to first hit $S_0$. However, QUICKER-$j$ lets us synchronize parallel processors with SIMD
architecture executing stochastically independent runs of simulated annealing. Each move at a given computer clock tick executes the same instruction, on processor-dependent data. When QUICKER-$j$ exits after $h \leq j$ iterations on some processor, then that processor executes $j - h$ additional dummy iterations in order to maintain synchronization. Such synchronization is not possible with either QUICKER or the naive move mechanism.

Stochastically, QUICKER-1 generates the same sample path as SLOW with the adaptive cooling schedule described in Section 2.1.

It seems intuitively plausible that convergence in probability would continue to hold under algorithm QUICKER-$j$ for the right choice of $j$. Formally, we have

**Theorem 3.1** If $j$ chosen so that the number of inner-loop iterations of QUICKER exceeds $j$ only finitely often, and if

$$\lim_{k \to \infty} \Pr(X_k \in S_0) = 1 \quad (3.1)$$

under algorithm QUICKER, then (3.1) also holds under algorithm QUICKER-$j$.

Theorem 3.1 is a corollary of a much more general result which we prove in this chapter. Broadly, the convergence properties of a probabilistic search algorithm are unaffected by any modification of its initial steps, provided the total number of modified steps is finite with probability one. That is, suppose the search strategy can have its standard move mechanism overridden up to a random but finite time $N$; just after $N$, the standard move mechanism takes over. We show that the current state of the Markov chain associated with the modified process converges in probability to the set $S_0$ of optimal states
whenever this is true in the unmodified process.

We do not require the state space \( S \) to be finite. Moreover, \( S_0 \) can be an arbitrary measurable subset of \( S \), not just the set of global optimizers.

In the next two sections, we give examples of applications of this general result, followed by a formal statement and proof in Section 3.3.

3.1 Simple Examples

3.1.1 Random Restart Begin the simulated annealing process in the usual fashion. Observe the process. If it appears to stall at a particular state or set of states, jump at random to a different part of the state space. Repeat this until a fixed number \( n \) of restarts have been made. With any reasonable definition of “stall”, the restart mechanism is in fact triggered \( n \) times. In this case, \( N \) is the transition number at which the \( n \)-th restart is made. Our results do not say that modifying simulated annealing in this fashion makes the algorithm either better or worse. However, we are assured the modified algorithm retains the convergence property (3.1).

In this example, \( N \) is a stopping time (Markov time), because we know its value after \( N \) transitions have occurred, but later we give important examples where this is not true. Perhaps surprisingly, our basic result continues to hold under certain natural conditions even when \( N \) is not a Markov time.

3.1.2 Greedy Overrides Upon arrival to a state, scan the energy levels of all its neighbors. If a neighbor is encountered with a lower energy level than any state yet seen, move immediately to the lowest such neighbor. On a finite state space, \( N \) is obviously finite. Under weak conditions, \( N \) is the transition number on which \( S_0 \) is first visited, and so is a Markov time.
Such a greedy override strategy is widely used in the tabu-search
community (see, for example, Glover [22],[23]). Under certain restrictions, Fox
[14] recasts tabu search as simulated annealing on a more elaborate state space.
It is easy to give instances where a greedy override strategy increases the time
to first visit a global optimum. Heuristically, however, it seems attractive unless
the state space is contrived.

3.1.3 Perverse Overrides Upon arrival to a state, scan the
neighbors. If one of the neighbors has higher energy level than any state
yet encountered, move immediately to the highest such neighbor. The analysis
of the previous subsection goes through, where this time $N$ is the transition
number on which the set of global maximizers is first encountered. We do not
claim this strategy is worthwhile, except perhaps on very unusual state spaces,
but it does show the robustness of the property (3.1).

3.2 Dynamic Neighborhoods

Consider a generic search algorithm of the form

\begin{algorithm}
\textbf{Algorithm Search$(x,t)$}
\begin{itemize}
  \item Generate a set $F(x,t)$ of candidates for the next state
  \item Optional: if $F(x,t)$ contains an element at least $\epsilon$
    better than any yet seen, move to it
  \item Until exit, repeat
    \begin{itemize}
      \item Select an element of $F(x,t)$
      \item If it passes the standard acceptance test, then
        \begin{itemize}
          \item Exit
        \end{itemize}
    \end{itemize}
\end{itemize}
\end{algorithm}

In ordinary simulated annealing, $F(x,t)$ is the fixed set of neighbors of $x$.
Here we allow the possiblity of using problem-specific information to choose a
smaller or larger $F$ at each step of the algorithm. Of course, we have to prove that convergence in probability still occurs; this is true in the two examples of this section.

Fox [14] shows that algorithm QUICKER, now depending on the triple $(x, t, F(x, t))$, can replace the “Until...repeat” loop in SEARCH. We want to further streamline SEARCH using the techniques of this chapter.

When $F$ is a small finite subset of $S$, it is practical to compute the acceptance probability $\alpha(x, k)$ needed for QUICKER or QUICKER-$j$, and also for their subroutine NEXT. If the method to pick an element of $F$ depends on the energy levels of all its elements, then $\alpha(x, k)$ can be computed with essentially no additional work.

In the setting of Fox ([14] and [15]), $S$ is finite. The set $F(x, t)$ is formed by supplementing the nominal fixed neighborhood of the current state $x$ with randomly generated elements from $S$. To show that the convergence theorems apply, Fox recasts this as standard annealing with fixed neighborhoods, but on a more elaborate state space.

To ensure a set $F(x, t)$ of reasonable size, the nominal neighborhood can be chosen relatively small, while the added elements are chosen to guarantee weak reversibility and irreducibility. Heuristically, the added elements diversify the search; [15] shows that they have theoretical value as well. We modify SEARCH in two ways: first, we perform the “Optional” step with $\epsilon = 0$ (this is the same as the greedy strategy in Section 3.1); second, we further speed up the “Until...repeat” loop by using QUICKER-$j$ instead of QUICKER. Our theorem applies to both modifications. To cool as rapidly as possible and still
take advantage of the convergence properties of Theorems 2.1 and 2.2, we want to ensure that the constant $\tilde{d}$ is not too large. This is avoided by making requiring the nominal neighborhood of each state to contain a state with at most a slightly higher energy level.

When $S$ is a compact subset of $\mathbb{R}^d$, Bélisle [2] implicitly performs SEARCH. He takes $F = S$, a natural choice when NEXT is not used. Under the strong assumptions of [2], there are no restrictions on the rate of cooling; in particular, convergence occurs with any adaptive schedule, so long as the temperatures decrease to zero. Moreover, any state can be reached from any other state, so there are no isolated local minima (unless there is only one optimal state; this special case is easily disposed of). Then we use QUICKER-1 (corresponding to an adaptive schedule) to speed up the “Until...repeat” loop. We also perform the “Optional” step here, this time setting $\epsilon > 0$.

It is straightforward to extend Bélisle’s results to the case where the neighborhoods $F$ are dynamically chosen, so long as there is a path with no uphill moves from any state to $S_0$. That is, every set $F(x, t)$ must be forced to include a state with the same or lower energy level than $x$. This works also when $S$ is finite, and it offers a way of proving convergence that is different from Fox’s recasting scheme, mentioned earlier in this section. Once again, there are no isolated local minima, so we use QUICKER-1 as before.

3.3 Overrides: The Main Result

Let $S_0$ be an arbitrary measurable subset of $S$, not necessarily finite. We require that the unmodified chain $(X_k)$ converge in probability to $S_0$ from
any starting state at any initial time. More precisely,

\[ \lim_{k \to \infty} \inf_{x \in S} \Pr\{X_k \in S_0 | X_n = x\} = 1, \]  

(3.2)

for all positive integers \( n \). When \( S \) is finite, we can replace (3.2) by the simpler condition

\[ \lim_{k \to \infty} \Pr\{X_k \in S_0 | X_n = x\} = 1, \]  

(3.3)

The epoch of the last override must be almost surely finite, but we do not require it to be a Markov time. It is enough that the override mechanism stops eventually and has no further influence on the destiny of the process. More precisely, denote by \( Y(k), k = 1, 2, \ldots, \) the Markov chain of moves with overrides, and by \( N \) the epoch of the last override. We assume that there exists a family \( \{Z(n, \cdot) : n \geq 0\} \) of Markov chains such that:

(i) The chains \( Z(n, \cdot) \) and \( Y(\cdot) \) coincide exactly up to time \( n \). That is,

\[ Z(n, k, \omega) = Y(k, \omega) \]

for \( k \leq n \) and every \( \omega \) in the probability space.

(ii) After time \( n \), the transition probabilities of \( Z(n, \cdot) \) and \( X \) are the same; that is,

\[ \Pr(Z(n, k + 1) = y | Z(n, k) = x) = \Pr(X_{k+1} = y | X_k = x) \]

for \( k > n \) and \( x, y \in S \).

(iii) There exists a random time \( N \), a.s. finite, after which the transition probabilities of \( Y(\cdot) \) and \( X \) are the same.

Here is the intuition for this structure. We start with the unmodified process and define a sequence of processes in which random overrides are allowed, but only up to a deterministic time \( n \). Each process \( Z(n, \cdot) \) couples with \( Y(\cdot) \) up to
time $n$ and behaves stochastically like $(X_k)$ after time $n$. We can characterize $Y(j)$ as $\lim_{n \to \infty} Z(n, j)$. The limit is attained for some finite $n$, which can be chosen uniformly in $j$.

**Theorem 3.2** Assume (3.2) and conditions (i)-(iii). Then (3.2) is also true for the chain $Y(\cdot)$.

**Proof:** Let $\epsilon > 0$ be arbitrary. For each positive integer $m$, let $\phi(m)$ be the smallest integer $j > m$ such that

$$\inf_{x \in S} \Pr \{X_k \in S_0 | X_m = x\} > 1 - \epsilon,$$

for all $k \geq j$. Condition (3.2) guarantees that $\phi(m)$ is well-defined and finite for every $m$.

Moreover, the function $\phi(\cdot)$ is non-decreasing, as we now show. Select $n \leq \phi(m)$. Then there exist $k \geq n$ and $x \in S$ such that

$$\Pr(X_k \in S_0 \mid X_m = x) \leq 1 - \epsilon.$$

Applying the Chapman-Kolmogorov equations, we have

$$\inf_{y \in S} \Pr(X_k \in S_0 \mid X_{m+1} = y) \leq \sum_{y \in S} \Pr(X_k \in S_0 \mid X_{m+1} = y) \Pr(X_{m+1} = y \mid X_m = x) = \Pr(X_k \in S_0 \mid X_m = x) \leq 1 - \epsilon,$$

which implies that $n \leq \phi(m + 1)$. Because $n$ was an arbitrary number between 0 and $\phi(m)$, $\phi(m) \leq \phi(m + 1)$.

Since $\phi(\cdot)$ is non-decreasing and grows without limit, there is, for $k \geq \phi(0)$, a unique integer $\phi^*(k)$ such that

$$\phi(\phi^*(k)) \leq k < \phi(\phi^*(k) + 1).$$  (3.4)
Note that the function $\phi^*(\cdot)$ is also non-decreasing and grows without limit.

Take any $k \geq \phi(0)$; write $\nu$ for $\phi^*(k)$. By condition (ii),

$$\Pr(Z(\phi^*(k), k) \in S_0) \geq \inf_{x \in S} \Pr(Z(\nu, k) \in S_0 \mid Z(\nu, \nu) = x)$$
$$= \inf_{x \in S} \Pr(X_k \in S_0 \mid X_\nu = x)$$
$$> 1 - \epsilon. \quad (3.5)$$

To finish the proof, we use a variant of the coupling inequality (see Asmussen[1], p. 143). By conditions (i) and (iii), the chains $Z(\nu, \cdot)$ and $Y(\cdot)$ coincide exactly for all time on the set \{\(N < \nu\}\}. Thus

$$|\Pr(Z(\nu, k) \in S_0) - \Pr(Y(k) \in S_0)|$$
$$= |\Pr(Z(\nu, k) \in S_0, N \geq \nu) - \Pr(Y(k) \in S_0, N \geq \nu)|$$
$$\leq \Pr(N \geq \nu). \quad (3.6)$$

We have proved that, for arbitrary $\epsilon$, for all large $k$,

$$\Pr(Y(k) \in S_0) > 1 - \epsilon - \Pr(N \geq \phi^*(k)).$$

As remarked above, the function $\phi^*(\cdot)$ is nondecreasing and grows without limit; by condition (iii), this completes the proof. \qed

33
CHAPTER 4

SIMULATED ANNEALING WITH A NOISY OBJECTIVE FUNCTION

We now discuss a significant application of the results in Chapter 3. Suppose we are interested in a state space where objective-function values are unknown and must be estimated by measurements subject to error. For example, each state might be a configuration of a queuing network. To run simulated annealing in the most naive fashion, pretend that the estimated objective-function values are the true values; calculate move probabilities and move based on the estimates. Meanwhile, every time a state is seen, measure its objective-function value and use the measurement to refine the estimates. In this chapter, we show that under the right conditions, this simple procedure works.

Yan and Mukai [38] have previously considered such situations. They assume that the state space satisfies $NS$, and that we know a finite interval $(a, b)$ surrounding all the objective-function values. In their algorithm, the probability of accepting a proposed move is itself a random quantity, typically estimated by multiple measurements of the tentative new state. A proposed move is rejected unless the result of every simulation is smaller than a uniform variate on $(a, b)$ drawn at the beginning of the move. The number of simulations required at each move is large and grows without limit as the algorithm
proceeds.

Gelfand and Mitter [19] propose a different method. They formulate the transition probability at time $t$ as

$$R(x, y) e^{[U(y) - U(x) + W(t)]^+},$$

where $W(t)$ represents a noise term which may depend on past history. This general formulation includes ours as a special case. However, their convergence proofs assume that $W(t)$ is Gaussian with variance asymptotically smaller than any linear multiple of $1/(-\log e(t))$. Since some states may be visited only rarely, it is unclear how to check the variance condition.

For technical reasons, we assume that the unknown objective function takes on values in a finite set, such as a subset of the computer-representable numbers. Neither [38] nor [19] need to do this. But since our treatment allows arbitrary precision, we do not consider this restriction important.

In the next section, we give a formal description and show that our conclusions follow from the results of Chapter 3. It turns out that we need estimators which are "asymptotically stable"; the remainder of the chapter discusses these and presents two examples.

4.1 Formal Setup

At the $j$-th entry to a state $x$, make an observation $T(x, j)$. Each of the observations $T(x, 1), T(x, 2), \ldots$, has expected value $h(x)$. The vector $h = (h(x) : x \in S)$ is unknown. Perform simulated annealing with acceptance probabilities formed from some estimate $H_n = (H(x, n_z) : x \in S)$, where each component $H(x, n_z)$ is based on history of observed values of $T(x, j)$, $j = 1, \ldots, n_z$, and $n_z$ is the number of times $x$ has been observed up to time $n$. 

35
Assume that there are only a finite number of possible values for \( h(x) \) at any state \( x \) (for example, integer multiples of 0.001 between 0 and 1). In effect, we accept finite precision in our knowledge of \( h(x) \). We say that the estimator \( H(x, \cdot) \) is **asymptotically stable** if there exists a random time \( N(x) \), almost surely finite, such that \( H(x, n) = h(x) \) for all \( n \geq N(x) \). We show later that such estimators exist.

In our setting, the \( x \)-th component of \( H_n \) takes account only of the \( x \)-th component of the observations \( (T(\cdot, j)) \). The observations of the other components and the current observation number are ignored. For every component, the observations are thus iid. Since there are only a finite number of states, we can consider asymptotic stability of each component separately. To initialize, for each component we might set \( H(x, 0) = -\infty \).

If \( H(x, \cdot) \) is asymptotically stable for each state \( x \), and if every state is visited infinitely often, then from some finite time \( N \) onwards, the vector \( H \) of \( H \)-estimates coincides with the vector \( h \) of true \( h \)-values. In this setting, the "standard" algorithm is simulated annealing using \( h \) in the computation of the move probabilities. The modification uses the estimates \( H_n \). To make this more precise, in both chains the original state space \( S \) is enlarged to include enough historical information to compute \( H_n \) at time \( n \). Only the moves in the second chain take account of that additional information. When \( h \) is unknown, the first chain is hypothetical.

Theorem 3.1 implies that if the first chain converges in probability to \( S_0 \) with \( h \), then so does the second chain.
This setup generalizes easily to any situation where the objective function \( h(x) \) depends on a finite-dimensional vector \( \theta \) of unknown parameters, and \( \theta \) takes on only a finite number \( v \) of values. The idea is to exploit structure in the state space so that \( v \) is much smaller than \( |S| \), although that is not necessary for our proofs. As above, we can interpret this as knowing the elements of \( \theta \) to finite precision.

4.2 Consistent and Stable Estimators

Generalizing the previous section, we want to run Monte Carlo simulation in an optimal way according to some criterion. The choice of simulation strategy depends on a parameter with an optimal value \( h \). We do not know \( h \).

However, at step \( n \) of the simulation we make an observation \( T(n) \), and then use the observations \( T(1), \ldots, T(n) \) to compute an estimate \( H(n) \) of \( h \). At time \( n \), the simulation strategy depends on \( H(n) \). Recall the definition of asymptotic stability in the previous section: there is a random time \( N \), almost surely finite, such that \( H(n) = h \) for all \( n \geq N \). Generally, to find such an \( H \) requires that we know a finite set \( F \) containing \( h \), and that the observations fall in a some set \( G \), generally a superset of \( F \). Thus, \( H(n) \) maps the \( n \)-fold Cartesian product of \( G \) to \( F \). We give below two examples of asymptotically stable estimators.

For simplicity, assume that \( h \) and \( T(j) \) are scalars and by rescaling that \( F \) is a set of consecutive integers. The set \( G \) is either a subset of the real numbers or a subset of the integer multiples of \( 1/d \), where \( d \) is an integer greater than 1. We can view the latter as a finite-precision approximation to the former. Let \( \langle x \rangle \) denote \( x \) rounded to the nearest integer, rounding upward.
in case of ties (i.e., \( x = [x + (1/2)] \)). Set \( t = \text{E}(T(j)) \) and let \( h = \{t\} \).

An example shows what can go wrong. Let the sequence of sample means be \( \overline{T}(n) = n^{-1} \sum_{j=1}^{n} T(j) \). The estimator \( H(n) = \langle \overline{T}(n) \rangle \) is not asymptotically stable when \( t \) is a half-integer; for large \( n \), it oscillates unpredictably between \([t] \) and \([t]\). If \( t \) is near a half-integer, then damped oscillation occurs; although there is eventual stability, arguably the rate of damping is very slow. This is the practical reason to consider the estimators in Sections 4.3 and 4.4.

Suppose we try to correct the situation by observing the entire sequence of sample means \( \overline{T}(1), \ldots, \overline{T}(n) \) (computed to double precision, perhaps), counting how many are above or below the half-integer line, and rounding up or down accordingly. This actually makes matters worse. For large \( n \), the sequence of sample means closely approximates Brownian motion, and the proportion of the \( \overline{T}(j) \) which lie above \( t \) has an arc sine distribution (see Feller [13], pp. 397–99.) when \( t \) is exactly a half-integer. When \( t \) is approximately a half-integer, this proportion forms an approximate arc sine distribution with large probabilities of being near 0 or 1. The refined estimator still oscillates, but the intervals between changes grow longer and longer (in fact the growth is exponential). To the careless observer, it may falsely appear to be stabilizing.

It is important to be clear about the difference between an estimator which is asymptotically stable and one which is merely consistent. If we had wanted, say, a static point estimator, restricted to the integers, of a parameter with true value 1.5, then we would be indifferent between an estimate of 1.0 and 2.0. Asymptotic stability is needed when we are interested, not in the parameter per se, but in a simulation process dependent on an estimate of the
parameter.

### 4.3 Asymptotically Stable Estimators: First Example

Let $M_k(n)$ be the number of observations equal to $k$ up to step $n$; that is,

$$M_k(n) = \sum_{j=1}^{n} I(T(j) = k).$$

Each observation $T(n)$ is formed by adding a random error $\xi_n$ to $t$, and then rounding to the nearest integer:

$$T(n) = \lfloor t + \xi_n \rfloor$$

Suppose that the errors $\xi_n$ are iid. Then the $T(n)$ have a common discrete distribution function with mass points $\ldots, p_{-1}, p_0, p_1, \ldots$, on the integers. Suppose also that the distribution of the $\xi_n$ is symmetric about zero and has a density with bounded support, strictly decreasing on positive numbers. Then when $t$ is not a half-integer the set $\{p_k > 0 : k \in \mathbb{Z}\}$ is finite and has $p_k = p(\ell_k)$ for its unique largest element. In the exceptional case where $t$ is a half-integer, there will be maxima at both $p(\ell_1)$ and $p(\ell_2)$. For large $n$, we expect $M_k(n)/n$ to approximately equal $p_k$, at least for those $k$ not in the tail of the distribution. These remarks form the intuitive basis for our estimator.

Let $f(n)$ be a function on the positive integers which is non-decreasing in $n$ and which satisfies both

$$f(n) = o(n) \quad \text{and} \quad \log(n) = o(f(n)). \quad (4.1)$$

(For example, $f(n) = \sqrt{n}$.) Below, break ties for the second-largest value arbitrarily.
ALGORITHM STABLE-1(x, t)

For each step $n$:
Make an observation $T(n)$;
Set $k \leftarrow T(n)$;
Set $M_k(n) \leftarrow M_k(n) + 1$;
If the set $N = \{M_k(n) : k \in F\}$ has a single maximum at $m$,
and $\arg \max N \setminus \{m\} = m + 1$,
and $|M_m(n) - M_{m+1}(n)| < f(n)$, then
Set $H(n) \leftarrow m + 1$.
Else if the set $\{M_k(n) : k \in F\}$ has a single maximum at $m$,
and $\arg \max N \setminus \{m\} = m - 1$,
and $|M_m(n) - M_{m-1}(n)| < f(n)$, then
Set $H(n) \leftarrow m$.
Else
Set $H(n) \leftarrow \arg \max\{M_k(n) : k \in F\}$.
End

The following theorem shows that $H(n)$ has the desired properties.

**Theorem 4.1** Assume that $f$ satisfies (4.1) and that the error terms
$\xi_n$ satisfy the conditions stated above. Then, almost surely, the estimator $H(n)$
of algorithm STABLE-1 exactly equals $h$ for all but finitely many $n$.

**Proof**: Suppose first that $t$ is not a half-integer. Assume that $m \leq t < m + (1/2)$; the other case is symmetric. By the conditions imposed on the
distribution of $\xi_n$, the set $\{p_k : k \in G\}$ has its unique largest value at $m$ and its
second largest value at $m + 1$. The strong law of large numbers and $|G| < \infty$
imply that, a.s. for large $n$, the set $\{M_k(n) : k \in G\}$ attains its unique largest
maximum at $m$ and its unique second largest value at $m + 1$. For arbitrary
$\delta > 0$, another application of the strong law shows that for all large $n$, on
almost all sample paths,

$$|M_m(n) - M_{m+1}(n)| > n(p_m - p_{m+1}) - \delta,$$
and the expression on the right is bounded below by $f(n)$ for all large $n$. This completes the proof if $t$ is not a half integer.

Now suppose that $t = m + (1/2)$. Then $p_m = p_{m+1}$ and these two are the only maxima in the set $\{p_k : k \in \mathbb{Z}\}$. Using the strong law as in the previous paragraph, the two largest elements of the set $\{M_k(n) : k \in \mathbb{Z}\}$ are $M_m(n)$ and $M_{m+1}(n)$ a.s. for all large $n$. Let $j(0) = 0$ and

$$j(i + 1) = \min_{n > j(i)} \{T(n) = m \text{ or } T(n) = m + 1\},$$

i.e., $\{j(1), j(2), \ldots\}$ is the sequence of hitting times to the set $\{m, m + 1\}$.

We claim that

$$|M_m(j(i)) - M_{m+1}(j(i))| \leq f(j(i))$$

a.s. for all but finitely many $i$. Consider the Bernoulli process $\{T(j(i)) : i \in \mathbb{Z}^+\}$. It is well known, and a simple consequence of the Borel-Cantelli lemmas (see, for example, Billingsley [3], pp. 53–54) that there exists a positive constant $K$ such that the events

$$\{T(j(i)) = T(j(i) + 1) = \cdots = T(j(i) + K \ln i) = m\}$$

and

$$\{T(j(i)) = T(j(i) + 1) = \cdots = T(j(i) + K \ln i) = m + 1\}$$

each occur for only finitely many values of $i$. Let $i_0$ be the largest $i$ for which either of these events occurs.

Since zero is a recurrent state in the random walk on the space of possible values of $M_m - M_{m+1}$, $M_m(j(i)) = M_{m+1}(j(i))$ for arbitrarily large
values of \( i \). Let \( i_t \) be the smallest integer larger than \( i_0 \) for which \( M_m(j(i_t)) = M_{m+1}(j(i_t)) \). Then for all \( i \geq i_t \),

\[
|M_m(j(i)) - M_{m+1}(j(i))| < K \ln i \leq K \ln j(i) < f(j(i)); \tag{4.2}
\]

where the last inequality is true for all large enough \( i \).

We finish by showing that (4.2) is still true when \( j(i) \) is replaced by an arbitrary time \( n \). Observing that \( j(\cdot) \) is a non-decreasing function on the positive integers, we define an inverse. If \( n < j(1) \), set \( j^*(n) = 0 \); otherwise, let \( j^*(n) \) be the largest integer \( i \) such that \( j(i) \leq n \). The function \( j^*(\cdot) \) has the following properties:

\[
j^*(n) \leq j(j^*(n)) \leq n;
\]

\[
M_m(j(j^*(n))) = M_m(n);
\]

\[
M_{m+1}(j(j^*(n))) = M_{m+1}(n).
\]

Thus, a.s. for all large \( n \),

\[
|M_{m+1}(n) - M_m(n)| = |M_{m+1}(j(j^*(n))) - M_m(j(j^*(n)))| \\
\leq K \ln(j^*(n)) \\
\leq K \ln n \\
\leq f(n). \quad \square
\]

4.4 Asymptotically Stable Estimators: Second Example

We drop previous assumptions about the error terms \( \xi_n \), and instead assume that they are iid with a zero mean and finite variance \( \sigma^2 \). We compute the sample means \( \overline{T}_n \) to a higher precision \( d \) than needed for our final estimator \( H_n \).
Let $g(n)$ be a function which satisfies
\[
\lim_{n \to \infty} \frac{\sqrt{2(\sigma^2/n) \ln \ln \sigma^2 n}}{g(n)} \leq 1 \quad \text{and} \quad \lim_{n \to \infty} g(n) = 0,
\]
almost surely. For example, we could take $g(n) = n^{-1/2} s^2 \ln \ln s^2 n$, where $s$ is the sample variance (this requires that the $\xi_n$ have a finite fourth moment).

More simply, we could take $g(n) = \ln n/n$.

**ALGORITHM STABLE2($x$, $t$)**

For each step $n$
- Make an observation $T(n)$
- Compute the sample mean $\bar{T}(n)$
- Set $J(n) \leftarrow \lfloor \bar{T}(n) \rfloor + (1/2)$
- If $|T_n - J(n)| \leq g(n)$, then
  - Set $H(n) \leftarrow \lfloor \bar{T}(n) \rfloor$
- Else
  - Set $H(n) \leftarrow \langle \bar{T}(n) \rangle$
End

By the law of the iterated logarithm,
\[
|\bar{T}(n) - t| \leq \phi(n) \sqrt{2/n} \leq g(n)
\]
for all but finitely many $n$, where $\phi(n) = \sqrt{\sigma^2 \ln \ln n \sigma^2}$. When $t = m + 1/2$ for some integer $m$, then $J(n) = t$ for all large $n$, so that STABLE-2 is stable by (4.3). When $t$ is not a half-integer,
\[
2g(n) < |J(n) - h|
\]
for all large $n$; combining this with (4.3),
\[
|J(n) - \bar{T}(n)| \geq |J(n) - t| - |\bar{T}(n) - t|
\]
\[
\geq |J(n) - h| - |\bar{T}(n) - t|
\]
\[ > 2g(n) - g(n) \]
\[ = g(n), \]

for all large enough \( n \), almost surely. We have proved

**Theorem 4.2** Algorithm 2 is stable and consistent.
CHAPTER 5

PATHWISE CONVERGENCE WITH ADAPTIVE COOLING

Up to this point, we have measured success of an algorithm by convergence in probability to the set of optimal states $S_0$. This standard may seem to say more than it does. Since there is no strong law of large numbers for the general nonstationary Markov chain, there is \textit{a priori} no guarantee that optimal states will dominate others in any sample run of the algorithm. In this chapter, we study pathwise convergence directly and show that under certain circumstances the proportion of optimal states visited does indeed go to one, almost surely. (However, in general the sequence of states visited has no almost sure limit—see, for example, [2].) We prove that this is true for the algorithms SLOW and QUICKER under Hajek’s conditions, and, under additional conditions different from those in previous chapters, for algorithm QUICKER-1. An example shows that the additional conditions cannot in general be discarded.

It is not clear that the pathwise-convergence results of this chapter either imply or are implied by convergence in probability; see, however, the discussion at the end of Section 5.1.

Our work was stimulated by that of Borkar [4], who applies the idea of “recurrence orders” first developed by Connors and Kumar [11] to obtain a sample-path result similar to ours on a deterministic cooling schedule satisfying

$$\sum_t c(t) = \infty$$
when $c$ is not less than a problem-specific constant $\hat{d}$ possibly different from, but no smaller than, Hajek's $d^\star$. We slightly improve Borkar's result, showing that it holds for $c \geq d^\star$.

Next, we exhibit conditions under which pathwise convergence occurs with an adaptive schedule (i.e., with QUICKER-1), and then present an example of a state space which fails to meet those conditions and in which pathwise convergence fails.

5.1 Main Results

Theorem 5.1 In simulated annealing with the deterministic schedule $\epsilon(t) = t^{-1/c}$, we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} I(X_t \in S_0) = 1 \quad a.s.,$$

for any $c \geq d^\star$.

Remark Borkar [4] proves the same result, except that in place of $d^\star$ he uses the constant

$$\hat{d} = \max_{x \in S} d(x),$$

where $d(\cdot)$ is the depth function of Section 1.1. Since $d^\star$ is the same maximum over the smaller set $S \setminus S_0$, the two constants are related by $d^\star \leq \hat{d}$, and our result is a slight improvement over Borkar's.

Now we describe what happens for algorithm QUICKER-1. Recall from Section 2.1 that $\bar{d}$ is the maximum over $x \in S$ of $\delta(x)$, where

$$\delta(x) = \min \{[U(y) - U(x)]^+ : y \in N(x)\}.$$  

Let $H = \min\{U(x) : x \notin S_0\}$ be the smallest objective-function value outside the set of global optimizers. Let $T(k)$ be the epoch of the $k$-th state change;
i.e., $T(1) = 1$ and

$$T(k + 1) = \min\{t > T(k) : X_t \neq X_{T(k)}\}. \quad (5.2)$$

Finally, let $\tilde{t}$ be the epoch of entry to the state $X_{\tilde{t}}$; i.e., $\tilde{t} = T(k)$, where $k$ is the unique integer such that $T(k) \leq t < T(k + 1)$.

**Theorem 5.2** In simulated annealing with the adaptive schedule $\epsilon(t) = \tilde{t}^{-1/c}$, (5.1) holds for any $c \geq d^*$, provided $\tilde{d} < H$.

We show in Section 5.3 that the condition $\tilde{d} < H$ cannot in general be omitted.

**Remark** Hajek [25] shows how to modify $(S, U, R)$, without changing its convergence properties, by inserting intermediate states so that all jumps are of a desired small size. He calls this the "continuous increase property." In principle, we could use this construction to force $\tilde{d}$ to be smaller than $H$. However, this expansion of the state space tends to degrade the performance of an adaptive algorithm, since it results in more frequent state changes and thus lower temperatures.

The results described in Theorems 2.1 and 2.2 of Section 1.2 imply an expected-value version of Theorem 5.1, since if $\Pr\{X_t \in S_0\} \to 1$, then

$$\lim E\left(\frac{1}{n} \sum_{t=1}^{n} I\{X_t \in S_0\}\right) = \lim \left(\frac{1}{n} \sum_{t=1}^{n} \Pr\{X_t \in S_0\}\right) = 1. \quad (5.3)$$

For deterministic cooling schedules, Connors and Kumar [11, 12] prove a result like (5.3) directly. (Their result involves a lim sup rather than a limit, but with cooling schedules satisfying (1.2), it is easy to generalize their proof.)

47
5.2 Sample-Path Convergence for Deterministic Cooling

Following Borkar [4], we associate with each state $x$ a random quantity $\beta(x)$, the (pathwise) recurrence order of state $x$, defined as follows:

$$\beta(x) = \begin{cases} 
-\infty & \text{if } \sum_t I[X_t = x] < \infty \\
p^{-} & \text{if } p = \sup\{c \geq 0 \mid \sum_t \epsilon(t)^c I[X_t = x] = \infty\} \text{ and } \sum_t \epsilon(t)^p I[X_t = x] < \infty \\
p & \text{if } p = \max\{c \geq 0 \mid \sum_t \epsilon(t)^c I[X_t = x] = \infty\}.
\end{cases}$$

Similarly, we can define the pathwise recurrence order of a transition from state $x$ to state $y$:

$$\beta(x, y) = \begin{cases} 
-\infty & \text{if } \sum_t I[X_t = x, X_{t+1} = y] < \infty \\
p^{-} & \text{if } p = \sup\{c \geq 0 \mid \sum_t \epsilon(t)^c I[X_t = x, X_{t+1} = y] = \infty\} \text{ and } \sum_t \epsilon(t)^p I[X_t = x, X_{t+1} = y] < \infty \\
p & \text{if } p = \max\{c \geq 0 \mid \sum_t \epsilon(t)^c I[X_t = x, X_{t+1} = y] = \infty\}.
\end{cases}$$

The recurrence order measures how often a state is visited, or a transition occurs, during a sample run. A state which is transient (in the ordinary sense) has recurrence order $-\infty$. Connors and Kumar ([11],[12]) work with the expected values of pathwise recurrence orders, quantities which we shall call mean recurrence orders. By replacing recurrence orders by mean recurrence orders in what follows, we get a result like (5.3) for QUICKER-1.

The following remarkable theorem shows that, to a certain extent, recurrence orders constitute a potential for objective-function values.

**Theorem 5.3** Assume that $S$ is finite and that weak reversibility holds. If $\beta(x) \geq 0$ and there is a path $x = x_0, x_1, \ldots, x_p = y$ such that $\max_{1 \leq i \leq p} (U(x_i) - U(x)) \leq \beta(x)$, then $\beta(y) = \beta(x) + U(x) - U(y)$, a.s.
Connors and Kumar prove Theorem 5.3 in [11] under the assumption of neighborhood symmetry and in [12] under the assumption of weak reversibility. Their result is stated in terms of mean recurrence orders, but the proof is algebraic, not probabilistic, and carries over without difficulty to sample-path recurrence orders.

The next three results are adapted from Borkar [4]. We sketch their proofs. The final theorem is slightly stronger than the corresponding result in [4], because we use $d^*$ rather than $\hat{d}$.

**Lemma 5.4** In simulated annealing with a deterministic cooling schedule implying $C_1 t^{-1/c} \leq \epsilon(t) \leq C_2 t^{-1/c}$ for $C_1 > 0$, $C_2 > 0$, $c > 0$, we have $\max_{x \in S} \beta(x) = c$, a.s.

**Proof:** For $b > c$, $\sum_t \epsilon(t)^b \leq C_2 \sum_t t^{-b/c} < \infty$; thus $\max_x \beta(x) \leq c$.

In addition,

$$\sum_x \sum_t \epsilon(t)^c I\{X_t = x\} = \sum_t \epsilon(t)^c \geq C_1 \sum_t t^{-1} = \infty;$$

therefore $\sum_t \epsilon(t)^c I\{X_t = x\} = \infty$ for some $x$ on each sample path, and so $\max_x \beta(x) \geq c$, a.s. □

**Lemma 5.5** If $\max_x \beta(x) \geq d^*$, then the set $\{x : \beta(x) = \max_x \beta(x)\}$ is a subset of $S_0$, almost surely.

**Proof:** Let $x$ be a state such that $\beta(x) = \beta_{\text{max}}$. Suppose $x \notin S_0$. By the definition of $d^*$, there is a path $x = x_0, x_1, \ldots, x_p = y$ from $x$ to some state $y \in S_0$ such that $\max_i U(x_i) \leq U(x) + d^* \leq U(x) + \beta(x)$. Applying Theorem 5.3, $0 < U(x) - U(y) = \beta(y) - \beta(x)$ a.s., which contradicts the choice of $x$. Thus, except on a set of probability 0, $x \in S_0$. □

**Proof of Theorem 5.1** From Lemma 5.4, $\max_x \beta(x) = c \geq d^*$. But
from Lemma 5.5, the only states $x$ satisfying $\beta(x) = c$ are in $S_0$. Thus
\[
\sum_t \epsilon(t)^c I(x \notin S_0) = \sum_t t^{-1} I(x \notin S_0) < \infty \quad \text{a.s.,}
\]
from which (5.1) follows by Kronecker’s lemma (see, for example, Chung [10], pp. 123–24). \qed

5.3 Sample-Path Convergence for Adaptive Cooling

In this section we prove Theorem 5.2. Assume that $\bar{\epsilon}(t) = \bar{t}^{-1/c}$, $c \geq d^*$, and that $\bar{d} < H$. To avoid repetition, we adopt the convention that all the statements in this proof are true in an almost-sure sense. The following assertion is proved below:

\[
c \leq \max_x \beta(x) \leq c + \bar{d}
\] (5.4)

Statement (5.4) and $\bar{d} < H$ together imply that $\max_x \beta(x) - H < c$, so that Theorem 5.3 implies that $\beta(x) < c$ for any state $x$ not in $S_0$. Thus
\[
\sum_t t^{-1} I(X_t \notin S_0) \leq \sum_t \bar{t}^{-1} I(X_t \notin S_0) = \sum_t \bar{\epsilon}(t)^c I(X_t \in S_0) < \infty,
\]
and (5.1) follows from Kronecker’s lemma as before.

To prove (5.4), observe first that
\[
\sum_t \bar{\epsilon}(t)^c = \sum_t \bar{t}^{-1} \geq \sum_t t^{-1} = \infty,
\]
so that $c \leq \max_x \beta(x)$. Pick $b > c + \bar{d}$. We need to show that
\[
\sum_t \bar{\epsilon}(t)^b = \sum_t \bar{t}^{-b/c} < \infty.
\] (5.5)

Under QUICKER-1, the sojourn time in each state has a geometric distribution and so is a.s. finite; thus, there are infinitely many state changes. Recalling
that \( T(k) \) is the transition number of the \( k \)-th state change, we rewrite (5.5) as
\[
\sum_t \dot{c}(t)^b = \sum_k (T(k + 1) - T(k)) T(k)^{-b/c} < \infty;
\] (5.6)
then choose \( \xi > 0 \) such that \( b > (c + \bar{d})(1 + \xi) \). We claim that
\[
A_k \overset{\text{def}}{=} (T(k + 1) - T(k)) T(k)^{-b/c} < k^{-(1+\xi)}
\] (5.7)
for all but finitely many \( k \), almost surely. This is enough to establish (5.6) and thus (5.4).

Fix \( k \), and say that \( X_{T(k)} = x \). Let \( N^*(x) \) be the (nonempty) set of neighbors of \( x \) at height \( \delta(x) \) above \( x \), i.e.
\[
N^*(x) = \{ y \in N(x) : U(y) = U(x) + \delta(x) \}. \tag{5.8}
\]
The sojourn time in \( x \) ends at or before a move to \( N^*(x) \) is proposed and accepted. Thus, it is stochastically bounded by a geometric variate with parameter
\[
\rho T(k)^{-\delta(x)/c} \geq \rho T(k)^{-\bar{d}/c},
\]
where
\[
\rho = \sum_{y \in N^*(x)} R(x,y) > 0
\]
is the probability of proposing a move to \( N^*(x) \).

Since \( T(k) \geq k \), we have
\[
\frac{\log T(k)}{c} - \frac{\log k}{c + d} \geq \log k \left( \frac{1}{c} - \frac{1}{c + d} \right) \geq \frac{d}{c + d} > 0
\]
for \( k \geq 3 \); exponentiating gives
\[
T(k)^{1/c} k^{-1/(c+d)} > e^{\bar{d}/(c+d)} > 1.
\]
This implies that

\[ T(k)^{\frac{b}{c}k^{-\frac{1+\xi}{c}}} > T(k)^{\frac{b}{c}k^{-\frac{b}{c}}(c+\bar{d})} > e^{\frac{b}{c}d(c+\bar{d})} > 1, \]

so that there is a constant \( \eta \) such that for all large \( k \),

\[ \left[ T(k)^{\frac{b}{c}k^{-\frac{1+\xi}{c}}} \right] \geq \eta \cdot \left( T(k)^{\frac{b}{c}k^{-\frac{1+\xi}{c}}} \right). \quad (5.9) \]

Thus,

\[ \Pr \left( A_k > k^{-\frac{1+\xi}{c}} \mid T(k) \right) = \Pr \left( T(k+1) - T(k) > k^{-\frac{1+\xi}{c}}T(k)^{\frac{b}{c}} \mid T(k) \right) \leq \left( 1 - \rho T(k)^{-\frac{\bar{d}}{c}} \right) \left[ k^{-\frac{(1+\xi)}{c}T(k)^{\frac{b}{c}}} \right]. \]

Since \( (1 - a)^r \leq e^{-ar} \) for any real \( a \) and positive real \( r \), this implies

\[ \Pr \left( A_k > k^{-\frac{1+\xi}{c}} \mid T(k) \right) \leq \exp \left( -\rho \left[ k^{-\frac{(1+\xi)}{c}T(k)^{\frac{b}{c}}} \right] T(k)^{-\frac{\bar{d}}{c}} \right) \leq \exp \left( -\rho \eta k^{-\frac{(1+\xi)}{c}T(k)^{\frac{b}{c}}(\frac{\bar{d}}{c})} \right). \]

Then the trivial bound \( T(k) \geq k \) implies

\[ \Pr \left( A_k > k^{-\frac{1+\xi}{c}} \mid T(k) \right) \leq \exp \left( -\eta \rho k^{\frac{(b+\bar{d})}{c} - (1+\xi)} \right). \quad (5.10) \]

The choice of \( \xi \) implies that the exponent of \( k \) on the right is strictly positive.

Now for any \( \alpha > 0 \),

\[ \sum_{j=1}^{\infty} \exp(-j^\alpha) < \int_0^{\infty} \exp(-x^\alpha) \, dx = \frac{1}{\alpha} \Gamma \left( \frac{1}{\alpha} \right) < \infty. \quad (5.11) \]

By the Borel-Cantelli lemma for adapted sequences (see Loève [36], pp. 398–99), the event \( A_k > k^{-\frac{1+\xi}{c}} \) occurs only finitely often, almost surely. This proves (5.6) and finishes the proof of Theorem 5.2. \( \square \)
5.4 An Example Where Pathwise Convergence Fails

Consider the state space $S = \{x, y, z\}$, with $U(x) = 1$, $U(y) = 2$, $U(z) = 0$, and tentative-move matrix

$$R = \begin{pmatrix}
  \cdot & 1 & \cdot \\
  r & 1 - r & \cdot \\
  \cdot & 1 & \cdot
\end{pmatrix},$$

where $r > 0$. Thus, $d^* = 1$, $\bar{d} = 2$, and $H = 1$. We show that pathwise Cesàro convergence fails for the adaptive schedule $\bar{c}(t) = t^{-1/d^*} = t^{-1}$. In fact, given any $\alpha < 1$, there are infinitely many transition numbers $t$ at which

$$t^{-1} \sum_{s=1}^{t} I(X_s = x) > \alpha, \quad \text{a.s.},$$

that is, the total proportion of visits to the strict local minimum $x$ gets arbitrarily close to 1, infinitely often.

Assume for simplicity that $X_1 = y$. Since we are using an adaptive schedule, there are infinitely many state changes, almost surely. In particular, there are infinitely many visits to state $y$. Let $S(j)$ denote the transition number immediately following the $j$-th visit to $y$; that is, $S(1) = 2$ and $S(k + 1) = 1 + \min\{t > S(k) : X_t = y\}$. Also, let $M(t)$ the remaining lifetime, at step $t$, of the current visit to state $x$; that is, $M(t) = \min\{s > 0 : X_{s+t} \neq x\}$ if $X_t = x$, and $M(t) = 0$ if $X_t \neq x$.

Since every proposed move out of state $y$ is accepted, we have

$$\Pr(X_{S(j)} = x) = r$$

for all $j$. Let $B_j$ be the event

$$\left\{X_{S(j)} = x, \ M(S(j)) > \left\lfloor -\ln 2 / \ln(1 - S(j)^{-1}) \right\rfloor \right\}.$$
then

\[
\Pr(B_j) = r \cdot \left(1 - S(j)^{-1}\right)^{\left[-\ln 2 / \ln(1 - S(j)^{-1})\right]}
\]

\[
= r \cdot \exp \left\{ \left[-\frac{\ln 2}{\ln(1 - S(j)^{-1})}\right] \ln(1 - S(j)^{-1}) \right\}
\]

\[
\geq r/2. \tag{5.12}
\]

The events \(B_j\) are not independent. However, \(B_1, \ldots, B_{j-1}\) influence \(B_j\) only through \(S(j)\); since (5.12) is true regardless of the value of \(S(j)\), we have

\[
\Pr(B_j \mid \Lambda) \geq r/2 > 0
\]

for any \(\Lambda \in \sigma(B_1, \ldots, B_{j-1})\). From Lemma 5.6 (see below) it now follows that given any integer \(l\), sequences of \(l\) consecutive events \(B_j, B_{j+1}, \ldots, B_{j+l-1}\) occur infinitely often. Since for any integer \(k\),

\[
\left[ -\frac{\ln 2}{\ln(1 - k^{-1})} \right] \geq |k \ln 2| > k/2,
\]

it follows that if \(B_j, \ldots, B_{j+2l-1}\) all occur, then at time \(S(j + 2l) - 1\), there will have occurred at least \(l \cdot S(j)\) visits to state \(x\) and at most \(S(j)\) visits to state \(z\). \(\Box\)

**Remark** A similar argument shows that (a.s.) the total proportion of visits to the global minimum \(z\) also gets arbitrarily close to one infinitely often. Thus, the Cesàro sums \(t^{-1} \sum_s I(X_s = x)\) and \(t^{-1} \sum_s I(X_s = z)\) oscillate forever without converging.

**Lemma 5.6** Let \(I_1, I_2, \ldots\) be a sequence of zero-one Bernoulli random variables adapted to an increasing sequence \(\mathcal{F}_1, \mathcal{F}_2, \ldots\) of \(\sigma\)-fields. Suppose that for all \(i\),

\[
\inf_{\Lambda \in \mathcal{F}_i} \mathbb{E}(I_{i+1} \mid \Lambda) > \alpha,
\]

54
for some positive constant $\alpha$ not depending on $i$. Then for any positive integer $l$, the sequence $(I_j)$ contains infinitely many strings of ones of length $l$.

**Proof:** Let $(I_j : j = 1, 2, \ldots), (\mathcal{F}_j : j = 1, 2, \ldots)$, and $\alpha$ be as in the statement of the lemma. We need to show that for any given integer $l > 0$, there is an infinite subsequence of indices $(i(j))$, $j = 1, 2, \ldots$, such that for each $j$, $I_{i(j)+1} = I_{i(j)+2} = \cdots = I_{i(j)+l-1} = 1$. The proof will show that we can take the $i(j)$ to be multiples of $l$.

Let $I_{jm} = \prod_{k=1}^{m} I_{j+k}$. We establish by induction on $m$ that

$$E(I_{jm} \mid \Lambda) > \alpha^m \quad (5.13)$$

for $\Lambda \in \mathcal{F}_j$. When $m = 1$, statement (5.13) is true by hypothesis. Assume (5.13). Then

$$E(I_{j,m+1} \mid \Lambda) = E(I_{j+1}I_{j+2}\cdots I_{j+m+1} \mid \Lambda)$$
$$= E(I_{j+1}\cdots I_{j+m+1} \mid I_{j+1} = 1, \Lambda)E(I_{j+1} \mid \Lambda)$$
$$= E(I_{j+1,n} \mid I_{j+1} = 1, \Lambda)E(I_{j+1} \mid \Lambda)$$
$$> \alpha^{m-1}\alpha,$$

since $\{I_{j+1} = 1\} \cap \Lambda \in \mathcal{F}_{j+1}$.

Now let $J_{jm} = I_{jm,j} = I_{j+1}I_{j+2}\cdots I_{j+m}$. If $j < k$, then the event \{\(J_{jm} = 0\)\} is in $\mathcal{F}_{(j+1)m} \subseteq \mathcal{F}_{km}$, so that (5.13) implies that for any $k > 0$ and $n > 0$,

$$\Pr\left(\bigcap_{j=k}^{n} (J_{jm} = 0)\right) = \Pr(J_{km} = 0, \ldots, J_{nm} = 0)$$
$$= \Pr(J_{km} = 0) \times \Pr(J_{k+1,m} = 0 \mid J_{km} = 0) \times \cdots$$
$$\times \Pr(J_{nm} = 0 \mid J_{km} = J_{k+1,m} = \ldots = J_{n-1,m} = 0)$$
$$\leq (1 - \alpha)^{n-k}.$$
letting $n \to \infty$ and then taking the union over all positive $k$ shows that

$$\Pr \left( \lim \inf \left( \frac{J_{jm}}{m} = 0 \right) \right) = \Pr(\forall j \exists m \geq j, J_{jm} = 0 \text{ all but finitely often})$$

$$= 0. \quad \square$$
CHAPTER 6

CONVERGENCE OF THE PRUNED CHAIN

It might be thought that the convergence results of [7], [11], [25], and [37], as well as those in previous chapters of this thesis, are an artifact, resulting from the algorithm spending long periods of time in optimal states. In this chapter we show that convergence occurs even if we do not take into account the length of time that the chain occupies a state, i.e., if we consider only the embedded chain of distinct states.

As in the previous chapter (see (5.2) on page 47), let the sequence \( \{T(k) : k = 1, 2, \ldots \} \) denote the epochs of transitions between distinct states. Then \( \{X_{T(k)} : k = 1, 2, \ldots \} \) is the induced chain of pairwise distinct states visited. Let \( S_1 \) be the set of global minimizers of the function \( U + \delta \). At the end of this chapter we describe the set \( S_1 \) for some typical problems. It often is a superset of \( S_0 \) and usually its component states have energy levels very close to the minimum. If there are no isolated local minima, as in the setting of [14], then \( S_1 = S_0 \).

**Theorem 6.1 (Convergence in Probability)** If \( (S, U, R) \) is weakly reversible and if the cooling schedule satisfies the conditions in [25], then

\[
\lim_{k \to \infty} \Pr(X_{T(k)} \in S_1) = 1.
\]
Theorem 6.2 (Pathwise Convergence) In simulated annealing using the adaptive schedule \( \epsilon(t) = \tilde{\epsilon}^{-1/c} \) for \( c \geq d^* + \tilde{d} \),

\[
\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} I(X_{T_k} \in S_1) = 1, \quad \text{a.s.}
\]

6.1 Convergence in Probability

Our procedure in both Theorem 6.1 and 6.2 is to show that the dynamics of the pruned chain are the same as those of ordinary simulated annealing on a triple \((S, U + \delta, \hat{R})\), where \( \hat{R}(x, y) > 0 \) iff \( R(x, y) > 0 \). The first task is to show that the modified problem satisfies \( WR \).

Lemma 6.3 \((S, U + \delta, \hat{R})\) has property \( WR \) if \((S, U, R)\) does.

Proof: The change from \( R \) to \( \hat{R} \) makes no difference, since any path under \( R \) is also a path under \( \hat{R} \) and vice versa. Suppose that state \( y \) is reachable from state \( x \) at \((U + \delta)\)-height \( h \). This means in particular that \( U(x) + \delta(x) \leq h \). Since the function \( \delta(\cdot) \) is nonnegative, it also means that \( y \) is reachable from \( x \) at \( U \)-height \( h \). Since \((S, U, R)\) satisfies \( WR \), there exists a path \( y = y_0, \ldots, y_p = x \) such that \( \max_{0 \leq i \leq p} U(y_i) \leq h \). Let \( y_i \) be a state on this path; we claim that \((U + \delta)(y_i) \leq h \). At the beginning of the proof we remarked that this is true for the state \( x \). Therefore, assume that \( 0 \leq i \leq p - 1 \); i.e., \( y_i \) is not the last state on the path. If \( U(y_i) > U(y_{i+1}) \), then \( \delta(y_i) = 0 \), so \((U + \delta)(y_i) = U(y_i) \leq h \), and if \( U(y_i) \leq U(y_{i+1}) \), then \( U(y_i) + \delta(y_i) \leq U(y_{i+1}) \leq h \). The claim is proved.

We have exhibited a path from \( y \) to \( x \) of \((U + \delta)\)-height not exceeding \( h \), which proves the lemma. \( \square \)
Lemma 6.4 If \((S, U, R)\) is weakly reversible, then

\[ [(U + \delta)(y) - (U + \delta)(x)]^+ = [U(y) - U(x) - \delta(x)]^+ \]

for all states \(x, y\) with \(y \in N(x)\).

**Proof:** If \(U(y) > U(x)\), then there must be a path \(y = y_0, y_1, \ldots, y_p = x\) leading from \(y\) to \(x\) at height not exceeding \(U(y)\). In particular, \(y\) has a neighbor not strictly uphill from it, namely \(y_1\), so that \(\delta(y) = 0\), and the conclusion of the lemma holds in this case. Suppose on the other hand that \(U(y) \leq U(x)\). Then there must be a path \(y = y_0, y_1, \ldots, y_p = x\) leading from \(y\) to \(x\) at a height not exceeding \(U(x)\). In particular, \(y_1 \in N(y)\) and \(U(y_1) \leq U(x)\), so that \(\delta(y) \leq U(x) - U(y)\). Together with \(\delta(x) \geq 0\), this implies

\[ (U + \delta)(y) - (U + \delta)(x) \leq U(y) + \delta(y) - U(x) \leq 0, \]

so that the conclusion of the lemma also holds in this case. \(\square\)

**Proof of Theorem 6.1:** It is sufficient to demonstrate that the condition (A.2) of [25] is satisfied; that is, that there exist constants \(c_1, c_2\) such that

\[ c_1 \epsilon(t)^{[(U + \delta)(y) - (U + \delta)(x)]^+} \leq \tilde{Q}(x, y, t, \epsilon) \leq c_2 \epsilon(t)^{[(U + \delta)(y) - (U + \delta)(x)]^+}, \quad (6.1) \]

where \(\tilde{Q}(x, y, t, \epsilon)\) denotes the transition probability, conditioned on acceptance, from state \(x\) to state \(y\) at time \(t\) under cooling schedule \(\epsilon(\cdot)\).

As in Section 2.2, we write the acceptance probability at time \(k\) from state \(x\) as

\[ \alpha(x, k) = A_0 + \sum_{i=1}^{n} A_i k^{-d_i / \epsilon}. \]
If $x$ is an isolated local minimum, then the probability of moving to $y$, conditioned on acceptance, is

$$\tilde{Q}(x, y, t, \epsilon) = \frac{cc(t)^{U(y) - U(x)}}{\sum_{i=1}^{n} A_i \epsilon(t)^{d_i(x)}}.$$ 

For some constant $K$ not dependent on $t$, the denominator satisfies the double inequality

$$A_1 \epsilon(t)^{\delta(x)} \leq \sum_{i=1}^{n} A_i \epsilon(t)^{d_i(x)} \leq K \epsilon(t)^{\delta(x)},$$

since the exponent in the numerator is strictly positive,

$$\frac{c}{K} \epsilon(t)^{U(y) - U(x) - \delta(x)} \leq \tilde{Q}(x, y, t, \epsilon) \leq \frac{c}{A_1} \lambda_i^{U(y) - U(x) - \delta(x)},$$

from which (6.1) follows by Lemma 6.4. This finishes the proof for the case where $x$ is an isolated local minimum. If $x$ is not an isolated local minimum, then

$$\tilde{Q}(x, y, t, \epsilon) = \frac{cc(t)^{U(y) - U(x)}}{A_0 + \sum_{i=1}^{n} A_i \epsilon(t)^{d_i(x)}}.$$ 

Here, the denominator is bounded below by $A_0$, which is strictly positive, and above by some (different) constant $K$, again not dependent on $t$. Therefore

$$\frac{c}{K} \epsilon(t)^{U(y) - U(x)} \leq \tilde{Q}(x, y, t, \epsilon) \leq \frac{c}{A_0} \epsilon(t)^{U(y) - U(x)}.$$ 

Since $\delta(x) = 0$ in this case, we also get (6.1). □

### 6.2 Pathwise Convergence for the Pruned Chain

In this section we prove Theorem 6.2. As in the proof of Theorem 6.1, we find expressions involving powers of $k$ to bound the transition probabilities of the $k$-th step in the pruned chain. The greatest difficulty lies in finding an asymptotic upper bound for $T(k)$, the time of the $k$-th transition; this is the
purpose of Lemmas 6.5 and 6.6. A modified version of the techniques used to prove Theorem 5.2 then establishes pathwise convergence.

**Lemma 6.5** On almost all sample paths, $T(k+1) - T(k) \leq T(k)^{\tilde{d}/\epsilon}$ for all except finitely many $k$.

**Proof:** Let $\xi > 0$ be arbitrary. We proceed as in the proof of inequality (5.10) on page 52. As in that proof, the exit time from any state is stochastically no greater than a geometric variate with parameter $T(k)^{-\tilde{d}/\epsilon}$; so for all large $k$,

$$
\Pr \{ T(k+1) - T(k) > T(k)^{\tilde{d}+\xi} \mid T(k) \} \leq (1 - T(k)^{-\tilde{d}/\epsilon})^{T(k)^{\tilde{d}/\epsilon} + \xi} \\
\leq \exp(-\eta T(k)^{\xi}) \\
\leq \exp(-\eta k^\xi),
$$

where $\eta$ is a positive constant small enough to compensate for the integer rounding. Applying (5.11), then the Borel-Cantelli lemma for adapted sequences as before, we have

$$
\Pr \{ T(k+1) - T(k) > T(k)^{\tilde{d}+\xi} \ i.o. \} = 0,
$$

for $\xi > 0$. Now for $j = 1, 2, \ldots$, let

$$
\Omega_j = \{ T(k+1) - T(k) > T(k)^{\tilde{d}+2^{-j}} \ i.o. \}.
$$

Then

$$
\Pr \{ T(k+1) - T(k) > T(k)^{\tilde{d}+2^{-j}} \ i.o. \} = \Pr \left\{ \bigcup_{j=1}^\infty \Omega_j \right\} \leq \sum_{j=1}^\infty \Pr \{ \Omega_j \} = 0. \quad \square
$$

**Lemma 6.6** $T(k) = O(k^{\epsilon/(\epsilon - \tilde{d})})$, almost surely.
Proof: For convenience, write $d = d/c$. Inductively define a function $f_k$ on the nonnegative integers by

$$f_k(0) = k; \quad f_k(j + 1) = f_k(j) + (f_k(j))^d.$$  

Now consider the related solution to the initial-value problem

$$g_k(0) = k; \quad g_k'(x) = (g_k(x))^d; \quad (6.2)$$

for $0 \leq d < 1$. It is easy to verify that this has the unique solution

$$g_k(x) = ((1 - d)x + k^{1-d})^{1/(1-d)}$$

on nonnegative real numbers $x$. Since the function $g_k$ is nondecreasing,

$$(g_k(j))^d \leq \int_j^{j+1} (g_k(x))^d \, dx = g_k(j + 1) - g_k(j),$$

and from this it easily follows by induction that

$$f_k(j) \leq g_k(j) = O(j)^{1/(1-d)}$$

for all nonnegative integers $j$.

On any sample path $\omega$, there is, by Lemma 6.5, a constant $K = K(\omega)$ such that $T(k + 1) - T(k) \leq T(k)^{\delta/c}$ for all $k \geq K$. Then

$$T(k) \leq f_K(k - K) = O(k)^{1/(1-d)} = O(k)^{\delta/(c-d)}. \quad \square$$

Proof of Theorem 6.2: As in Section 5.3, let $N^*(x)$ be the set of neighbors at height $\delta(x)$ above $x$. The transition probabilities for the pruned chain are

$$\Pr \left( X_{T(k+1)} = y \mid X_{T(k)} = x \right) = \frac{R(x,y)e(T(k)[U(y)-U(x)])^+}{\sum_{z \in N(x)} R(x,z)e(T(k)[U(z)-U(y)])^+}$$

62
\[
\begin{align*}
\sim & \quad \frac{\hat{R}(x, y)}{\sum_{z \in N^{*}(x)} \hat{R}(x, z)} \cdot \frac{\epsilon(T(k))^{(U(y) - U(x))^+}}{\epsilon(T(k))^{\hat{\beta}(x)}} \\
&= \hat{R}(x, y) \epsilon(T(k))^{[U(y) - (U + \hat{\delta})(x)]^+} \\
&= \hat{R}(x, y) \epsilon(T(k))^{[U + \hat{\delta})(y) - (U + \hat{\delta})(x)]^+},
\end{align*}
\]

where

\[
\hat{R}(x, y) = \frac{R(x, y)}{\sum_{z \in N^{*}(x)} R(x, z)}.
\]

Given \( \xi > 0 \), we have

\[
(1 - \xi) \hat{R}(x, y) \epsilon(T(k))^{[U + \hat{\delta})(y) - (U + \hat{\delta})(x)]^+ \leq \Pr \left( X_{T(k+1)} = y \mid X_{T(k)} = x \right)
\]
and

\[
\Pr \left( X_{T(k+1)} = y \mid X_{T(k)} = x \right) \leq \hat{R}(x, y) \epsilon(T(k))^{[U + \hat{\delta})(y) - (U + \hat{\delta})(x)]^+,
\]

for all large \( k \). Applying Lemma 6.6 and the trivial inequality \( T(k) \geq k \),

\[
k^{-1/(c - \delta)} \leq \epsilon(T(k)) \leq k^{-1/c}, \quad \text{a.s.}
\]

Therefore, the transition probabilities for the pruned chain are bounded below by

\[
(1 - \xi) \hat{R}(x, y) (k^{-1/(c - \delta)})^{[U + \hat{\delta})(y) - (U + \hat{\delta})(x)]^+}
\]
and above by

\[
\hat{R}(x, y) (k^{-1/c})^{[U + \hat{\delta})(y) - (U + \hat{\delta})(x)]^+}.
\]

This is equivalent to bounding the cooling schedule for the pruned chain by

\((1 - \xi)(k^{-1/(c - \delta)})\) and \( k^{-1/c} \).

To finish the proof, we need one more lemma. Recall the recurrence orders \( \beta(x) \) defined on page 48.
Lemma 6.7 If there exist constants $b, B, c, C$ such that

$$bt^{-1/c} \leq \epsilon(t) \leq Bt^{-1/C},$$

then $c \leq \max_{x \in S} \beta(x) \leq C$.

This is proved in exactly the same way as Lemma 5.4.

Applying Lemma 6.7, $c - \tilde{d} \leq \max_{x \in S} \hat{\beta}(x) \leq c$, where $\hat{\beta}(x)$ is the recurrence order of the pruned chain. An application of Lemma 5.5 completes the proof. □

6.3 $S_1$ Compared to $S_0$

We briefly discuss the set $S_1$ of global minimizers of the function $U + \delta$ and its relationship to the set $S_0$. The following two examples show that this relationship can attain either extreme—the sets can be completely coincident or completely disjoint.

Example 6.1 If there are no isolated global minima then $S_0 = S_1$. This is not surprising; in the unpruned chain, the expected waiting time in any individual state is $O(1)$; its asymptotic behavior should not differ greatly from the pruned chain, where we force the waiting time to be exactly 1.

Example 6.2 Let $S = \{w, x, y, z\}$, $N(w) = \{x\}$, $N(x) = \{w, y\}$, $N(y) = \{y, z\}$, $N(z) = \{y\}$, with $U(w) = 1.5$, $U(x) = 1$, $U(y) = 2$, and $U(z) = 0$. (See Figure 6.1).

In this case, the probability distribution of $X_{T(k)}$ converges to a unit mass on $\{w, x\}$, even though neither of these states is a global minimizer. Again, this is reasonable, since a visit to $z$ is always followed by a visit to $y$, but the chain is very likely to contain long oscillations between states $x$ and $w$.
Despite Example 6.2, we believe that convergence to the set $S_1$ would in most settings be an acceptable outcome. Some typical applications of simulated annealing to "real world" optimization problems illustrate this.

**Example 6.3** In Hajek's [25] formulation of the maximum-matching problem, a situation like Example 6.2 could not occur, since every move changes the objective function (number of edges) by one. In fact, any problem formulation which satisfies the "continuous increase assumption" of [25] has $S_0 \subseteq S_1$; moreover, the energy level of every state in $S_1$ is either minimal or one step away from minimal.

**Example 6.4** One of the earliest papers to propose simulated annealing as an optimization technique was [32]. In its formulation of the travelling salesman problem, a move between states (tours) consists of reversing one section of the tour. A situation like Example 6.2 could occur only if there were two tours such that

(i) each could be reached from the other by a single move;

(ii) no shorter tours could be reached from either by a single move; and
(iii) the length of each tour is shorter than the length of any tour one move away from an optimum.

Unless the arrangement of cities is very regular, or the number of cities very few, this situation would be rare. Ordinarily, elements of $S_1$ would be tours at most one move away from optimal, from which a true optimum could be found by a short post-processing routine.
CHAPTER 7

DIRECTIONS FOR FUTURE WORK

The nonstationary Markov chains of simulated annealing are a rich source of unanswered questions; we present a small selection.

7.1 Hitting Time

A useful complement to the convergence results of this thesis would be information about the first hitting time \( T \) to \( S_0 \). In a recent paper, Chiang and Chow [7] used Wentzell and Freidlin's theory of \( W \)-graphs to show that the second eigenvalue of \( P_t \), the transition matrix at time \( t \), is asymptotically

\[
\lambda(t) = 1 - e(t)^{d^*},
\]

where \( d^* \) is Hajek's constant, defined in Chapter 1. This is suggestive for the following reason. We know that

\[
\max_{x \in S} \Pr(T > t \mid X_k = x) = \left\| \prod_{j=k}^{t} \tilde{P}_j \right\|_{\infty}
\]

where \( \tilde{P}_j \) denotes the matrix \( P_j \) with all rows and columns corresponding to \( S_0 \) deleted. For large \( j \), it is well known (see, for example, Section 1 of [25]) that the stationary distribution (i.e., the principal eigenspace) of \( P_j \) is almost entirely concentrated on \( S_0 \). Therefore, \( \lambda(j) \) is nearly the same as the principal eigenvalue of \( \tilde{P}_j \). If there were a constant \( A \)

\[
\left\| \prod_{j=k}^{t} \tilde{P}_j \right\|_{\infty} \cong A \cdot \prod_{j=k}^{t} \lambda(j), \quad (7.1)
\]
at least for large $k$, then we could write

$$\Pr(T > t \mid X_k = x) \cong A \cdot \exp \left( - \sum_{j=1}^{t} \epsilon(j)^{d^*} \right),$$

and in principle estimate all the moments of $T$. (Of course, we would also like to compute or find good bounds for $A$.) However, a relationship like (7.1), which must depend on the special structure of the matrices $P_j$, has thus far eluded us.

### 7.2 Tabu Penalties and Simulated Annealing

Fox [14] incorporates $k$-step tabu penalties into the simulated annealing algorithm. Roughly, he contracts a new space with states having multiple components; candidates for the latter are $k$-tuples in $S^k$ which correspond to possible paths, without self-loops, of length $k$ in $S$. Unfortunately, the modified problem does not in general satisfy $WR$. To avoid this difficulty, Fox imposes additional structure. Among other things, he requires that from any state there is a path with no uphill moves leading to $S_0$; this is achieved by incorporating a generalization of the “neighborhood enrichment” mentioned in Section 3.2.

Neighborhood enrichment has heuristic appeal as a way to diversify the search. Also, an adaptation of the proof of Bélisle [2] shows that Fox’s procedure guarantees convergence in probability to $S_0$ with any cooling schedule that goes to zero, no matter how quickly—even in the extreme case $\epsilon(t) = 0$ for all $t$. Nevertheless, we feel that insight could be gained by reconsidering the general problem without neighborhood enrichment. Are there other, weaker conditions which guarantee $WR$ in the expanded state space? An answer to this question would likely increase our understanding of the property of weak
reversibility. Moreover, it might allow us to quantify in some sense the benefits of tabu search; this would have both theoretical and practical interest.

7.3 Noisy Objective Function Revisited

In Chapter 4, we assumed that the objective function values were unknown but confined to a known finite set. Even though we can force an arbitrary degree of precision, the assumption is still somewhat unsatisfying, at least from the point of view of traditional mathematics. Dropping the assumption would allow us to replace the asymptotically stable estimators of Chapter 4 by estimators which are merely consistent.

Recall the procedure: every time we visit a state, we observe its energy level with a measurement subject to error. Past observations are combined to form a vector of estimates of the objective function $U(x)$ at each state $x$. If the estimator(s) are consistent, then after a random but a.s. finite time, the estimated value of every $U(x)$ will be within any prescribed $\xi > 0$ of its true value.

The first difficulty is that this is not a stopping time. We could get around this with a stronger version of Theorem 3.2. The theorem would need to hold when conditions (ii) and (iii) imply merely approximate equality of the respective transition probabilities. Under those weakened conditions, equalities (3.5) and (3.6) are no longer valid, and it is not immediately clear that they could be replaced with appropriate inequalities.

The next step would be to show that knowing the energy levels to within an arbitrary $\xi$ suffices. We are looking at a Markov chain where the
transition probabilities at time $t$ satisfy some inequality

$$A \xi(t)^{-\xi+U(y)-U(x)^+} \leq P(x, y, t) \leq B \xi(t)^{\xi+U(y)-U(x)^+}. \quad (7.2)$$

It might turn out that (7.2) is enough to guarantee convergence. Alternatively, assuming, as do Gelfand and Mitter [19], that the error term $\xi$ does not depend on $x$ or $y$, a system satisfying (7.2) can be viewed as ordinary simulated annealing with a random cooling schedule $\bar{\xi}(\cdot)$. Assuming the original cooling schedule satisfies (1.2),

$$\bar{\xi}(t) = t^{-(1+\hat{\xi})}, \quad (7.3)$$

where $\hat{\xi}$ is a random variable in the interval $(a, b)$. In effect, we move the randomness from the objective function to the cooling schedule. To get convergence in probability, we need to show that convergence results such as those in [7] or [25] continue to be true if the cooling schedule under the weaker condition (7.3). It is not yet clear whether this is possible.

### 7.4 Refinement of Algorithm QUICKER-1

Fox has suggested replacing the geometric variate with its expectation (rounded upward to the nearest integer) in algorithm QUICKER-1. A proof like that in in Section 5.3 (but easier) shows that the modified algorithm exhibits the same pathwise convergence discussed there. We do not know if Fox’s modified QUICKER-1 gives convergence in probability; this would be worth investigating.

Fox’s modified QUICKER-1 would be slightly simpler to implement. It seems to reduce variance in the random cooling schedule. However, more work is needed to determine whether this translates into reduced variance in the
sequence of states generated. The modified algorithm would prevent anomalous rapid cooling. Heuristically, it can be argued that this is desirable. Once again, though, more work is needed to determine whether this improves the performance of the algorithm.
BIBLIOGRAPHY


