EFFICIENCY OF SIMULATED ANNEALING: ANALYSIS BY RAPIDLY-MIXING MARKOV CHAINS AND RESULTS FOR FRACTAL LANDSCAPES

by

Gregory B. Sorkin

Memorandum No. UCB/ERL M91/12

12 February 1991

EFFICIENCY OF SIMULATED ANNEALING: ANALYSIS BY RAPIDLY-MIXING MARKOV CHAINS AND RESULTS FOR FRACTAL LANDSCAPES

by

Gregory B. Sorkin

Memorandum No. UCB/ERL M91/12

12 February 1991

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

EFFICIENCY OF SIMULATED ANNEALING: ANALYSIS BY RAPIDLY-MIXING MARKOV CHAINS AND RESULTS FOR FRACTAL LANDSCAPES

by

Gregory B. Sorkin

Memorandum No. UCB/ERL M91/12

12 February 1991

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

[1] The author is completing his doctoral studies with Prof. Alberto Sangiovanni-Vincentelli. He is on leave from the I.B.M. Thomas J. Watson Research Center.

Abstract

We present a new theoretical framework for analyzing simulated annealing. The behavior of simulated annealing depends crucially on the "energy landscape" associated with the optimization problem: the landscape must have special properties if annealing is to be efficient.

We prove that certain fractal or linearly-separable properties of the energy landscape are sufficient for simulated annealing to be efficient in the following sense: A solution of relative energy no more than ε – that is, a solution whose expected energy differs from the minimum by no more than ε times the full energy range of the problem – can be found in time polynomial in $1/\varepsilon$, where the exponent of the polynomial depends on certain parameters of the fractal. Higher-dimensional versions of the problem can be solved with almost identical efficiency.

The cooling schedule used to achieve this result is the familiar geometric schedule of annealing practice, rather than the logarithmic schedule of previous theory. Our analysis is more realistic than those of previous studies of annealing in the constraints we place on the problem space and the conclusions we draw about annealing's performance.

The techniques used are also new in this field. Annealing is modeled as a random walk on a graph, and recent theorems relating the "conductance" of a graph to the mixing rate of its associated Markov chain generate both a new conceptual approach to annealing and new analytical, quantitative methods. Another component in the analysis is an original and fundamental result: the expected energy cannot increase during annealing with monotonically decreasing temperatures. Surprisingly, with arbitrarily low but non-monotonic temperatures, the expected energy can be made arbitrarily high. An original analysis of annealing with any monotonically decreasing cooling schedule provides a strong, fundamental result which is one key component in our analysis of annealing on special landscapes.

The efficiency of annealing is compared with that of random sampling and descent algorithms. While these algorithms are more efficient for some cases, their run times increase exponentially with the number of dimensions, making annealing better for problems of high dimensionality.

We find that a number of circuit placement problems have energy landscapes with fractal properties, thus giving for the first time a reasonable explanation of the successful application of simulated annealing to problems in the VLSI domain.

Key words: Simulated annealing; Combinatorial optimization; Fractals; Rapidly-mixing Markov chains.

Contents

List of Figures

Chapter 1

Introduction

Simulated annealing is a general algorithm for finding good solutions to a wide variety of combinatorial optimization problems [13]. A finite graph $G = (V, E)$ and a function $f : V \rightarrow \mathbb{R}$ are given. The goal is to find a vertex $v \in V$ such that $f(v)$ is small. Without loss of generality, for convenience we assume f is scaled so that $\min_{v \in V} f(v) = 0$ and $\max_{v \in V} f(v) = 1$. V is called the set of "states" of the problem, E the set of "moves" between states, and $f(v)$ the "energy" of state v. Thinking of the graph as being drawn on the plane and the energies represented as elevations, we refer to the graph and energy function together as the "energy landscape". The annealing algorithm is shown in Figure 1.1.

What is meant by setting x to x' "with probability p" is this: A sequence of independent, uniformly distributed random real numbers between 0 and 1 is given. Each time an update decision is to be made, the next number in the sequence is compared with p. If the number is smaller than p then the update is made ("accepted").

K is the number of generations used for annealing, t_k is the number of attempts ("time") spent in generation k, and T_k is the temperature used in that generation. The sequence $\{(T_k, t_k)\}_{k=1}^{K}$ and the value K are referred to as the "cooling schedule". The choices of these parameters, and a corresponding estimate of the final energy $f(x)$, are the outstanding open problems in the practice and theory of simulated annealing.

Previous theoretical results describing the behavior of simulated annealing are few. For convenience we describe a cooling schedule by the set of pairs (T_k, t_k).

In 1982 it was independently shown by two groups [21, 16] that for "infinite-time" cooling schedules $t_k \equiv \infty$ and $T_k \rightarrow 0$, annealing "reaches equilibrium" at each temperature,

```
choose a sequence $(T_k, t_k)$ and value $K$
$x \leftarrow$ random state
for ``generation'' $k = 1$ to $K$ {
    for $t = 1$ to $t_k$ {
        $x' \leftarrow$ random neighbor of $x$
        if $f(x') < f(x)$ then $x \leftarrow x'$
        else
            $x \leftarrow x'$ (``accept'') with probability $e^{-(f(x')-f(x))/T_k}$
            otherwise $x$ unchanged (``reject'')
    }
}
output $x$
```

Figure 1.1: Annealing algorithm. (T_k, t_k) is a sequence of pairs representing respectively the kth temperature to be used for annealing, and the number of moves to be attempted at that temperature. We say t is the "time" within the kth "generation" of annealing.

and therefore as $T \to 0$ annealing finds a global minimum with probability 1. The proof is simple, and relies only on the most elementary theory of Markov chains.

A stronger result, proved in various forms by several groups including [6, 19, 7], is that for "logarithmic" cooling schedules $t_k \equiv 1$ and $T_k = a/\ln k$, with a sufficiently large, annealing approaches equilibrium ever more closely as time goes on and temperature decreases, and so again finds a global minimum with probability approaching 1.

While the latter theories give a sound basis for simulated annealing, their shortcomings are significant. The theorems as stated presume infinite time[1] and slow cooling schedules[2]. There is no hope of being able to weaken these suppositions significantly, since in general finding a global minimum with high probability *requires* that each state be examined and therefore that the time spent be at least the number of states. This is unrealistic, both because the number of states is typically exponential in the size of the problem, and because if each state is to be explored anyway we may as well use a simpler technique such

[1] The finite-time results of [19] are an exception.

[2] Practical cooling schedules are either geometric by construction or are "automatic", which come out to be approximately geometric anyway [22]. Any decreasing geometric sequence asymptotically approaches 0 faster than any logarithmic one.

as exhaustive enumeration or random sampling.

We seek a theory which applies to a restricted class of problems and which in exchange for this restriction allows a realistic cooling schedule (fast and perhaps geometric), yields strong finite-time results, and provides a mathematical description of the tradeoff between the time spent annealing and the expected energy of the solution produced. That is, we want a theory which explains why, for many practical problems such as those of VLSI layout, simulated annealing produces good results with fast cooling schedules. Previous theories do not explain this.

In this paper, we explore the relationship between energy landscapes and simulated annealing's performance. We begin in Chapter 2 with an intuitive argument that landscapes which are susceptible to annealing are self-similar, or fractal. In Chapter 3 we show that some practical landscapes – in particular those for circuit placement – do have random self-similarity properties. In Chapter 4 we derive some basic mathematical results concerning annealing, focusing on stationary properties and on the quality–time tradeoff for annealing at a constant temperature. The efficiency results for annealing in this general setting are poor, as expected. (Appendix A uses these tools to rederive the standard theoretical, asymptotic results for annealing with a "logarithmic" cooling schedule.) Chapter 5 shows that annealing with a cooling schedule whose temperatures do not decrease monotonically can produce surprising and undesirable results; it then gives a general and mathematically elegant description of the behavior of annealing with monotonically decreasing schedules. The conclusions of these two chapters are applied in Chapters 6 and 7 to show that annealing is comparatively efficient on a class of linearly separable functions. Similarly, Chapter 8 shows annealing to be efficient on a carefully constructed class of deterministic fractals. The efficiency of annealing on these fractals is compared with that of other similarly general algorithms in Chapter 9. We think the basis of these analyses is applicable to more realistic function models, and we hope to be able to extend it formally to landscapes including fractional Brownian motions.

Ideally we would like to identify a set of properties – perhaps fractal properties – satisfied by all practical problems on which annealing has been found successful, and then to prove that annealing works efficiently on any function possessing these properties. Unfortunately, such a result is some way off.

To keep them from interrupting the flow of the paper, most of the proofs have been deferred to appendices; they are indicated by the symbol $\boxed{\text{pf} \rightarrow \text{appendix}}$ in the text. We will be using throughout some notation that is not completely standard. By $f(x) \sim g(x)$ (read "$f(x)$ asymptotically equal to $g(x)$") we mean that under a limit to be inferred from the context, $\lim f(x)/g(x) = 1$. By $f(x) \lesssim g(x)$, "$f(x)$ asymptotically less than or equal to $g(x)$", we mean $\limsup f(x)/g(x) \leq 1$; of course $f(x) \gtrsim g(x)$ is to be interpreted symmetrically. The appearance of "const" in an equation indicates that the equation holds for some constant undeserving of a name of its own. Similarly "poly(x)" indicates a polynomial function.

Chapter 2

Dependence of Simulated Annealing on the Energy Landscape

In this chapter we wish to introduce the underlying ideas informally, showing why the performance of simulated annealing is strongly dependent on the "energy landscape" of the problem, and suggesting the properties the landscape must have in order for annealing to be efficient. In Chapters 3 and 8 respectively we will exhibit fractal properties observed in some real problems, and formally prove the efficient performance of annealing on a class of problems with qualitatively similar but more restrictive properties.

Suppose that the energy landscape is as sketched in the "bad" landscape of Figure 2.1(a). States are represented along the x axis, with adjacent states being neighbors. On this landscape, the energy differences between the low-energy states (the valley bottoms) are fairly small, while the energy barriers separating them (the mountains) are large. It is well known that the time required to cross a barrier of height h at temperature T is exponential in h/T, while the probability of a state of energy f is exponential in $-f/T$. So crossing the high barriers in reasonable time demands T large, while favoring the better valleys over the less good ones requires T small, implying that annealing cannot work both well and quickly on this space.

On the other hand, annealing should work well on a function like that of the "good" landscape of Figure 2.1(b). In this case, annealing can work hierarchically. Initially,

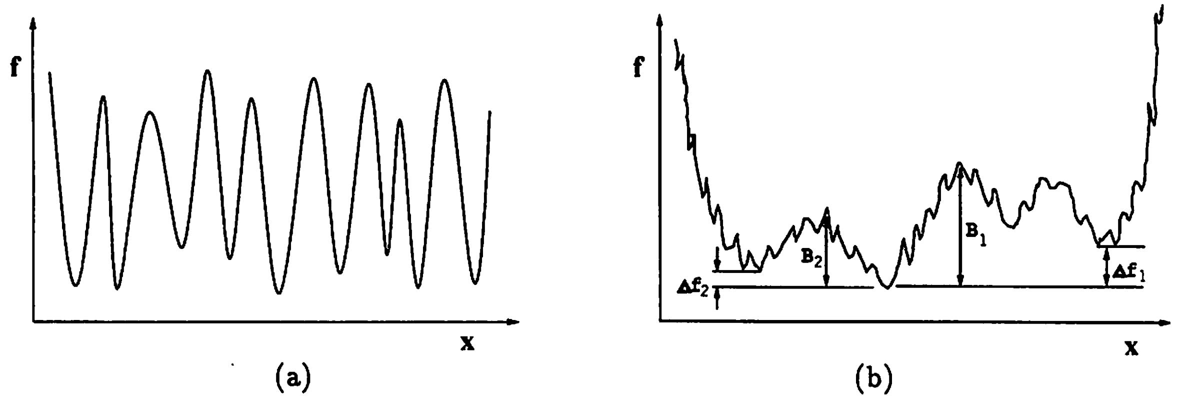


Figure 2.1: Sketch of landscapes which are bad (a) and good (b) for annealing.

the goal can be just to choose the better (left) of the two valleys separated by the tallest barrier B_1. While B_1 is large, the overall energy difference Δf_1 of the two valleys is also fairly large, so with a comparably sized value of T_1 we can be in the lower-energy value with high probability in short time.

Next we can aim to settle in the better (right) of the two valleys separated by the smaller barrier B_2. While the energy difference Δf_2 between these valleys is smaller than Δf_1 was, B_2 is also smaller than B_1, and so using T_2 similarly smaller than T_1 again allows us to be in the lower-energy valley with high probability in reasonable time. This process is repeated for smaller and smaller energy scales.

The success of annealing relies on the overall energy difference of collections of states being large compared with the barriers dividing these collections. So we would argue that either all the energy barriers and energy differences are of the same scale (which does not seem to be the case in practice), or else in smaller and smaller areas of the landscape the energy barriers must scale down along with the energy differences, giving a general "self-similarity"or "fractalness" like that of the "good" landscape of Figure 2.1(b).

Chapter 3

Fractalness of Circuit Placement Problem Landscapes

To support the relevance of the paper, we wish to show that some problems on which annealing works well are fractal.

We take the following as our basic definition of fractalness (see [24]):

Definition 3.0.1 *A random function* $f : \mathbb{R}^n \rightarrow \mathbb{R}$ *is a* **fractional Brownian motion (fBm)** *if the distribution of* $f(X')$ *conditional on* $f(X)$*,* X*, and* X'*, is normal with mean 0 and variance proportional to* $\|X' - X\|^{2H}$*, where* H *is a parameter in* $(0, 1)$.

As noted in [32], such functions have the statistical scaling property that for any r, $f(rX)$ is statistically indistinguishable from $r^H f(X)$. This may be thought of as the intuition underlying the definition. Figure 3.1 shows an example of this rescaling for Brownian motion. Brownian motion is the special case of fBm with $H = 1/2$, for which reducing the x scale by any factor r and the y scale by the factor $r^{1/2}$ yields a function statistically similar to the original one. Note that the rescaling $X \mapsto rX$ can be done in $\mathbb{R}^n$ but not in a general combinatorial setting, where X is just the vertex of a graph. However, in the latter setting the distance between two points can be defined as the length of a shortest path between them. So allowing distance functions $d(X, X')$ other than the Euclidean norm, we can naturally generalize Definition 3.0.1:

Definition 3.0.2 *A random function* f *on a metric space is* **fractal** *if the distribution of* $f(X')$ *conditional on* $f(X)$*,* X*, and* X'*, is normal with mean 0 and variance proportional to* $d(X', X)^{2H}$*, where* H *is a parameter in* $(0, 1)$.

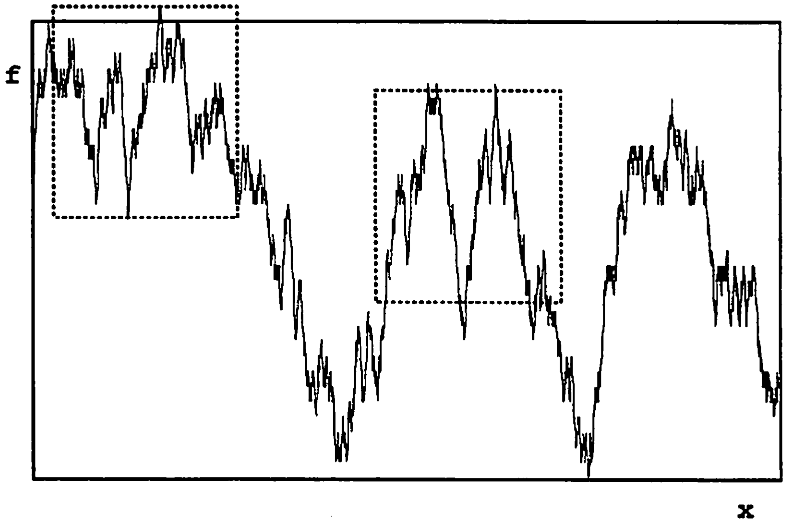


Figure 3.1: A sample of Brownian motion, a special case of fractional Brownian motion. When the portion of the graph in either small rectangle is expanded by 4X along the x axis and 2X along the y axis (to make it the same size as the whole figure), the resulting functions are statistically similar to the original one.

The distribution in Definition 3.0.2 (like that in Definition 3.0.1) is over the probability space from which f is drawn, but experimentally we will be sampling over random values of X and X' for a given sample function f. The equivalence of these two procedures is referred to as "ergodicity", and we take it for granted.

Analysis of the energy landscapes of a number of placement problems shows that they do indeed have fractal properties, and we believe these problems to be representative of VLSI layout problems.

We experimented on three placement problems. One was a real-world example with 87 objects; one was a random graph with 225 objects; and the last was also random with 225 objects, but was hierarchically constructed in conformance with Rent's Rule [9] with an exponent of 2/3. In each case placements in both one- and two-dimensional arrays were performed.

A state in any of these problems is given by an ordering of the objects, *i.e.* by a permutation of the numbers $1, \ldots, N$. The moves used were swaps of pairs of objects, making the distance between two states the minimum number of swaps required to turn one permutation into the other.

Because Definition 3.0.2 involves a set of the distributions (one for each value of distance), it is difficult to check. It is much simpler to measure the expectation of $[f(X') - f(X)]^2$, whose behavior is described by the following lemma:

Lemma 3.0.3 *If $f(X)$ is a fractal (per Definition 3.0.2), then*

$$\mathbf{E}[(f(X') - f(X))^2] \propto d(X', X)^{2H}. \tag{3.1}$$

We refer to any relationship of this form as a *power-law* relation.

Although as a condition on f equation (3.1) is weaker than Definition 3.0.2, we take satisfaction of (3.1) to be preliminary evidence of fractalness; we will discuss satisfaction of Definition 3.0.2 itself in a moment.

To check whether (3.1) is satisfied, random point pairs at various distances apart are generated, and the average squared energy difference of pairs a given distance apart is plotted against that distance. A typical experimental result is shown in Figure 3.2, derived from the real netlist placed in one dimension.

To check Definition 3.0.2 fully we need to verify that the distribution of energies for each distance is actually normal, in addition to having the predicted mean and variance. To do so, for each distance we constructed a quantile-quantile plot of the sampled energy differences against a normal distribution. In such a plot, normally-distributed data would show up as a straight line, excepting sampling deviations. A typical result is shown in Figure 3.3.

Instead of sampling random point pairs, fractal properties can also be inferred by taking *random walks* on the energy landscape. A random walk is a sequence of points $X(t)$ ($t = 0, 1, 2, \ldots$) where $X(t+1)$ is generated by a random move from $X(t)$. We study the energy timeseries $f(X(t))$.

This technique offers a number of advantages over the mean square energy difference versus distance method just discussed. First, the practical difficulty of finding point pairs separated by small distances is obviated. Second, interpretation of the energy versus distance data required parameter estimation (to fit a power-law function), and it is unclear what constitutes a statistically sensible estimation procedure in this context. By contrast, the energy versus time timeseries produced by the random walk can be subjected to established timeseries analysis techniques, including methods for estimating parameters and confidence intervals.

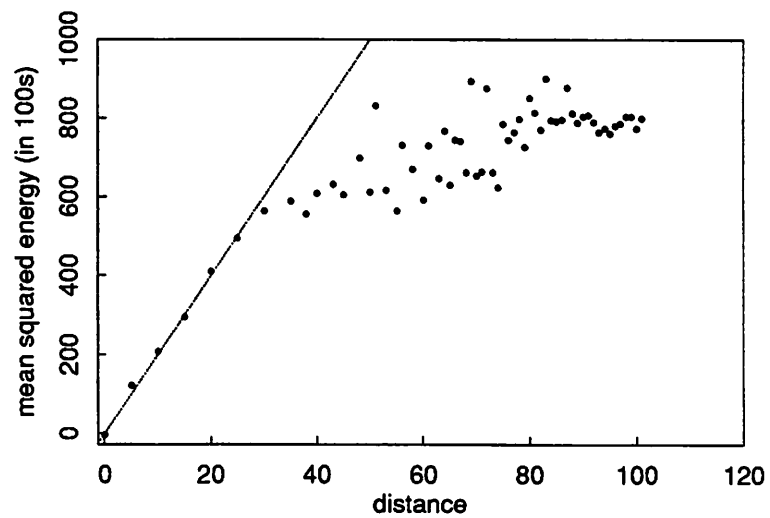


Figure 3.2: Mean squared energy differences of point pairs against the distances separating them. The good fit of the straight line to the small-distance portion of the data indicates that mean squared energy is approximately proportional to distance in this realm, and supports the interpretation that f is fractal in the sense of Definition 3.0.2.

The theoretical basis for the random walk method is contained in the following two lemmas.

Lemma 3.0.4 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Given any space where a random walk $X(t)$ satisfies $d(X(t), X(0)) = ct$, for some constant c. Let f be a fractal (per Definition 3.0.2) with parameter H on this space. Then $f(X(t))$ is a fBm (on $\mathbb{R}^1$) and has parameter H.*

Lemma 3.0.5 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Given a fBm f on $\mathbb{R}^n$ with parameter H. Make a random walk $X(t)$ on $\mathbb{R}^n$. Then $f(t) = f(X(t))$ satisfies equation (3.1) with parameter $\frac{1}{2}H$. Furthermore, for n large, $f(t)$ is approximately a fBm (on $\mathbb{R}^1$) with parameter $\frac{1}{2}H$.*

Here the "random walk" we make on $\mathbb{R}^n$ must be a Brownian motion – the limiting case of a random walk taking a large number of small steps – but this is a minor technicality.

Analysis of the combinatorial structure of permutation spaces with pairwise interchange moves shows that in such spaces, the expected distance $d(X(t), X(0))$ traversed by a

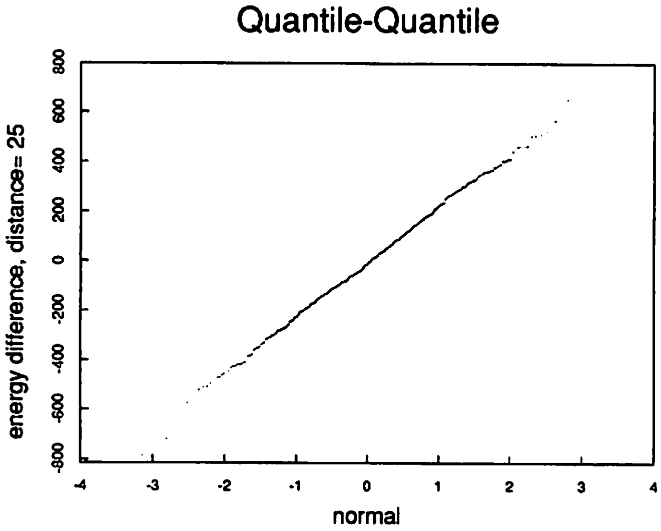
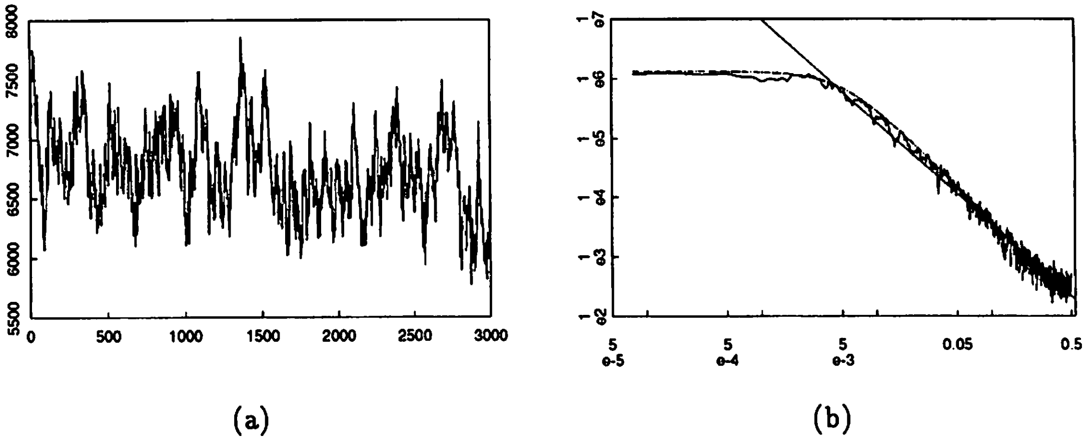


Figure 3.3: Quantile-Quantile plot of energy differences against normal distribution. These energy differences were obtained from pairs of points 25 moves apart. A straight line would indicate normally-distributed data. Together with Figure 3.2 where the mean and variance of the data were compared with a linear predictor, this figure confirms that Definition 3.0.2 is statistically satisfied by the observed data.

random walk of length t is almost exactly equal to t, for t significantly smaller than the diameter of the space (say for $t \leq \hat{t}$). In this case the fBm scaling law $\mathbf{E}[(f(t') - f(t))^2] \propto |t' - t|^{2H}$ applies only for $|t' - t| \leq \hat{t}$.

It can be shown that the spectral density of a fBm has spectral energy which is power-law in frequency [32]. If the scaling of the fBm breaks down for large times ($|t' - t| > \hat{t}$ in the case just described) then the spectral energy is power-law in frequency only down to corresponding frequencies (frequencies above $1/\hat{t}$).

Figure 3.4 shows $f(X(t))$, the energy timeseries for a random walk of 15,000 steps on the "real" netlist, along with its spectral density. Using the statistics package S-Plus (a Statistical Sciences, Inc. enhancement of Bell Telephone Laboratories' "S", [1]) the spectrum was computed by applying a cosine-bell taper to the first and last 10% of the data, computing the periodogram, and smoothing with a moving-average filter of length 20 [3]. Power-law behavior is observed to hold down to frequencies of about .01 – which, as



(a)

(b)

Figure 3.4: (a) Energy versus time during a random walk on the state space. (b) Spectrum, with a fitted power law (the straight line), and the spectrum of a fitted first-order autoregressive process.

the reciprocal of the diameter of the space, is what we would expect.

The superposed dashed curve is the spectrum of a first-order autoregressive (AR(1)) process fitted to the energy timeseries. A stationary AR(1) process is one in which $\mathbf{E}[f(t+1)|f(t), f(t-1), \ldots] = \alpha f(t) + (1-\alpha)c_0$ where α is a constant with $|\alpha| < 1$, and $c_0 = \mathbf{E}[f(0)]$ is the mean of the process. The excellent fit between the fitted and observed spectra suggests that the AR(1) model is a good one for annealing.

An AR(1) process is locally fractal (*i.e.* approximately satisfies Definition 3.0.1 for small time periods) with parameter $H = 1/2$. A fBm with $H = 1/2$ is called simple Brownian motion (Bm), and arises when $f(t+1)$ is $f(t)$ plus a random increment. True Brownian motion is nonstationary, but the factor α in the AR(1) process adds a "pull" towards c_0 which gives stationarity. So the combinatorial landscapes of interest may actually be relatively simple, the energy of a state being given by that of a neighbor plus a random quantity. While the case $H = 1/2$ may be unworthy of the name "fractal", it is just as easy to treat the general case, with an arbitrary fractal parameter; we address this in the following chapters.

The landscape analysis techniques described here are widely applicable and may be useful in other contexts, including the study of optimization techniques other than annealing. Other landscape analysis tools and details of the experiments summarized above are presented in [14, 28, 29, 30].

Chapter 4

Mathematics of Annealing

4.1 Annealing as a Markov Chain

Because the next state in simulated annealing depends probabilistically only on the current state and not on previous states (see Figure 1.1), simulated annealing is a Markov chain. If the temperature is kept fixed, the transition rule is time invariant, and the chain is said to be *homogeneous.* We now summarize and apply a few standard facts about homogeneous Markov chains (see for example [23]).

The state space of the chain corresponding to annealing is V, the same as the set of states of the optimization problem. Let the Markov chain itself be the sequence $(v_t)_{t=0}^{\infty}$, with transition matrix $M = (m_{uv})$ where $m_{uv} = \mathrm{P}[v_{t+1} = v \mid v_t = u]$. Denote by P_t the probability distribution (state probability vector) of v_t, so $P_t(v) = \mathrm{P}[v_t = v]$. By standard Markov chain theory, $P_t = P_0 M^t$.

A *stationary vector* or *equilibrium distribution* is a distribution π such that $\pi M = \pi$; this is equivalent to saying that if $P_t = \pi$ then $P_{t+1} = \pi$. A Markov chain is called *ergodic* if for all state pairs (u, v), $\lim_{t\to\infty} \mathrm{P}[v_t = v \mid v_0 = u] > 0$. For any ergodic Markov chain, there exists a unique stationary distribution π, and it has the property that for any initial distribution P_0, $P_t \xrightarrow{t\to\infty} \pi$. The Markov chain is said to be stationary at time t if $P_t = \pi$, because then for all $t' > t$, $P_{t'} = \pi$ as well. Consistent with the notation $P_t(v)$, by $\pi(v)$ we mean the stationary probability of the state v.

In any practical application of annealing to combinatorial optimization, the state space is finite and is connected by the move set. As long as this is so and not all states have equal energies, the Markov chain is ergodic. (In [22] this is proved for a class of hill-climbing algorithms which includes annealing.)

One can also look at the *reversed chain, i.e.* the sequence $v_t, v_{t-1}, v_{t-2}, \ldots$. If the chain is stationary it turns out that the reversed chain is also a Markov chain. A Markov chain is said to be *time-reversible* if the transition probabilities of the reversed chain are identical to those of the original chain. Again, [22] proves that annealing (at fixed temperature) is reversible.

4.2 Graph Model for Markov Chains

In recent work of Sinclair and Jerrum [26], a homogeneous time-reversible Markov chain is identified with a weighted undirected graph containing self-loops, and the rate of convergence of P_t to π is bounded via the "conductance" of this "underlying graph".[1] We now set forth this technique and apply it to simulated annealing.

For a general time-reversible Markov chain on state space V, the *underlying graph* G also has vertex set V and has edge weights

$$w(u,v) = \text{const} \cdot \pi(u) m_{uv} = \text{const} \cdot \pi(v) m_{vu} \tag{4.1}$$

where "const" is an arbitrary positive constant. If the constant is 1 then $\sum_{u,v \in V} w(u,v) = 1$. Allowing other constants means we do not have to have this normalization, which is sometimes convenient (for example in Lemma 4.2.2). Reversibility for a Markov chain is equivalent to the condition that, in stationarity, $\mathrm{P}[v_{t+1} = v \,\text{and}\, v_t = u] = \mathrm{P}[v_{t+1} = u \,\text{and}\, v_t = v]$. This condition is known as *detailed balance* because for each edge $\{u,v\}$ the "probability flux" from u to v is equal to that from v to u. Since in stationarity $\pi(u)m_{uv} = \mathrm{P}[v_{t+1} = v \,\text{and}\, v_t = u]$, the detailed balance condition for time-reversible Markov chains means that the definition of $w(u,v)$ in equation (4.1) is consistent. Zero-weight edges may be thought of as being eliminated from the graph, though this is of no mathematical consequence.

A *random walk* on a graph G with nonnegative edge weights is a sequence $(v_t)_{t=0}^{\infty}$, where

$$\mathrm{P}[v_{t+1} = v \mid v_t = u] = \frac{w(u,v)}{\sum_{u' \in V} w(u,u')}. \tag{4.2}$$

[1] The Sinclair and Jerrum paper has a much broader scope, using this convergence rate to show that solutions to a class of combinatorial counting problems can be approximated in polynomial time.

That is, from the vertex v_t, v_{t+1} is determined by randomly choosing an edge incident to v_t in proportion to its weight. (In equation (4.2), nonexistent edges are simply considered to have weight 0.)

It is straightforward to verify that the transition probabilities of a time-reversible Markov chain and those of the random walk on the corresponding underlying graph are equal, and so these two random processes are equivalent. The stationary properties of random walks are given by the following lemma:

Lemma 4.2.1 *$\pi(u)m_{uv} \propto w(u,v)$ and $\pi(u) \propto \sum_{v \in V} w(u,v)$.*

That is, the stationary probability of traversing an edge in a given direction is proportional to the weight of the edge, and the stationary probability of a vertex is proportional to the sum of the weights of the incident edges. The lemma follows from equation (4.1).

For simulated annealing, the edge weights for the underlying graph are given by the following lemma.

Lemma 4.2.2 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let an annealing problem be given by the undirected unweighted graph G_A and the energy function f on its vertices. Then the underlying edge-weighted graph G corresponding to annealing on G_A at temperature T has the same structure (vertices and edges) as G_A, with the addition of self-loops at each vertex. It has edge weights given by*

$$w(v,u) = e^{-\max(f(v),f(u))/T} \tag{4.3}$$

for edges $\{v,u\}$ present in G_A, and

$$w(v,v) = e^{-f(v)/T} \sum_{f(u)>f(v)} [1 - e^{-(f(u)-f(v))/T}] \tag{4.4}$$

for the added self-loops $\{v,v\}$, with the sum taken over pairs $\{u,v\}$ which are edges of G.

This lemma and Corollary 4.2.4 are proved together in the appendix. The proofs consist of straightforward verification that the transition probabilities for the random walk on G are the same as the transition probabilities for annealing on G_A at temperature T.

Definition 4.2.3 *Given an annealing graph, the* **partition function** *is $Z(T) = \sum_v deg(v)e^{-f(v)/T}$.*

Corollary 4.2.4 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *At temperature T, the stationary probability $\pi_T(v)$ of state v is* $\deg(v)e^{-f(v)/T}/Z(T)$.

We define $\pi_0 = \lim_{T \to 0} \pi_T$.

Definition 4.2.5 *A graph G is* **regular** *with degree d if all its vertices have degree d.*

Corollary 4.2.6 *For a regular annealing graph G_A, $\pi_T(v) \propto e^{-f(v)/T}$.*

By analogy with statistical physics this is often referred to as "Boltzmann's law".[2]

Note that in taking a random walk on a graph, edges are randomly selected in proportion to their weights, but there are no "rejected moves". Rather, the "move edges" referred to in equation (4.3) represent the generation *and acceptance* of a simulated annealing move, while the "reject loops" of (4.4) represent the rejection of an annealing move.

4.3 Observations on the Stationary Distributions for Annealing

For intuitive purposes we present a few observations on the nature of the stationary distributions at various temperatures, and the associated partition functions. Only Corollary 4.3.3 and Theorem 4.3.6 are referred to in the sequel.

Proposition 4.3.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ $Z(T) = \sum_v \deg(v) e^{-f(v)/T}$ *is monotonically increasing. For f ranging from 0 to 1 and $T \geq 0$, $Z(T) \geq 1$, and for regular graphs of degree d, $Z(T) \geq d$.*

Theorem 4.3.2 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *$\pi_T(v)$ is a bitonic function of T: there is a value $T_{\text{crit}}(v)$ (the* **critical temperature for** *v) such that $\pi_T(v)$ increases with increasing T for $T < T_{\text{crit}}(v)$ and decreases with increasing T for $T > T_{\text{crit}}(v)$. Further, $T_{\text{crit}}(v)$ is the value of temperature at which $f(v) = \mathbf{E}_{\pi_T}[f]$, i.e. at which the expected energy equals the energy of v.*

Corollary 4.3.3 *For states v where $f(v)$ is minimal, $\pi_T(v)$ decreases monotonically as T increases; and for states where $f(v)$ is maximal, $\pi_T(v)$ increases monotonically.*

Proof $\mathbf{E}_{\pi_{T_{\text{crit}}}}[f] = f(v)$ gives critical temperatures of 0 and ∞ respectively for these two types of vertices. ■

[2] The "partition function" $Z(T)$ is also found in the physics literature. There the $\deg(v)$ term does not appear, presumably because the physical systems studied always have associated graphs which are regular.

Proposition 4.3.4 $\boxed{\text{pf} \to \text{appendix}}$ *For f ranging from 0 to 1, and $T > 0$, $\pi_T(v) > \pi_0(v)$ if and only if $f(v) > 0$*

While this follows from Corollary 4.3.3, a direct proof is provided in the appendix.

Definition 4.3.5 *For a given annealing problem, the* **critical temperature** *T_{crit} is the largest value such that for $T < T_{\text{crit}}$, for each state v, $\pi_T(v)$ is monotonic in T.*

Theorem 4.3.6 $\boxed{\text{pf} \to \text{appendix}}$ *For any finite annealing problem, $T_{\text{crit}} > 0$.*

4.4 Graph Conductance and Mixing of Markov Chains

For a general Markov chain, the rate at which the chain converges to the stationary distribution is related to the dominant eigenvalue of the transition matrix M, and a major contribution of [26] is to relate the dominant eigenvalue to a structural property, the "conductance", of the underlying graph.

Definition 4.4.1 *The* **conductance** *of an ordered partitioning $(S, \bar{S})$ ($0 < |S| < |V|$) of an edge-weighted undirected graph G, is*

$$\Phi_S(G) = \frac{\sum_{u \in S, v \in \bar{S}} w(u,v)}{\sum_{u \in S, v \in V} w(u,v)}. \tag{4.5}$$

This value is equal to the stationary probability of making a transition out of S, conditional upon starting in S.

Definition 4.4.2 *The* **(global) conductance** *of an edge-weighted undirected graph G is*

$$\Phi(G) = \min_S \Phi_S(G) \tag{4.6}$$

where the minimum is taken over subsets $S \subset V$ with $0 < |S| < |V|$ and $\sum_{v \in S} \pi(v) \leq 1/2$.

Roughly speaking, the conductance is a measure of the probability of escaping in one step from any "small" subset.

Proposition 4.4.3 *If G is a connected unweighted graph (equivalently if all edge weights are 1), $\Phi(G) \geq 1/n^2$.*

This follows directly from the definition.

To study the convergence of a probability distribution to equilibrium, we also need a measure of the distance of one distribution from another.

Definition 4.4.4 *For a probability distribution P on V with nonzero mass on every point, the* **relative pointwise distance (rpd)** *of another distribution P' from P is*

$$\|P' - P\|_{\mathrm{rpd}} = \max_{v \in V} \frac{|P'(v) - P(v)|}{P(v)}. \tag{4.7}$$

Notice that the rpd is not symmetric. It will sometimes be more convenient to use a different measure:

Definition 4.4.5 *For probability distributions P and P' on the set V, the* **total variation distance (tvd)** *between P' and P is*

$$\|P' - P\|_{\mathrm{tvd}} = \frac{1}{2} \sum_{v \in V} |P'(v) - P(v)|. \tag{4.8}$$

We note that tvd is symmetric in its arguments, and is bounded by 0 and 1. In addition, the following two propositions hold. First,

Proposition 4.4.6

$$\|P' - P\|_{\mathrm{tvd}} = \sum_{v:\, P'(v) > P(v)} [P'(v) - P(v)] \tag{4.9}$$

We will sometimes use this in lieu of the definition. Also,

Proposition 4.4.7 *For any P' and P,*

$$\|P' - P\|_{\mathrm{tvd}} \leq \frac{1}{2} \|P' - P\|_{\mathrm{rpd}}. \tag{4.10}$$

Proof Any weighted average of a set of values is less than the maximum, so

$$\begin{aligned} \sum |P'(v) - P(v)| &= \frac{1}{2} \sum P(v) \frac{|P'(v) - P(v)|}{P(v)} &(4.11)\\ &\leq \frac{1}{2} \max \frac{|P'(v) - P(v)|}{P(v)}. &(4.12)\end{aligned}$$

■

Definition 4.4.8 *A* **strongly aperiodic** *Markov chain is one in which the probability of each self-transition is at least 1/2.*

Theorem 4.4.9 (conductance and mixing) *Let P_t represent the probability distribution at time t of a strongly aperiodic Markov chain with underlying graph G and stationary distribution π. Then*

$$\|P_t - \pi\|_{\mathrm{rpd}} \leq \frac{1}{\min_u \pi(u)} \cdot (1 - \Phi(G)^2/2)^t \tag{4.13}$$

The theorem is due to Sinclair and Jerrum [26]. Related inequalities can be found in [5].

Annealing does not itself give rise to strongly aperiodic Markov chains, but a variant does: At each step, toss a fair coin. If the outcome is heads, make a standard annealing move (generate a neighboring state and accept or reject it). If the outcome is tails, simply remain in the current state.

This modified process has an underlying graph whose conductance is exactly half that for normal annealing, corresponding to the need to spend twice as much time if half the moves are nullified.[3] Because the idea of discarding half the moves is plainly ridiculous in practice, we will work with the unmodified annealing process. While this comes at the expense of mathematical rigor, multiplying by 4 all time bounds derived in this paper gives rigorously justified results for the modified annealing process.

Rigor could also be achieved in two other ways. One comes from noting that the "coin-tossing" version of annealing just described can be simulated more efficiently: If a large number t of coin-tossing moves are to be attempted at temperature T, then the number of actual annealing moves attempted is a binomially distributed random number, $t' \sim B(t, 1/2)$. We could simply directly generate such a number (in lieu of the coin tossing) and then attempt that many moves.

The second way is to go to a continuous-time annealing. Where t moves would be made at temperature T in standard simulated annealing, we now make a Poisson-distributed number of moves $t' \sim \text{Pois}(t)$. This may be thought of as waiting an infinitesimally small time dt, tossing a coin which comes up heads with probability dt, and making a move if the coin comes up heads. This is just an extreme case of the coin-tossing version of annealing.

By Theorem 4.4.9, $||P_t - \pi||_{\text{rpd}}$ approaches 0 exponentially in t. The "time constant" – the time required for decay by a factor of e – is no more than $2/\Phi^2$. [26] also states that this time constant is at least $1/2\Phi$.

[3] While Theorem 4.4.9 requires spending four times as much time if Φ is halved, this can be attributed to a weakness of the bound; a corresponding lower bound in [26] is linear rather than quadratic in Φ. Also, under the intuitive view that discarding half the moves requires trying twice as many, if we measure time in the number of moves actually attempted, the expected time required for the modified annealing is exactly equal to the time required for normal annealing.

4.5 Rapid Mixing

Of particular interest to us is under what conditions simulated annealing approaches its stationary distribution quickly. We will begin with a simpler question, namely under what conditions annealing at *infinite* temperature quickly approaches stationarity. In our standard paradigm, where we are given a graph whose vertices correspond to some combinatorial configurations and whose edges denote some method of moving from one configuration to another, we are simply asking about the mixing rate for a random walk on this graph.

In the combinatorial problems of interest, the description of a problem instance ("the input to a program for solving the optimization problem") has size which is polynomial in a natural parameter N, while the size of the annealing graph ("the number of potential solutions") is exponential in N. For example, an N-object placement problem can be described by the set of pairs of objects which need to be connected (no more than N^2 of them), and so can be described in space proportional to $N^2 \ln N$. However, this placement problem has $N!$ possible solutions.

In general, suppose that there is a family $\{G(N)\}_{N=1}^{\infty}$ of (possibly edge-weighted) graphs such that $G(N)$ has a number of vertices which is exponential in N and such that the convergence to equilibrium of a random walk on $G(N)$ has characteristic time no more than poly(N). The random walks are then said to be "rapidly mixing". A sufficient condition for this to be true is that $G(N)$ has "large" conductance, *i.e.* $\Phi(G) \geq 1/\text{poly}(N)$, for then the characteristic convergence time is no more than poly$(N)^2$.

In [4, 5, 10, 11], random walks on a number of combinatorially motivated graphs are shown to be rapidly mixing, using a variety of techniques.

The studies of Chapter 3 focused on circuit placement problems, where a vertex of the state graph is a permutation of $\{1, \ldots, N\}$ and an edge connects two vertices if the two permutations differ only by the interchange of two numbers. Theorem 4.5.1 shows that such a graph has large conductance. Since this is the underlying graph for annealing at temperature $T = \infty$ (where all edges have equal weight), annealing at $T = \infty$ is rapidly mixing for placement problems. The consequence for finite temperatures is given by Theorem 4.5.2.

Theorem 4.5.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let G be the graph whose vertex set V consists of permutations of the objects $1, \ldots, N$, and in which two vertices (permutations) are connected by an edge if some single pairwise exchange of two objects takes one permutation to the* *other. Then G has conductance* $\geq 1/2N^2$.

The proof, contained in Appendix C, is of independent interest. It relies on the construction of "canonical paths" between vertices in the two partitions, with the property that a given edge is used in a few paths at most. Since there are many paths, there are also many cut edges, and the conductance is large. This style of proof is most elegant and is applicable to a variety of other combinatorial problems: see for example [10, 11], and a random-path variant in [4].

Henceforth we will assume that we are given an annealing problem whose energies range from 0 to 1, as mentioned in the introduction. Let $G(T)$ denote the corresponding underlying graph, whose edge weights depend on T, but whose structure is invariant over temperature.

Theorem 4.5.2 *Let a graph $G(T)$ as above be given. Abbreviating $\Phi(G(T))$ as $\Phi(T)$,*

$$\Phi(T) \geq e^{-1/T}\Phi(\infty). \tag{4.14}$$

Proof The theorem follows directly from the edge weights for annealing (equations (4.3) and (4.4)), and the definition of conductance (definitions 4.4.1 and 4.4.2). ■

4.6 Asymptotic Convergence to Global Minima

To date the only formally justifiable theories of simulated annealing have pertained to convergence to a global minimum with probability approaching 1. We do not feel that this is the best approach to understanding why annealing works in practice, but the result can be duplicated using the tools presented here. The full derivation is contained in Appendix A. The basic result is that for the "logarithmic" cooling schedule $T_t = 1/a\ln t$ with $a < 1$, for any P_0, $||P_t - \pi_{T_t}|| \rightarrow 0$. That is, the actual probability distribution at time t approaches the equilibrium distribution at the corresponding temperature T.

The result is similar to those of [6, 7, 19], which represent the strongest asymptotic theory of annealing. Our proof has a very simple intuitive basis, though unfortunately the calculations become tedious. The underlying notion is this: we wish to show that $||P_{t+1} - \pi_{T_{t+1}}|| < ||P_t - \pi_{T_t}||$. In principle we use the two inequalities $||P_{t+1} - \pi_{T_{t+1}}|| \leq (1 - \Phi(T_{t+1}))||P_t - \pi_{T_{t+1}}||$ and $||P_t - \pi_{T_{t+1}}|| \leq ||P_t - \pi_{T_t}|| + ||\pi_{T_t} - \pi_{T_{t+1}}||$. Then we need only ensure that for the logarithmic cooling schedule the reduction by the factor $1 - \Phi$ in the first inequality sufficiently outweighs the additive $||\pi_{T_t} - \pi_{T_{t+1}}||$ in the second to force $||P_t - \pi_{T_t}||$ to 0 eventually. The difficulty is that the first inequality requires the use of an L_2-like distance function [18] while the second only applies for total variation distance, and so must be replaced with a more complex inequality. Appendix A presents the details, constructing a valid triangle inequality and bounds for the quantities involved to prove that for the logarithmic cooling schedule there is indeed convergence of $||P_t - \pi_{T_t}||$ to 0.

4.7 Annealing at Constant Temperature

How long does it suffice to run annealing to produce a solution whose expected energy is no more than a given value? The remainder of this section is devoted to the proof and discussion of Theorem 4.7.6.

Without loss of generality, throughout most of this paper we assume that G has n vertices and that f ranges from 0 to 1. (The notable exceptions to this rule are Chapters 6 and 7.) For all cases of interest to us, G will also be regular, but all derivations will be for the general case.

Lemma 4.7.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *For a function f on V ranging from $f_{\min}$ to $f_{\max}$ with $f_{\text{range}} = f_{\max} - f_{\min}$; an arbitrary distribution P_t on V; and π_T and π_0 the stationary distributions at temperatures T and 0:*

$$\mathbf{E}_{P_t}[f(v)] \leq \mathbf{E}_{\pi_T}[f(v)] + ||P_t - \pi_T||_{\text{tvd}} f_{\text{range}}. \tag{4.15}$$

As suggested by the notation, what we have in mind is that P_t be the distribution after annealing for t steps starting from an arbitrary initial distribution P_0.

Reasoning from inequality (4.15), to force $\mathbf{E}_{P_t}$ small we will need to make both $\mathbf{E}_{\pi_T}[f(v)]$ small and $||P_t - \pi_T||_{\text{tvd}}$ small – say less than a given amount ε. We begin with the first term.

Definition 4.7.2 *For an annealing graph G with n vertices let $\hat{T}(\varepsilon, \Delta, n) = \Delta / \ln(n^2/\varepsilon)$. If G is regular, use n in lieu of n^2.*

Lemma 4.7.3 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Given Δ, let $T \leq \hat{T}(\varepsilon, \Delta, n)$. If $\Delta \leq \varepsilon$ or $\Delta \leq \min\{f(v) : f(v) > \varepsilon\}$ then*

$$\mathbf{E}_{\pi_T}[f] \leq 2\varepsilon, \tag{4.16}$$

and if $\Delta \leq \Delta f$ then

$$\mathbf{E}_{\pi_T}[f] \leq ||\pi_T - \pi_0||_{\text{tvd}} \leq \varepsilon. \tag{4.17}$$

This bound is tight in the sense that there is a class of annealing problems for which $\mathbf{E}_{\pi_T}[f] \sim \varepsilon$.

We also wish to make the second term of inequality (4.15), $||P_t - \pi_T||_{\text{tvd}}$, less than ε.

Definition 4.7.4 *For an annealing graph G with n vertices and for a given temperature T, let*

$$\hat{t}(T, \varepsilon, n) = 2\left(\ln\left(\frac{n^2}{\varepsilon}\right) + \frac{1}{T}\right)\frac{1}{\Phi(\infty)^2}e^{2/T} \tag{4.18}$$

If G is regular, use n in lieu of n^2. Following Proposition 4.4.3 we may also use $1/n^2$ in lieu of $\Phi(\infty)$.

Lemma 4.7.5 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *If $t \geq \hat{t}(T, \varepsilon, n)$ then beginning from any distribution P_0 and annealing at temperature T for time t, the final distribution satisfies*

$$||P_t - \pi_T||_{\text{tvd}} \leq \varepsilon. \tag{4.19}$$

The inequalities in the following theorem are immediate consequences of Lemmas 4.7.3, 4.7.5, and 4.7.1 respectively.

Theorem 4.7.6 *Let a graph G with n vertices, an energy function f from G to $\mathbb{R}$ having minimum 0 and maximum 1, and a small value ε be given. Let Δ be any of Δf, ε, or* $\min\{f(v) : f(v) > \varepsilon\}$. *With the functions $\hat{T}$ and $\hat{t}$ as defined above, begin from an arbitrary initial distribution P_0 and anneal at temperature $T \leq \hat{T}(\varepsilon, \Delta, n)$ for $t \geq \hat{t}(T, \varepsilon, n)$ steps. If the final distribution is denoted P_t, we have:*

$$\begin{aligned} \mathbf{E}_{\pi_T}[f(v)] &\leq 2\varepsilon && (4.20) \\ ||P_t - \pi_T||_{\text{tvd}} &\leq \varepsilon && (4.21) \\ \mathbf{E}_{P_t}[f(v)] &\leq 3\varepsilon. && (4.22) \end{aligned}$$

If $\Delta = \Delta f$ the bounds in (4.20) and (4.22) can be improved to ε and 2ε respectively.

Remark 4.7.7 *If we use $T = \hat{T}$ and $t = \hat{t}$, the value of $\hat{t}$ in Theorem 4.7.6 comes to*

$$\hat{t} = 2(1 + \frac{1}{\Delta})\ln(n^2/\varepsilon)\frac{1}{\Phi(\infty)^2}(n^2/\varepsilon)^{2/\Delta}. \tag{4.23}$$

If G is regular then each n^2 may be replaced by n.

Due to its central role in our analysis of annealing, some commentary on Theorem 4.7.6 and Remark 4.7.7 is in order.

Taking $\Delta = \Delta f$ in Remark 4.7.7 indicates that having a smaller Δf could make annealing more time-consuming. While it might seem that lowering the energy of any state should only help annealing, so that in particular a small Δf should be no worse than a large one, that intuition is flawed. First, we would think that raising the energy of a state would necessarily raise the equilibrium energy of the system, but this is not so:

Remark 4.7.8 *For a system in equilibrium at temperature T, raising the energy of a given state may lower the expected energy of the system.*

Example A small counterexample is given in the appendix. ■

But the real reason the $2/\Delta f$ appears is this. At low temperatures a state of energy Δf contributes an amount proportional to $\Delta f \cdot e^{-\Delta f/T}$ to the expected energy. The first term, the linear Δf, would indeed make this contribution smaller if Δf is smaller. But for small temperatures that effect is more than offset by the second, exponential, term $e^{-\Delta f/T}$, which is larger if Δf is smaller. Asymptotically, then, having a large minimum nonzero energy (a large Δf) is indeed advantageous for achieving a lower expected energy at a given temperature.

4.8 Energy vs. Time

Because it is the measure of the efficiency of annealing, it is important to think about the relationship between the time $\hat{t}(\hat{T}(\varepsilon, \Delta, n))$ and the corresponding upper bound $\hat{\mathbf{E}} = 3\varepsilon$ (or 2ε) on expected energy.

Roughly speaking, the final $(n/\varepsilon)^{2/\Delta}$ in $\hat{t}$ comes from the "time constant" for the process: the time t required to make $(1 - \Phi(T)^2/2)^t \leq 1/e$. The leading terms $2\left(1 + \frac{1}{\Delta}\right)\ln(n/\varepsilon)$ just represent the additional time factor required to raise this $1/e$ to a sufficiently large power to overwhelm the value $1/\pi_{\min}$, and to bring $\mathbf{E}_{\pi_T}[f]$ to ε rather than just $1/e$. The fact that the leading terms are only logarithmic in $(n/\varepsilon)^{2/\Delta}$ means that the primary difficulty is to bring the process anywhere near equilibrium; after that, bringing it extremely close to equilibrium takes only a few (logarithmically many) times as long.

The remaining factor, $1/\Phi(\infty)^2$, can be thought of as a scaling factor for time. Since it may take this long to approach equilibrium in a random walk on G even at infinite temperature, it is reasonable to compare all other times to this value. This $1/\Phi(\infty)^2$ is generally not a problem. We have already pointed out that for combinatorial problems of "size" N (with a number of states n which is exponential in N), the conductance $\Phi(\infty)$ is generally inverse polynomial in N, and $1/\Phi(\infty)$ is not too large even in practice. By Proposition 4.4.3, $1/\Phi(\infty)$ is at worst $1/n^2$, which like the leading terms is only logarithmic in $(n/\varepsilon)^{2/\Delta}$.

What does this say about the running time t sufficient to generate a solution of given "quality" – a solution whose expected energy is less than 3ε?

The time is dominated by the final $(n/\varepsilon)^{2/\Delta}$. Here n is large, we are interested in ε small, and we would expect Δ to be small, making the time $\hat{t}$ of Theorem 4.7.6 extremely large. In particular, both ε and Δ are less than 1, giving $\hat{t} \geq (n/\varepsilon)^{2/\Delta} \gg n$.

This should not be any surprise. We have already argued that, in general, finding a state of low energy requires searching all the states. Since the analysis above is completely general, the run times it requires must be at least n.

The asymptotic dependence of run time on "quality" q depends on precisely what question one is asking.

A first formulation would be the nature of the function $\hat{t}(q)$ sufficient to guarantee $\mathbf{E}[f] \leq 1/q$, assuming Δf is known. Asymptotically for $q \to \infty$ with $q = 1/2\varepsilon$, letting $\Delta = \Delta f$ in Theorem 4.7.6 and Remark 4.7.7 gives

$$\frac{\ln \hat{t}}{\ln q} \leq \frac{\frac{2}{\Delta f}\ln(n/\varepsilon) + \cdots}{\ln(1/\varepsilon) + \ln(1/3)} \sim \frac{2}{\Delta f}. \tag{4.24}$$

Equivalently,

$$\hat{t} = q^{2/\Delta f + w(q)} \tag{4.25}$$

where w is some function approaching 0 as $q \to \infty$. That is, the sufficient running time $\hat{t}$ is a "power-law" function of the "quality".

While we will refer to this as a "polynomial" time-quality tradeoff, we do so with the caveat that the algorithm is not itself "polynomial-time" in the standard sense. An algorithm is polynomial-time if the run time is bounded by a polynomial function of the input size. The present case fails to conform for several reasons:

- The degree of the polynomial depends on Δf, and therefore varies with the problem instance.

- Typically the number of states n is exponential in the "problem size". Since the run time includes a factor of $n^{2/\Delta f}$, it is exponential in problem size.

- For algorithms with a quality-time tradeoff, the input to the algorithm includes the desired quality q. The size of this input is naturally $\log_2 q$, and for the algorithm to be "polynomial-time" it should be polynomial in $\log_2 q$ rather than in q itself.

A second view of the run time versus quality tradeoff would be the amount of time Theorem 4.7.6 and Remark 4.7.7 dictate if Δf is unknown, as is likely in practice. In that case we must use $\Delta = \varepsilon$ and $q = 1/3\varepsilon$, for

$$\frac{\ln \hat{t}}{\ln q} \sim \frac{2}{\varepsilon} = 6q \tag{4.26}$$

and in analogy to (4.25),

$$\hat{t}(q) = q^{6q+w(q)}. \tag{4.27}$$

So if Δf is unknown the time used to guarantee a solution of given quality increases exponentially.

There is a third, intermediate, view: we could spend time $t(q)$ for which $\mathbf{E}[f] \leq 1/q$ is guaranteed *only for q sufficiently large.* Again taking $q = 1/3\varepsilon$ we could use

$$t = q^{g(q)} \tag{4.28}$$

for any function $g(q) \to \infty$, since then for q sufficiently large $g(q) \geq 2/\Delta f + w(q)$ and the desired expected energy is assured by (4.25). For example, $t = g(q) = q^{\ln q}$ will do. Thus t can be made "quasi-polynomial" in q, where $f(n)$ is quasi-polynomial if $f(n) = O(2^{\log^k n})$.

It is worth noting that if we have a bound of the form $\mathbf{E}[f] \leq 1/q = c\varepsilon$ after time $\hat{t}(q)$, it is only in the strict, second view that the constant c plays a role in the asymptotics of $\hat{t}(q)$, *i.e.* in the asymptotic dependence of $\hat{t}$ on the bound for $\mathbf{E}[f]$.

Theorem 4.7.6 and Remark 4.7.7 express the tradeoff between the expected energy produced by annealing at fixed temperature and the run time required. We have made two basic observations: time is roughly polynomial in inverse expected energy, with dominant power $2/\Delta$; and in practical cases where n is large and Δ small, both the multiplicative constant and the exponent of the polynomial make the time prohibitively large.

Could we do better with a time-varying cooling schedule, *i.e.* by true annealing rather than annealing at fixed temperature? We do not know. For the logarithmic cooling schedule common to theoretical studies of annealing the expected energy approaches 0 over time, but there is no reason to believe that a given expected energy is achieved more rapidly this way than by annealing at a fixed, appropriately-chosen temperature. In fact a comparison of Theorem 4.7.6 with the logarithmic cooling schedule results derived here in Appendix A indicates that the two methods have the same asymptotic behavior; the constants are difficult to compare.

In Chapters 7 and 8 we shall see how for some classes of problems the relationship of Theorem 4.7.6 can be "bootstrapped" to give an efficient annealing algorithm, by applying Theorem 4.7.6 to pieces of the problem (where n is smaller and Δf is larger) rather than to the whole problem at once. This will be done by exploiting the linear separability of the energy functions of Chapter 7, and the self-similarity, or "fractal", properties of the energy landscapes of Chapter 8.

In future, our analysis of the dependence of run time on solution quality will for the most part be limited to computation of the ratio $\ln t / \ln q$; that ratio can always be interpreted in any of the three fashions discussed above.

Chapter 5

Annealing at Non-Constant Temperature

In the previous chapter we analyzed the behavior of simulated annealing being run at a fixed temperature, and in the following chapters we will be exploiting these results. Essentially, we will attack the problem in various "generations", during each of which the temperature is kept fixed.

While we will not try to take *advantage* of the fact that temperature is decreasing, it will be necessary to show that decreasing the temperature cannot make matters *worse* – for example, that the expected energy cannot increase. This is not so simple a matter as it appears, as shown by the following remark.

Remark 5.0.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Begin in equilibrium at T_0. Apply the schedule $T_1, T_2, \ldots, T_n$, where each $T_i \leq T_0$. Then the probability distribution following this schedule may give* less *weight to the global minimum than it had in the initial equilibrium distribution: in fact, may give it arbitrarily small weight.*

The "probability pump" example in the appendix is particularly amusing.

Despite this demonstration of how bad matters can be in general, for monotonic cooling schedules the situation is much better. Theorem 5.1.16 shows that for monotonic cooling schedules starting in equilibrium, the expected energy cannot increase. That theorem, the cornerstone of section 5.1, is based on entropy arguments from statistical physics.

Theorem 5.2.3 further shows that for such schedules which begin well below the critical temperature, the system always remains near equilibrium. This stronger result (requiring stronger hypotheses) is needed in section 8.4. It is proved by elementary probabilistic arguments in section 5.2.

5.1 Monotonic Cooling: Entropy-Based Approach

Borrowing from statistical physics (see for example [27]), we define:

Definition 5.1.1 *The* **inverse temperature** β *is* $1/T$.

All expressions of interest involve $1/T$ rather than T itself, and using β is not only a notational convenience but also has the advantage of preserving continuity when negative temperatures are allowed.

Whenever β is used as a function argument we will interpret it to refer to the corresponding temperature, for instance by π_β we mean π_T with $T = 1/\beta$. We will also assume a correspondence between various β's and T's, so β_t will mean T_t and so on.

It appears impossible to prove directly that expected energy decreases with time for monotonic cooling schedules, because energy is not a very well-behaved quantity. For instance, even if the expected energy is larger than its T-equilibrium value, a move at temperature T can increase it. The quantity which is well behaved (is a Liapunov function for annealing at fixed temperature) is the *relative entropy.*

As usual, let the state probability vector at a given time be P, with probabilities $P(v)$ on the individual states.

Definition 5.1.2 *Let* π *be a probability measure on* V *such that for all* $v \in V$, $\pi(v) > 0$. *If* P *is any probability distribution on* V, *the* **entropy of** P **relative to** π *is*

$$H(P,\pi) = \sum_v P(v) \ln\left(\frac{P(v)}{\pi(v)}\right) \tag{5.1}$$

$$= \sum_v \pi(v)\varphi\left(\frac{P(v)}{\pi(v)}\right) \tag{5.2}$$

where $\varphi(x) = x \ln x$ *for* $x > 0$ *and* $\varphi(0) = 0$. *When* π *is clear from context we may write* $H(P)$ *in lieu of* $H(P,\pi)$.

This definition and Theorem 5.1.4 and its proof are taken directly from [15].

Lemma 5.1.3 $\boxed{\text{pf} \rightarrow \text{appendix}}$ $H(P) \geq 0$, *with equality iffP* $\equiv \pi$.

Theorem 5.1.4 $\boxed{\text{pf} \to \text{appendix}}$ *Let M be the transition matrix for a Markov chain on V. Suppose π is a strictly positive measure on V and is stationary for M, i.e. $\pi M = \pi$ or equivalently, for all u, $\sum_v \pi(v)M(v,u) = \pi(u)$. Then for any distribution P on V,*

$$H(P \cdot M) \le H(P). \tag{5.3}$$

Corollary 5.1.5 *If P_t is the state probability vector at time t for annealing at fixed temperature T with equilibrium distribution π_T, the relative entropies $H_t = H(P_t)$ with respect to π_T are monotonically nonincreasing.*

Proof Immediate from 5.1.4 with M equal to the transition matrix for annealing at temperature T. ■

Definition 5.1.6 *The* **Gibbs entropy** *or simply the* **entropy**[1] *of the distribution P on the finite set V is*

$$S(P) = -\sum_{v \in V} \varphi(v). \tag{5.4}$$

While we will not be using it, for the reader's intuition we note that:

Lemma 5.1.7 *For distributions P on state space V with $|V| = n$, $0 \le S(P) \le \ln n$, with $S(P) = 0$ iff$P(v) = 1$ for some v, and $S(P) = \ln n$ iff$P(v) = 1/n$ for all v.*

The proof is simple and may be found in [17].

Lemma 5.1.8 $\boxed{\text{pf} \to \text{appendix}}$ *The entropy of P relative to π_β may be written*

$$H(P, \pi_\beta) = -S(P) - L(P) + \ln Z(\beta) + \beta F(P) \tag{5.5}$$

where the entropy $S(P)$ and partition function $Z(\beta)$ are per definitions 5.1.6 and 4.2.3, and we define $F(P) = \mathbf{E}_P[f(v)]$, and $L(P) = \mathbf{E}_P[\ln \deg(v)]$.

The derivatives of Z, F, and $H + L$ with respect to β will be important, and are particularly simple:

Lemma 5.1.9 $\boxed{\text{pf} \to \text{appendix}}$ *The derivative of the partition function with respect to inverse temperature β is*

$$\frac{dZ(\beta)}{d\beta} = -Z(\beta)\mathbf{E}_{\pi_\beta}[f]. \tag{5.6}$$

[1] Statistical physics [20] uses the term "Gibbs entropy" and the symbol S, while information theory [17] uses the term "entropy" and the symbol H. Since we are using H for relative entropy, to minimize confusion we will use S for Gibbs entropy.

Lemma 5.1.10 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *The derivative of the expected energy in equilibrium at inverse temperature β is*

$$\frac{dF(\pi_\beta)}{d\beta} = -\mathrm{Var}_{\pi_\beta}[f]. \tag{5.7}$$

Corollary 5.1.11 *$F(\pi_\beta)$ is a monotonically decreasing function of β.*

Lemma 5.1.12 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *The derivative of the Gibbs entropy plus the expected log of the degree in equilibrium at inverse temperature β is*

$$\frac{d(S+L)(\pi_\beta)}{d\beta} = -\beta\mathrm{Var}_{\pi_\beta}[f]. \tag{5.8}$$

Corollary 5.1.13 *$(S+L)(\pi_\beta)$ is increasing for $\beta < 0$ and decreasing for $\beta > 0$.*

This means that both F and $S+L$ are monotonically increasing functions of T in the "physical" regime $T \geq 0$.

Lemma 5.1.14 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *If $F(P) = F(\pi_\beta)$ then $(S+L)(P) \leq (S+L)(\pi_\beta)$.*

That is, for a given value of expected energy, entropy plus expected degree is maximized by the Gibbs distribution.

Lemma 5.1.15 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let $P_0 = \pi(\beta_0)$ with $\beta_0 > 0$. For any distribution P, if*

$$(S+L)(P) - (S+L)(P_0) \geq 0 \tag{5.9}$$

then

$$\frac{1}{\beta_0}[(S+L)(P) - (S+L)(P_0)] < F(P) - F(P_0). \tag{5.10}$$

Theorem 5.1.16 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Beginning from the distribution $P_0 = \pi_{T_0}$, anneal with cooling schedule $T_1, T_2, \ldots$ where $T_0 \geq T_1 \geq T_2 \cdots$. If the intermediate distributions are $P_1, P_2, \ldots$, then at any time t, $\mathbf{E}_{P_t}[f] \leq \mathbf{E}_{P_0}[f]$.*

This is the main result of this chapter. The following corollaries simply cast it into a form which is more convenient in the later chapters.

Corollary 5.1.17 (monotonic cooling) $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Begin from a distribution P_0 : $||P_0 - \pi_{T_0}||_{\mathrm{tvd}} \leq \varepsilon$ and anneal with cooling schedule $T_1, T_2, \ldots$ where $T_0 \geq T_1 \geq T_2 \cdots$. Then at any time t, $\mathbf{E}_{P_t}[f] \leq \mathbf{E}_{\pi_{T_0}}[f] + \varepsilon$.*

Definition 5.1.18 *The cooling schedule $T_1, \ldots, T_t$ is a* **subschedule** *of cooling schedule $T'_1, T'_2, \ldots$ with offset c if for all $\tau \in \{1, \ldots, t\}$, $T_\tau = T'_{c+\tau}$. The schedule T' is a* **superschedule** *of the schedule T.*

Corollary 5.1.19 (monotonic superschedule) $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let the temperature T and the cooling schedule $\{T(\tau)\}_{\tau=1}^{t}$ be such that for any P_0, the final distribution P_t satisfies $\|P_t - \pi_{T_t}\|_{\text{tvd}} \leq \varepsilon$. If $T(\tau)$ is a subschedule of a monotonically nonincreasing schedule $\{T'(\tau)\}_{\tau=1}^{t'}$, then $\mathbf{E}_{P_{t'}}[f] \leq \mathbf{E}_{\pi_{T_{t'}}}[f] + \varepsilon$.*

If the subsequence is good enough to achieve some desired result, why bother with higher and lower temperatures at all? In Chapters 7 and 8 we will consider model problems which can be decomposed in some way into subproblems of varying energy scales. The early, high temperatures in the cooling schedule will "solve" the subproblems with large energy scales, and the later, lower temperatures will solve the smaller-scale subproblems. It is crucial that the work done at the later low temperatures not harm the results obtained at the earlier higher temperatures.

5.2 Monotonic Cooling at Low Temperatures: Probabilistic Approach

Starting from equilibrium at a low temperature T_0 and annealing at non-increasing temperatures less than or equal to T_0, it is not only true that the expected energy does not increase (as was proved in the previous section), but in fact the distribution remains near stationarity. This stronger result will be needed in Section 8.4, and relies on two hypotheses. First, that T_0 is so low that $\pi_{T_0} \approx \pi_0$: that is, the equilibrium distribution at T_0 is close to the temperature-0 equilibrium. And second, that $T_0 \leq T_{\text{crit}}$, so that for any state v, $\pi_T(v)$ is monotonic in T for the values $T_0 \geq T \geq 0$ of interest.

We already showed (Theorem 4.3.6) that $T_{\text{crit}} > 0$. The first hypothesis above is also satisfiable:

Proposition 5.2.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *For any δ there exists $0 < T_\delta \leq T_{\text{crit}}$ such that for all $T \leq T_\delta$, $\|\pi_T - \pi_0\|_{\text{tvd}} \leq \delta$.*

Theorem 5.2.2 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let the n-vector P_t be the state probability vector at time t of a time-variant Markov random process, starting at time 0 with state probability vector*

P_0. Let M_t be the transition matrix applied before time t ($t = 1, 2, \ldots$), so $P_t = P_{t-1}M_t$. Let π_t be the stationary distribution corresponding to M_t, so $\pi_t M_t = \pi_t$. Then for any distribution π_0,

$$\|P_t - \pi_t\|_{\mathrm{tvd}} \leq \|P_0 - \pi_0\|_{\mathrm{tvd}} + \|\pi_0 - \pi_1\|_{\mathrm{tvd}} + \cdots + \|\pi_{t-1} - \pi_t\|_{\mathrm{tvd}}. \tag{5.11}$$

In practice π_0 will be taken to be the equilibrium distribution corresponding to another transition matrix M_0, and P_0 will be near π_0.

Corollary 5.2.3 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let an annealing problem with critical temperature T_{crit} and a sequence of temperatures $\{T_t\}$ satisfying $T_{\text{crit}} \geq T_0 \geq T_1 \geq T_2 \geq \cdots$ be given. From an initial state probability vector P_0 anneal at temperatures $T_1, T_2, \ldots$. Then the distribution P_t at any time t satisfies*

$$\|P_t - \pi_{T_t}\|_{\mathrm{tvd}} \leq \|P_0 - \pi_{T_0}\|_{\mathrm{tvd}} + \|\pi_{T_0} - \pi_{T_t}\|_{\mathrm{tvd}} \tag{5.12}$$

and

$$\|P_t - \pi_0\|_{\mathrm{tvd}} \leq \|P_0 - \pi_{T_0}\|_{\mathrm{tvd}} + \|\pi_{T_0} - \pi_0\|_{\mathrm{tvd}}. \tag{5.13}$$

Chapter 6

Annealing on Functions of Unknown Energy Range

Until now it has been assumed that the functions of interest have energies ranging from 0 to 1. In generalizing this to energies ranging from $f_{\min}$ to $f_{\max}$ there are two distinct issues. First, the matter of rescaling – which heretofore we have taken for granted – will be addressed explicitly. Second, we will consider the case where the energy range is known only approximately, and then extend this to the case where only very crude bounds on the range are known.

6.1 Arbitrary but Known Energy Range

Definition 6.1.1 *Let G and G' be graphs with functions f and f' respectively mapping their vertices to the reals. We say (G, f) and (G', f') are* similar *energy graphs if there is an isomorphism $\sigma : G \rightarrow G'$, and an affine transform T on the reals ($T(y) = ay + b$ with $a \neq 0$), such that for all vertices $v \in G$, $T(f(v)) = f'(\sigma(v))$.*

Lemma 6.1.2 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *If (G, f) and (G', f') are similar with scale factor a, then annealing on G at temperature T is equivalent to annealing on G' at temperature aT. That is, if $v_0' = \sigma(v_0)$, the Markov chain v_t' defined by annealing on G' is identical to $\sigma(v_t)$ – the image in G' of the Markov chain defined by annealing on G.*

6.2 Unknown Energy Range

Throughout this chapter and Chapter 7 the arguments ε and n of $\hat{T}$ and $\hat{t}$ will be implicit when not written explicitly. In particular, by $\hat{T}(\varepsilon)$ or $\hat{T}(\varepsilon, n)$ we shall mean $\hat{T}(\varepsilon, \varepsilon, n)$; and by $\hat{t}(T)$ we shall mean $\hat{t}(T, \varepsilon, n)$. We will also be assuming the worst-case conductance $\Phi = 1/n^2$, giving $\hat{t}(\varepsilon) = 2(1 + \frac{1}{\varepsilon}) \ln(n^2/\varepsilon) \frac{1}{\Phi(\infty)^2} (n^2/\varepsilon)^{2/\varepsilon}$.

Lemma 6.2.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let an annealing graph G with n vertices and energy function $f : G \rightarrow \mathbb{R}$ be given. If f ranges from exactly $f_{\min}$ to $f_{\max}$, let $f_{\text{range}} = f_{\max} - f_{\min}$. Let a value $0 < \varepsilon < 1$ be given, as well as values $0 < r < 1$, $c > 0$, and $k \in \mathbb{Z}$ satisfying $cr^k \leq f_{\text{range}} \leq cr^{k-1}$. Anneal at temperature $T = cr^k \hat{T}(\varepsilon) = cr^k \varepsilon / \ln(n^2/\varepsilon)$ for time*

$$\begin{aligned} t &= \hat{t}(r\hat{T}) && (6.1) \\ &= 2\left(\ln(n^2/\varepsilon) + \frac{1}{r\hat{T}}\right) n^4 e^{2/(r\hat{T})}. && (6.2) \end{aligned}$$

Then $\mathbf{E}_{\pi_T}[f] \leq f_{\min} + 2\varepsilon\, f_{\text{range}}$, and regardless of the initial state probability vector P_0 the final distribution P_t satisfies $\|P_t - \pi_T\|_{\text{tvd}} \leq \varepsilon$. It follows that $\mathbf{E}_{P_t}[f] \leq f_{\min} + 3\varepsilon\, f_{\text{range}}$.

Theorem 4.7.6 gave us a way of setting temperature and time to guarantee a low expected energy, given energy ranging from exactly 0 to exactly 1. Rescaling, it can be applied to functions with any energy scale, but the scale must still be known exactly. Lemma 6.2.1 above allows a similar result if the maximum energy is known to lie in some interval: it simply chooses a temperature small enough to guarantee low T-equilibrium energy even if the energy range is at the low end of this interval, and a time long enough to adequately approximate equilibrium even if the energy range is at the high end of the interval.

Combining Lemma 6.2.1 and Corollary 5.1.19 (cooling with a monotonic superschedule), we can bracket the energy into intervals $(cr, c], (cr^2, cr], (cr^3, cr^2], \ldots$, solve the annealing sufficiently well in each interval, and so guarantee a solution to the problem as a whole:

Theorem 6.2.2 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let us be given a value $0 < \varepsilon < 1$, and an annealing graph G which is known to have no more than n vertices and whose energy function f is known to have range $c_1 \leq f_{\text{range}} \leq c_2$. Let $r < 1$ and $K \in \mathbb{Z}$ be such that $r^K c_2 \leq c_1$. For $k = 1, \ldots, K$ in turn, anneal at temperature $T_k = c_2 r^k \hat{T}(\varepsilon, n)$ for time $t = \hat{t}(r\hat{T})$. Then the* *distribution P_K after the K'th "generation" of annealing satisfies*

$$\mathbf{E}_{P_K}[f] \leq f_{\min} + 3\varepsilon\, f_{\mathrm{range}}. \tag{6.3}$$

6.3 Efficiency

Corollary 6.3.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Under the conditions of Theorem 6.2.2, let $r = 1 - \hat{T}(\varepsilon)/2$ and $K = \lceil \ln\left(\frac{c_2}{c_1}\frac{1}{\varepsilon}\right)/\ln(1/r)\rceil$. Then the cooling schedule specified by Theorem 6.2.2 uses total running time asymptotically equal to*

$$K\hat{t}(r\hat{T}(\varepsilon)) \lesssim \frac{2e}{\varepsilon}\ln^2(1/\varepsilon)\hat{t}(\hat{T}(\varepsilon)). \tag{6.4}$$

to yield its solution of "quality" (relative expected energy) no more than 3ε.

Recall that

$$\hat{t}(\hat{T}(\varepsilon)) = 2(1+\frac{1}{\varepsilon})\ln(n^2/\varepsilon)n^4(n^2/\varepsilon)^{2/\varepsilon}. \tag{6.5}$$

Focusing on the expression $\left[(n^2/\varepsilon)^{2/\varepsilon}\right]$, we see that

$$\hat{t}(\hat{T}(\varepsilon)) \sim n^4\left[(n^2/\varepsilon)^{2/\varepsilon}\right]\ln\left[(n^2/\varepsilon)^{2/\varepsilon}\right] \tag{6.6}$$

while

$$K\hat{t}(r\hat{T}(\varepsilon)) \lesssim en^4\left[(n^2/\varepsilon)^{2/\varepsilon}\right]\ln^3(\left[(n^2/\varepsilon)^{2/\varepsilon}\right]). \tag{6.7}$$

The principal term $\left[(n^2/\varepsilon)^{2/\varepsilon}\right]$ is the same in both cases, and the factor increase in time for annealing a function whose energy range is known only loosely compared with annealing a function whose range is known exactly is only logarithmic in this dominant term.

Taking the ratio of the logarithms of run time and "quality" $q = 1/3\varepsilon$,

$$\frac{\ln t_{\mathrm{tot}}}{\ln q} = \frac{\ln K\hat{t}(r\hat{T}(\varepsilon))}{\ln 1/3\varepsilon} \sim 2/\varepsilon = 6q, \tag{6.8}$$

which is precisely the same result as given by equation (4.26) for functions with energy range equal to 1.

In contrast to the specification of Corollary 6.3.1, the following remark describes an inefficient choice of r and K.

Remark 6.3.2 *Under the conditions of Theorem 6.2.2, it suffices to use $K = 1$ and $r = c_1/c_2$. However, the total running time specified in that theorem would then be $t_{\mathrm{tot}} = e^{2/r\hat{T}} = (n^2/\varepsilon)^{2/r\varepsilon}$.*

In this case the ratio of the logarithms would be

$$\frac{\ln t_{\text{tot}}}{\ln q} = \frac{\ln \hat{t}(r\hat{T}(\varepsilon))}{\ln 1/3\varepsilon} = 2/r\varepsilon = (c_2/c_1) \cdot 6q. \tag{6.9}$$

Since there is no *a priori* bound on c_2/c_1, this is considerably worse than the ratio achieved by Corollary 6.3.1 (equation (6.8)).

The key Theorem 6.2.2 and Corollary 6.3.1 relied only on achieving relative expected energies less than 3ε in each generation. As in Theorems 4.7.6, *etc*, this can be done with $\hat{t}$ and $\hat{T}$ dependent on any of a number of values of Δ. While the presentation above has used $\Delta = \varepsilon$, this is not the only possibility. If we generalize our definition of Δf to be the gap between the smallest and second-smallest energies relative to the energy range, it is clear that this too will suffice.

With that choice, the total time required for annealing on functions of unknown energy range is polynomial in the solution quality desired, $t_{\text{tot}} = q^{2/\Delta f + w(q)}$, just as for the known-range case (equation (4.25)). Even if Δf is not known, time t_{tot} which is quasi-polynomial in q suffices to guarantee that quality *for q sufficiently large*, again just as for the known-range case (equation (4.28)).

Chapter 7

Application of Annealing to Linearly Separable Functions

The results of the previous chapter may be applied to "linearly separable" functions: many-argument functions expressible as a sum of functions of single arguments, $f(\bar{v}) = f_1(x_1) + \cdots + f_d(x_d)$. We now make this more precise.

7.1 Linearly Separable Energy Functions

Definition 7.1.1 *Given graphs G_i with vertex sets V_i and edge sets E_i, the* **product graph** *$G = \prod_i G_i$ is the graph with vertex set $V = \prod_i V_i$ and edge set E consisting of all pairs (v, v') such that v and v' differ in a single coordinate i and $(v_i, v'_i) \in E_i$.*

Definition 7.1.2 *The* **energy product graph** *$(G, f) = \prod(G_i, f_i)$ has $G = \prod_i G_i$ and $f(v_1, \ldots, v_d) = f_1(v_1) + \cdots + f_d(v_d)$.*

For example, Figure 7.1 shows the product of a 3-chain and a 4-chain, each having self-loops on its endpoints; the energies (not shown) would simply be formed as an addition table of the energies of the vertices of the chains.

Definition 7.1.3 *The function $f : G \rightarrow \mathbb{R}$ is* **linearly separable** *into $f_1, \ldots, f_d$ if (G, f) is the product of (G_1, f_1) through (G_d, f_d).*

Equivalently to Definition 7.1.3, any linearly separable function can be expressed as $f(\bar{v}) = \sum_{i=1}^{d} r_i f_i(v_i)$, where the f_i's have minimum 0 and maximum 1, and the r_i's are nonnegative scale factors. Without loss of generality we will suppose that $r_1 \geq r_2 \geq \cdots \geq r_d$.

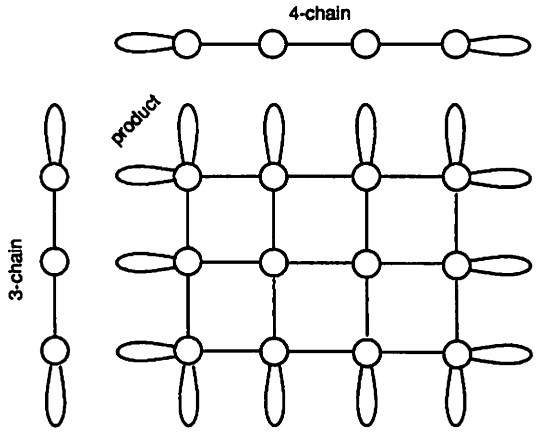


Figure 7.1: Product of 3-chain and 4-chain.

7.2 Annealing on Separable Functions

Simulated annealing can be applied efficiently to linearly separable functions, even though the annealing algorithm has no explicit knowledge of the separability.

Lemma 7.2.1 *On an annealing graph $G = \prod_{i=1}^{d} G_i$, anneal with schedule $\{(T_k, t_k)\}_{k=1}^{K}$ with total number of moves $s = \sum t_k$. Let P_t denote the distribution of the state x of G after t moves, and P_t^k the corresponding marginal distribution of x_k. Then P_s^k is equal to the distribution of x_k that would result from annealing on G_k with initial distribution P_0^k and cooling schedule $\{(T_k, t'_k)\}$, where $t'_k \sim B(t_k, 1/d)$ (i.e. t'_k is random with binomial distribution $B(t_k, 1/d)$).*

Proof Each move has probability $1/d$ of being applied to coordinate k. Only then can x_k change, and then (by Definition 7.1.2) the move generation and acceptance laws are just those for annealing on G_k. This theorem could also be considered a special case of Theorem 8.4.4, which is proved in gory detail. ■

Definition 7.2.2 *Let $\Upsilon(n, p, \varepsilon)$ be the smallest integer n' such that for the random variable X chosen from $B(n', p)$,* $\mathrm{P}[X < np] \le \varepsilon$.

In the cases of interest we will find $\Upsilon(n, p, \varepsilon) \sim n$.

Lemma 7.2.3 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let* $(G,f) = \prod_{i=1}^{d}(G_i, r_i f_i)$ *where without loss of generality each* f_i *ranges from exactly* 0 *to* 1 *and each* $r_i \geq 0$. *Let* n *be an upper bound on the order of each* G_i.

Let $0 < r < 1$, $c > 0$, *and* $k \in \mathbb{Z}$ *satisfy* $r^k c < r_i < r^{k-1}c$.

Given $\varepsilon > 0$, *let* $\hat{T} = \hat{T}(\varepsilon)$ *and* $T = r^k c\hat{T}$.

Let $\hat{t} = \hat{t}(r\hat{T})$ *and* $t = \Upsilon(d\hat{t}, 1/d, \varepsilon)$.

From any initial distribution anneal on G *at temperature* T *for time* t. *Let the final distribution have corresponding i-marginal distribution* P. *Then* $\mathbf{E}_{\pi_T}[f_i] \leq 2\varepsilon$ *and* $\|P - \pi_T\|_{\text{tvd}} \leq 2\varepsilon$. *Consequently* $\mathbf{E}_P[r_i f_i] \leq 4\varepsilon r_i$.

Proposition 7.2.4 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let* $t(\varepsilon) > (1/\varepsilon)^2$. *Then for any fixed* d, $\Upsilon(dt(\varepsilon), 1/d, \varepsilon) \sim dt(\varepsilon)$.

Since $\hat{t}(\varepsilon) > (1/\varepsilon)^2$, a consequence of Proposition 7.2.4 is that in Lemma 7.2.3, $t \sim d\hat{t}$ as $\varepsilon \rightarrow 0$.

Theorem 7.2.5 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Let* $(G,f) = \prod_{i=1}^{d}(G_i, r_i f_i)$ *with* f_i *ranging from exactly* 0 *to* 1 *and* $r_i \geq 0$. *Let* n *be a known upper bound on the order of each* G_i, *and let* c_1 *and* c_2 *be known lower and upper bounds for the largest* r_i.

Given $\varepsilon > 0$, *let* $0 < r < 1$ *and* $K \in \mathbb{Z}$ *satisfy* $r^K c_2 \leq c_1 \varepsilon / d$.

Let $\hat{T} = \hat{T}(\varepsilon)$ *and* $T_k = r^k c_2 \hat{T}$. *Let* $\hat{t} = \hat{t}(r\hat{T})$ *and* $t = \Upsilon(d\hat{t}, 1/d, \varepsilon)$.

From any initial distribution anneal on G *with cooling schedule* $\{(T_k, t)\}_{k=0}^{K}$. *Then the final distribution* P *gives relative expected energy*

$$\mathbf{E}_{\text{sep}} \equiv \frac{\mathbf{E}_P[f(v)]}{\max_v f(v)} \leq 5\varepsilon; \tag{7.1}$$

that is, the expected final cost relative to the full range of the cost function is no more than 5ε.

7.3 Efficiency

An assessment of the efficiency of annealing for linearly separable functions follows the line of the discussion for functions of unknown range in Chapter 6.3.

Corollary 7.3.1 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *Under the conditions of Theorem 7.2.5, let* $r = 1 - \hat{T}(\varepsilon)/2$ *and* $K = \lceil \ln\left(\frac{c_2}{c_1}\frac{d}{\varepsilon}\right) / \ln(1/r) \rceil$. *Since* $\Upsilon(d\hat{t}, 1/d, \varepsilon) \sim d\hat{t}$, *the cooling schedule specified by that theorem uses total running time*

$$t_{\mathrm{sep}} \sim dK\hat{t}(r\hat{T}(\varepsilon)) \lesssim d\frac{2e}{\varepsilon}\ln^2(1/\varepsilon)\hat{t}(\hat{T}(\varepsilon)). \tag{7.2}$$

to yield its solution of "quality" (relative expected energy) $\mathbf{E}_{\mathrm{sep}} \leq 5\varepsilon$.

Again we focus on $\left\lceil (n^2/\varepsilon)^{2/\varepsilon} \right\rceil$, the principal term in $\hat{t}(\hat{T}(\varepsilon))$. The time given in (7.2) for annealing a "d-dimensional" linearly separable function whose energy range is known only loosely, compared with annealing a single function whose range is known exactly, is increased by a factors only logarithmic in $\left\lceil (n^2/\varepsilon)^{2/\varepsilon} \right\rceil$ and linear in d. With $q = 1/\,5\varepsilon$, our standard measure of efficiency is

$$\frac{\ln t_{\mathrm{tot}}}{\ln q} = \frac{\ln dK\hat{t}(r\hat{T}(\varepsilon))}{\ln 1/\,5\varepsilon} \sim 2/\varepsilon = 10q, \tag{7.3}$$

only a constant factor worse than the results for one-dimensional functions of known or unknown energy range ((4.26) and (6.8)).

By contrast, annealing this in a single shot (in parallel to Remark 6.3.2 and equation (6.9)) would mean using $K = 1$ and $r = c_1\,/\,c_2 d$, for the poor result

$$\frac{\ln t_{\mathrm{tot}}}{\ln q} = \frac{\ln \hat{t}(r\hat{T}(\varepsilon))}{\ln 1/\,5\varepsilon} = 2/r\varepsilon = (c_2/c_1)\cdot 10q. \tag{7.4}$$

Again, the values of $\hat{t}$ and $\hat{T}$ could be based on values of Δ other than ε. In this case we would have to use $\Delta = \min_i \Delta f_i$, the worst-case relative second-smallest energy of any dimension. This results in the same conclusion as in Chapter 6.3, *i.e.* that t_{tot} is polynomial in solution quality if Δ is known, or quasi-polynomial (for q sufficiently large) if Δ is unknown.

7.4 Hypercube Model

A special case of possible interest is linearly separable functions f defined on "hypercubes". The domain of such a function (the annealing state space) will be the space $\{0, \ldots, b-1\}^d$, that is, a $b \times \cdots \times b$ hypercube. A state is a vector $\bar{x} = \langle x_1, \ldots, x_d \rangle$. Of course, the separability of f means that we may express f as $f(\bar{x}) = \sum_{i=1}^{d} f_i(x_i)$.

To introduce the move set it is simplest to first consider just the space $\{0,\ldots,b-1\}$ itself. A move from $x \in \{0,\ldots,b-1\}$ consists of changing x by plus or minus 1. The terminal values, $x=0$ and $x=b-1$, can be handled in either of two ways. One is to consider x as in integer modulo b, and treat the terminal values just like the others; in this case the graph of the states and moves is a b-cycle. The other way is to let the graph be a b-chain with self-loops at the endpoints; so from $x=0$ there is a move to 1 and also a (null) move to 0, and from $x=b-1$ the moves are to $b-2$ and to $b-1$. Figure 7.2 illustrates the two move sets for $b=4$.

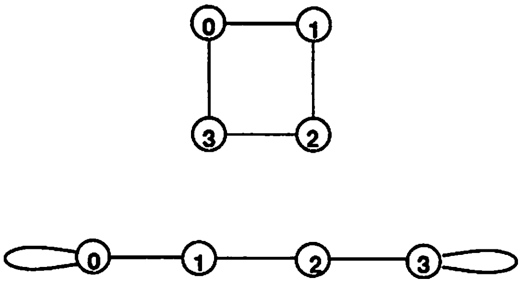


Figure 7.2: Move sets for the space $\{0,1,2,3\}$.

A move on the d-dimensional hypercube $\{0,\ldots,b-1\}^d$ will consist simply of a move on any one of the x_i's. That is, whichever move graph is chosen for a single coordinate, the overall move graph is the product graph. All results from the general case carry over with $n=b$. They may be improved trivially by replacing the worst-case conductance of $1/n^2$ with the actual $1/b$ true for this case.

7.5 Discussion

An advantage of Corollary 7.3.1 is that it may be a practical basis for constructing a cooling schedule, as the construction it presents relies only on loose bounds on relatively few parameters. The bound 5ϵ on the "relative quality" of the solution produced is a natural parameter for a user to specify. The number of dimensions d of the problem may well be known, and otherwise it is likely that an upper bound is; similarly for the maximum order n of the dimensions. The most difficult parameters to estimate may be the bounds c_1 and c_2 on the range r_1 of the most-variable component function. The range of f itself could serve as an upper bound for r_1, and a crude estimate of it obtained by subtracting the minimum of f from the maximum of f, as obtained by any quick optimization method. Similarly the range of f divided by d is a lower bound for r_1. We speculate that a random walk of length dn on G might also suffice for estimating r_1.

We should also point out that the cooling schedules specified by Theorem 7.2.5 and its Corollary 7.3.1 are simple, but at the price of being strongly constrained and not quite optimal. The constraint imposed by the theorem is that the schedule is exactly geometric: the temperature is reduced by the same factor r each time and the same amount of time $\hat{t}$ is spent in each generation. We would expect that it would be better to spend more time on the more significant components of f, *i.e.* more time at higher temperatures, resulting in a cooling schedule in which temperature decreases faster than geometrically.

Also, for the geometric schedule the optimization of r was done only in the asymptotic limit as $\varepsilon \rightarrow 0$, and even then we rounded the resulting expression for r.

In [31] we expect to take up these issues again. Keeping the linearly separable model, we hope to construct an adaptive cooling schedule which is as efficient as possible (as opposed to the asymptotically efficient schedules of this chapter); relies as little as possible on detailed formulas such as $\hat{t}(\hat{T}(\varepsilon))$; and is not limited to cooling schedules which are geometric.

Meanwhile in Chapter 8 we discuss a fractal energy function model which resembles the hypercube model. The primary difference is that the rather than combining the arguments x_i into a vector argument $\bar{x}$, the values $x_1, \ldots, x_d$'s are (roughly speaking) interpreted as the digits of a single value x; and the move set is modified correspondingly. At the same time, we simplify matters by setting $f_i = F$ and $r_i = r^i$ for a fixed function F and fixed value r.

Chapter 8

Application of Annealing to a Class of Deterministic Fractals

We will define a class of "fractal" energy functions on the "state space" $[0,1]$, and show that for any function f in this class, a variant of simulated annealing finds a solution of low energy quite rapidly.

Definition 8.0.1 *We are given an integer $b > 0$, a real number $r \in (0,1)$, and a function $F : \{0,\ldots,b-1\} \to \mathbb{R}$. Write x base b as $.x_1x_2\ldots$. (If $b^{-k}|x$, so that x has two base b representations, use the terminating one.) Then the* **deterministic fractal** *$f(x)$ based on b, r, and F, is given by $f(1) = f(0)$ and otherwise*

$$f(x) = F(x_1) + rf(\mathrm{comp}_{x_1}(.x_2x_3\ldots)) \tag{8.1}$$

where "comp" is the complement function defined as

$$\mathrm{comp}_a(x) = \begin{cases} x & \text{if } a \text{ is even,} \\ 1-x & \text{if } a \text{ is odd.} \end{cases} \tag{8.2}$$

The discontinuities of f will not be an issue. Without loss of generality we will assume F ranges from 0 to 1, so that f ranges from 0 to 1/(1-r).

8.1 Properties of the Fractal Functions

Before proceeding to a formal analysis we will give an intuitive description of f and some of its properties.

Generations 1, 2, and 6 of an iterative construction of the fractal energy function f are illustrated in Figure 8.1. In this and all our examples we take $b = 3$, $F(0) = 5/7, F(1) = 0, F(2) = 1$, and $r = .3$.

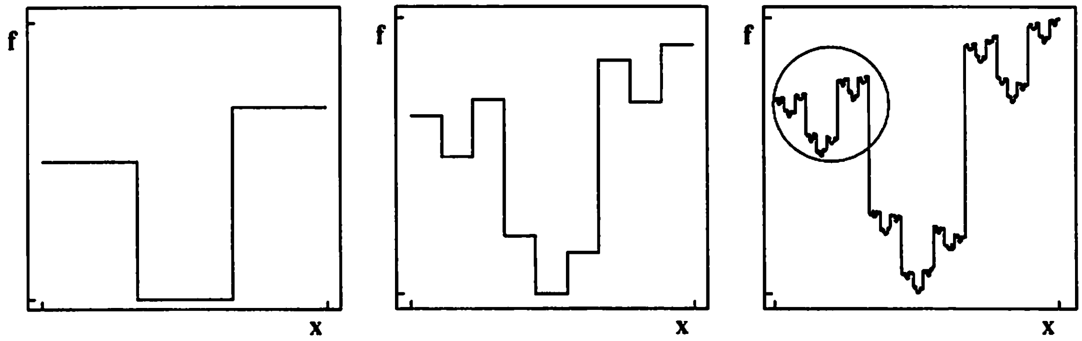


Figure 8.1: Fractal generations 1, 2, and 6.

The "generation 0" approximation of f is just the function 0 on the interval $[0, 1]$.

In "generation 1", the line segment from 0 to 1 is broken into b pieces and each piece is raised by some amount (here 5/7, 0, and 1 from left to right).

In generation 2 each 1-piece is itself split into b "2-pieces", which are raised by r times the values used in generation 1. There is an additional twist though: every other 1-piece (here just the middle one) is "mirrored" so that all the operations are now done from right to left rather than left to right.

A similar process is performed for generation 3, mirroring every other 2-piece. The infinite-generation limit of the process is the graph of the function f from states (x axis) to energies (y axis). Note that the 1-piece of f circled in the last frame of Figure 8.1 is "similar" to f itself: if its domain and codomain axes are expanded by b and $1/r$, the result is a translation of f on $[0, 1]$. Any other k-piece of f is also similar to the whole, which is the essential "self-similarity" property of f.

What does this have to do with the random fractals (fractional Brownian motions) of Chapter 3? A characteristic property of fBm is that taking a "piece" of the fractal (a subset of the domain and its image in the codomain) and expanding its horizontal and vertical axes by factors r_{horiz} and r_{vert} yields a function "statistically similar" to the original one. Here r_{horiz} and r_{vert} are arbitrary, subject to the constraint ${r_{vert}}^2 = {r_{horiz}}^{2H}$ where H is the parameter of the fBm. Where fBm's had a statistical scaling property, the deterministic fractals we are now considering scale exactly.

Some notation will be useful. Label the b "1-pieces" $0, \ldots, b-1$, from left to right. (See Figure 8.2. Since the 1-piece numbered "1" is mirrored, we have depicted it as a vector pointing backwards.) Let the state space S_1 consist of the midpoints of these 1-pieces, *i.e.* the 3 points evenly spaced in $[0, 1]$.

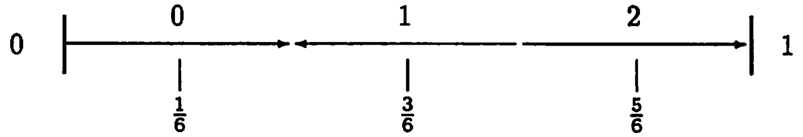


Figure 8.2: The three 1-intervals with their orientations and labels, and the points of S_1 with their coordinates.

The k-labels for the k-pieces can be extended to $(k+1)$-labels for the $(k+1)$-pieces: Any k-piece contains b $(k+1)$-pieces. The first k digits of their labels match the label of the k-piece, and the last digits run from 0 to $b-1$, in the direction the k-piece is oriented. In our example (Figure 8.3), the 2-pieces are labeled $\langle 0,0\rangle, \langle 0,1\rangle, \langle 0,2\rangle$, $\langle 1,2\rangle, \langle 1,1\rangle, \langle 1,0\rangle$, and $\langle 2,0\rangle, \langle 2,1\rangle, \langle 2,2\rangle$.

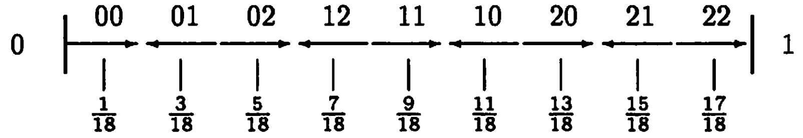


Figure 8.3: The nine 2-intervals with their orientations and labels, and the points of S_2 with their coordinates.

In general, let state space S_k consist of the midpoints of the k-pieces. With any value x associate a k-code $\mathrm{s}_k(x)$ which is the label of the k-piece containing x (this is unique except at f's points of discontinuity). There is a natural four-way correspondence between the k-pieces, their k-labels, the states of S_k, and their k-codes. Because the $(k+1)$-code of a point is an extension of its k-code, we can define the infinite-length code $\mathrm{s}(x)$ which is the "limit" of the k-codes. Note that the codes resemble the base b representation of x

but incorporate the mirroring used in defining f, which will make the expression of $f(x)$ in terms of $\mathrm{s}(x)$ particularly simple. We will return to this point in the formal treatment of the state spaces S_k, the k-codes $\mathrm{s}(x)$, and their properties in relation to f, to which we now proceed.

Definition 8.1.1 *A* **k-piece** *of the interval* $[0,1]$ *is an open interval* $(\frac{j}{b^k}, \frac{j+1}{b^k})$, *for* $j \in \{0,\dots,b^k-1$.

Note that we can also make an iterative construction: the 1-pieces are $(0,1/b)$, $(1/b,2/b)$, $\dots$, $((b-1)/b,1)$. (Roughly, each piece contains all numbers which agree in the first digit of their base-b representations.) Any k-piece is $(x_1 + S) \div b$ where S is a $(k-1)$-piece and $x_1 \in \{0,\dots,b-1\}$. (x_1 is the first digit in the base-b representation of any x in that k-piece.)

Definition 8.1.2 *The* **kth state space** S_k *is* $\{x : x \in [0,1]$ *and* $b^k x - \frac{1}{2} \in \mathbb{Z}\}$.

Definition 8.1.3 *The* **(infinite) code** $\mathrm{s}(x)$ *is a sequence* $\langle s_1, s_2, \dots\rangle$ *given by* $\mathrm{s}(1) = \mathrm{s}(0)$ *and otherwise*

$$\mathrm{s}(.x_1x_2\dots) = \langle x_1\ ,\ \mathrm{s}(\mathrm{comp}_{x_1}(.x_2x_3\dots))\rangle, \tag{8.3}$$

where the comma denotes concatenation. The **k-code** $\mathrm{s}(x)$ *is the vector* $\langle s_1, s_2, \dots, s_k\rangle$, *the first k components of* $\mathrm{s}(x)$.

The following set of lemmas describes some salient properties of the codes and their relationship with the k-pieces and the function f. The proofs are all relatively uninteresting proofs by induction.

Lemma 8.1.4 $\boxed{\text{pf} \to \text{appendix}}$ *For x, x' not divisible by b^{-k}, $\mathrm{s}(x)$ and $\mathrm{s}(x')$ agree in components 1 through k if and only if x and x' lie in a common k-piece.*

Lemma 8.1.5 (Gray code property) *As a function on S_k, the k-code is a generalized Gray code: it is invertible, and has the property that the codes for geometrically adjacent x's differ in only one component, and differ by 1 there.*[1] *Specifically, the component is the ith if x and x' lie in a common $(i-1)$-piece but in different i-pieces.*

[1] When $b = 2$ the k-codes form the standard length-k binary Gray codes. See [8] for a description of Gray codes.

Proof May be done by induction, or by applying Lemmas 8.1.4 and 8.1.6. ■

Lemma 8.1.6 (Additional Gray code property) $\boxed{\text{pf} \rightarrow \text{appendix}}$ *For any vector* $\vec{s} = \langle s_1, s_2, \ldots \rangle$*, let* $k{\downarrow}\vec{s} = \langle s_{k+1}, s_{k+2}, \ldots \rangle$*. Then for any integers* $k > 0$ *and* $j \in \{0, \ldots, b^k - 1\}$*, for all* $x \in (0,1)$*,*

$$k{\downarrow}\mathrm{s}\left(\frac{j+x}{b^k}\right) = \mathrm{s}(\mathrm{comp}_j(x)). \tag{8.4}$$

This means that if we imagine scanning through the values $x' = (j+x)/b^k$ in the jth k-piece and reading off their codes $\mathrm{s}(x')$ from component $k+1$ onwards, this "sequence" of codes is the same as that obtained by scanning along $(0,1)$ (scanning from right to left if j is odd) and reading off the codes from component 1 onwards.

Lemma 8.1.7 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *For* $\mathrm{s}(x) = \langle s_1, s_2, \ldots \rangle$*,*

$$f(x) = F(s_1) + rF(s_2) + r^2F(s_3) + \cdots. \tag{8.5}$$

Lemma 8.1.8 *Going from left to right through the values* $x \in S_k$ *in a given* $(k-1)$*-piece (or right to left if the piece is negatively oriented), the value* $f(x)$ *of the ith point* x *is* $\mathrm{const} + r^{k-1}F(i)$.

Proof Follows immediately from Lemmas 8.1.6 and 8.1.7. ■

Lemma 8.1.9 *For any* $(k-1)$*-piece of* S_k*, define a graph whose vertices are the* b *points of* S_k *within this* $(k-1)$*-piece; whose edges connect geometrically adjacent points, with self-loops added at the two endpoints; and whose energies are given by the deterministic fractal* $f(x)$*. Any such graph is similar (per Definition 6.1.1) to the one for the single 0-piece of* S_1*, with energies scaled down by* r^{k-1}*.*

Proof Both graphs are chains on b vertices, with self-loops added at the endpoints. The isomorphism maps the ith smallest point in one chain to that in the other, or else the ith smallest point in one chain to the ith largest point in the other. As i goes from 0 to $b-1$ the energy values in S_1 are $F(i) + \mathrm{const}$, while (by Lemma 8.1.8) those in the $(k-1)$-piece of S_k are $r^{k-1}F(i) + \mathrm{const}$. ■

8.2 Annealing on Fractals

The sequence of state spaces S_k gives us a way to anneal on the infinite state space $[0,1]$: we begin by annealing on S_1, then map its final state to an initial state in S_2; anneal on S_2 and map that final state to an initial state in S_3; and so forth.

A "move" in state space S_k goes from one state to a geometrically adjacent one. So that the two end states will have degree 2 like the others, we add moves from each end state to itself. We note that by the Gray code property of the k-codes (Lemma 8.1.5), in moving from one point in S_k to an adjacent one, exactly one element of the k-code changes, and it changes by 1.

The idea of changing the move set as the algorithm proceeds is well established in the literature [25], and changing the state space is a natural way to allow an infinite state space, making the theoretical analysis more interesting by removing any minimum granularity (*e.g.* a minimum nonzero energy).

The modified annealing algorithm for the changing state spaces and move sets is given in Figure 8.4. Identifying an x in S_{k-1} with a $(k-1)$-interval, the function "*fudge* " maps $x \in S_{k-1}$ to an arbitrary[2] $x' \in S_k$ such that x "contains" x'. We might think of *fudge* as simply being the identity map, but unfortunately if b is even then the values in S_{k-1} are

```
choose a sequence $(T_k, t_k)$ and value $K$
$x_0^{0,t_0}$ = random state in $S_0$
for $k = 1$ to $K$ {
    $x_k^{k,0} = fudge(x_{k-1}^{k-1,t_{k-1}})$
    From $x_k^{k,0}$ anneal on $S_k$ at temperature $T_k$ for time $t_k$, ending at
    state $x_k^{k,t_k}$.
}
```

Figure 8.4: Algorithm for annealing on fractals with changing state space and move set.

[2] In Chapter 8.4, we will have to restrict *fudge* a bit. But even there, allowing *fudge* to be random over the values in its codomain is perfectly acceptable.

not in S_k, so *fudge* allows assignment of "nearby" values instead.

Roughly speaking, the algorithm first anneals on S_1. Beginning from that final state (modulo *fudge*), it anneals on S_2, and so forth. We expect the generation-1 annealing to pick a good (low-energy) 1-piece of the unit interval, the generation-2 annealing to home in on a good 2-piece within that 1-piece, and so forth. Figure 8.5 illustrates the change from generation 1 to generation 2.

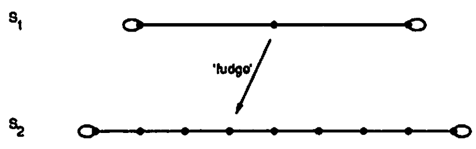


Figure 8.5: Unconfined annealing: generations 1 and 2. Initially we anneal on S_1, at temperature T_1. States are indicated by bullets, and legal moves by lines and loops. In this case, the final state in generation 1 happens to be the center one. The function *fudge* maps it to any of the 3 center states of S_2, in this case the left one. This is the initial state for annealing on S_2, at temperature T_2.

In analogy with the notation Δf, we define ΔF to be the smallest nonzero value of ΔF. Assuming that b is small and that ΔF is not too small, Theorem 4.7.6 shows that annealing on S_1 is efficient – after annealing for a short time, the current state has low expected energy. In particular:

Theorem 8.2.1 *Given any $\varepsilon > 0$, starting from an arbitrary initial distribution and annealing on S_1 at temperature*

$$\hat{T}(\varepsilon) = \Delta F / \ln(b^2/\varepsilon) \tag{8.6}$$

for time

$$\hat{t}(\hat{T}) = 2\left(\ln\left(\frac{b^2}{\varepsilon}\right) + \frac{1}{T_1}\right) b^4 e^{2/T_1}, \tag{8.7}$$

the final distribution satisfies

$$\mathbf{E}[F(x)] \leq 2\varepsilon. \tag{8.8}$$

This theorem is simply an application of Theorem 4.7.6. In analogy with equation (4.25) which followed from Theorem 4.7.6, Theorem 8.2.1 means that a solution of quality $q = 1/2\varepsilon$ can be found in time $t_1 = (1/q)^{2/\Delta F + w(q)}$ where w is some function approaching 0 as $q \to \infty$.

To analyze the behavior of simulated annealing in the later generations, we will temporarily switch to a "confined" version of the annealing algorithm which is easier to analyze; we will then go back and complete the analysis for the "unconfined" annealing as already defined.

8.3 Confined Annealing

In "confined" annealing, during generation k we insist that the state remain within the $(k-1)$-piece in which it started. Thus, the kth-generation state space is made up of b^{k-1} disconnected pieces, each having b states. So that all vertices will still have degree 2, wherever we disallowed a move between two states we now add self loops to each of them.

That is, the graph of a component of S_k is a chain on k points, with self-loops on the endpoints, just as S_1 is. By Lemma 8.1.9 any such component with the energy function f on it is similar to f on S_1, so by Lemma 6.1.2, annealing at temperature $r^{k-1}T_1$ on a component of S_k is equivalent to annealing at temperature T_1 on S_1.

Another way to view this is to identify the state x during annealing with its infinite code $\mathrm{s}(x) = \langle s_1, s_2, \ldots \rangle$. In confined annealing during generation k, only s_k may change. If temperature $T_k = r^{k-1}T_1$ is used during generation k, the transition rule acting on s_k in generation k is the same as the transition rule acting on s_1 in generation 1.

Let $s_k^{i,t}$ be the value of s_k after t moves during generation i. Let t_k be the total number of moves attempted in generation k.

Proposition 8.3.1 *If confined annealing is run for K generations, $s_k^{K,t_K} = s_k^{k,t_k}$.*

Proof After generation k, *fudge* (by its definition) never changes s_k, nor, by the "confinement" premise, does the annealing itself. ■

Theorem 8.3.2 *Let a value ε and a deterministic fractal with energy scale parameter r be given. Let $\hat{T} = \hat{T}(\varepsilon)$ and $\hat{\imath} = \hat{\imath}(\hat{T})$. Apply confined annealing with cooling schedule $(T_k, t_k) = (r^{k-1}\hat{T}, \hat{\imath})$, for $k = 1, \ldots, K$, with $K = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil$. Then the state returned has relative expected energy*

$$\mathbf{E}_{\text{con}} \equiv \frac{\mathbf{E}[f]}{f_{\text{range}}} \leq 3\varepsilon \tag{8.9}$$

and the algorithm consumes run time

$$t_{\text{con}} = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil \cdot \hat{t}(\varepsilon). \tag{8.10}$$

Just as Theorem 8.2.1 showed that for generation 1, t_1 was roughly polynomial in quality $q = 1/2\varepsilon$ with exponent $2/\Delta F$, Theorem 8.3.2 shows that over the entire annealing run, the total running time t_{con} is still polynomial in the relative quality $q = 1/3\varepsilon$ with the same exponent $2/\Delta F$, and with constants not much worse. (As usual the logarithmic terms can be hidden in a vanishingly small addition to the exponent.)

Noting that the deterministic fractal $f(x)$ can be thought of as a linearly separable function $f(x) = \sum r^k F(s_k)$, confined annealing can be thought of as almost a special case of annealing for linearly separable functions (Chapter 7). In fact, the results for the two cases, represented by Theorem 8.3.2 and Corollary 7.3.1, are very similar.

One difference between the two cases is that for the confined annealing we have presumed that both r and ΔF are known, saving a factor of poly$(1/\varepsilon)$ in Theorem 8.3.2 compared with Corollary 7.3.1.

The second difference is that in confined annealing, during generation k we always make a move to s_k, the coordinate of interest. For a linearly separable function in d dimensions, we had to make d moves to assure that one move was made to relevant coordinate. This explains why in Theorem 8.3.2 there is no parallel to the factor of d present in Corollary 7.3.1.

8.4 Unconfined Annealing

The "confined" annealing algorithm just discussed is most tractable but is not realistic. In practice there is no easy way to describe the "location" of a point in a combinatorial space, and restricting oneself to points in some region appears impossible.

However, the "confinement" rule may be dropped with only minor effects on the process of annealing on the fractal. Lemma 8.1.5 shows that during generation k, an "unconfined" move changes exactly one element of the k-code and changes it by exactly 1. Because of this it will be possible to treat the components of the k-code independently: the kth component will behave exactly as it did in confined annealing, and the other components (treated all together) will behave like an annealing problem where the temperature is reduced monotonically starting from low-temperature near-equilibrium.

The definition of the unconfined algorithm is again that given in Figure 8.4, but now we must restrict the function *fudge* . Representing x by its code s(x), one satisfactory definition is

$$\mathit{fudge}(\langle s_1, s_2, \ldots, s_k\rangle) = \langle\langle s_1, s_2, \ldots, s_k\rangle\ ,\ s_{\text{rand}}\rangle \tag{8.11}$$

where s_{rand} is a random variable over $\{0, \ldots, b-1\}$, and at each application of *fudge* an independent s_{rand} is chosen. Some natural cases of this definition are the following:

- If b is odd, let $s_{\text{rand}} = (b-1)/2$ (deterministically), meaning $\mathit{fudge}(x) = x$.
- If b is even, let $s_{\text{rand}} = (b/2) - 1$ or $b/2$, meaning $\mathit{fudge}(x) = x \pm 1/2b^{k+1}$ (randomly).
- For any b, let s_{rand} be uniform over $\{0, \ldots, b-1\}$, so *fudge* maps x uniformly randomly to one of the b points in S_{k+1} in the same k-interval as x.

Intuitively, we focus on annealing on the kth component, and imagine the other $k-1$ components as defining b^{k-1} "copies" of this process. A move which changes s_k is a move in the basic process; a move changing any other component is a switch from one "copy" to another. The following definition formalizes the idea of "copies" of a basic structure, and the succeeding theorems show how annealing on such a structure can be analyzed. (Viewing Figure 8.6 may be helpful.)

Definition 8.4.1 *An energy graph $\bar{G}$ with vertex set $\bar{V}$, edge set $\bar{E}$, and energy function $\bar{f}$ is a* **replica** *with index set I and energy scale factor c of the basic graph G with vertex set V, edge set E, and energy function f if there is a function σ from $\bar{V}$ to $I \times V$ with the following properties:*

1. *$\sigma : \bar{V} \to I \times V$ is a bijection.*
2. *Let $\{\bar{v}, \bar{v}'\} \in \bar{E}$ be an edge in $\bar{G}$, and let $(i, v) = \sigma(\bar{v})$ and $(i', v') = \sigma(\bar{v}')$. Then either $i = i'$ or $v = v'$, or both.*
3. *Define a "neighborhood" in G by $N(v) = \{v' : \{v', v\} \in E\}$. Consider this to be a multiset, with v' appearing as many times as there are edges between v and v'. Similarly let $\bar{N}(\bar{v}) = \{\bar{v}' : \{\bar{v}', \bar{v}\} \in \bar{E}\}$.*
   Define σ_I and σ_V by $\sigma(\bar{v}) = (\sigma_I(\bar{v}), \sigma_V(\bar{v}))$.
   If $v = \sigma_V(\bar{v})$ then σ_V is a bijection between $\bar{N}(\bar{v})$ and $N(v)$.

4. *There is a function f_I on I such that for all $\bar{v} \in \bar{V}$, taking $(i, v) = \sigma(\bar{v})$, $\bar{f}(\bar{v}) = f_I(i) + cf(v)$.*

Intuitively, a replica $\bar{G}$ can be formed from a graph G as follows: Make a number of copies of G, indexing copies with values i, and using the same vertex labels v within each copy. If a vertex v has a self-loop, the self-loops on (i, v) and (i', v) can be replaced by an edge from (i, v) to (i', v). The energy of a vertex (i, v) in $\bar{G}$ is the sum of the energy of v in G and the arbitrary function f_I of i. Note that the degree of (i, v) in $\bar{G}$ is equal to the degree of v in G.

We often omit mention of the function σ and simply think of a vertex of $\bar{G}$ as a pair (i, v).

Theorem 8.4.2 $\boxed{\text{pf} \rightarrow \text{appendix}}$ *For $i, k \geq 1$, S_{i+k} with the unconfined move set and deterministic fractal energy function f is a replica of S_k (also with unconfined move set and energy f). The index set is $\{0, \ldots, b-1\}^i$ and the energy scale factor is r^i.*

Figure 8.6 provides an illustration.

Theorem 8.4.3 $\boxed{\text{pf} \rightarrow \text{appendix}}$ **(product density on replica graphs)** *Let $\bar{G}$ be a replica of G with scale factor c. Let P_t be an arbitrary probability distribution on the states V of G, and π_T be the "equilibrium" distribution on I given by $\pi_T(i) \propto e^{-f_I(i)/T}$. Let $\bar{P}_t$ be the distribution on states of $\bar{G}$ given by $\bar{P}_t = \pi_{cT} \times P_t$, i.e. $\bar{P}_t(i, v) = \pi_{cT}(i) \cdot P_t(v)$. Let $\bar{P}_{t+1}(i, v)$ be the distribution on $\bar{G}$ after a single annealing move at temperature cT starting from $\bar{P}_t$, and let $P_{t+1}(v)$ be the distribution on G after a single annealing move at temperature T starting from distribution P_t. Then $\bar{P}_{t+1} = \pi_{cT} \times P_{t+1}$.*

Roughly speaking, the theorem says that if the distribution on a replica graph $I \times V$ is a product distribution which is the equilibrium distribution on I and arbitrary on V, then after an annealing move i is still in equilibrium, and the change to v is just as it would be for annealing on V itself.

Theorem 8.4.4 $\boxed{\text{pf} \rightarrow \text{appendix}}$ **(marginal density on replica graphs)** *Let $\bar{G}$ be a replica of G with scale factor c. Let $\bar{P}_t(i, v)$ be an arbitrary probability distribution on the states $I \times V$ of $\bar{G}$, and let $P_t(v) = \sum_i \bar{P}_t(i, v)$ be the corresponding marginal distribution of v. Let $\bar{P}_{t+1}(i, v)$ be the distribution after a single annealing move on $\bar{G}$ at temperature*

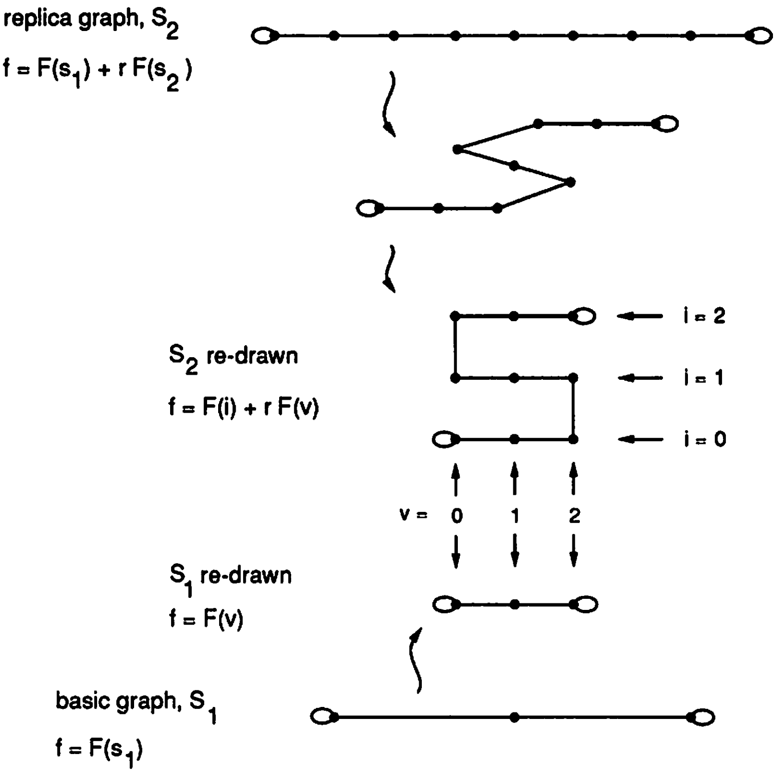


Figure 8.6: With unconfined move set, S_2 is a replica of S_1. S_2 is shown at top, with the energy function f expressed in terms of coordinates s_1 and s_2 of the 2-codes of its points. S_1 is shown at bottom, with f expressed in terms of the sole code coordinate s_1. "Folding" S_2 and shrinking S_1 provides an obvious correspondence between the points of S_2 and S_1: The points of S_1 are labeled $v = s_1$. The 1-pieces of S_2 become the "copies", labeled $i = 0, 1, 2$ (i is the value of s_1 in the 2-code), while the individual points are labeled $v = 0, 1, 2$ (v is s_2).

cT starting from $\bar{P}_t(i, v)$; and let $P_{t+1}(v)$ be the distribution after a single annealing move on G at temperature T starting from $P_t(v)$. Then $P_{t+1}(v)$ is the marginal distribution of v corresponding to $\bar{P}_{t+1}(i, v)$.

The theorem means that regardless of the nature of the distribution on $I \times V$, the evolution of v is just as it would be for annealing on V itself.

Theorem 8.4.3 allows us to treat the distribution of s_1 during generations after the first. Theorem 8.4.4 is used to show that the behavior of the component s_k in generation $k+i$ is identical to that of the component s_1 in generation $1+i$. Together they describe the distribution of each code component.

Theorem 8.4.3 may be thought of as saying that for annealing on a replica graph at a fixed temperature, if the value of i is in exact equilibrium for that temperature then it remains in equilibrium. We now prove a slightly stronger result to the effect that if i is *near* equilibrium at a low temperature T, then it remains near equilibrium, even as temperature is lowered. The statement for replica graphs is given by Theorem 8.4.5, and is an extension of Corollary 5.2.3.

Theorem 8.4.5 $\boxed{\text{pf} \to \text{appendix}}$ *Let $\bar{G}$ be a replica graph with vertices $I \times V$, and let the distributions π_T on I be as per Theorem 8.4.3. Let T_{crit} be the critical temperature for I, and $\{T_t\}$ be a sequence of temperatures satisfying $T_{\text{crit}} \geq T_0 \geq T_1 \geq T_2 \geq \cdots$. Anneal on $\bar{G}$ at temperatures $T_1, T_2, \ldots$ beginning from the distribution $\bar{P}_0$. If there exists a distribution P_0 on V such that the initial distribution $\bar{P}_0$ on $I \times V$ satisfies*

$$\|\bar{P}_0 - (\pi_{T_0} \times P_0)\|_{\text{tvd}} \leq \varepsilon, \tag{8.12}$$

then the distribution at time t satisfies

$$\|\bar{P}_t - (\pi_{T_t} \times P_t)\|_{\text{tvd}} \leq \varepsilon + \|\pi_{T_0} - \pi_{T_t}\|_{\text{tvd}}. \tag{8.13}$$

In the current application we extend this slightly:

Corollary 8.4.6 $\boxed{\text{pf} \to \text{appendix}}$ *Let π_T be the equilibrium distribution of s_1 for annealing on S_1 at temperature T, and let T_{crit} be the associated critical temperature. Run the unconfined annealing algorithm with cooling schedule $\{(T_k, t_k)\}_{k=1}^{K}$, where T_k is monotonically nonincreasing in k and $T_{\text{crit}} \geq T_1$. If s_1^{1,t_1} and s_1^{k,t_k} are the values of s_1 at the end of generation 1 and at the end of generation k respectively, then*

$$\|s_1^{k,t_k} - \pi_{T_k}\|_{\text{tvd}} \leq \|s_1^{1,t_1} - \pi_{T_1}\|_{\text{tvd}} + \|\pi_{T_1} - \pi_{T_k}\|_{\text{tvd}}. \tag{8.14}$$

(This is a slight abuse of notation in the use of the random variables s themselves rather than their distributions.)

Corollary 8.4.7 *Let a value $\varepsilon > 0$ be given. Run unconfined annealing with cooling schedule (T_k, t_k), with $T_1 = \hat{T}(\varepsilon)$ and $t_1 = \hat{t}(\hat{T}(\varepsilon))$ as in Theorem 8.2.1, and T_k monotonically nonincreasing. Assume ε is small enough that $T_1 \leq T_{\text{crit}}$. Then at any generation k,*

$$\|s_1^{k,t_k} - \pi_{T_k}\|_{\text{tvd}} \leq \varepsilon + \|\pi_{T_1} - \pi_{T_k}\|_{\text{tvd}}. \tag{8.15}$$

Proof Direct application of Corollary 8.4.6, using the fact $\|s_1^{1,t_1} - \pi_{T_1}\|_{\text{tvd}} \leq \varepsilon$ (which was the basis of Theorem 8.2.1). ∎

This corollary allows us to "limit the damage" done by lowering the temperature below the value of interest, *i.e.* to ignore the effect on the variable s_1 of annealing in generations $i > 1$. We now develop a complementary result which will allow us to treat each variable like s_1.

Theorem 8.4.8 $\boxed{\text{pf} \to \text{appendix}}$ *Consider two different unconfined annealing processes. The first uses cooling schedule (T_{1+i}, t_{1+i}), runs for generations with $i \geq 0$, and at time t into generation $1+i$ has distribution $P^{1+i,t}(\langle s_1, s_2, \ldots, s_{1+i}\rangle)$ on state space S_{1+i}. For a fixed integer $k > 0$, the second uses cooling schedule $(\bar{T}_{k+i}, \bar{t}_{k+i}) = (r^{k-1}T_{1+i}, t_{1+i})$, again runs for generations with $i \geq 0$, and at time t into generation $k+i$ has distribution $\bar{P}^{k+i,t}(\langle s_1, s_2, \ldots, s_{k+i}\rangle)$ on state space S_{k+i}, with corresponding marginal distribution $\bar{P}^{k+i,t}_{k,\ldots,k+i}(\langle s_k, \ldots, s_{k+i}\rangle)$. If the "initial" distributions satisfy $\bar{P}^{k,0}_k \equiv P^{1,0}$, then for all i, t representing positive time,*

$$\bar{P}^{k+i,t}_{k,\ldots,k+i} \equiv P^{1+i,t}. \tag{8.16}$$

It follows that $\bar{P}^{k+i,t}_k \equiv P^{1+i,t}_1$.

Proof The proof is given in the appendix, but is a straightforward application of Theorems 8.4.2 and 8.4.4. ∎

Thus, for a geometric cooling schedule, the (marginal) distribution of s_k once generation k has started follows precisely the same law as the (marginal) distribution of s_1 from the start of generation 1.

This leads us to the following theorem:

Theorem 8.4.9 $\boxed{\text{pf} \to \text{appendix}}$ *Let a value $0 < \varepsilon < 1$ and a fractal with energy scale parameter r be given. Let $\hat{T} = \hat{T}(\varepsilon)$ and $\hat{t} = \hat{t}(\varepsilon)$. Assume ε is sufficiently small that $\hat{T} \leq T_{\text{crit}}$, the critical temperature for annealing on S_1. Apply unconfined annealing with cooling schedule $(T_k, t_k) = (r^{k-1}\hat{T}, \hat{t})$, and $k = 1, \ldots, K$ with $K = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil$. The state returned has relative expected energy satisfying*

$$\mathbf{E}_{\text{uncon}} \equiv \frac{\mathbf{E}[f]}{f_{\text{range}}} \leq 3\varepsilon \tag{8.17}$$

and the algorithm has run time

$$t_{\text{uncon}} = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil \cdot \hat{t}. \tag{8.18}$$

The theorem is identical to Theorem 8.3.2, but applies to unconfined annealing rather than confined annealing.

To sum up, we have now shown the efficiency of a fairly natural model of annealing on a deterministic fractal (per definition 8.0.1). The intuitive basis of the proof, corresponding to the analysis for confined annealing (Theorem 8.3.2), is to apply in a hierarchical manner the energy-time tradeoff for annealing on arbitrary landscapes.

What makes the proof difficult for the more realistic "unconfined" model of annealing is that the levels of the hierarchy (the various components s_i) cannot be separated entirely: in generation k we may make a change to a component s_i with $i < k$. We can think of this as having "global" consequences (the effect on the distribution of s_i) and "local" consequences (the effect on the distribution of s_k). By Theorem 8.4.5 and its Corollary 8.4.6, assuming that s_1 starts near temperature-0 equilibrium, the effect on the global distribution is minor. And because of the mirror symmetry between adjacent pieces of f, changing a component other than s_k is no different from traversing a self-loop on s_k in the confined case (Lemma 8.4.4), the local distribution evolves exactly as it does for confined annealing.

That we can prove annealing is efficient on these deterministic fractals raises the hope that we can do the same for random fractals more closely resembling the problems encountered in practice. We touch on this again in the Chapter 10.

8.5 Efficiency

In both confined and unconfined annealing we treat the state variables s_k much as we treated the independent variables of a linearly separable function in Chapter 7, albeit with much additional labor for the unconfined annealing. It is then natural that the efficiency results for annealing on fractals should parallel those for linearly separable functions. The essential measure of the efficiency of annealing on fractals comes from Theorem 8.4.9.

$\mathrm{E}_{\mathrm{uncon}}$, the expected energy of the solution returned by unconfined annealing divided by the full range of the fractal f, is no more than 3ε, which we have been considering a solution of "quality" $q = 1/\,3\varepsilon$. Since

$$\hat{t} = \hat{t}(\hat{T}(\varepsilon)) = 2\left(\ln\left(\frac{b^2}{\varepsilon}\right) + \frac{1}{\hat{T}}\right) b^4 e^{2/\hat{T}} \tag{8.19}$$

$$= 2(1 + \frac{1}{\Delta F})\ln(b^2/\varepsilon)\frac{1}{\Phi(\infty)^2}(b^2/\varepsilon)^{2/\Delta F} \tag{8.20}$$

and the total time $t_{\text{uncon}} = K\hat{t}$ with $K = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil$, the ratio of the logarithms of run time and quality is

$$\begin{aligned} \frac{\ln t_{\text{uncon}}}{\ln q} &\sim \frac{\frac{2}{\Delta F}\ln(b/\varepsilon) + O(\ln\ln(1/\varepsilon))}{\ln(1/\varepsilon)} && (8.21) \\ &\sim 2/\Delta F. && (8.22) \end{aligned}$$

While the results for confined annealing (Theorem 8.3.2) generalize to $\Delta = \varepsilon$ (and the other values suggested in Section 4.7), those for unconfined annealing do not. This is because Theorem 8.4.9 requires that $\|\pi_{T_1} - \pi_0\|_{\text{tvd}} \leq \varepsilon$, which is guaranteed by $\Delta = \Delta F$ but not by $\Delta = \varepsilon$.

Chapter 9

Annealing Compared with Other Algorithms

While it is valuable to know about the time versus quality tradeoff of annealing itself, knowing whether annealing is a good algorithm to use depends on its relative performance compared with other algorithms.

The most significant set of experimental comparisons is that of [12], where for each of a number of problems annealing was compared against the best algorithm known for that problem. Generally speaking the specialized algorithms tailored to the particular problem at hand ran far faster than annealing, but annealing's results were often as good or better. Over all, if annealing was not always the best algorithm, it was at least competitive.

Our perspective is somewhat different. We are studying annealing in a fairly general context, and would therefore like to compare it with a similarly general algorithm. There are few candidates: *random search, descent*, and *steepest descent* are the only ones that come to mind. In all cases we consider the running time of the algorithm to be the number of states searched. Random search (or "random sampling") consists simply of generating states at random and keeping track of the one with the best energy so far. Descent is most easily described as annealing at temperature 0, so that uphill moves are never accepted. Steepest descent is similar, but instead of moving to *any* neighbor of lower or equal energy, the move is to a neighbor of *lowest* energy. Because descent and steepest descent may get stuck quickly, we will actually consider "repeated" variants of them, where after getting stuck the algorithm is restarted from a random initial state.

All these algorithms are easy to evaluate. In making a comparison with annealing on the fractal functions of Chapter 8 the real difficulty is in deciding on what makes for a fair comparison.

Specifically, a great deal of regularity has been built into the fractal energy function. While our analysis of annealing depends on this regularity, the algorithm itself does not.[1] A version of random sampling which was allowed to use the regularity explicitly could be extremely efficient: like confined annealing, such a sampling algorithm could locate a good point in S_1, then search only in the corresponding interval to find a good point in S_2, and so forth. Just as we did not propose confined annealing as a serious algorithm (it was merely a stepping stone to unconfined annealing), we do not allow this version of random sampling.

9.1 Random Sampling

We begin with analysis of random sampling because the descent algorithms can then be viewed in the same framework.

Theorem 9.1.1 *For a deterministic fractal f with c_0 zeros, if $t_{\text{rand}} = (b/c_0)^K$ points are randomly chosen from $[0,1]$, the relative expected energy of the lowest-energy point is*

$$\mathbf{E}_{\text{rand}}[f] \equiv \frac{\mathbf{E}[f]}{f_{\text{range}}} \gtrsim (1/e) r^K \Delta F (1-r). \tag{9.1}$$

Proof Choosing x uniformly at random in $[0,1]$ is equivalent to choosing each component s_k of its state code uniformly at random from $\{0,\ldots,b-1\}$. Then $f(x) < r^K \Delta F$ only if $F(s_1) = \cdots = F(s_K) = 0$; call such an x "K-good". If F has c_0 zeros, this happens with probability $(c_0/b)^K$. Consequently, if $(b/c_0)^K$ x's are chosen, the probability that none is K-good is $[1-(c_0/b^K)]^{(b/c_0)^K} \sim 1/e$, in which case the cost is at least $r^K \Delta F$. Thus for run time $t_{\text{rand}} = (b/c_0)^K$, the absolute expected cost is $\mathbf{E}[f] \gtrsim (1/e) r^K \Delta F$. Recalling that the range of the deterministic fractal is $f_{\text{range}} = 1/(1-r)$ gives the relative expected cost stated. ■

For the same t_{rand} it is possible to derive a similar upper bound for the expected energy, proving that $\mathbf{E}_{\text{rand}}[f] = \Theta\left(r^K\right)$. The proof comes from looking at k-goodness for all values $k \leq K$.

[1] This is not entirely true. In constructing the state spaces S_i we are using knowledge of b, and in reducing the temperature we are using r. But these objections do not apply to the very similar process of annealing on linearly separable functions.

This energy-time relationship is similar in form to that for unconfined annealing, as given by Theorem 8.4.9. But in the case of unconfined annealing, $\ln t_{\mathrm{uncon}} / \ln(1/\mathbf{E}_{\mathrm{uncon}}) \lesssim 2/\Delta F$; in this case $\ln t_{\mathrm{rand}} / \ln(1/\mathbf{E}_{\mathrm{rand}}) \sim \ln(b/c_0)/\ln(1/r)$.

Thus for both annealing and random sampling, run time is roughly polynomial in inverse expected energy. Unfortunately, the exponents for the two cases are not comparable, as they depend on different parameters. However, a much more dramatic contrast exists for an interesting generalization of the class of fractals we have been discussing so far, and we take this up in Section 9.2.

Meanwhile it is interesting to note that even though annealing and random sampling both require run time which is power-law in inverse expected energy, the reasons are quite different. We know that if $f(x)$ is small then the early components of its code $\langle s_1, s_2, \ldots \rangle$ must all be zeros of F. Consider the search through x's as a search through codes.

For simplicity suppose F has a unique zero. In expected time only b, random sampling finds an x with $F(s_1) = 0$. But it takes time b^2 to find an x where both s_1 and s_2 are zeros of F, and in general requires time b^K to find an x with $\langle s_1, \ldots, s_K \rangle$ all zeros of F. In short, the later "generations" of random sampling take much longer than the early ones.

But for annealing, s_1 is a random variable. Getting a small expected energy requires making s_1 a zero of F with high probability, and making that probability less than ε takes time something like $(1/\varepsilon)^{2/\Delta F}$. The same probability is adequate for the other s_i's, which is to say that the later generations of annealing take the same amount of time.

9.2 Multidimensional Fractals

For a given value d and one-dimensional fractal f satisfying Definition 8.0.1, define a d-dimensional fractal $\vec{f} : [0,1]^d \to \mathbb{R}$ by $\vec{f}(\langle x_1, \ldots, x_d \rangle) = \sum_{i=1}^{d} f(x_i)$. Annealing on $\vec{f}$ is defined as an extension of unconfined annealing on f: Define a sequence of state spaces $\vec{S}_k = {S_k}^d$. Moves in $\vec{S}_k$ can be defined as follows: given two vertices $\vec{x} = \langle x_1, \ldots, x_d \rangle$ and $\vec{x}' = \langle x'_1, \ldots, x'_d \rangle$ in $\vec{S}_k$, they are connected by an edge if they differ in only a single component (say the ith), and if x_i and x'_i are connected by an edge in S_k.

In other words, annealing on $\vec{f}$ is like annealing independently on d copies of f and adding the results together. At each time step, the copy on which a move is to be made (the value "i" of the previous paragraph) is selected at random. In fact, these functions are a special case of the linearly separable functions considered in Chapter 7.

Compared with annealing on the fractal f, spending d times as long in each generation yields a final result of the same relative expected energy:

Theorem 9.2.1 *Let a value ε and d-dimensional fractal $\vec{f}$ based on a deterministic fractal f with energy scale parameter r be given. Let $\hat{T} = \hat{T}(\varepsilon)$ and $\hat{t} = \hat{t}(\hat{T})$, and assume that ε is sufficiently small that $\hat{T} \leq T_{\text{crit}}$. Let $t = \Upsilon(d\hat{t}, 1/d, \varepsilon)$ (Υ is defined in 7.2.2 and its use illustrated by Theorem 7.2.5). Apply unconfined annealing with cooling schedule $\{(T_k, t)\}_{k=0}^{K}$, where $(T_k, t_k) = (r^{k-1}\hat{T}, t)$ and $K = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil$. This algorithm returns a state whose relative expected energy is $\vec{\mathbf{E}}_{\text{uncon}} \leq 4\varepsilon$ and it consumes run time $\vec{t}_{\text{uncon}} \sim d \cdot t_{\text{uncon}}$.*

Proof That $\vec{t}_{\text{uncon}} \sim d \cdot t_{\text{uncon}}$ follows from Proposition 7.2.4; this property was exploited in the analysis of linearly separable functions (Theorem 7.2.5). By definition of Υ, with probability at least $1-\varepsilon$, for any i, at least $\hat{t}$ moves are made on f_i. Conditional upon this, by Theorem 8.4.9, $\mathbf{E}[f_i] \leq 3\varepsilon f_{\text{range}}$. Taking into account the possibility that fewer than $\hat{t}$ moves are made, $\mathbf{E}[f_i] \leq (1-\varepsilon)(3\varepsilon f_{\text{range}}) + (\varepsilon)(f_{\text{range}}) \leq 4\varepsilon f_{\text{range}}$, giving total expected energy $\mathbf{E}[\vec{f}] \leq d \cdot 4\varepsilon f_{\text{range}}$. Since the range of $\vec{f}$ is $d \cdot f_{\text{range}}$, $\vec{\mathbf{E}}_{\text{uncon}} \leq 4\varepsilon$. ■

Noting that $\vec{t}_{\text{uncon}} = d \cdot t_{\text{uncon}}$ and $\vec{\mathbf{E}}_{\text{uncon}} = \mathbf{E}_{\text{uncon}}$,

$$\frac{\ln \vec{t}_{\text{uncon}}}{\ln(1/\vec{\mathbf{E}}_{\text{uncon}})} = \frac{\ln d + \ln t_{\text{uncon}}}{\ln(1/\mathbf{E}_{\text{uncon}})} \sim \frac{\ln t_{\text{uncon}}}{\ln(1/\mathbf{E}_{\text{uncon}})}. \tag{9.2}$$

It follows that $\vec{t}_{\text{uncon}}$ and $\vec{\mathbf{E}}_{\text{uncon}}$ have the same asymptotic power-law relationship as did t_{uncon} and $\mathbf{E}_{\text{uncon}}$, *i.e.* $\vec{t}_{\text{uncon}}$ is roughly $(1/\vec{\mathbf{E}}_{\text{uncon}})^{2/\Delta F}$.

What about the performance of random sampling on the same d-dimensional problem?

Theorem 9.2.2 *In time $\vec{t}_{\text{rand}} = (b/c_0)^{Kd}$, random sampling on a d-dimensional fractal produces a solution whose relative expected energy is $\vec{\mathbf{E}}_{\text{rand}} \gtrsim (1/e)r^K\Delta F(1-r)/d$.*

Proof Consider searching for a point of absolute energy no more than $r^K\Delta F$. For a point $\vec{x} = \langle x_1, \ldots, x_d \rangle$, if $\vec{f}(\vec{x}) \leq \varepsilon$ then certainly each $f(x_i) \leq \varepsilon$. In particular, if $\vec{x}$ is "K-good" in the same sense as before (having energy no more than $r^K\Delta F$), then each x_i must be K-good. A single x_i is good only if the first K components of its code are all zeros of F, which has probability $(c_0/b)^K$. Then a random $\vec{x}$ is good only if each x_i is good, which has probability $(c_0/b)^{K\cdot d}$. Sampling $(b/c_0)^{Kd}$ points, the probability that none is good is asymptotically $1/e$, in which case the energy is at least $r^K \Delta F$. Therefore the relative energy is at least $(1/e) \cdot r^K \Delta F \cdot (1-r)/d$. ■

It follows that for d-dimensional random sampling,

$$\frac{\ln \vec{t}_{\mathrm{rand}}}{\ln(1/\vec{\mathbf{E}}_{\mathrm{rand}})} \sim \frac{K d \ln(b/c_0)}{K \ln(1/r)} = \frac{d \ln(b/c_0)}{\ln(1/r)}, \tag{9.3}$$

which is d times what it is for the one-dimensional case. Thus time is once again power-law in energy, but the exponent is increased d-fold.

These comparative results for annealing and random sampling in d-dimensional cases can be explained in an intuitive manner. For annealing, the dimensions can be treated separately, leading to a factor d increase in both time and (absolute) energy. Another way to view this case is that when the number of dimensions increases, the conductance of the underlying graph does not change significantly, and it is the conductance that dominates how long annealing takes.

For random sampling the situation is entirely different. Suppose for convenience that F has a unique zero. To produce a good solution, random sampling must happen upon the best of b^K bins. In the d-dimensional case, there is still only one good bin, but the total number of bins is $(b^K)^d$, requiring time b^{Kd}.

The upshot is that annealing is relatively insensitive to the dimensionality, while random sampling is very sensitive to it. In particular, for both algorithms the run time and quality are characterized by the ratio $\ln t / \ln(1/\mathbf{E}[f])$. For annealing this ratio is $2/\Delta F$ regardless of the dimension d, while for random sampling it is $d \ln(c_0/b)/\ln(1/r)$. Even if for $d = 1$ random sampling is asymptotically more efficient (*i.e.* its ratio is smaller), as d grows the ratio for random sampling must become larger than that for annealing. For high-dimensional problems, annealing will be more efficient.

9.3 Descent Algorithms

It might appear that in comparing annealing to random sampling, we have pitted it against a straw man: even the name "random sampling" does not conjure up images of a powerful algorithm. But as we said, there are not many algorithms as general as annealing to which we can compare it, and we will now argue that in general, neither repeated descent nor repeated steepest descent is significantly better than random sampling. In fact, both behave quite similarly to random sampling.

Like annealing (but unlike random search) the descent algorithms depend on a move set: we move from a state x to a neighboring state x' of lower energy. For annealing, the move set we used varied with the generation number. If we are looking for solutions of expected energy $r^K \Delta F$, the natural choices of move set for descent algorithms are: (1) always move on S_K, or (2) move on S_1 until stuck (until all neighboring states have higher energy), then on S_2, and so on.

A quick glance at Figure 8.1 should make it clear that move set (1) is doomed: the energy function is jagged, and any descent algorithm cannot go long before being trapped in a local minimum. While we do not present a rigorous analysis of this case, it can be performed in the same manner as the analysis of move set (2) which follows now.

Theorem 9.3.1 *Let c_R be the number of states in S_1 which are "zero-reachable" – reachable from a zero of F by uphill moves only. Assume $c_R \neq b$. If repeated descent or repeated steepest descent is applied for $t_{\text{desc}} = (b/c_R)^K$ restarts, implying run time (number of samples) at least this large, the relative expected energy is $\mathbf{E}_{\text{desc}} \gtrsim (1/e) r^K \Delta F(1-r)$.*

Proof For the final state to have energy less than $r^K \Delta F$, each component s_k $(k \leq K)$ must be a zero of F. Thus each component must end up at a zero, having started at a random value in $\{0, \ldots, b-1\}$ and gone through a sequence of downhill moves. For this to be possible, the initial value of s_k must be reachable by a sequence of uphill moves from a zero of F; the number of such values was defined as c_R.

The probability that the random initial value of a single s_k is zero-reachable is c_R/b, so the probability that all K of them are zero-reachable is $(c_R/b)^K$. Since this is a necessary condition for the descent algorithm to reach a K-good state, the probability that a run of the algorithm reaches a K-good state is no more than $(c_R/b)^K$. Thus, repeated descent run $(b/c_R)^K$ times fails to reach a K-good state with probability $[1-(c_R/b)^K]^{(b/c_R)^K} \sim 1/e$. ■

In short, except for the improvement of the constant c_0 to c_R, the performance of descent algorithms is no better than that of random sampling. In particular, Theorem 9.2.1 (which describes the performance of random sampling on d-dimensional fractals) also holds for descent algorithms, with c_0 replaced by c_R.

Chapter 10

Conclusions

10.1 Summary and Overview

In the Introduction we argued that, despite its proven success on a range of practical problems, annealing cannot be particularly efficient for arbitrary problems. Specifically, annealing's behavior depends on the energies of the states and on the edges connecting these states. Regardless of the edge structure, if the energies are assigned randomly, annealing can do no better than random sampling. Similarly, for any energy assignment, if the edges are those of the complete graph, annealing is worse than random sampling. Thus, if annealing is to work well the energy function and move set must be well-matched, in some undetermined sense. We gave an intuitive argument that "fractalness" (self-similarity of the energy function with respect to the move set) was the required property.

In Chapter 3 we gave experimental evidence of the fractalness of the energy landscapes of a number of realistic problems. Complementary experiments, to see if problems on which annealing performs poorly have landscapes which fail these fractalness tests, have not yet been performed.

In Chapter 4 we presented a "random walk on a graph" model of simulated annealing. It is well-known that a (time-reversible) Markov chain can be modeled as a random walk on an (undirected) graph, but to the best of our knowledge this technique has not previously been applied to simulated annealing. The underlying graph for annealing corresponds to the move set, and inevitably has a natural structure.

The real power of the graph model comes from a recent theorem of Jerrum and Sinclair (and extensions by Mihail) which relate the "mixing" time of the Markov chain (the time required to approach stationarity) to the "conductance" of the underlying graph. Because of their regular structures, it is often possible to calculate a useful bound on the conductance of the underlying graphs for annealing.

Even before using the conductance quantitatively, the definition is helpful from an intuitive standpoint. It is natural to think of annealing as being essentially a descent algorithm, with the added power to escape from energy "valleys": most previous studies of annealing, including those proving its asymptotic properties, have focused on such valleys. The usual definition of an energy valley is a set of points the escape from which requires crossing some energy threshold. While this definition is sufficient for asymptotic arguments, it is not intuitively satisfying. For example, even if there was a single low-energy path leading out of a set of points, for all practical purposes it would still be an isolated energy valley: the probability of happening upon just the right path is negligible. In statistical physics there is a definition of "free energy", which combines the notions of energy and the number of elements having that energy. While statistical physics is sometimes invoked in the study of simulated annealing, apart from some arguments by analogy the free energy concept does not seem applicable.

What we really have in mind by an energy valley is a set of states from which it takes a long time to escape, either because there are few edges out of it or because those edges have high energy (low probability). The conductance of the partitioning of a graph measures precisely the summed probability of these edges, and the Jerrum and Sinclair result relates it to the escape time from the corresponding "valley". The conductance of the graph itself is that corresponding to the valley from which it is hardest to escape, which is the valley of interest for annealing. In short, conductance leads to a definition of energy valleys that seems to conform precisely to our intuition.

In annealing, the temperature is gradually lowered as time goes on. The asymptotic success of annealing requires convergence of the actual state probability vector to the stationary vector at the current temperature. When the temperature is lowered, the stationary vector changes, potentially getting farther from the actual vector. But when an annealing move is made, the actual vector gets closer to the stationary vector by an amount which can be bounded in terms of the conductance. Annealing asymptotically converges to a global minimum, for temperature sequences with the property that the movement towards stationarity exceeds the movement away from it. Quick "back of the envelope" estimates of the conductance and the change in the stationary vectors lead to the logarithmic cooling schedule well known to give asymptotic convergence; Appendix 4.6 presents the detailed calculations to derive that result rigorously.

It is simpler to analyze the case where a fixed temperature is used, and Chapter 4.7 showed that in the general case, achieving a low final expected energy with constant-temperature annealing takes running time which is very long – the governing parameters are the total number of states and the inverse of the minimum nonzero energy, both of which are huge.

In Chapters 5 and 6 we extended these basic results, respectively to annealing schedules where the temperature is lowered after some desirable distribution has already been reached, and to functions whose energy range is not known.

In Chapter 7 these tools were applied to "linearly separable" functions, and it was shown that these multi-variable functions can be minimized in time comparable to that for single-variable functions. In this sense annealing is an efficient algorithm for high-dimensional linearly separable functions.

In Chapter 8 we introduced a class of tame fractals. On these, the arguments of Chapter 4.7 were applied in a hierarchical manner, again showing annealing to be a comparatively efficient algorithm.

In all cases – the general case of Chapter 4, the separable functions of Chapter 7, or the fractals of Chapter 8, the cooling schedules we construct require run time which is a power of the solution quality desired. The constants in this relationship make annealing feasible in the special cases but not in the general case. In particular, the power does not change, and the constants do not become much worse, for separable functions with more variables.

Chapter 9 showed the same constancy for the efficiency of annealing on a higher-dimensional version of the deterministic fractals. In the same chapter it was shown that for random search, descent, and steepest descent time also increases as a power of solution quality, but that as the number of dimensions increases the power increases linearly, making these algorithms inefficient for high-dimensional problems. While not explicitly considered, the run time for the sampling and descent algorithms on separable functions parallels that on multi-dimensional fractals, so the same conclusion applies: these algorithms are inferior to annealing for linearly separable functions of many variables.

In conclusion, we have presented a new framework for simulated annealing, in two senses. First, we have considered restricted problem domains (linearly separable functions and a class of fractals) for which annealing is efficient, which it is not for arbitrary problems. Second, we have modeled annealing as a random walk on a graph and used knowledge about the conductance of these underlying graphs, to provide both a useful conceptual framework and powerful quantitative methods. The formal results are so far confined to quite simple models, but even this is well beyond what has been done before.

10.2 Future Goals

Of course, we would like to extend our theoretical analysis to where it comes closer to describing real problems. The framework in which we have set annealing is extremely general, and we can imagine generalizing the approaches of Chapters 7 and 8 to graphs and energy functions with weaker properties. We would presumably partition the graph G into two pieces giving minimum conductance, partition each of those, and so forth. Then we could follow the same approach as before, annealing to get into the better top-level cluster, lowering the temperature to get into the better sub-cluster, and so on. However, in this more general setting, substantial (if not insurmountable) new difficulties arise for each step of the analysis.

Some natural cases to try first would be extending the fractal model to fractional Brownian motions, and extending the linearly separable function model to functions where there are interactions between the variables but the interactions are either weak or rare. Another candidate would be a graph partitioning or Ising model problem, where it might be possible to directly analyze the temperature-dependent conductance of the graph representing the full state space and the subgraphs representing the valleys in which one may be "stuck" at lower temperatures.

In addition to these goals in the theoretical domain, experimental work remains as well. The same landscape analysis experiments performed on problems where annealing works well should be performed on some where annealing works poorly (or where simple descent methods work better) and the results compared. Simple simulations of unconfined annealing on the regular fractals should be performed: the bounds dictated by theory may be overly conservative, or the simulations may simply suggest something new. Simulations with relaxed rules – say changing the temperature by a factor other than the energy scale parameter of the fractal, and making moves of random lengths – should also be performed to see if the algorithm is robust. Experimental comparisons of annealing on the deterministic fractals with annealing on fractional Brownian motions could indicate which properties carry over, and could be used to guide the theoretical analysis.

Finally, the important matter of constructing efficient cooling schedules is now poorly understood in practice as well as being an open theoretical problem. It might be possible to address this issue by applying to actual problems the techniques developed for the simple models, even though these techniques will no longer be formally justified. For example, an N-object circuit placement problem resembles the linearly separable function model of Chapter 7, albeit with occasional interactions between the variables. The similarity suggests checking for a polynomial relationship between solution quality and run time for annealing on this problem; if such a relationship is found it could be useful for specifying in advance the run time and the cooling schedule required to get a solution of given quality. Also, while for simplicity we have only considered geometric cooling schedules, the energy bound we derive as a function of run time could be improved by spending less time in the later generations, where the energy variations are smaller. Exploiting this observation could lead to a new class of more efficient, hyper-geometric, cooling schedules.

Appendix A

Asymptotic Convergence to Global Minima

As mentioned in Section 4.6, while we do not feel that analysis of the logarithmic cooling is the most useful view of annealing, duplication of those standard results using the tools we rely upon seems worthwhile both as a vindication of the method and as independent derivation of the result.

We are given an annealing problem, where without loss of generality we assume the energy f is scaled to have minimum 0 and maximum 1. Let $a < 1$ be any constant near 1.

We define the metric $||\cdot||_2$ and show that for the cooling schedule $T_t = 1/a \ln t$, for any P_0, $||P_t - \pi_{T_t}||_2 \to 0$. It follows that for any state u, $P_t(u) \to \pi_0(u)$; this also implies convergence of the expected energy to 0.

Definition A.0.1 *For a probability vector π_T on state space V, with π_T nowhere 0,*

$$||P_t - \pi_T||_2 = \sum_{u \in V} \frac{[P_t(u) - \pi_T(u)]^2}{\pi_T(u)}. \tag{A.1}$$

This may be thought of as a squared L_2 norm, with axes rescaled by factors $\pi_T(u)$. This norm allows the following theorem.

Theorem A.0.2 *Let a strongly aperiodic Markov chain with stationary distribution π and underlying graph of conductance Φ be given. Let the probability distribution P_{t+1} be that after a single step, starting from distribution P_t. Then*

$$||P_{t+1} - \pi||_2 \leq (1 - \Phi)||P_t - \pi||_2. \tag{A.2}$$

This version of Sinclair and Jerrum's result [26] is given by Mihail [18].

To show that $d_t = ||P_t - \pi_t||_2 \to 0$, we will bound $||P_{t+1} - \pi_{t+1}||_2$ in terms of $||P_t - \pi_t||_2$.

Lemma A.0.3 *Let M_{t+1} be an upper bound for $|\pi_t(u) - \pi_{t+1}(u)| \,/\, \pi_{t+1}(u)$. Then*

$$d_{t+1} \quad \leq \quad (1 - \Phi(T_{t+1}))[(1 + M_{t+1})\, d_t + 3n \cdot M_{t+1}]. \tag{A.3}$$

Proof From Theorem A.0.2,

$$||P_{t+1} - \pi_{t+1}||_2 \leq (1 - \Phi(T_{t+1}))||P_t - \pi_{t+1}||_2. \tag{A.4}$$

The second factor may be expanded:

$$\begin{aligned}
||P_t - \pi_{t+1}||_2 &= \sum_{u \in V} \frac{[P_t(u) - \pi_{t+1}(u)]^2}{\pi_{t+1}(u)} \qquad \text{(now drop the } u\text{'s for clarity)} && \text{(A.5)} \\
&= \sum \frac{[(P_t - \pi_t) + (\pi_t - \pi_{t+1})]^2}{\pi_{t+1}} && \text{(A.6)} \\
&\leq \sum \frac{(P_t - \pi_t)^2}{\pi_t} \frac{\pi_t}{\pi_{t+1}} + 2 \sum |P_t - \pi_t| \frac{|\pi_t - \pi_{t+1}|}{\pi_{t+1}} && \text{(A.7)} \\
&\quad + \sum |\pi_t - \pi_{t+1}| \frac{|\pi_t - \pi_{t+1}|}{\pi_{t+1}}.
\end{aligned}$$

As differences of probabilities, $|\pi_{t+1}(u) - \pi_t(u)|$ and $|P_t(u) - \pi_t(u)|$ are no more than 1, so

$$\begin{aligned}
||P_t - \pi_{t+1}||_2 &\leq \sum \frac{(P_t - \pi_t)^2}{\pi_t}(1 + M_{t+1}) + 2 \sum 1 \cdot M_{t+1} + \sum 1 \cdot M_{t+1} && \text{(A.8)} \\
&= (1 + M_{t+1}) \cdot ||P_t - \pi_t||_2 + 3n \cdot M_{t+1}. && \text{(A.9)}
\end{aligned}$$

Combining inequalities (A.4) and (A.9) yields the lemma. ■

It remains to construct the bound M_{t+1} and a bound for $\Phi(T)$. To do so we now assume a cooling schedule of the form $T_t = 1/a \ln t$ for some constant a. Equivalently, the inverse temperature is $\beta_t = a \ln t$.

For Φ we simply apply Theorem 4.5.2: annealing at inverse temperature β on a problem with n states,

$$\Phi(\beta) \geq \frac{1}{n^2} e^{-\beta}, \tag{A.10}$$

which for the cooling schedule $\beta_t = a \ln t$ means

$$1 - \Phi(\beta_{t+1}) \geq 1 - \frac{1}{n^2} t + 1^{-a} \sim 1 - \frac{1}{n^2} t^{-a}. \tag{A.11}$$

Lemma A.0.4 *For any $c_2 > a$,*

$$M_{t+1} = c_2 t^{-1}. \tag{A.12}$$

is an upper bound for $|\pi_t(u)/\pi_{t+1}(u) - 1|$.

Proof Express $\pi_t(u)$ as

$$\pi_t(u) = \deg(u) e^{-\beta_t f(u)} / Z(\beta_t), \tag{A.13}$$

for

$$\frac{\pi_t(u)}{\pi_{\beta_{t+1}}(u)} = \frac{\deg(u) e^{-\beta_t f(u)}}{\deg(u) e^{-\beta_{t+1} f(u)}} \cdot \frac{Z(\beta_{t+1})}{Z(\beta_t)}. \tag{A.14}$$

Beginning with the first term,

$$\frac{\deg(u)e^{-\beta_t f(u)}}{\deg(u)e^{-\beta_{t+1} f(u)}} = e^{(\beta_{t+1}-\beta_t)f(u)} \tag{A.15}$$

$$= \left(\frac{t+1}{t}\right)^{af(u)} \tag{A.16}$$

$$\leq (1+1/t)^a. \tag{A.17}$$

For t large, to first order this is $1+a/t$ so for any constant $c_1 > a$ (we will also make $c_1 < c_2$) and t sufficiently large,

$$\frac{\deg(u)e^{-\beta_t f(u)}}{\deg(u)e^{-\beta_{t+1} f(u)}} < c_1 t^{-1}. \tag{A.18}$$

To bound the second term, sort the values of f in increasing order, so $0 = f_0 < f_1 < \cdots < f_l = 1$. For each index $i \leq l$, let $k_i = \sum_{u:f(u)=f_i} \deg(u)$. Then

$$\frac{Z(\beta_{t+1})}{Z(\beta_t)} = \frac{k_0 + k_1 e^{-f_1\beta_{t+1}} + \cdots + k_l e^{-\beta_{t+1}}}{k_0 + k_1 e^{-f_1\beta_t} + \cdots + k_l e^{-\beta_t}} \tag{A.19}$$

For t, and therefore β_t, sufficiently large, the term $k_1 e^{-f_1\beta_{t+1}}$ dominates all other $k_i e^{-f_i\beta_{t+1}}$ in the numerator; similarly $k_1 e^{-f_1\beta_t}$ dominates the denominator. Both tend to 0, so

$$\frac{Z(\beta_{t+1})}{Z(\beta_t)} - 1 \sim \frac{k_1}{k_0}e^{-f_1\beta_{t+1}} - \frac{k_1}{k_0}e^{-f_1\beta_t}. \tag{A.20}$$

Substituting $\beta_t = a\ln t$,

$$\frac{Z(\beta_{t+1})}{Z(\beta_t)} - 1 \sim \frac{k_1}{k_0}\left[(t+1)^{-af_1} - (t)^{-af_1}\right] \tag{A.21}$$

$$\sim \frac{k_1}{k_0}t^{-1-af_1}. \tag{A.22}$$

It follows from (A.18) and (A.22) that

$$\left|\frac{\pi_t(u)}{\pi_{t+1}(u)} - 1\right| \lesssim \left|1 + c_1 t^{-1}\right| \left|1 + \frac{k_1}{k_0}t^{-1-af_1}\right| - 1 \sim c_1 t^{-1}. \tag{A.23}$$

Since $c_2 > c_1$, for t sufficiently large

$$\left|\frac{\pi_t(u)}{\pi_{t+1}(u)} - 1\right| < c_2 t^{-1}. \tag{A.24}$$

■

Lemma A.0.5 *For some constant c_3, for all t sufficiently large, if $d_t \leq c_3(t)^{a-1}$ then $d_{t+1} \leq c_3(t+1)^{a-1}$.*

Proof Substituting (A.11) and (A.12) into (A.3),

$$d_{t+1} \lesssim \left(1 - \frac{1}{n^2}t^{-a}\right)\left[\left(1 + c_2 t^{-1}\right) d_t + 3nc_2 t^{-1}\right] \tag{A.25}$$

By the mean value theorem, there is some $t' \in [t, t+1]$ such that

$$(t+1)^{a-1} - (t)^{a-1} = \frac{d}{dt}(t^{a-1})\bigg|_{t'} = (a-1)(t')^{a-2} \geq (a-1)(t)^{a-2}, \tag{A.26}$$

so

$$(t+1)^{a-1} \geq (t)^{a-1} + (a-1)(t)^{a-2}. \tag{A.27}$$

It is clear that the larger d_t is, the larger the bound on d_{t+1} given by (A.25). Therefore we can assume d_t is as large as possible, *i.e.* $d_t = c_3(t)^{a-1}$. Under this assumption,

$$\begin{aligned} d_{t+1} - d_t &\lesssim \left(-\frac{1}{n^2}t^{-a} + c_2 t^{-1} - \frac{1}{n^2}c_2 t^{-a-1}\right) c_3(t)^{a-1} && \text{(A.28)} \\ &\quad +3n\, c_2\, t^{-1} - \frac{1}{n^2} 3n\, c_2\, t^{-a-1} \\ &\sim -\frac{1}{n^2}t^{-a}\, c_3 t^{a-1} + 3nc_2 t^{-1} && \text{(A.29)} \\ &= \left(-\frac{c_3}{n^2} + 3nc_2\right) t^{-1}. && \text{(A.30)} \end{aligned}$$

Choose any $c_3 > 3n^3 c_2$, so that $-c_3/n^2 + 3nc_2 < 0$. Since t^{-1} dominates t^{a-2}, for t sufficiently large

$$\left(-\frac{c_3}{n^2} + 3nc_2\right) t^{-1} < c_3(a-1)t^{a-2}, \tag{A.31}$$

and it follows from (A.30) that

$$d_{t+1} < c_3(t)^{a-1} + c_3(a-1)t^{a-2}. \tag{A.32}$$

But (A.27) means

$$c_3(t+1)^{a-1} \geq c_3(t)^{a-1} + c_3(a-1)t^{a-2}, \tag{A.33}$$

so together these two imply that

$$d_{t+1} < c_3(t+1)^{a-1}. \tag{A.34}$$

■

While Lemma A.0.5 provides an inductive step on t, we have not shown that for an initial t, $d_t < c_3(t+1)^{a-1}$. But by inspection of (A.25), if $d_t < f(t)$ implies $d_{t+1} < f(t)$, then for any constant $c_4 > 1$, $d_t < c_4 f(t)$ implies $d_{t+1} < c_4 f(t)$. Whatever initial value of t we wish to work with, simply choose c_4 large enough that $d_t < c_4 c_3(t+1)^{a-1}$ and then use the inductive step indicated above.

Pulling all this together, the proof that $d_t \to 0$ is as follows. Choose any constants $0 < a < c_2 < 1$. Choose any $c_3 > 0$ such that $-c_3/n^2 + 3nc_2 < 0$. Choose a τ such that all the "for t sufficiently large" conditions (specifically the inequalities (A.25) and (A.31)) hold for any $t \geq \tau$. Finally choose a $c_4 \geq 1$ large enough that $d_\tau < c_4 c_3(\tau)^{a-1}$. Inductive application of (A.34) shows that $d_t < c_4 c_3(t)^{a-1}$ for all $t \geq \tau$. It follows that $d_t \rightarrow 0$.

We have now shown that for any positive constant a less than 1, letting $T_t = 1/a \ln t$, if π_t is the stationary distribution at T_t and P_t is the actual probability distribution after annealing for one move at each temperature $T_1, \ldots, T_t$, then regardless of the initial distribution P_0, $\|P_t - \pi_t\|_2 \rightarrow 0$ as $t \rightarrow \infty$.

Immediately, for any u, $[P_t(u) - \pi_t(u)]^2/\pi_t(u) \rightarrow 0$, whence $P_t(u) - \pi_t(u) \rightarrow 0$. Since $\pi_t(u) = \pi_{T_t}(u) \rightarrow \pi_0(u)$, the actual time-$t$ state distributions satisfy $P_t(u) \rightarrow \pi_0(u)$. In particular, the probability of being in some global minimum at time t approaches 1 as t approaches ∞, and the expected energy approaches 0.

Appendix B

Proofs for Chapter 3

Lemma 3.0.4 *Given any space where a random walk $X(t)$ satisfies $d(X(t), X(0)) = ct$, for some constant c. Let f be a fractal (per Definition 3.0.2) with parameter H on this space. Then $f(X(t))$ is a fBm (on $\mathbb{R}^1$) and has parameter H.*

Proof We need to show that $f(X(t_2)) - f(X(t_1))$ is normally distributed with variance proportional to $|t_2 - t_1|$:

$$\begin{aligned} f(X(t_2)) - f(X(t_1)) &\sim N(0, d(X(t_2), X(t_1))^{2H}) && \text{(B.1)} \\ &= N(0, c^{2H}|t_2 - t_1|^{2H}). && \text{(B.2)} \end{aligned}$$

■

Lemma 3.0.5 *Given a fBm f on $\mathbb{R}^n$ with parameter H. Make a random walk $X(t)$ on $\mathbb{R}^n$. Then $f(t) = f(X(t))$ satisfies equation (3.1) with parameter $\frac{1}{2}H$. Furthermore, for n large, $f(t)$ is approximately a fBm (on $\mathbb{R}^1$) with parameter $\frac{1}{2}H$.*

Proof Without loss of generality assume that $X(0) = \vec{0}$ and $f(X(0)) = 0$. Write $X(t) = \langle X_1(t), \ldots, X_n(t)\rangle$. The "random walk" here consists of taking t/δ^2 steps of size $\pm\delta$ on randomly-chosen coordinates X_i, in the limit $\delta \to 0$. Any X_i may be thought of as a sum of t/δ^2 independent random variables, which take the values 0 (if the step is to a coordinate other than the ith), $+\delta$ (if the step is in the positive direction on the ith coordinate), or $-\delta$ (if the step is in the negative direction); these values occur with probabilities $1 - 1/n$, $1/2n$, and $1/2n$ respectively. As the sum of a large number (t/δ) of i.i.d. random variables X_i is normally distributed with mean 0 and variance t/n.

Let N_i be the number of steps composing $X_i(t)$. Each N_i has distribution $B(t/\delta^2, 1/n)$. For $\delta \to 0$ this is almost exactly $(t/\delta^2)/n$, with probability approaching 1. Since the N_i are almost deterministic, they are also almost independent. Each X_i is the sum of N_i random variables which take the values δ and $-\delta$ with probability 1/2. Then the X_i's are also almost independent, and have distributions $N(0, t/n)$.

Let $Z_i = X_i/\sqrt{t/n}$, so the Z_i are independent standard normal variables.

$$\begin{aligned} \|X(t)\|^2 &= \frac{t}{n}[Z_1(t)^2 + \cdots + Z_n(t)^2] && \text{(B.3)} \\ &\sim \frac{t}{n}\chi_n^2, && \text{(B.4)} \end{aligned}$$

where χ_n^2 denotes a chi-square distribution with n degrees of freedom.

Immediately,

$$\begin{aligned} \mathbf{E}[f(X(t))^2] &= \mathbf{E}[\|X(t)\|^{2H}] && \text{(B.5)} \\ &= \mathbf{E}[(\frac{t}{n}\chi_n^2)^H] && \text{(B.6)} \\ &\propto t^H, && \text{(B.7)} \end{aligned}$$

confirming equation (3.1).

Also, this chi-square distribution has mean n and variance $2n$, so for n large it approaches a point mass at n. In that case, with probability approaching 1 as $n \to \infty$, $\|X(t)\|$ is almost exactly $(t/n) \cdot n = t$, and $\|X(t)\|^{2H}$ is almost exactly t^H. It follows that $f(X(t))^2$, whose exact distribution is $N(0, \|X(t)\|^{2H})$, is almost exactly distributed as $N(0, t^H)$. It is in this sense that for n large, $f(X(t))$ is a fBm on $\mathbb{R}^1$ with parameter H. ■

Appendix C

Proofs for Chapter 4

Lemma 4.2.2 *Let an annealing problem be given by the undirected unweighted graph G_A and the energy function f on its vertices. Then the underlying edge-weighted graph G corresponding to annealing on G_A at temperature T has the same structure (vertices and edges) as G_A, with the addition of self-loops at each vertex. It has edge weights given by*

$$w(v,u) = e^{-\max(f(v),f(u))/T} \tag{C.1}$$

for edges $\{v,u\}$ present in G_A, and

$$w(v,v) = e^{-f(v)/T} \sum_{f(u)>f(v)} [1 - e^{-(f(u)-f(v))/T}] \tag{C.2}$$

for the added self-loops $\{v,v\}$, with the sum taken over pairs $\{u,v\}$ which are edges of G.

Corollary 4.2.4 *At temperature T, the stationary probability $\pi_T(v)$ of state v is* $\deg(v)e^{-f(v)/T}/Z(T)$.

Proof (Lemma 4.2.2 and Corollary 4.2.4) The corollary follows from application of Lemma 4.2.1 to the edge weights defined by the lemma in equations (4.3) and (4.4):

$$\begin{aligned}
\pi(v) &\propto \sum_u w(v,u) && \text{(C.3)}\\
&= f(v,v) + \sum_{f(u)>f(v)} w(v,u) + \sum_{f(u)\leq f(v)} w(v,u) && \text{(C.4)}\\
&= \sum_{f(u)>f(v)} e^{-f(v)/T}[1 - e^{-(f(u)-f(v))/T}] + \sum_{f(u)>f(v)} e^{-f(u)/T} && \text{(C.5)}\\
&\quad + \sum_{f(u)\leq f(v)} e^{-f(v)/T} &&\\
&= \sum_{f(u)>f(v)} \Big[(e^{-f(v)/T} - e^{-f(u)/T}) + e^{-f(u)/T})\Big] + \left[\sum_{f(u)\leq f(v)} e^{-f(v)/T}\right] && \text{(C.6)}\\
&= \sum_u e^{-f(v)/T} && \text{(C.7)}\\
&= \deg(v)e^{-f(v)/T}. && \text{(C.8)}
\end{aligned}$$

(All sums are restricted to edges (v,u) in G_A.) Since the $\pi(v)$ must sum to 1, the normalizing constant is clearly $1/Z(T)$.

To prove Lemma 4.2.2 we apply equation (4.2). In this case, for edges (v,u) in G_A,

$$\begin{aligned}
\mathrm{P}[v_{t+1}=v \mid v_t=u] &= \frac{w(u,v)}{\sum_{u'\in V} w(u,u')} && \text{(C.9)}\\
&= \frac{e^{-\max(f(v),f(u))/T}}{\deg(u)e^{-f(u)/T}} && \text{(C.10)}\\
&= \frac{1}{\deg(v)}\cdot\begin{cases} 1 & \text{if } f(v)\le f(u)\\ e^{-(f(v)-f(u))/T} & \text{if } f(v)>f(u)\end{cases}. && \text{(C.11)}
\end{aligned}$$

This is precisely the probability of generating v ($1/\deg(u)$) times the probability of accepting v (1 or exponentially small in the energy increase) specified for simulated annealing. Since the "accepted" move probabilities in the random walk model match those of annealing, the probability of staying in u (walking the added self-loop on u) must be equal to the probability that annealing stays in u by rejecting a move. ■

Proposition 4.3.1 *$Z(T)=\sum_v \deg(v)e^{-f(v)/T}$ is monotonically increasing. For f ranging from 0 to 1 and $T\ge 0$, $Z(T)\ge 1$, and for regular graphs of degree d, $Z(T)\ge d$.*

Proof Each summand of $Z(T)$ is increasing. For the second part,

$$\sum_v \deg(v)e^{-f(v)/T} \ge \sum_{v:f(v)=0} \deg(v), \tag{C.12}$$

which is always at least 1, and is at least d if the graph is regular. ■

Theorem 4.3.2 *$\pi_T(v)$ is a bitonic function of T: there is a value $T_{\text{crit}}(v)$ (the* **critical temperature for** v*) such that $\pi_T(v)$ increases with increasing T for $T<T_{\text{crit}}(v)$ and decreases with increasing T for $T>T_{\text{crit}}(v)$. Further, $T_{\text{crit}}(v)$ is the value of temperature at which $f(v)=\mathbf{E}_{\pi_T}[f]$, i.e. at which the expected energy equals the energy of v.*

Proof

$$\begin{aligned}
\frac{d\pi_\beta(v)}{d\beta} &= \frac{d}{d\beta}\frac{1}{Z(\beta)}\deg(v)e^{-\beta f(v)} && \text{(C.13)}\\
&= -\frac{Z'}{Z^2}\cdot\deg(v)e^{-\beta f(v)}+\frac{1}{Z}\deg(v)e^{-\beta f(v)}(-f(v)) && \text{(C.14)}\\
&= \frac{Z\mathbf{E}_{\pi_\beta}[f]}{Z^2}\cdot Z\pi_\beta(v)-\pi_\beta(v)f(v) && \text{(C.15)}\\
&= \left[\mathbf{E}_{\pi_\beta}[f]-f(v)\right]\pi_\beta(v). && \text{(C.16)}
\end{aligned}$$

Note that $\pi_\beta(v)>0$ and that $\frac{d\pi}{dT}$ and $\frac{d\pi}{d\beta}$ have opposite sign, so the sign of $\frac{d\pi}{dT}$ is the same as that of $f(v)-\mathbf{E}_{\pi_T}[f]$. Clearly $\mathbf{E}_{\pi_T}[f]$ is monotonically increasing in T (this is formally proved later as Corollary 5.1.11). Thus for $T<T_{\text{crit}}(v)$, $\frac{d\pi}{dT}$ is positive and $\pi_T(v)$ is an increasing function of T; similarly for $T<T_{\text{crit}}(v)$, $\pi_T(v)$ is an increasing function of T. ■

Proposition 4.3.4 *For f ranging from 0 to 1, and $T>0$, $\pi_T(v)>\pi_0(v)$ if and only if $f(v)>0$*

Proof For $f(v) > 0$ and $T > 0$, $\pi_T(v) = \deg(v)e^{-f(v)/T}/Z > 0$, while $\pi_0(v) = 0$.

For $f(v) = 0$, $\pi_T(v) = \deg(v)/Z(T)$. This is decreasing with T since Z is increasing with T. ■

Theorem 4.3.6 *For any finite annealing problem, $T_{\text{crit}} > 0$.*

Proof The critical temperature is simply the minimum of the critical temperatures of all states other then global minima. In our usual scaling where the minimum nonzero energy is Δf, T_{crit} is the critical temperature of a state v of energy Δf. ■

Theorem 4.5.1 *Let G be the graph whose vertex set V consists of permutations of the objects $1, \ldots, N$, and in which two vertices (permutations) are connected by an edge if some single pairwise exchange of two objects takes one permutation to the other. Then G has conductance $\geq 1/2N^2$.*

Proof Despite the fact that the proof is somewhat involved, we present it in full because we think it is of significant interest. The "canonical path" style of argument is most elegant, and is applicable to a variety of other graphs of interest.

Let $(S, \bar{S})$ be a partition defining the conductance. We will exhibit a set of $|S||\bar{S}|$ paths between S and $\bar{S}$ with the property that the paths are "almost independent" – *i.e.* any edge is used in at most 2 $|V|$ paths; this will act as a bound on the multiple-counting of edges when paths are counted. In particular, each path going from S to $\bar{S}$ must contain at least one cut edge, so

$$\begin{aligned} \text{\# cut path edges} &\geq \text{\# paths} \;/\; (2\;|V|) && \text{(C.17)} \\ &= |S||\bar{S}| \;/\; (2\;|V|). && \text{(C.18)} \end{aligned}$$

Since each vertex has degree $N(N-1)/2$,

$$2 \cdot \text{\# edges within } S \leq |S|\;[N(N-1)/2]. \qquad \text{(C.19)}$$

Then the conductance is

$$\begin{aligned} \Phi(G) &= \frac{\text{\# edges from } S \text{ to } \bar{S}}{2 \text{ \# edges within } S + \text{ \# edges from } S \text{ to } \bar{S}} && \text{(C.20)} \\ &= \frac{1}{1 + \frac{2 \text{ \# edges within } S}{\text{\# edges from } S \text{ to } \bar{S}}} && \text{(C.21)} \\ &\geq \frac{1}{1 + \frac{|S|\; N(N-1)/2}{|S||\bar{S}|/(2\;|V|)}} && \text{(C.22)} \\ &\geq \frac{1}{2N^2}, && \text{(C.23)} \end{aligned}$$

using the fact that $|\bar{S}| \geq |V|/2$ per the definition of conductance.

We now proceed to construct the set of paths described above. There will be one path defined between each permutation A in S and each permutation B in $\bar{S}$. Defining the canonical path from any permutation A to any B requires some review of permutations.

Let $\sigma = (\sigma_1, \ldots, \sigma_N)$ be a permutation of the numbers $1, \ldots, N$. Another way to express σ is as a set of cycles,

$$\sigma = (\sigma_1^1, \ldots, \sigma_{k_1}^1) \cdots (\sigma_1^m, \ldots, \sigma_{k_m}^m), \qquad \text{(C.24)}$$

meaning that if we begin with the identity permutation $(1, \ldots, N)$ and move σ^1_1 to position σ^1_2, σ^1_2 to position σ^1_3, *etc.*, completing the first cycle by moving $\sigma^1_{k_1}$ to position σ^1_1, and then repeat this process for all the other cycles, the result is the permutation σ. For instance, the permutation $(2, 5, 4, 3, 1)$ is expressible as $(1, 5, 2)(3, 4)$. The cyclic form is unique except for rotation of the elements within a cycle and changes in the order of the cycles. We will assume a canonical cyclic form where each cycle begins with its smallest value and the cycles themselves are sorted by their first (smallest) values. The example just given is canonical because 1 is the smallest and first element of one cycle, 3 is the smallest and first element of the other, and the cycle containing 1 precedes that containing 3.

Now let σ be the permutation mapping A into B. For example if $A = (2, 5, 4, 3, 1)$ and $B = (4, 3, 1, 5, 2)$, to map A to B we must move A's 1 to slot 3, which is occupied by A's 4; A's 4 must move to slot 1, which is occupied by A's 2; and A's 2 moves to the original position of A's 1. Thus the first cycle in σ is $(1, 4, 2)$. Similarly A's 3 must move to the position occupied by A's 5 and vice-versa, so the second cycle in σ is $(3, 5)$, and $\sigma = (1, 4, 2)(3, 5)$. We will write $B = \sigma(A)$.

Let the points A_j of the canonical path be generated by applying in turn the cycles of σ, creating each cycle from pairwise interchanges. We demonstrate with the same A and B (and therefore the same $\sigma = (1, 4, 2)(3, 5)$) as above. After each step, the objects shown in boldface are those which were just exchanged.

$$\begin{aligned} A = A_0 &= (2, 5, 4, 3, 1) && \text{(C.25)} \\ A_1 &= (2, 5, \mathbf{1}, 3, \mathbf{4}) && \text{(C.26)} \\ A_2 &= (\mathbf{4}, 5, 1, 3, \mathbf{2}) \text{ (completing cycle 1)} && \text{(C.27)} \\ B = A_3 &= (4, \mathbf{3}, 1, \mathbf{5}, 2) \text{ (completing cycle 2).} && \text{(C.28)} \end{aligned}$$

That is, A is turned into B by applying the cycles of σ in turn, where a cycle of length k is executed as $k - 1$ pairwise interchanges of successive pairs of elements in the cycle.

Given the vertex A_j (but *not* the value of j itself), define $B_j = \sigma(A_j)$. Then if we are given A_j and the extra information B_j, we can determine σ, and therefore the sequence of object interchanges used in the canonical path from A to B. If in addition we are given the extra information A_{j+1}, the next vertex along the path from A to B, the difference between A_j and A_{j+1} tells us which pair of objects was just swapped, allowing determination of where we are in the sequence of moves, and thus j itself. By reversing the moves already performed we can recover A, and by performing the remaining moves we can construct B. In short, from the directed edge (A_j, A_{j+1}) and extra information B_j we can determine A, B, and the canonical path between them.

Suppose we are only given the edge $\{A_j, A_{j+1}\}$. How many canonical paths could it belong to? There are 2 possible orientations of the edge, and $|V|$ possible values of extra information B_j (B_j is just a permutation – *i.e.* a vertex of G), and each choice determines a full canonical path, including A and B themselves. Thus each edge belongs to no more than $2|V|$ canonical paths. ■

Lemma 4.7.1 *For a function f on V ranging from $f_{\min}$ to $f_{\max}$ with $f_{\text{range}} = f_{\max} - f_{\min}$; an arbitrary distribution P_t on V; and π_T and π_0 the stationary distributions at temperatures T and 0:*

$$\mathbf{E}_{P_t}[f(v)] \leq \mathbf{E}_{\pi_T}[f(v)] + \|P_t - \pi_T\|_{\text{tvd}} f_{\text{range}}. \tag{C.29}$$

Proof For any two distributions P and P' on V,

$$\begin{aligned}
\mathbf{E}_{P'}[f(v)] &= f_{\min} + \sum_{v\in V} P'(v)(f(v) - f_{\min}) \qquad (C.30)\\
&= f_{\min} + \sum_{v\in V}[P'(v) - P(v)](f(v) - f_{\min}) + \sum_{v\in V} P(v)(f(v) - f_{\min}) \qquad (C.31)\\
&\leq f_{\min} + \sum_{v:P'(v)>P(v)} [P'(v) - P(v)]f_{\mathrm{range}} + \mathbf{E}_P[f(v) - f_{\min}] \qquad (C.32)\\
&= \mathbf{E}_P[f] + \|P' - P\|_{\mathrm{tvd}} f_{\mathrm{range}}. \qquad (C.33)
\end{aligned}$$

■

Lemma 4.7.3 *Given Δ, let $T \leq \hat{T}(\varepsilon, \Delta, n)$. If $\Delta \leq \varepsilon$ or $\Delta \leq \min\{f(v) : f(v) > \varepsilon\}$ then*

$$\mathbf{E}_{\pi_T}[f] \leq 2\varepsilon, \tag{C.34}$$

and if $\Delta \leq \Delta f$ then

$$\mathbf{E}_{\pi_T}[f] \leq \|\pi_T - \pi_0\|_{\mathrm{tvd}} \leq \varepsilon. \tag{C.35}$$

Proof Begin with the case $\Delta \leq \Delta f$, so $T \leq \Delta f / \ln(n^2/\varepsilon)$. If $f(v) > 0$, then $f(v) \geq \Delta f$, and

$$\begin{aligned}
\pi_T(v) &= \frac{1}{Z}\deg(v)e^{-f(v)/T} \qquad (C.36)\\
&\leq 1 \cdot n \cdot e^{-\Delta f \cdot \ln(n^2/\varepsilon)/\Delta f} \qquad (C.37)\\
&\leq \varepsilon/n \qquad (C.38)
\end{aligned}$$

(the inequality $1/Z \leq 1$ is from Proposition 4.3.1).

By Proposition 4.3.4, $\pi_T(v) > \pi_0(v)$ if and only if $f(v) > 0$, so

$$\begin{aligned}
\|\pi_T - \pi_0\|_{\mathrm{tvd}} &= \sum_{v:f(v)>0} [\pi_T(v) - \pi_0(v)] \qquad (C.39)\\
&\leq n \cdot [\varepsilon/n - 0] \qquad (C.40)\\
&= \varepsilon. \qquad (C.41)
\end{aligned}$$

And by Lemma 4.7.1,

$$\mathbf{E}_{\pi_T}[f] \leq \mathbf{E}_{\pi_0}[f] + \|\pi_T - \pi_0\|_{\mathrm{tvd}} = \|\pi_T - \pi_0\|_{\mathrm{tvd}}, \tag{C.42}$$

concluding the proof for this first case.

If G is regular then Proposition 4.3.1 gives $Z \geq \deg$, and for states v of nonzero energy we again have

$$\pi_T(v) \leq \frac{1}{Z}\deg(v)e^{-\Delta f \cdot \ln(n/\varepsilon)/\Delta f} \leq \varepsilon/n, \tag{C.43}$$

leading to the same conclusions regarding $\|\pi_T - \pi_0\|_{\mathrm{tvd}}$ and $\mathbf{E}_{\pi_T}[f]$.

Now consider the other given values of Δ, all of which are less than or equal to $\min\{f(v) : f(v) > \varepsilon\}$. If $f(v) \geq \varepsilon$ then $f(v) \geq \min\{f(v) : f(v) > \varepsilon\} \geq \Delta$. So for states with $f(v) \geq \varepsilon$, and at temperature $T \leq \Delta / \ln(n^2/\varepsilon)$,

$$\begin{aligned} \pi_T(v) &= \frac{1}{Z}\deg(v)e^{-f(v)/T} && \text{(C.44)} \\ &\leq 1 \cdot n \cdot e^{-\Delta \cdot \ln(n^2/\varepsilon)/\Delta} && \text{(C.45)} \\ &\leq \varepsilon/n. && \text{(C.46)} \end{aligned}$$

Consequently

$$\begin{aligned} \mathbf{E}_{\pi_{\dot{T}}}[f] &= \sum_{v:f(v)\leq\varepsilon} \pi_T(v)f(v) + \sum_{v:f(v)>\varepsilon} \pi_T(v)f(v) && \text{(C.47)} \\ &\leq \sum_{v:f(v)\leq\varepsilon} \pi_T(v)\cdot\varepsilon + \sum_{v:f(v)>\varepsilon} (\varepsilon/n)\cdot 1 && \text{(C.48)} \\ &\leq 2\varepsilon. && \text{(C.49)} \end{aligned}$$

If G is regular then

$$\pi_T(v) \leq \frac{1}{Z}\deg(v)e^{-\Delta\cdot\ln(n/\varepsilon)/\Delta} \leq \varepsilon/n, \tag{C.50}$$

again supporting inequalities (C.48) and (C.49). ■

Lemma 4.7.5 *If $t \geq \hat{t}(T, \varepsilon, n)$ then beginning from any distribution P_0 and annealing at temperature T for time t, the final distribution satisfies*

$$\|P_t - \pi_T\|_{\mathrm{tvd}} \leq \varepsilon. \tag{C.51}$$

Proof By Corollary 4.2.4, the probability $\pi_{\min}$ of the least likely state satisfies $\pi_{\min} \geq \frac{1}{n^2}e^{-1/T}$. Theorem 4.5.2 states $\Phi(T) \geq e^{-1/T}\Phi(\infty)$. By Proposition 4.4.7 and Theorem 4.4.9,

$$\begin{aligned} \|P_t - \pi_T\|_{\mathrm{tvd}} &\leq \frac{1}{\pi_{\min}}\left(1 - \Phi(T)^2/2\right)^t && \text{(C.52)} \\ &\leq (n^2 e^{1/T})e^{-t\Phi(T)^2/2}, && \text{(C.53)} \end{aligned}$$

or

$$\ln\|P_t - \pi_T\|_{\mathrm{tvd}} \leq \ln(n^2) + \frac{1}{T} - \frac{t}{2}e^{-2/T}\Phi(\infty)^2. \tag{C.54}$$

Then to make $\|P_t - \pi_T\|_{\mathrm{tvd}} \leq \varepsilon$ it suffices to use

$$t \geq 2\left(\ln(n^2/\varepsilon) + \frac{1}{T}\right)\frac{1}{\Phi(\infty)^2}e^{2/T}. \tag{C.55}$$

For regular graphs, Corollary 4.2.6 gives $\pi_{\min} \geq \frac{1}{n}e^{-1/T}$, giving an n rather than n^2 in the final expression as well. ■

Remark 4.7.8 *For a system in equilibrium at temperature T, raising the energy of a given state may lower the expected energy of the system.*
Example Consider a system with 3 states, of energies 0, 1, and x. In equilibrium at temperature T, letting $\beta = 1/T$, by Boltzmann's law the expected energy is

$$\mathbf{E} = \frac{xe^{-x\beta} + e^{-\beta}}{1 + e^{-x\beta} + e^{-\beta}}. \tag{C.56}$$

Differentiation with respect to x shows an extremal energy at $\bar{x}$ such that

$$e^{-\bar{x}\beta} + (e^{-\beta} + 1)(-\bar{x}\beta + 1) + \beta e^{-\beta} = 0. \tag{C.57}$$

To take an example, $T = 1/2$ yields $\bar{x} \approx .722934887$ and $\mathbf{E} \approx .222934887$. Larger values of x (as well as smaller ones) give smaller expected energy.

This example is particularly good since we are often interested in the temperature relative to the range of energies of the states rather than in T itself. In this case the value of x minimizing $\mathbf{E}$ fell between 0 and 1, and so did not affect the energy range. We find that with $\bar{x} = 0$ equation (C.57) has no positive solutions for β, implying that $\bar{x}$ is always greater than 0. Substituting $\bar{x} = 1$ into (C.57) yields $2e^{-\beta} - \beta + 1 = 0$, or $\beta \approx 1.463055514$. Thus $\bar{x}$ is less than 1 for β larger than this, or equivalently for temperatures less than about .683501064.

The computation is even simpler if we take a system with just 2 states, of energies 0 and x, at temperature $T = 1$. In this case the expected energy is $\mathbf{E} = \frac{xe^{-x/T}}{1+e^{-x/T}}$. Note that $\mathbf{E}[f] \to 0$ for $x \to \infty$ and for $x \to 0$. The maximum of $\mathbf{E}$ occurs when $0 = (e^x + 1) - xe^x$, which has numerical solution $x \approx 1.278464543$. Since the energy range in this case is x itself, the behavior of the "relative temperature", T/x, is not immediately evident. ■

Appendix D

Proofs for Chapter 5

Remark 5.0.1 *Begin in equilibrium at T_0. Apply the schedule $T_1, T_2, \ldots, T_n$, where each $T_i \leq T_0$. Then the probability distribution following this schedule may give* less *weight to the global minimum than it had in the initial equilibrium distribution: in fact, may give it arbitrarily small weight.*

Example ("Probability Pump") The probability of the global minimum can be made arbitrarily close to 0 independent of the initial distribution and of the upper bound on the T_i. Consider the 4-state example of Figure D.1. Apply a schedule of the form

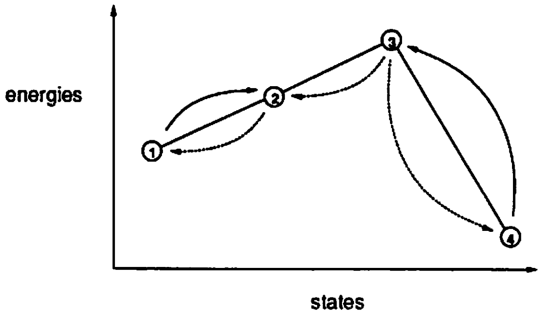


Figure D.1: Small but non-constant temperature does not imply nearness to small-temperature equilibrium

$0, \ldots, 0, T, 0, \ldots, 0, T, \ldots$: the number of zeros separating two T's will be large but constant. After each application of temperatures $0, \ldots, 0$, virtually all the probability mass is in states 1 and 4 (dashed arrows in figure). In particular, after one step starting from state 3, we are either in state 2 or state 4. From state 2 the move to state 3 is never accepted. Thus the probability of remaining in state 2 after c attempted moves is precisely $1/2^c$ – the probability that all attempted moves happen to be to state 3 rather than state 1. So the probability of being in either state 2 or 3 after $c + 1$ moves is less than $1/2^c$ times the probability of beginning there.

When temperature T is applied (solid arrows), some fixed fraction of the probability mass on state 4 ($e^{-(f_3-f_4)/T}$ of it) is transferred to state 3. When the $c+1$ temperature-0 iterations are made, half the transferred amount reverts to state 4, but the rest (but for the fraction $1/2^c$) shifts to state 2 and then to state 1. Thus with each application of $T, 0, \ldots, 0$, a fixed fraction of the probability on state 4 (at least $e^{-(f_3-f_4)/T}[1 - 1/2^c]$) is transferred to state 1, and the probability of state 4, the global minimum, decays to 0 exponentially!
■

This indicates that monotonicity of the cooling schedule is essential to its tractability.

Lemma 5.1.3 *$H(P) \geq 0$, with equality iff$P \equiv \pi$.*

Proof

$$\begin{aligned} H(P) &= \sum \pi(v)\varphi\left(\frac{P(v)}{\pi(v)}\right) && \text{(D.1)} \\ &\geq \varphi\left(\sum \pi(v)\frac{P(v)}{\pi(v)}\right) && \text{(D.2)} \\ &= \varphi(1) = 0, && \text{(D.3)} \end{aligned}$$

where the inequality follows from convexity of φ. Since π is strictly positive and φ is strictly convex, equality holds iff$P(v)/\pi(v)$ is constant. ■

Theorem 5.1.4 *Let M be the transition matrix for a Markov chain on V. Suppose π is a strictly positive measure on V and is stationary for M, i.e. $\pi M = \pi$ or equivalently, for all u, $\sum_v \pi(v)M(v,u) = \pi(u)$. Then for any distribution P on V,*

$$H(P \cdot M) \leq H(P). \tag{D.4}$$

Proof

$$\begin{aligned} H(P \cdot M) &= \sum_u \pi(u)\varphi\left(\frac{1}{\pi(u)}\sum_v P(v)M(v,u)\right) && \text{(D.5)} \\ &= \sum_u \pi(u)\varphi\left(\sum_v \frac{P(v)}{\pi(v)}\frac{\pi(v)M(v,u)}{\pi(u)}\right) && \text{(D.6)} \\ &\leq \sum_u \pi(u)\sum_v \varphi\left(\frac{P(v)}{\pi(v)}\right)\frac{\pi(v)M(v,u)}{\pi(u)} && \text{(D.7)} \\ &= \sum_v \pi(v)\varphi\left(\frac{P(v)}{\pi(v)}\right) && \text{(D.8)} \\ &= H(P), && \text{(D.9)} \end{aligned}$$

where the inequality follows from the convexity of φ and the fact that $\sum_v \frac{\pi(v)M(v,u)}{\pi(u)} = 1$.
■

Lemma 5.1.8 *The entropy of P relative to π_β may be written*

$$H(P, \pi_\beta) = -S(P) - L(P) + \ln Z(\beta) + \beta F(P) \tag{D.10}$$

where the entropy $S(P)$ and partition function $Z(\beta)$ are per definitions 5.1.6 and 4.2.3, and we define $F(P) = \mathbf{E}_P[f(v)]$, and $L(P) = \mathbf{E}_P[\ln \deg(v)]$.

Proof

$$\begin{aligned} H(P, \pi_T) &= \sum_v P(v) \ln \left(\frac{P(v)}{\pi_T(v)} \right) && \text{(D.11)} \\ &= \sum_v \{ P(v) \ln P(v) - [\ln \deg(v) - \ln Z(T) - \beta f(v)] \} && \text{(D.12)} \\ &= -S(P) - L(P) + \ln Z(\beta) + \beta F(P). && \text{(D.13)} \end{aligned}$$

■

Lemma 5.1.9 *The derivative of the partition function with respect to inverse temperature β is*

$$\frac{dZ(\beta)}{d\beta} = -Z(\beta) \mathbf{E}_{\pi_\beta}[f]. \qquad \text{(D.14)}$$

Proof

$$\begin{aligned} \frac{dZ}{d\beta} &= \frac{d}{d\beta} \left(\sum \deg(v) e^{-\beta f(v)} \right) && \text{(D.15)} \\ &= \sum \deg(v) e^{-\beta f(v)} (-f(v)) && \text{(D.16)} \\ &= -Z(\beta) \sum \frac{1}{Z(\beta)} \deg(v) e^{-\beta f(v)} f(v) && \text{(D.17)} \\ &= = -Z(\beta) \mathbf{E}_{\pi_\beta}[f]. && \text{(D.18)} \end{aligned}$$

We have shown every step in the proof because the same tricks are repeated throughout this chapter. ■

Lemma 5.1.10 *The derivative of the expected energy in equilibrium at inverse temperature β is*

$$\frac{dF(\pi_\beta)}{d\beta} = -\mathrm{Var}_{\pi_\beta}[f]. \qquad \text{(D.19)}$$

Proof

$$\begin{aligned} \frac{dF(\pi_\beta)}{d\beta} &= \frac{d}{d\beta} \sum \frac{1}{Z(\beta)} \deg(v) f(v) e^{-\beta f(v)} && \text{(D.20)} \\ &= \sum \frac{1}{Z(\beta)} \deg(v) f(v) (-f(v)) e^{-\beta f(v)} && \text{(D.21)} \\ &\quad + \sum \frac{-Z'(\beta)}{Z(\beta)^2} \deg(v) f(v) e^{-\beta f(v)}. \end{aligned}$$

Substituting for $Z'(\beta)$ from Lemma 5.1.9,

$$\begin{aligned} \frac{dF(\pi_\beta)}{d\beta} &= -\mathbf{E}_{\pi_\beta}[f(v)^2] + \frac{Z(\beta)\mathbf{E}_{\pi_\beta}[f]}{Z(\beta)^2} \cdot Z(\beta) \mathbf{E}_{\pi_\beta}[f] && \text{(D.22)} \\ &= -\mathbf{E}_{\pi_\beta}[f(v)^2] + \mathbf{E}_{\pi_\beta}[f(v)]^2 && \text{(D.23)} \\ &= -\mathrm{Var}_{\pi_\beta}[f]. && \text{(D.24)} \end{aligned}$$

■

Lemma 5.1.12 *The derivative of the Gibbs entropy plus the expected log of the degree in equilibrium at inverse temperature β is*

$$\frac{d(S+L)(\pi_\beta)}{d\beta} = -\beta \mathrm{Var}_{\pi_\beta}[f]. \tag{D.25}$$

Proof From Definition 5.1.2 it is evident that the entropy of any distribution relative to itself is 0. Using Lemma 5.1.8 to expand $H(\pi_\beta, \pi_\beta)$,

$$(S+L)(\pi_\beta) = \ln Z(\beta) + \beta F(\pi_\beta). \tag{D.26}$$

The derivative is

$$\frac{d}{d\beta}(S+L)(\pi_\beta) = \frac{Z'(\beta)}{Z(\beta)} + \beta F'(\pi_\beta) + F(\pi_\beta). \tag{D.27}$$

Substituting for $Z'(\beta)$ and $F'(\beta)$ from Lemmas 5.1.9 and 5.1.10 respectively,

$$\begin{aligned} \frac{d}{d\beta}(S+L)(\pi_\beta) &= \frac{-Z(\beta)F(\pi_\beta)}{Z(\beta)} + \beta\left(-\mathrm{Var}_{\pi_\beta}[f]\right) + F(\pi_\beta) && \text{(D.28)} \\ &= -\beta \mathrm{Var}_{\pi_\beta}[f]. && \text{(D.29)} \end{aligned}$$

■

Lemma 5.1.14 *If $F(P) = F(\pi_\beta)$ then $(S+L)(P) \leq (S+L)(\pi_\beta)$.*

Proof By Lemmas 5.1.3 and 5.1.8 respectively,

$$0 \leq H(P, \pi_\beta) = -(S+L)(P) + \ln Z(\beta) + \beta F(P), \tag{D.30}$$

and by the same reasoning

$$0 = H(\pi_\beta, \pi_\beta) = -(S+L)(\pi_\beta) + \ln Z(\beta) + \beta F(\pi_\beta). \tag{D.31}$$

Since $F(P) = F(\pi_\beta)$, $(S+L)(P) \leq (S+L)(\pi_\beta)$. ■

Lemma 5.1.15 *Let $P_0 = \pi(\beta_0)$ with $\beta_0 > 0$. For any distribution P, if*

$$(S+L)(P) - (S+L)(P_0) \geq 0 \tag{D.32}$$

then

$$\frac{1}{\beta_0}[(S+L)(P) - (S+L)(P_0)] < F(P) - F(P_0). \tag{D.33}$$

Proof Let β_1 be such that $F(\pi_{\beta_1}) = F(P)$. (Such a β_1 exists - and is unique - since $F(\beta)$ is continuous and is monotonically decreasing from $f_{\max}$ to $f_{\min}$ as β goes from $-\infty$ to $+\infty$.) From Lemma 5.1.14, $(S+L)(P) \leq (S+L)(\beta_1)$ so

$$(S+L)(P) - (S+L)(P_0) \leq (S+L)(\pi_{\beta_1}) - (S+L)(\pi_{\beta_0}). \tag{D.34}$$

By the hypothesis of the lemma $(S+L)(P)-(S+L)(P_0)>0$. With (D.34) this implies $(S+L)(\pi_{\beta_1})>(S+L)(\pi_{\beta_0})$. Since $\beta_0>0$ and (by Corollary 5.1.13) $(S+L)(\pi_\beta)$ is decreasing for $\beta>0$, if it were the case that $\beta_1\geq\beta_0$ then $(S+L)(\pi_{\beta_1})\leq(S+L)(\pi_{\beta_0})$, which would be a contradiction. It follows that $\beta_1<\beta_0$.

Now, by Lemmas 5.1.12 and 5.1.10,

$$\begin{aligned}\frac{d}{d\beta}\left(\frac{1}{\beta_0}(S+L)(\pi_\beta)-F(\pi_\beta)\right) &= \frac{1}{\beta_0}\left(-\beta\mathrm{Var}_{\pi_\beta}[f]\right)-\left(-\mathrm{Var}_{\pi_\beta}[f]\right) &\text{(D.35)}\\ &= (1-\frac{\beta}{\beta_0})\mathrm{Var}_{\pi_\beta}[f], &\text{(D.36)}\end{aligned}$$

which is strictly positive for $\beta<\beta_0$. Since $\beta_1<\beta_0$ it follows that

$$\frac{1}{\beta_0}(S+L)(\pi_{\beta_1})-F(\pi_{\beta_1})<\frac{1}{\beta_0}(S+L)(\pi_{\beta_0})-F(\pi_{\beta_0}). \qquad \text{(D.37)}$$

From (D.34 and rearrangement of the inequality above,

$$\begin{aligned}\frac{1}{\beta_0}(S+L)(P)-(S+L)(P_0) &\leq \frac{1}{\beta_0}(S+L)(\pi_{\beta_1})-(S+L)(\pi_{\beta_0}) &\text{(D.38)}\\ &< F(\pi_{\beta_1})-F(\pi_{\beta_0}). &\text{(D.39)}\end{aligned}$$

■

Theorem 5.1.16 *Beginning from the distribution $P_0=\pi_{T_0}$, anneal with cooling schedule $T_1,T_2,\dots$ where $T_0\geq T_1\geq T_2\cdots$. If the intermediate distributions are $P_1,P_2,\dots$, then at any time t, $\mathbf{E}_{P_t}[f]\leq\mathbf{E}_{P_0}[f]$.*

Proof From a distribution P_t we make a move at temperature T_{t+1} to yield distribution P_{t+1}. Let

$$H_t=\sum_v P_t(v)\ln\left(\frac{P_t(v)}{\pi_{t+1}(v)}\right) \qquad \text{(D.40)}$$

and

$$H_t'=\sum_v P_{t+1}(v)\ln\left(\frac{P_{t+1}(v)}{\pi_{t+1}(v)}\right). \qquad \text{(D.41)}$$

By Lemma 5.1.4 we know that

$$H_t'-H_t\leq 0, \qquad \text{(D.42)}$$

and by Lemma 5.1.8,

$$H(P,\pi_T)=-(S+L)(P)+\ln Z(T)+\frac{1}{T}F(P). \qquad \text{(D.43)}$$

So,

$$\begin{aligned}0 &\geq H_t'-H_t &\text{(D.44)}\\ &= \frac{1}{T_{t+1}}[F(P_{t+1})-F(P_t)]-[(S+L)(P_{t+1})-(S+L)(P_t)] &\text{(D.45)}\\ &\quad +[Z(T_{t+1})-Z(T_t)]\end{aligned}$$

which implies

$$F(P_{t+1}) - F(P_t) \leq T_{t+1}[(S+L)(P_{t+1}) - (S+L)(P_t)]. \tag{D.46}$$

Now define $a_\tau = (S+L)(P_\tau)$ so

$$F(P_{t+1}) - F(P_t) \leq T_{t+1}(a_{\tau+1} - a_\tau). \tag{D.47}$$

Summing from $\tau = 0$ to $t-1$,

$$F(P_t) - F(P_0) \leq \sum_{\tau=0}^{t-1} T_{\tau+1}(a_{\tau+1} - a_\tau). \tag{D.48}$$

But

$$\begin{aligned} & \sum_{\tau=0}^{t-1} T_{\tau+1}(a_{\tau+1} - a_\tau) \\ &= T_1(a_1 - a_0) + T_2(a_2 - a_1) + \cdots + T_{t-1}(a_{t-1} - a_{t-2}) + T_t(a_t - a_{t-1}) && \text{(D.49)} \\ &= -T_1 a_0 + (T_1 - T_2)a_1 + \cdots + (T_{t-1} - T_t)a_{t-1} + T_t a_t && \text{(D.50)} \\ &= T_1 \left(\sum_{\tau=1}^{t-1} \lambda_\tau a_\tau - a_0 \right) && \text{(D.51)} \end{aligned}$$

where $\lambda_t = T_t / T_1$, and for all $\tau \neq t$, $\lambda_\tau = (T_\tau - T_{\tau+1})/T_1$. Note that since the T_t's are decreasing and non-negative, all $\lambda_\tau \geq 0$. Also, $\sum \lambda_\tau = 1$. That is, the λ_τ's yield a convex combination of the a_τ's, and

$$\begin{aligned} \sum_{\tau=0}^{t-1} T_{\tau+1}(a_{\tau+1} - a_\tau) &= T_1 \left(\sum_{\tau=1}^{t-1} \lambda_\tau a_\tau - a_0 \right) && \text{(D.52)} \\ &\leq T_1 \left(\max_{\tau=1}^{t-1} a_\tau - a_0 \right). && \text{(D.53)} \end{aligned}$$

Substituting inequality (D.47) back in for the left side of (D.53), for all t,

$$F(P_{t+1}) - F(P_t) \leq T_1 \left(\max_{\tau=1}^{t-1} a_\tau - a_0 \right). \tag{D.54}$$

Letting

$$k(t) = \arg \max_{\tau=1}^{t-1} a_\tau, \tag{D.55}$$

for all t

$$F(P_{t+1}) - F(P_t) \leq T_1 \left(a_{k(t)} - a_0 \right). \tag{D.56}$$

We wish to show that $F(P_t) - F(P_0) \leq 0$, so it suffices to prove that $a_{k(t)} - a_0 \leq 0$.

Note that $k(k(t)) = k(t)$, since maximality of a_k in $\{a_1, \ldots, a_t\}$ implies maximality of a_k in the subset $\{a_1, \ldots, a_{k(t)}\}$. Thus as a special case of inequality (D.56),

$$F(P_{k(t)}) - F(P_0) \leq T_1(a_{k(k(t))} - a_0) = T_1(a_{k(t)} - a_0). \tag{D.57}$$

We need only show that $a_{k(t)} - a_0 \leq 0$.

Suppose $a_{k(t)} - a_0 \geq 0$. Then

$$T_1(a_{k(t)} - a_0) \leq T_0(a_{k(t)} - a_0) \leq F(P_{k(t)}) - F(P_0) \tag{D.58}$$

where the second inequality is from Lemma 5.1.15. But inequalities (D.58) and (D.57) are in contradiction, so $a_{k(t)} - a_0$ must be strictly less than 0. ■

Corollary 5.1.17 (monotonic cooling) *Begin from a distribution* $P_0 : \|P_0 - \pi_{T_0}\|_{\text{tvd}} \leq \varepsilon$ *and anneal with cooling schedule* $T_1, T_2, \ldots$ *where* $T_0 \geq T_1 \geq T_2 \cdots$. *Then at any time* t, $\mathbf{E}_{P_t}[f] \leq \mathbf{E}_{\pi_{T_0}}[f] + \varepsilon$.

Proof Let $\delta^+(v) = \max\{0, (P_0 - \pi_{T_0})(v)\}$, *i.e.* $\delta^+(v)$ is the vector of excesses of P_0 over π_{T_0}. Similarly let $\delta^-(v) = \max\{0, (\pi_{T_0} - P_0)(v)\}$, the vector of excesses of π_{T_0} over P_0. $\|P_0 - \pi_{T_0}\|_{\text{tvd}} \leq \varepsilon$ implies that the summed entries of δ^+ and δ^- are each no more than ε, so they may be written as $\delta^+ = \varepsilon'\Delta^+$ and $\delta^- = \varepsilon'\Delta^-$ where Δ^+ and Δ^- are probability vectors and $\varepsilon' \leq \varepsilon$.

So $P_0 = \pi_{T_0} + \varepsilon'\Delta^+ - \varepsilon'\Delta^-$. Because annealing is a Markov process, with $M^t = M_1 \cdots M_t$,

$$P_t = P_0 M^t = [\pi_{T_0} + \varepsilon'\Delta^+ - \varepsilon'\Delta^-]M^t \tag{D.59}$$

and

$$\begin{aligned}
\mathbf{E}_{P_t}[f] &= \sum_v f(v)\left[(\pi_{T_0}M^t)(v) + \varepsilon'(\Delta^+ M^t)(v) - \varepsilon'(\Delta^- M^t)(v)\right] && \text{(D.60)}\\
&= \mathbf{E}_{\pi_{T_0}}[f] + \varepsilon'\mathbf{E}_{\Delta^+ M^t}[f] - \varepsilon'\mathbf{E}_{\Delta^- M^t}[f] && \text{(D.61)}\\
&\leq \mathbf{E}_{\pi_{T_0}}[f] + \varepsilon' \cdot 1 - \varepsilon' \cdot 0 && \text{(D.62)}\\
&\leq \mathbf{E}_{\pi_{T_0}}[f] + \varepsilon && \text{(D.63)}
\end{aligned}$$

■

Corollary 5.1.19 (monotonic superschedule) *Let the temperature* T *and the cooling schedule* $\{T(\tau)\}_{\tau=1}^{t}$ *be such that for any* P_0, *the final distribution* P_t *satisfies* $\|P_t - \pi_{T_t}\|_{\text{tvd}} \leq \varepsilon$. *If* $T(\tau)$ *is a subschedule of a monotonically nonincreasing schedule* $\{T'(\tau)\}_{\tau=1}^{t'}$, *then* $\mathbf{E}_{P_{t'}}[f] \leq \mathbf{E}_{\pi_{T_{t'}}}[f] + \varepsilon$.

Proof Let the offset of the schedules be c. Whatever the distribution P_c, the distribution P_{t+c} satisfies $\|P_{t+c} - \pi_T\|_{\text{tvd}} \leq \varepsilon$. Taking this to be the initial distribution for Lemma 5.1.17, that lemma implies $\mathbf{E}_{P_{t'}}[f] \leq \mathbf{E}_{\pi_T}[f] + \varepsilon$. ■

Proposition 5.2.1 *For any* δ *there exists* $0 < T_\delta \leq T_{\text{crit}}$ *such that for all* $T \leq T_\delta$, $\|\pi_T - \pi_0\|_{\text{tvd}} \leq \delta$.

Proof By Corollary 4.3.3, for every $v \in V$, as T decreases from T_{crit} towards 0, $\pi_T(v)$ tends monotonically towards $\pi_0(v)$; that is, $|\pi_T(v) - \pi_0(v)|$ goes to 0 monotonically. It follows that

$$\|\pi_T - \pi_0\|_{\text{tvd}} = \frac{1}{2}\sum_{v \in V} |\pi_T(v) - \pi_0(v)| \to 0 \tag{D.64}$$

monotonically. The proposition follows from taking T_δ to be the largest temperature less than or equal to T_{crit} such that $\|\pi_{T_\delta} - \pi_0\|_{\text{tvd}} \leq \delta$. ■

Theorem 5.2.2 *Let the n-vector* P_t *be the state probability vector at time* t *of a time-variant Markov random process, starting at time 0 with state probability vector* P_0. *Let* M_t

be the transition matrix applied before time t ($t = 1, 2, \ldots$), so $P_t = P_{t-1}M_t$. Let π_t be the stationary distribution corresponding to M_t, so $\pi_t M_t = \pi_t$. Then for any distribution π_0,

$$\|P_t - \pi_t\|_{\text{tvd}} \le \|P_0 - \pi_0\|_{\text{tvd}} + \|\pi_0 - \pi_1\|_{\text{tvd}} + \cdots + \|\pi_{t-1} - \pi_t\|_{\text{tvd}}. \tag{D.65}$$

Proof Let $\zeta(w)$ represent an unknown vector the sum of the absolute values of whose elements is no more than $2w$. Note a few properties of this notation:

- If P and Q are probability vectors (positive elements summing to 1) then $\|P - Q\|_{\text{tvd}} \le w$ is equivalent to $P - Q = \zeta(w)$.
- The sum of two vectors of masses w_1 and w_2 is a vector whose mass is no more than $w_1 + w_2$. We write $\zeta(w_1) + \zeta(w_2) = \zeta(w_1 + w_2)$.
- When a transition matrix acts on a vector of mass w the result is a vector of mass no more than w. For a transition matrix M (positive elements, each row sums to 1), we write $\zeta(w) \cdot M = \zeta(w)$.

By induction on t we show that

$$P_t = \pi_t + \zeta(\|P_0 - \pi_0\|_{\text{tvd}}) + \sum_{i=0}^{t-1} \zeta(\|\pi_i - \pi_{i+1}\|_{\text{tvd}}), \tag{D.66}$$

from which the theorem statement follows immediately. The statement is trivially true for $t = 0$. Then,

$$\begin{aligned}
P_{t+1} &= P_t M_{t+1} && \text{(D.67)}\\
&= \left(\pi_t + \zeta(\|P_0 - \pi_0\|_{\text{tvd}}) + \sum_{i=0}^{t-1} \zeta(\|\pi_i - \pi_{i+1}\|_{\text{tvd}})\right) M_{t+1} && \text{(D.68)}\\
&= \left([\pi_{t+1} + (\pi_t - \pi_{t+1})] + \zeta(\|P_0 - \pi_0\|_{\text{tvd}}) + \sum_{i=0}^{t-1} \zeta(\|\pi_i - \pi_{i+1}\|_{\text{tvd}})\right) M_{t+1} && \text{(D.69)}\\
&= \left(\pi_{t+1} + \zeta(\|P_0 - \pi_0\|_{\text{tvd}}) + \sum_{i=0}^{t} \zeta(\|\pi_i - \pi_{i+1}\|_{\text{tvd}})\right) M_{t+1} && \text{(D.70)}\\
&= \pi_{t+1} + \zeta(\|P_0 - \pi_0\|_{\text{tvd}}) + \sum_{i=0}^{t} \zeta(\|\pi_i - \pi_{i+1}\|_{\text{tvd}}). && \text{(D.71)}
\end{aligned}$$

■

Corollary 5.2.3 *Let an annealing problem with critical temperature T_{crit} and a sequence of temperatures $\{T_t\}$ satisfying $T_{\text{crit}} \ge T_0 \ge T_1 \ge T_2 \ge \cdots$ be given. From an initial state probability vector P_0 anneal at temperatures $T_1, T_2, \ldots$. Then the distribution P_t at any time t satisfies*

$$\|P_t - \pi_{T_t}\|_{\text{tvd}} \le \|P_0 - \pi_{T_0}\|_{\text{tvd}} + \|\pi_{T_0} - \pi_{T_t}\|_{\text{tvd}} \tag{D.72}$$

and

$$\|P_t - \pi_0\|_{\text{tvd}} \le \|P_0 - \pi_{T_0}\|_{\text{tvd}} + \|\pi_{T_0} - \pi_0\|_{\text{tvd}}. \tag{D.73}$$

Proof Follows from theorem 5.2.2 and the fact that, because the π_{T_t}'s are monotonic in t,

$$\|\pi_{T_0} - \pi_{T_1}\|_{\text{tvd}} + \cdots + \|\pi_{T_{t-1}} - \pi_{T_t}\|_{\text{tvd}} = \|\pi_{T_0} - \pi_{T_t}\|_{\text{tvd}}. \tag{D.74}$$

The second statement of the corollary follows from the first:

$$\begin{aligned}
\|P_t - \pi_O\|_{\text{tvd}} &\leq \|P_t - \pi_{T_t}\|_{\text{tvd}} + \|\pi_{T_t} - \pi_O\|_{\text{tvd}} && \text{(D.75)}\\
&\leq \|P_0 - \pi_{T_0}\|_{\text{tvd}} + \|\pi_{T_0} - \pi_{T_t}\|_{\text{tvd}} + \|\pi_{T_t} - \pi_O\|_{\text{tvd}} && \text{(D.76)}\\
&= \|P_0 - \pi_{T_0}\|_{\text{tvd}} + \|\pi_{T_0} - \pi_O\|_{\text{tvd}}. && \text{(D.77)}
\end{aligned}$$

■

Appendix E

Proofs for Chapter 6

Lemma 6.1.2 *If (G, f) and (G', f') are similar with scale factor a, then annealing on G at temperature T is equivalent to annealing on G' at temperature aT. That is, if $v_0' = \sigma(v_0)$, the Markov chain v_t' defined by annealing on G' is identical to $\sigma(v_t)$ - the image in G' of the Markov chain defined by annealing on G.*

Proof Because G and G' are isomorphic, the probability of generating u from $v \in G$ is equal to the probability of generating $\sigma(u)$ from $\sigma(v) \in G'$. The energies are related by an affine transform, $f'(\sigma(u)) - f'(\sigma(v)) = a[f(u) - f(v)]$, so the probabilities of accepting these moves are also equal. Thus the two chains have equal transition probabilities, and from isomorphic initial states are identical Markov chains. ■

Lemma 6.2.1 *Let an annealing graph G with n vertices and energy function $f : G \to \mathbb{R}$ be given. If f ranges from exactly $f_{\min}$ to $f_{\max}$, let $f_{\mathrm{range}} = f_{\max} - f_{\min}$. Let a value $0 < \varepsilon < 1$ be given, as well as values $0 < r < 1$, $c > 0$, and $k \in \mathbb{Z}$ satisfying $cr^k \le f_{\mathrm{range}} \le cr^{k-1}$. Anneal at temperature $T = cr^k\hat{T}(\varepsilon) = cr^k\varepsilon/\ln(n^2/\varepsilon)$ for time*

$$t = \hat{t}(r\hat{T}) \tag{E.1}$$

$$= 2\left(\ln(n^2/\varepsilon) + \frac{1}{r\hat{T}}\right) n^4 e^{2/(r\hat{T})}. \tag{E.2}$$

Then $\mathbf{E}_{\pi_T}[f] \le f_{\min} + 2\varepsilon\, f_{\mathrm{range}}$, and regardless of the initial state probability vector P_0 the final distribution P_t satisfies $\|P_t - \pi_T\|_{\mathrm{tvd}} \le \varepsilon$. It follows that $\mathbf{E}_{P_t}[f] \le f_{\min} + 3\varepsilon\, f_{\mathrm{range}}$.

Proof By Lemma 4.7.3, if f ranged from 0 to 1 then using $T = \hat{T}(\varepsilon)$ would ensure $\mathbf{E}_{\pi_T}[f] \le \varepsilon$. The lemma stipulates $T = cr^k\hat{T} \le f_{\mathrm{range}}\hat{T}$. By the simple scaling result of Theorem 6.1.2, using $T \le f_{\mathrm{range}}\hat{T}(\varepsilon)$ ensures $\mathbf{E}_{\pi_T}[f] \le f_{\min} + 2\varepsilon\, f_{\mathrm{range}}$.

Rescaling Lemma 4.7.5 says that using time $t = \hat{t}(T/f_{\mathrm{range}})$ guarantees $\|P_t - \pi_T\|_{\mathrm{tvd}} \le \varepsilon$. In this case, $T/f_{\mathrm{range}} \ge T/cr^{k-1} = r\hat{T}$, so time $t = \hat{t}(r\hat{T})$ is sufficient to guarantee $\|\pi_T - P_t\|_{\mathrm{tvd}} \le \varepsilon$.

Lemma 4.7.1 yields $\mathbf{E}_{P_t}[f] \le f_{\min} + 3\varepsilon\, f_{\mathrm{range}}$. ■

Theorem 6.2.2 *Let us be given a value $0 < \varepsilon < 1$, and an annealing graph G which is known to have no more than n vertices and whose energy function f is known to have range $c_1 \le f_{\mathrm{range}} \le c_2$. Let $r < 1$ and $K \in \mathbb{Z}$ be such that $r^K c_2 \le c_1$. For $k = 1, \ldots, K$ in turn, anneal at temperature $T_k = c_2 r^k \hat{T}(\varepsilon, n)$ for time $t = \hat{t}(r\hat{T})$. Then the distribution P_K after* *the K'th "generation" of annealing satisfies*

$$\mathbf{E}_{P_K}[f] \le f_{\min} + 3\varepsilon\, f_{\text{range}}. \tag{E.3}$$

Proof Let k be the value for which

$$r^k c_2 \le f_{\text{range}} \le r^{k-1} c_2. \tag{E.4}$$

By Lemma 6.2.1, annealing at T_k for time t_k results in distribution P_k with

$$\mathbf{E}_{\pi_{T_k}}[f] \le f_{\min} + 2\varepsilon\, f_{\text{range}} \tag{E.5}$$

and

$$\|P_k - \pi_{T_k}\|_{\text{tvd}} \le \varepsilon. \tag{E.6}$$

The cooling schedule specified is monotonically nonincreasing and includes this (T_k, t_k) as a subschedule, so by a rescaling of Corollary 5.1.19, the final distribution P_K satisfies

$$\mathbf{E}_{P_k}[f] \;\le\; \mathbf{E}_{\pi_{T_k}}[f] + \varepsilon f_{\text{range}} \;\le\; f_{\min} + 3\varepsilon\, f_{\text{range}}. \tag{E.7}$$

■

Corollary 6.3.1 *Under the conditions of Theorem 6.2.2, let $r = 1 - \hat{T}(\varepsilon)/2$ and $K = \lceil \ln\left(\frac{c_2}{c_1}\frac{1}{\varepsilon}\right) / \ln(1/r) \rceil$. Then the cooling schedule specified by Theorem 6.2.2 uses total running time asymptotically equal to*

$$K\hat{t}(r\hat{T}(\varepsilon)) \lesssim \frac{2e}{\varepsilon}\ln^2(1/\varepsilon)\hat{t}(\hat{T}(\varepsilon)). \tag{E.8}$$

to yield its solution of "quality" (relative expected energy) no more than 3ε.

Proof We will be a little bit sloppy here to avoid wasting effort on something so straightforward. We will take it for granted that the optimal value of r will tend to 1 from below as $\varepsilon \to 0$, and that the associated K will go to infinity.

Making $r^K c_2 \le c_1\varepsilon$ means using

$$\begin{aligned} K &= \lceil \ln\left(\frac{c_2}{c_1}\frac{1}{\varepsilon}\right) / \ln(1/r) \rceil & \text{(E.9)} \\ &= \ln\left(\frac{c_2}{c_1}\frac{1}{\varepsilon}\right) / \ln(1/r) + O(1). & \text{(E.10)} \end{aligned}$$

The run time t in each "generation" is

$$t \sim 2\left(\ln(n^2/\varepsilon) + \frac{1}{r\hat{T}}\right) n^4 e^{2/(r\hat{T})}, \tag{E.11}$$

for

$$\ln t \sim \frac{2}{r\hat{T}} + \ln\frac{1}{r\hat{T}} \sim \frac{2}{r\hat{T}} \tag{E.12}$$

under the continuing assumption that as $\varepsilon \to 0$, $r \to 1$, while we know that $\hat{T}(\varepsilon) \to 0$ and that $1/\hat{T}(\varepsilon) = \frac{1}{\varepsilon}\ln(n^2/\varepsilon) \geq \ln(n^2/\varepsilon)$. Thus the log of the total time is

$$\ln Kt \sim \left(\frac{2}{r\hat{T}}\right) + \left(\ln\ln\frac{1}{\varepsilon} - \ln\ln\frac{1}{r}\right). \tag{E.13}$$

$\hat{T} = \varepsilon/\ln(n^2/\varepsilon)$, so $\ln\ln(1/\varepsilon) = o(\ln(1/\varepsilon)) = o(1/\hat{T})$ for ε small, and

$$\ln Kt \sim \frac{2}{r\hat{T}} - \ln\ln\frac{1}{r}. \tag{E.14}$$

Letting $r = 1 - \delta$ (and presuming $\delta \to 0$ as $\varepsilon \to 0$),

$$\ln Kt \sim \frac{2}{\hat{T}}(1+\delta) - \ln\delta. \tag{E.15}$$

The minimum of this expression occurs at

$$0 = \frac{\ln Kt}{\delta} = \frac{2}{\hat{T}} - \frac{1}{\delta}, \tag{E.16}$$

leading to the asymptotically optimal choice $\delta = \hat{T}/2 = \frac{1}{2}\varepsilon/\ln(n^2/\varepsilon)$.

Noting that $1/r = 1+\delta+o(\delta) = 1+\hat{T}/2+o(\hat{T}/2)$ leads to the interesting conclusion that

$$\frac{1}{r\hat{T}} = \frac{1}{\hat{T}} + \frac{1}{2} + o(1) \tag{E.17}$$

and so

$$e^{2/r\hat{T}} = e^{2/\hat{T}+2/2+o(1)} \sim e \cdot e^{2/\hat{T}}. \tag{E.18}$$

It follows that

$$\frac{\hat{t}(r\hat{T})}{\hat{t}(\hat{T})} = \frac{\ln(n^2/\varepsilon) + \frac{1}{r\hat{T}}\, e^{2/r\hat{T}}}{\ln(n^2/\varepsilon) + \frac{1}{\hat{T}}\, e^{2/\hat{T}}} \sim e. \tag{E.19}$$

Also, since $\ln(1/r) \sim \delta = \hat{T}/2$,

$$K \sim \frac{\ln(1/\varepsilon)}{\ln(1/r)} \sim \frac{\ln(1/\varepsilon)}{\frac{1}{2}\varepsilon/\ln(n^2/\varepsilon)} \sim \frac{2}{\varepsilon}\ln^2(1/\varepsilon). \tag{E.20}$$

We conclude that

$$\frac{K\hat{t}(r\hat{T})}{\hat{t}(\hat{T})} \sim eK \sim \frac{2e}{\varepsilon}\ln^2(1/\varepsilon). \tag{E.21}$$

■

Appendix F

Proofs for Chapter 7

Proposition 7.2.4 *Let $t(\varepsilon) > (1/\varepsilon)^2$. Then for any fixed d, $\Upsilon(dt(\varepsilon), 1/d, \varepsilon) \sim dt(\varepsilon)$.*
Proof Chebyshev's inequality (see for example [2]) says that for a random variable X with mean μ and variance σ^2, for all $a > 0$,

$$\mathrm{P}[|X - \mu| \geq a\sigma] \leq 1/a^2. \tag{F.1}$$

Let $X \sim B(dt', 1/d)$ and $a = \sqrt{1/\varepsilon}$. Then Chebyshev's inequality says that

$$\begin{aligned} \mathrm{P}\left[X \leq t' - \frac{1}{\sqrt{\varepsilon}}\sqrt{t'(1-1/d)}\right] &\leq \mathrm{P}\left[|X - t'| \geq \frac{1}{\sqrt{\varepsilon}}\sqrt{t'(1-1/d)}\right] & \text{(F.2)} \\ &\leq \varepsilon. & \text{(F.3)} \end{aligned}$$

If

$$t \leq t' - \frac{1}{\sqrt{\varepsilon}}\sqrt{t'(1-1/d)} \tag{F.4}$$

then this will imply the desired result

$$\mathrm{P}[X < t] \leq 1 - \varepsilon. \tag{F.5}$$

We claim that $t' = t + 2\sqrt{t/\varepsilon}$ suffices. First, note that $t' = t(1 + 2\sqrt{1/\varepsilon t})$ and that $1/(\varepsilon t) = \varepsilon(1/\varepsilon^2)/t(\varepsilon) \leq \varepsilon \to 0$. Thus

$$\begin{aligned} \frac{1}{\sqrt{\varepsilon}}\sqrt{t'(1-1/d)} &= \sqrt{\frac{1}{\varepsilon}(1-1/d)t(1+2\sqrt{1/\varepsilon t})} & \text{(F.6)} \\ &= \sqrt{t/\varepsilon}\left(1 - \frac{1}{2d} + \sqrt{\frac{1/\varepsilon}{t}} + o\left(\sqrt{\frac{1/\varepsilon}{t}}\right)\right) & \text{(F.7)} \\ &< 2\sqrt{t/\varepsilon} & \text{(F.8)} \end{aligned}$$

for

$$\begin{aligned} t' - \frac{1}{\sqrt{\varepsilon}}\sqrt{t'(1-1/d)} &> \left(t + 2\sqrt{\frac{t}{\varepsilon}}\right) - 2\sqrt{\frac{t}{\varepsilon}} & \text{(F.9)} \\ &= t. & \text{(F.10)} \end{aligned}$$

We have now shown that for $X = B(dt', 1/d)$ with $t' = t(1 + 2\sqrt{1/\varepsilon t}) \sim t$, $\mathrm{P}[X \geq t] \geq 1 - \varepsilon$. Thus $\Upsilon(dt(\varepsilon), 1/d, \varepsilon) \leq dt' \sim dt$. Conversely, $\Upsilon(dt(\varepsilon), 1/d, \varepsilon) \geq dt$, so $\Upsilon(dt, 1/d, \varepsilon) \sim dt$. ■

Lemma 7.2.3 *Let $(G, f) = \prod_{i=1}^{d}(G_i, r_i f_i)$ where without loss of generality each f_i ranges from exactly 0 to 1 and each $r_i \geq 0$. Let n be an upper bound on the order of each G_i.*

Let $0 < r < 1$, $c > 0$, and $k \in \mathbb{Z}$ satisfy $r^k c < r_i < r^{k-1}c$.

Given $\varepsilon > 0$, let $\hat{T} = \hat{T}(\varepsilon)$ and $T = r^k c \hat{T}$.

Let $\hat{t} = \hat{t}(r\hat{T})$ and $t = \Upsilon(d\hat{t}, 1/d, \varepsilon)$.

From any initial distribution anneal on G at temperature T for time t. Let the final distribution have corresponding i-marginal distribution P. Then $\mathbf{E}_{\pi_T}[f_i] \leq 2\varepsilon$ and $\|P - \pi_T\|_{\text{tvd}} \leq 2\varepsilon$. Consequently $\mathbf{E}_P[r_i f_i] \leq 4\varepsilon r_i$.

Proof With probability at least $1 - \varepsilon$, at least t moves are made on G_i, Write the final distribution P as $(1 - \varepsilon)P' + \varepsilon P''$, where P' and P'' are (respectively) the distributions conditional upon making at least t moves or fewer than t.

Since G_i is a graph with energy function $r_i f_i$ and energy range r_i, Lemma 6.2.1 guarantees that $\mathbf{E}_{\pi_T}[r_i f_i] \leq 2\varepsilon r_i$ and that $\|P_i' - \pi_T\|_{\text{tvd}} \leq \varepsilon$.

For *any* distributions π, P', and P'', and any $0 \leq \varepsilon \leq 1$, if $P = (1 - \varepsilon)P' + \varepsilon P''$ then:

$$\begin{aligned}
\|P - \pi\|_{\text{tvd}} &= \frac{1}{2}\sum_v |P(v) - \pi(v)| && \text{(F.11)}\\
&= \frac{1}{2}\sum_v \left|(1-\varepsilon)\left(P'(v) - \pi(v)\right) + \varepsilon\left(P''(v) - \pi(v)\right)\right| && \text{(F.12)}\\
&\leq \frac{1}{2}\sum_v \left\{(1-\varepsilon)\left|P'(v) - \pi(v)\right| + \varepsilon\left|P''(v) - \pi(v)\right|\right\} && \text{(F.13)}\\
&= (1-\varepsilon)\|P' - \pi\|_{\text{tvd}} + \varepsilon\|P'' - \pi\|_{\text{tvd}}. && \text{(F.14)}
\end{aligned}$$

The consequence for this case is that

$$\begin{aligned}
\|P - \pi\|_{\text{tvd}} &\leq (1-\varepsilon)\cdot\|P' - \pi_T\|_{\text{tvd}} + \varepsilon\cdot 1 && \text{(F.15)}\\
&\leq 2\varepsilon. && \text{(F.16)}
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbf{E}_P[r_i f_i] &\leq \mathbf{E}_{\pi_T}[r_i f_i] + \|P - \pi\|_{\text{tvd}}\cdot r_i && \text{(F.17)}\\
&\leq 4\varepsilon r_i. && \text{(F.18)}
\end{aligned}$$

■

Theorem 7.2.5 *Let $(G, f) = \prod_{i=1}^{d}(G_i, r_i f_i)$ with f_i ranging from exactly 0 to 1 and $r_i \geq 0$. Let n be a known upper bound on the order of each G_i, and let c_1 and c_2 be known lower and upper bounds for the largest r_i.*

Given $\varepsilon > 0$, let $0 < r < 1$ and $K \in \mathbb{Z}$ satisfy $r^K c_2 \leq c_1\varepsilon/d$.

Let $\hat{T} = \hat{T}(\varepsilon)$ and $T_k = r^k c_2 \hat{T}$. Let $\hat{t} = \hat{t}(r\hat{T})$ and $t = \Upsilon(d\hat{t}, 1/d, \varepsilon)$.

From any initial distribution anneal on G with cooling schedule $\{(T_k, t)\}_{k=0}^{K}$. Then the final distribution P gives relative expected energy

$$\mathbf{E}_{\text{sep}} \equiv \frac{\mathbf{E}_P[f(v)]}{\max_v f(v)} \leq 5\varepsilon; \qquad \text{(F.19)}$$

that is, the expected final cost relative to the full range of the cost function is no more than 5ε.

Proof First, consider a component i for which $r_i > c_1\varepsilon/d$ (**case I**). From $r^K c_2 \le c_1\varepsilon/d$ it follows that $r^K c_2 < r_i$. Also, by definition of c_2, $r^0 c_2 \ge r_i$. Then there is some $k(i)$ in $0, \ldots, K$ such that $r^{k(i)} c_2 < r_i \le r^{k(i)-1} c_2$.

Lemma 7.2.3 says that from any initial distribution, annealing on G at temperature $T_{k(i)}$ for time t yields a final distribution P_k satisfying

$$\mathbf{E}_{\pi_{T_k}}[r_i f_i] \le 2\varepsilon r_i \tag{F.20}$$

and

$$\|P_k - \pi_{T_k}\|_{\text{tvd}} \le 2\varepsilon. \tag{F.21}$$

The "schedule" with temperature fixed at $T_{k(i)}$ for time t is a subschedule of the monotonically nonincreasing schedule described by this theorem. Applying Corollary 5.1.19 and inequalities (F.20) and (F.21), the distribution P following the full cooling schedule has the property that $\mathbf{E}_P[f_i] \le 2\varepsilon + 2\varepsilon$, or

$$\mathbf{E}_P[r_i f_i] \le 4\varepsilon r_i. \tag{F.22}$$

Inequality (F.22) holds whenever $r_i > c_1\varepsilon/d$. The alternative is that $r_i \le c_1\varepsilon/d$ (**case II**), in which case of course $\mathbf{E}_P[r_i f_i] \le r_i \le c_1\varepsilon/d$.

Summing over all i,

$$\begin{aligned}
\mathbf{E}_P[f] &= \sum_{i=1}^{d} \mathbf{E}_P[r_i f_i] && \text{(F.23)}\\
&\le \sum_{\text{case I}} 4\varepsilon r_i + \sum_{\text{case II}} c_1\varepsilon/d && \text{(F.24)}\\
&\le 4\varepsilon \sum_{i=1}^{d} r_i + \varepsilon c_1 && \text{(F.25)}\\
&\le 5\varepsilon \sum_{i=1}^{d} r_i. && \text{(F.26)}
\end{aligned}$$

Since $\max_v f(v) = \sum r_i$ the conclusion of the theorem follows. ■

Corollary 7.3.1 *Under the conditions of Theorem 7.2.5, let* $r = 1 - \hat{T}(\varepsilon)/2$ *and* $K = \lceil \ln\left(\frac{c_2}{c_1}\frac{d}{\varepsilon}\right) / \ln(1/r) \rceil$. *Since* $\Upsilon(d\hat{t}, 1/d, \varepsilon) \sim d\hat{t}$, *the cooling schedule specified by that theorem uses total running time*

$$t_{\text{sep}} \sim dK\hat{t}(r\hat{T}(\varepsilon)) \lesssim d\frac{2e}{\varepsilon}\ln^2(1/\varepsilon)\hat{t}(\hat{T}(\varepsilon)). \tag{F.27}$$

to yield its solution of "quality" (relative expected energy) $\mathbf{E}_{\text{sep}} \le 5\varepsilon$.

Proof The computations are just the same as those for Corollary 7.3.1, which yielded an efficient choice of r for Theorem 7.2.5, the only difference being the presence of the factor d here.

In this case we need

$$K = \ln\left(\frac{c_2}{c_1}\frac{d}{\varepsilon}\right)/\ln(1/r) + O(1). \tag{F.28}$$

The run time t in each "generation" is

$$t \sim d \cdot 2\left(\ln(n^2/\varepsilon) + \frac{1}{r\hat{T}}\right) n^4 e^{2/(r\hat{T})}, \tag{F.29}$$

for

$$\ln t \sim \frac{2}{r\hat{T}} + \ln\frac{1}{r\hat{T}} \sim \frac{2}{r\hat{T}} \tag{F.30}$$

– the d has disappeared after the logarithm is taken.

Writing $r = 1 - \delta$,

$$\ln dKt \sim \frac{2}{\hat{T}}(1+\delta) - \ln\delta \tag{F.31}$$

just as in the proof of Corollary 6.3.1. Thus the asymptotically optimal choice of r and K is the same as in that single-variable case and the rest of the theorem follows immediately. ■

Appendix G

Proofs for Chapter 8

Lemma 8.1.4 *For x, x' not divisible by b^{-k}, $\mathrm{s}(x)$ and $\mathrm{s}(x')$ agree in components 1 through k if and only if x and x' lie in a common k-piece.*

Proof Induction on k. Straightforward for $k = 1$. Suppose x and x' lie in different $(k+1)$-pieces but in a common k-piece, with $k \neq 0$. Since they lie in a common k-piece, they also lie in a common 1-piece, and so they have a common first digit x_1 (this uses the fact that the pieces are *open* intervals). $bx - x_1$ and $bx' - x_1$ lie in different k-pieces but in a common $(k-1)$-piece, so by the inductive hypothesis $\mathrm{s}(bx - x_1)$ and $\mathrm{s}(bx' - x_1)$ first differ in the kth component. The same is true for their complements: $\mathrm{s}(1 - [bx - x_1])$ and $\mathrm{s}(1 - [bx' - x_1])$ first differ in the kth component. Since $\mathrm{s}(x) = \langle x_1\ ,\ \mathrm{s}(\mathrm{comp}_{x_1}(bx - x_1))\rangle$ and $\mathrm{s}(x') = \langle x_1\ ,\ \mathrm{s}(\mathrm{comp}_{x_1}(bx' - x_1))\rangle$, these sequences first differ in the $(k+1)$st component.

Conversely, suppose $\mathrm{s}(x)$ and $\mathrm{s}(x')$ differ in component $k+1$, $k \neq 0$. By definition, $\mathrm{s}(x) = \langle x_1\ ,\ \mathrm{s}(\mathrm{comp}_{x_1}(bx - x_1))\rangle$ and $\mathrm{s}(x) = \langle x_1\ ,\ \mathrm{s}(\mathrm{comp}_{x_1}(bx' - x_1))\rangle$. Since these differ in component $k+1$, the latter portions differ in component k. By the inductive hypothesis, $\mathrm{comp}_{x_1}(bx - x_1)$ and $\mathrm{comp}_{x_1}(bx' - x_1)$ lie in different k-pieces. Thus x and x' lie in different $(k+1)$-pieces.

■

Lemma 8.1.6 (Additional Gray code property) *For any vector $\vec{s} = \langle s_1, s_2, \ldots\rangle$, let $k{\downarrow}\vec{s} = \langle s_{k+1}, s_{k+2}, \ldots\rangle$. Then for any integers $k > 0$ and $j \in \{0, \ldots, b^k - 1\}$, for all $x \in (0,1)$,*

$$k{\downarrow}\mathrm{s}\left(\frac{j+x}{b^k}\right) = \mathrm{s}(\mathrm{comp}_j(x)). \tag{G.1}$$

Proof The proof is by induction on k. For $x' = (j+x)/b^{k+1}$, rewrite j as $j_1 b + j_2$, with $j_1 \in \{0, \ldots, b^k - 1\}$ and $j_2 \in \{0, \ldots, b-1\}$. Then

$$x' = \frac{j+x}{b^{k+1}} \quad = \quad \frac{j_1 + (j_2 + x)/b}{b^k}. \tag{G.2}$$

By the inductive hypothesis,

$$k{\downarrow}\mathrm{s}(x') \quad = \quad \mathrm{s}\left(\mathrm{comp}_{j_1}(\frac{j_2 + x}{b})\right) \tag{G.3}$$

$$= \mathrm{s}\left(\frac{\mathrm{comp}_{j_1}(j_2)+\mathrm{comp}_{j_1}(x)}{b}\right) \tag{G.4}$$

$$= \left\langle \mathrm{comp}_{j_1}(j_2)\,,\ \mathrm{s}\left(\mathrm{comp}_{\mathrm{comp}_{j_1}(j_2)}(\mathrm{comp}_{j_1}(x))\right)\right\rangle. \tag{G.5}$$

so

$$(k+1)\!\downarrow\!\mathrm{s}(x) = \mathrm{s}\left(\mathrm{comp}_{\mathrm{comp}_{j_1}(j_2)+j_1}(x)\right). \tag{G.6}$$

We wished to show that $(k+1)\!\downarrow\!\mathrm{s}(x') = \mathrm{s}(\mathrm{comp}_j(x))$. Since the complement function relies only on the parity of its subscripted argument, and $j = j_1 b + j_2$, we need only show that $\mathrm{comp}_{j_1}(j_2) + j_1$ has the same parity as $j_1 b + j_2$: Using "≡" to denote congruence modulo 2,

$$\mathrm{comp}_{j_1}(j_2) = \begin{cases} j_2 & \text{if } j_1 \equiv 0 \\ (b-1) - j_2 & \text{if } j_1 \equiv 1 \end{cases} \tag{G.7}$$

$$\equiv \begin{cases} j_2 & \text{if } j_1 \equiv 0 \\ (b-1) + j_2 & \text{if } j_1 \equiv 1 \end{cases} \tag{G.8}$$

$$\equiv (b-1)j_1 + j_2. \tag{G.9}$$

Thus $\mathrm{comp}_{j_1}(j_2) + j_1 \equiv [(b-1)j_1 + j_2] + j_1 \equiv bj_1 + j_2$. ■

Lemma 8.1.7 *For* $\mathrm{s}(x) = \langle s_1, s_2, \ldots\rangle$,

$$f(x) = F(s_1) + rF(s_2) + r^2F(s_3) + \cdots. \tag{G.10}$$

Proof Let $\langle s_1, s_2, \ldots\rangle = \mathrm{s}(x)$, and $f'(x) = \sum_{k=1}^{\infty} r^{k-1}F(s_k)$. We prove that $f(x)$ and $f'(x)$ are identical as formal power series in r, and therefore $f(x) \equiv f'(x)$. Specifically, by induction on k we will prove that f and f' agree in terms 1 through k for every k.

Writing x as $.x_1x_2\ldots$, by definition of s, $s_1 = x_1$, so the statement is true for $k = 1$.

Also by definition of s, $\langle s_2, s_3, \ldots\rangle = \mathrm{s}(\mathrm{comp}_{x_1}(bx - x_1))$, so $f'(x) = x_1 + rf'(\mathrm{comp}_{x_1}(bx - x_1))$. But by definition of f, $f(x) = x_1 + rf(\mathrm{comp}_{x_1}(bx - x_1))$. By the inductive hypothesis, $f'(\mathrm{comp}_{x_1}(bx - x_1))$ and $f(\mathrm{comp}_{x_1}(bx - x_1))$ agree through the kth terms, therefore $f'(x)$ and $f(x)$ agree through the $k+1$st terms.

■

Theorem 8.3.2 *Let a value* ε *and a deterministic fractal with energy scale parameter* r *be given. Let* $\hat{T} = \hat{T}(\varepsilon)$ *and* $\hat{t} = \hat{t}(\hat{T})$. *Apply confined annealing with cooling schedule* $(T_k, t_k) = (r^{k-1}\hat{T}, \hat{t})$, *for* $k = 1, \ldots, K$, *with* $K = \lceil \ln(1/\varepsilon)/\ln(1/r)\rceil$. *Then the state returned has relative expected energy*

$$\mathbf{E}_{\mathrm{con}} \equiv \frac{\mathbf{E}[f]}{f_{\mathrm{range}}} \leq 3\varepsilon \tag{G.11}$$

and the algorithm consumes run time

$$t_{\mathrm{con}} = \lceil \ln(1/\varepsilon)/\ln(1/r)\rceil \cdot \hat{t}(\varepsilon). \tag{G.12}$$

Proof By Theorem 8.2.1, annealing with $(\hat{T}, \hat{t})$ in generation 1 gives $\mathbf{E}[F(s_1^{1,t_1})] \leq 2\varepsilon$. By the similarity of the connected components of S_k to S_1, annealing in generation k with

$$(T_k, t_k) = (r^{k-1}T_1, t_1) \tag{G.13}$$

results in a distribution for s_k satisfying

$$\mathbf{E}[F(s_k^{k,t_k})] \leq 2\varepsilon. \tag{G.14}$$

By Proposition 8.3.1, $s_k^{K,t_K} = s_k^{k,t_k}$, so equation (G.14) also applies to s_k^{K,t_K}.

The energy of the final state x produced by this confined annealing algorithm therefore satisfies

$$\begin{aligned} \mathbf{E}[f] &= \sum_{k=1}^{K} r^{k-1}\mathbf{E}[F(s_k^{K,t_K})] + \sum_{k=K+1}^{\infty} r^{k-1}\mathbf{E}[F(s_k^{K,t_K})] && \text{(G.15)} \\ &\leq \frac{1}{1-r}(2\varepsilon + r^K \cdot 1). && \text{(G.16)} \end{aligned}$$

Since f ranges from 0 to $1/(1-r)$, the relative expected energy achieved is

$$\mathbf{E}_{\text{con}} \equiv \frac{\mathbf{E}[f]}{f_{\text{range}}} \leq 2\varepsilon + r^K. \tag{G.17}$$

The running time is obviously $t_{\text{con}} = Kt_1$.

Choosing $K = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil$, for $r^K \leq \varepsilon$, yields the theorem as stated. ■

Theorem 8.4.2 *For $i, k \geq 1$, S_{i+k} with the unconfined move set and deterministic fractal energy function f is a replica of S_k (also with unconfined move set and energy f). The index set is $\{0, \ldots, b-1\}^i$ and the energy scale factor is r^i.*

Proof Write $\bar{x} \in S_{k+i}$ as $(j + x)/b^i$, for $j \in \{0, \ldots, b^i - 1\}$ and $x \in S_k$. Thus j numbers the i-piece containing $\bar{x}$, while x is the offset of $\bar{x}$ from the left end of that interval. Then $\sigma(\bar{x}) = (j, \text{comp}_j(x))$ fulfills the conditions of Definition 8.4.1:

Property 1 follows from the invertibility of σ: $\sigma^{-1}(j, x) = (j + \text{comp}_j(x))/b^i$.

We now verify properties 2 and 3. With a point $\bar{x} \in S_{i+k}$ associate $x \in S_k$ and $j \in \{0, \ldots, b^i - 1\}$ with $\bar{x} = (j + x)/b^i$; similarly for its neighbors $\bar{x}' = \bar{x} - 1/b^{i+k}$ and $\bar{x}'' = \bar{x} + 1/b^{i+k}$.

Case I: x is neither $1/(2b^k)$ nor $1 - 1/(2b^k)$. In this case x' and x'' are just $x \pm 1/b^k$ (yielding property 3), and $j' = j'' = j$ (property 2).

Case II: $x = 1/(2b^k)$ or $x = 1 - 1/(2b^k)$. We treat only $x = 1 - 1/(2b^k)$; the other case follows symmetrically. Then $(j', x') = (j, 1 - 3/(2b^k))$, and $(j'', x'') = (j + 1, 1/(2b^k))$. Property 2 is satisfied for the edge $\{\bar{x}, \bar{x}'\}$ because $j = j'$, and for the edge $\{\bar{x}, \bar{x}''\}$ because

$$\sigma_V(\bar{x}'') = \text{comp}_{j''}(x'') = \text{comp}_{j+1}(1 - x) = \text{comp}_j(x) = \sigma_V(\bar{x}). \tag{G.18}$$

As for property 3, in addition to equation (G.18) we have

$$\sigma_V(\bar{x}') = \text{comp}_{j'}(x') = \text{comp}_j(1 - 3/(2b^k)). \tag{G.19}$$

Together these mean that σ_V maps $\bar{N}(\bar{x})$ to $\{\text{comp}_j(1 - 1/(2b^k)), \text{comp}_j(1 - 3/(2b^k))\}$. Whether j is even or odd, these are the two neighbors in S_k of $\sigma_V(\bar{x}) = \text{comp}_j(x) = \text{comp}_j(1 - 1/(2b^k))$.

We now prove property 4. First, $f(\bar{x}) = \sum_{l=1}^{i+k} r^{l-1} F(s_l(\bar{x}))$. Define $k{\uparrow}\langle s_1, s_2, \ldots\rangle$ to be $\langle s_1, s_2, \ldots, s_k\rangle$. By Lemmas 8.1.4 and 8.1.6, for $\bar{x} = (j + x)/b^i$ (with the usual conventions for j and x), $i{\uparrow}\text{s}(\bar{x}) = i{\uparrow}\text{s}((j + 1/2)/b^i)$ and $i{\downarrow}\text{s}(\bar{x}) = \text{s}(\text{comp}_j(x))$, so

$$f(\bar{x}) = \sum_{l=1}^{i} r^{l-1} F(s_l((j + 1/2)/b^i)) + r^i \sum_{l=1}^{\infty} r^{l-1} F(s_l(\text{comp}_j(x))). \tag{G.20}$$

Using $\text{comp}_j(x) = \sigma_V(\bar{x})$, the above expression for $f(\bar{x})$ is of the form $f_I(j) + r^i f(\sigma_V(\bar{x}))$. ■

Theorem 8.4.3 (product density on replica graphs) *Let $\bar{G}$ be a replica of G with scale factor c. Let P_t be an arbitrary probability distribution on the states V of G, and π_T be the "equilibrium" distribution on I given by $\pi_T(i) \propto e^{-f_I(i)/T}$. Let $\bar{P}_t$ be the distribution on states of $\bar{G}$ given by $\bar{P}_t = \pi_{cT} \times P_t$, i.e. $\bar{P}_t(i, v) = \pi_{cT}(i) \cdot P_t(v)$. Let $\bar{P}_{t+1}(i, v)$ be the distribution on $\bar{G}$ after a single annealing move at temperature cT starting from $\bar{P}_t$, and let $P_{t+1}(v)$ be the distribution on G after a single annealing move at temperature T starting from distribution P_t. Then $\bar{P}_{t+1} = \pi_{cT} \times P_{t+1}$.*

Proof First, for any Markov chain, $P_{t+1}(v)$ is $P_t(v)$ plus the "probability flux" summed over edges incident to v. Mathematically, for a Markov chain X_t, the flux from u to v is

$$P(u \to v) = P(X_t = u \,\text{and}\, X_{t+1} = v), \tag{G.21}$$

the probability of going from u to v at step t. It can be decomposed as

$$P(u \to v) = P_t(u) \cdot P(u \to v \mid u) \tag{G.22}$$

where $P_t(u) = P(X_t = u)$ and $P(u \to v \mid u) = P(X_{t+1} = v \mid X_t = u)$. Note that $P(u \to v \mid u)$ is time-independent for a homogeneous Markov chain, while $P(u \to v)$ implicitly depends on t, through $P_t(u)$. Then we are merely saying that

$$\begin{aligned} P_{t+1}(v) &= P_t(v) + \sum_{u \in V} [P(u \to v) - P(v \to u)] & \text{(G.23)} \\ &= P_t(v) + \sum_{u \in V} [P_t(u) P(u \to v \mid u) - P_t(v) P(v \to u \mid v)]. & \text{(G.24)} \end{aligned}$$

Now, in the current context we are dealing with chains on $\bar{G}$ and G, so we use $\bar{P}$ for probabilities in the former and P for the latter. We often abbreviate a vertex (i, v) of $\bar{G}$ as just (iv), and write $\deg(i, v)$ for the degree of vertex (i, v) in $\bar{G}$ and $\deg(v)$ for the degree of vertex v in G. Also, we let $A_T(y) = \min(1, e^{-y/T})$, the probability of accepting at temperature T an annealing move which raises the energy by y.

Since all edges in $\bar{G}$ incident to (i, v) are either of the form (i, v') or (i', v) (i and v do not both change along a single edge), application of the "flux" argument to $\bar{G}$ means that

$$\bar{P}_{t+1}(iv) = \bar{P}_t(iv) + \sum_{i'} [\bar{P}(i'v \to iv) - \bar{P}(iv \to i'v)] + \sum_{v'} [\bar{P}(iv' \to iv) - \bar{P}(iv \to iv')]. \tag{G.25}$$

(Double-counting the case $i' = i$, $v' = v$ is irrelevant since the net flux of such terms is 0.) For any edge of the type $\{(iv),(i'v)\}$ in $\bar{G}$,

$$\begin{aligned}
\bar{P}(iv \to i'v) &= \bar{P}_t(iv) \cdot \bar{P}(iv \to i'v \,|\, iv) && \text{(G.26)}\\
&= \bar{P}_t(iv) \cdot \frac{1}{\deg(i,v)} \cdot A_{cT}(\bar{f}(i'v) - \bar{f}(iv)) && \text{(G.27)}\\
&= \pi_{cT}(i) P_t(v) \frac{1}{\deg(v)} A_{cT}(f_I(i') - f_I(i)). && \text{(G.28)}
\end{aligned}$$

Symmetrically,

$$\bar{P}(i'v \to iv) = \pi_{cT}(i') P_t(v) \frac{1}{\deg(v)} A_{cT}(f_I(i) - f_I(i')). \qquad \text{(G.29)}$$

By definition of π_T,

$$\pi_{cT}(i) \cdot A_{cT}(f_I(i') - f_I(i)) = \pi_{cT}(i') \cdot A_{cT}(f_I(i) - f_I(i')), \qquad \text{(G.30)}$$

so for any edge $\{(iv),(i'v)\}$,

$$\bar{P}(i'v \to iv) - \bar{P}(iv \to i'v) = 0. \qquad \text{(G.31)}$$

For any edge of the type $\{(iv),(iv')\}$ in $\bar{G}$,

$$\begin{aligned}
\bar{P}(iv \to iv') &= \bar{P}_t(iv) \cdot \bar{P}(iv \to iv' \,|\, iv) && \text{(G.32)}\\
&= \bar{P}_t(iv) \cdot \frac{1}{\deg(i,v)} \cdot A_{cT}(\bar{f}(iv') - \bar{f}(iv)) && \text{(G.33)}\\
&= \pi_{cT}(i) P_t(v) \cdot \frac{1}{\deg(v)} \cdot A_{cT}(c[f(v') - f(v)]) && \text{(G.34)}\\
&= \pi_{cT}(i) P_t(v) \frac{1}{\deg(v)} A_T(f(v') - f(v)) && \text{(G.35)}\\
&= \pi_{cT}(i) P(v \to v'). && \text{(G.36)}
\end{aligned}$$

Symmetrically,

$$\bar{P}(iv' \to iv) = \pi_{cT}(i) P(v' \to v). \qquad \text{(G.37)}$$

Substituting these expressions into equation (G.25),

$$\begin{aligned}
\bar{P}_{t+1}(iv) &= \bar{P}_t(iv) + \sum_{i'} [\bar{P}(i'v \to iv) - \bar{P}(iv \to i'v)] && \text{(G.38)}\\
&\quad + \sum_{v'} [\bar{P}(iv' \to iv) - \bar{P}(iv \to iv')] \\
&= \pi_{cT}(i) P_t(v) + \sum_{i'} [0] + \pi_{cT}(i) \sum_{v'} [P(v' \to v) - P(v \to v')] && \text{(G.39)}\\
&= \pi_{cT}(i) \left\{ P_t(v) + \sum_{v'} [P(v' \to v) - P(v \to v')] \right\} && \text{(G.40)}\\
&= \pi_{cT}(i) P_{t+1}(v), && \text{(G.41)}
\end{aligned}$$

where the final equality comes from recognizing that (in G) $P_t(v)$ plus the net flux into v is $P_{t+1}(v)$. ■

Theorem 8.4.4 (marginal density on replica graphs) *Let $\bar{G}$ be a replica of G with scale factor c. Let $\bar{P}_t(i,v)$ be an arbitrary probability distribution on the states $I \times V$ of $\bar{G}$, and let $P_t(v) = \sum_i \bar{P}_t(i,v)$ be the corresponding marginal distribution of v. Let $\bar{P}_{t+1}(i,v)$ be the distribution after a single annealing move on $\bar{G}$ at temperature cT starting from $\bar{P}_t(i,v)$; and let $P_{t+1}(v)$ be the distribution after a single annealing move on G at temperature T starting from $P_t(v)$. Then $P_{t+1}(v)$ is the marginal distribution of v corresponding to $\bar{P}_{t+1}(i,v)$.*

Proof By the flux argument used in the proof of Theorem 8.4.3,

$$\begin{aligned}\bar{P}_{t+1}(iv) &= \bar{P}_t(iv) + \sum_{i' \in I}[\bar{P}(i'v \to iv) - \bar{P}(iv \to i'v)] \\ &\quad + \sum_{v' \in V}[\bar{P}(iv' \to iv) - \bar{P}(iv \to iv')]\end{aligned} \tag{G.42}$$

We will be substituting

$$\bar{P}(iv \to iv') = \bar{P}_t(iv) \cdot \frac{1}{\deg(i,v)} \cdot A_{cT}(\bar{f}(iv') - \bar{f}(iv)) \tag{G.43}$$

$$= \bar{P}_t(iv) \cdot \frac{1}{\deg(v)} \cdot A_T(f(v') - f(v)), \tag{G.44}$$

and the symmetric form (swapping v and v').

The marginal distribution for v corresponding to $\bar{P}_{t+1}(iv)$ is

$$\begin{aligned}\sum_i \bar{P}_{t+1}(i,v) &= \sum_{i \in I} \bar{P}_t(iv) + \sum_{i,i' \in I}[\bar{P}(i'v \to iv) - \bar{P}(iv \to i'v)] \\ &\quad + \sum_{v' \in V}\left[\sum_{i \in I} \bar{P}_t(iv') \cdot \frac{1}{\deg(v')} \cdot A_{cT}(c[f(v) - f(v')])\right. \\ &\qquad \left. - \sum_{i \in I} \bar{P}_t(iv) \cdot \frac{1}{\deg(v)} \cdot A_{cT}(c[f(v') - f(v)])\right]\end{aligned} \tag{G.45}$$

$$\begin{aligned}&= P_t(v) + 0 + \sum_{v' \in V}\left[P_t(v') \cdot \frac{1}{\deg(v')} \cdot A_T(f(v) - f(v'))\right. \\ &\qquad \left. - P_t(v) \cdot \frac{1}{\deg(v)} \cdot A_T(f(v') - f(v))\right]\end{aligned} \tag{G.46}$$

$$= P_t(v) + \sum_{v' \in V}[P(v' \to v) - P(v \to v')] \tag{G.47}$$

$$= P_{t+1}(v). \tag{G.48}$$

$$P_{t+1}(v) = \sum_{i \in I} \bar{P}_{t+1}(iv) \tag{G.49}$$

$$= \sum_{i,i' \in I} \{\text{antisymmetric function}(i,i')\} + \tag{G.50}$$

$$
\begin{aligned}
& \sum_{v'\in V}\left[\sum_{i\in I}\bar{P}_t(iv')\cdot\frac{1}{\deg(v')}\cdot A_T(\bar{f}(iv')-\bar{f}(iv))-\right. && \text{(G.51)}\\
& \qquad \left.\sum_{i\in I}\bar{P}_t(iv)\cdot\frac{1}{\deg(v)}\cdot A_T(\bar{f}(iv)-\bar{f}(iv'))\right] && \text{(G.52)}\\
= {} & 0+\sum_{v'\in V}\left[P_t(v')\cdot\frac{1}{\deg(v')}\cdot A_T(\bar{f}(iv')-\bar{f}(iv))-\right. && \text{(G.53)}\\
& \qquad \left.P_t(v)\cdot\frac{1}{\deg(v)}\cdot A_T(\bar{f}(iv)-\bar{f}(iv'))\right] && \text{(G.54)}\\
= {} & P_{t+1}(v). && \text{(G.55)}
\end{aligned}
$$

■

Theorem 8.4.5 *Let $\bar{G}$ be a replica graph with vertices $I\times V$, and let the distributions π_T on I be as per Theorem 8.4.3. Let T_{crit} be the critical temperature for I, and $\{T_t\}$ be a sequence of temperatures satisfying $T_{\text{crit}}\geq T_0\geq T_1\geq T_2\geq\cdots$. Anneal on $\bar{G}$ at temperatures $T_1,T_2,\ldots$ beginning from the distribution $\bar{P}_0$. If there exists a distribution P_0 on V such that the initial distribution $\bar{P}_0$ on $I\times V$ satisfies*

$$\|\bar{P}_0-(\pi_{T_0}\times P_0)\|_{\text{tvd}}\;\leq\;\varepsilon, \tag{G.56}$$

then the distribution at time t satisfies

$$\|\bar{P}_t-(\pi_{T_t}\times P_t)\|_{\text{tvd}}\;\leq\;\varepsilon+\|\pi_{T_0}-\pi_{T_t}\|_{\text{tvd}}. \tag{G.57}$$

Proof The reasoning is just as for Theorem 5.2.2. Inductively, assume

$$\|\bar{P}_t-(\pi_{T_t}\times P_t)\|_{\text{tvd}}\leq\varepsilon+\|\pi_{T_0}-\pi_{T_t}\|_{\text{tvd}}. \tag{G.58}$$

Truth for $t=0$ is given by the hypothesis (8.12).

As in the proof of Theorem 5.2.2, let $\zeta(w)$ be an unknown vector the sum of the absolute value of whose elements is no more than $2w$; in this case the components of ζ will be indexed by i alone or (i,v) depending on the case. Rewrite $\bar{P}_t(i,v)$:

$$
\begin{aligned}
\bar{P}_t(i,v) &= \pi_{T_t}(i)\cdot P_t(v)+[\zeta(\varepsilon+\|\pi_{T_0}-\pi_{T_t}\|_{\text{tvd}})](i,v) && \text{(G.59)}\\
&= [\pi_{T_{t+1}}+\zeta(\|\pi_{T_t}-\pi_{T_{t+1}}\|_{\text{tvd}})](i)\cdot P_t(v) && \text{(G.60)}\\
&\quad +[\zeta(\varepsilon+\|\pi_{T_0}-\pi_{T_t}\|_{\text{tvd}})](i,v) && \\
&= \pi_{T_{t+1}}(i)\cdot P_t(v)+[\zeta(\|\pi_{T_t}-\pi_{T_{t+1}}\|_{\text{tvd}})](i,v) && \text{(G.61)}\\
&\quad +[\zeta(\varepsilon+\|\pi_{T_0}-\pi_{T_t}\|_{\text{tvd}})](i,v) && \\
&= \pi_{T_{t+1}}(i)\cdot P_t(v)+[\zeta(\varepsilon+\|\pi_{T_0}-\pi_{T_{t+1}}\|_{\text{tvd}})](i,v). && \text{(G.62)}
\end{aligned}
$$

The final inequality follows from

$$\|\pi_{T_0}-\pi_{T_t}\|_{\text{tvd}}+\|\pi_{T_t}-\pi_{T_{t+1}}\|_{\text{tvd}}=\|\pi_{T_0}-\pi_{T_{t+1}}\|_{\text{tvd}}, \tag{G.63}$$

a consequence of $T_{\text{crit}}\geq T_0\geq T_t\geq T_{t+1}$.

Applying Theorem 8.4.3, after one step at temperature T_{t+1} we get

$$\bar{P}_{t+1}(i,v) \quad = \quad \pi_{T_{t+1}}(i) \cdot P_{t+1}(v) + [\zeta(\varepsilon + \|\pi_{T_0} - \pi_{T_{t+1}}\|_{\text{tvd}})](i,v). \tag{G.64}$$

■

Corollary 8.4.6 *Let π_T be the equilibrium distribution of s_1 for annealing on S_1 at temperature T, and let T_{crit} be the associated critical temperature. Run the unconfined annealing algorithm with cooling schedule $\{(T_k, t_k)\}_{k=1}^{K}$, where T_k is monotonically nonincreasing in k and $T_{\text{crit}} \geq T_1$. If s_1^{1,t_1} and s_1^{k,t_k} are the values of s_1 at the end of generation 1 and at the end of generation k respectively, then*

$$\|s_1^{k,t_k} - \pi_{T_k}\|_{\text{tvd}} \;\leq\; \|s_1^{1,t_1} - \pi_{T_1}\|_{\text{tvd}} + \|\pi_{T_1} - \pi_{T_k}\|_{\text{tvd}}. \tag{G.65}$$

Proof The proof is virtually identical to that of Theorem 8.4.5. The only difference is that in this case we must consider changes of generation as well as annealing moves. Suppose that at the end of generation k the state s^{k,t_k} has probability

$$P(\langle s_1, s_2, \ldots, s_k\rangle) = \pi(s_1)P'(\langle s_2, \ldots, s_k\rangle) + [\zeta(\varepsilon)](\langle s_1, s_2, \ldots, s_k\rangle). \tag{G.66}$$

Here $P'(\langle s_2, \ldots, s_k\rangle)$ is an arbitrary function (obviously we are thinking of it as the marginal distribution on $\langle s_2, \ldots, s_k\rangle$), and by the last term we mean an element (indexed by $\langle s_1, s_2, \ldots, s_k\rangle$) of an unknown vector ζ of magnitude ε. Since the change of generations maps s^{k,t_k} to $s^{k+1,0}$ by

$$\langle s_1, s_2, \ldots, s_k\rangle \mapsto \langle\langle s_1, s_2, \ldots, s_k\rangle\ ,\ s_{\text{rand}}\rangle, \tag{G.67}$$

$$\begin{aligned} P(\langle s_1, s_2, \ldots, s_{k+1}\rangle) \quad &= \quad \{\pi(s_1)P'(\langle s_2, \ldots, s_k\rangle) & \text{(G.68)} \\ &\quad\quad +[\zeta(\varepsilon)](\langle s_1, s_2, \ldots, s_k\rangle)\} \cdot \mathrm{P}[s_{k+1} = s_{\text{rand}}] \\ &= \quad \pi(s_1)P''(\langle s_2, \ldots, s_{k+1}\rangle) + [\zeta(\varepsilon)](\langle s_1, s_2, \ldots, s_{k+1}\rangle), & \text{(G.69)} \end{aligned}$$

where of course

$$P''(\langle s_2, \ldots, s_{k+1}\rangle) = P'(\langle s_2, \ldots, s_k\rangle) \cdot \mathrm{P}[s_{k+1} = s_{\text{rand}}]. \tag{G.70}$$

That is, if s^{k,t_k} has a distribution which is approximately a product distribution, then $s^{k+1,0}$ does also, and the error term is unchanged in magnitude. Thus the analysis of Theorem 8.4.5 applies to this case as well. ■

Theorem 8.4.8 *Consider two different unconfined annealing processes. The first uses cooling schedule (T_{1+i}, t_{1+i}), runs for generations with $i \geq 0$, and at time t into generation $1+i$ has distribution $P^{1+i,t}(\langle s_1, s_2, \ldots, s_{1+i}\rangle)$ on state space S_{1+i}. For a fixed integer $k > 0$, the second uses cooling schedule $(\bar{T}_{k+i}, \bar{t}_{k+i}) = (r^{k-1}T_{1+i}, t_{1+i})$, again runs for generations with $i \geq 0$, and at time t into generation $k+i$ has distribution $\bar{P}^{k+i,t}(\langle s_1, s_2, \ldots, s_{k+i}\rangle)$ on state space S_{k+i}, with corresponding marginal distribution $\bar{P}^{k+i,t}_{k,\ldots,k+i}(\langle s_k, \ldots, s_{k+i}\rangle)$. If the "initial" distributions satisfy $\bar{P}^{k,0}_k \equiv P^{1,0}$, then for all i, t representing positive time,*

$$\bar{P}^{k+i,t}_{k,\ldots,k+i} \equiv P^{1+i,t}. \tag{G.71}$$

It follows that $\bar{P}_k^{k+i,t} \equiv P_1^{1+i,t}$.

Proof We show by induction on "time" i, t that $\bar{P}_{k,\ldots,k+i}^{k+i,t} \equiv P^{1+i,t}$. There are two cases: either the generation number changes (along with the state space), or a move is made in the current generation. The condition that $\bar{t}_{k+1} \equiv t_{1+i}$ assures that the same case applies to both processes.

Case I: A new random variable is concatenated onto the previous state codes, which has the same effect on the two distributions under consideration.

Case II: As per Theorem 8.4.2, regard S_{k+i} as a replica of the basic graph whose vertices are identified by $\langle s_k, \ldots, s_{k+i} \rangle$, *i.e.* regard S_{k+i} as a replica of S_i. The preservation of the equality of the two distributions follows immediately from Theorem 8.4.4 and the assumption $\bar{T}_{k+i} \equiv r^{k-1} T_{1+1}$.

Taking marginals with respect to the first variables of both sides of equation (8.16) shows that $\bar{P}_k^{k+i,t} \equiv P_1^{i,t}$. ■

Theorem 8.4.9 *Let a value* $0 < \varepsilon < 1$ *and a fractal with energy scale parameter* r *be given. Let* $\hat{T} = \hat{T}(\varepsilon)$ *and* $\hat{t} = \hat{t}(\varepsilon)$. *Assume* ε *is sufficiently small that* $\hat{T} \leq T_{\text{crit}}$, *the critical temperature for annealing on* S_1. *Apply unconfined annealing with cooling schedule* $(T_k, t_k) = (r^{k-1}\hat{T}, \hat{t})$, *and* $k = 1, \ldots, K$ *with* $K = \lceil \ln(1/\varepsilon)/\ln(1/r) \rceil$. *The state returned has relative expected energy satisfying*

$$\mathbf{E}_{\text{uncon}} \equiv \frac{\mathbf{E}[f]}{f_{\text{range}}} \leq 3\varepsilon \tag{G.72}$$

and the algorithm has run time

$$t_{\text{uncon}} = [\ln(1/\varepsilon)/\ln(1/r)] \cdot \hat{t}. \tag{G.73}$$

Proof Initially s_k is $s_k^{k,0}$. By Theorem 8.4.7, an unconfined annealing beginning from a like-distributed s_1 (or any other distribution) would at generation $1+i$ $(i \geq 0)$ have

$$||s_1^{1+i,t_{1+i}} - \pi_{T_{1+i}}||_{\text{tvd}} \leq \varepsilon + ||\pi_{T_1} - \pi_{T_{1+i}}||_{\text{tvd}}. \tag{G.74}$$

Thus by Theorem 8.4.8, at generation $k+i$ we have

$$||s_k^{k+i,t_{k+i}} - \pi_{T_{1+i}}||_{\text{tvd}} \leq \varepsilon + ||\pi_{T_1} - \pi_{T_{1+i}}||_{\text{tvd}}. \tag{G.75}$$

Now apply the triangle inequality for total variation distance, and the monotonicity of entries of π_T below T_{crit}:

$$\begin{aligned}
||s_k^{k+i,t_{k+i}} - \pi_0||_{\text{tvd}} &\leq ||s_k^{k+i,t_{k+i}} - \pi_{T_{1+i}}||_{\text{tvd}} + ||\pi_{T_{1+i}} - \pi_0||_{\text{tvd}} && \text{(G.76)}\\
&\leq \varepsilon + ||\pi_{T_1} - \pi_{T_{1+i}}||_{\text{tvd}} + ||\pi_{T_{1+i}} - \pi_0||_{\text{tvd}} && \text{(G.77)}\\
&= \varepsilon + ||\pi_{T_1} - \pi_0||_{\text{tvd}} && \text{(G.78)}\\
&\leq \varepsilon + \varepsilon. && \text{(G.79)}
\end{aligned}$$

The final inequality above uses $||\pi_{T_1} - \pi_0||_{\text{tvd}} \leq \varepsilon$. Since we are using $T_1 = \hat{T}(\varepsilon, \Delta F, b)$, this is a consequence of Lemma 4.7.3.

For $i = K - k$, inequality (G.79) yields $||s_k^{K,t_K} - \pi_0||_{\text{tvd}} \leq 2\varepsilon$. This is precisely what had in the proof of Theorem 8.3.2, so the same parameter choices lead to the same result. ■

Acknowledgments

I am very grateful to Alberto Sangiovanni-Vincentelli, Umesh Vazirani, Richard M. Karp, David Brillinger, and Fabio Romeo for their invaluable thoughts and criticisms. I also thank Steve White for the close collaboration and many stimulating discussions of annealing that initiated this research. The work was performed while the author was on leave from IBM Research, and was also sponsored by DARPA and monitored by SNWSC under contract numbers N00039-87-C-0182 and N00039-88-C-0292.

Bibliography

[1] R. A. Becker and J. M. Chambers. *S: An Interactive Environment for Data Analysis.* Wadsworth, Belmont, CA, 1984.

[2] B. Bollobás. *Random Graphs.* Academic Press, London, 1985.

[3] D. R. Brillinger. *Time Series: Data Analysis and Theory.* Holden-Day, San Francisco, CA, 1981.

[4] P. Dagum, M. Luby, M. Mihail, and U. Vazirani. Polytopes, permanents, and graphs with large factors. In *Proceedings of the 29th Annual Symposium on Foundations of Computer Science*, pages 412–421, 1988.

[5] P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of Markov chains. Unpublished manuscript, 1990.

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

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

[8] R. W. Hamming. *Coding and Information Theory.* Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1986.

[9] W. R. Heller, W. F. Mikhail, and W. E. Donath. Prediction of wire space requirements for LSI. *Journal of Design Automation and Fault-Tolerant Computing*, 2(2):117–144, May 1978.

[10] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. Technical Report CSR-1-90, University of Edinburgh, Feb. 1990.

[11] M. R. Jerrum and A. Sinclair. Conductance and the rapid mixing property for Markov chains: The approximation of the permanent resolved. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pages 235–244, 1988.

[12] D. S. Johnson, C. R. Aragon, L. A. McGeoch, and C. Schevon. Optimization by simulated annealing: An experimental evaluation; part I, graph partitioning. *Operations Research*, 37(6):865–892, 1989.

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

[14] S. Kirkpatrick and G. Toulouse. Configuration space analysis of traveling salesman problems. RC 10972 (#49218), I.B.M., Jan. 1985.

[15] T. M. Liggett. *Interacting Particle Systems*, volume 276 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, New York, NY, 1985.

[16] M. Lundy and A. Mees. Convergence of the annealing algorithm. *Mathematical Programming*, 34:111–124, 1986.

[17] R. J. McEliece. *The Theory of Information and Coding*, volume 3 of *Encyclopedia of Mathematics and Its Applications*. Addison-Wesley, Reading, MA, 1977.

[18] M. Mihail. Conductance and convergence of Markov chains: A combinatorial treatment of expanders. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pages 526–531, 1989.

[19] D. Mitra, F. Romeo, and A. Sangiovanni-Vincentelli. Convergence and finite-time behavior of simulated annealing. *Advances in Applied Probability*, 18:747–771, 1986.

[20] F. Reif. *Fundamentals of Statistical and Thermal Physics*. McGraw-Hill, New York, NY, 1965.

[21] F. Romeo and A. Sangiovanni-Vincentelli. Probabilistic hill climbing algorithms. In *1985 Chapel Hill Conference on Very Large Scale Integration*, pages 393–417, 1985.

[22] F. I. Romeo. *Simulated Annealing: Theory and Applications to Layout Problems*. PhD thesis, University of California at Berkeley, Mar. 1989. Memorandum No. UCB/ERL M89/29.

[23] S. R. Ross. *Stochastic Processes*. Wiley, New York, NY, 1946.

[24] D. Saupe. Algorithms for random fractals. In H.-O. Peitgen and D. Saupe, editors, *The Science of Fractal Images*, chapter 2, pages 71–136. Springer-Verlag, New York, 1988.

[25] C. Sechen and A. Sangiovanni-Vincentelli. Timberwolf3.2: A new standard cell placement and global routing package. In *Proceedings of the 23rd Design Automation Conference*, pages 432–439, 1986.

[26] A. Sinclair and M. Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82:93–133, 1989.

[27] A. D. Sokal. New numerical algorithms for critical phenomena (multi-grid methods and all that). In D. P. Landau, K. K. Mon, and H.-B. Schüttler, editors, *Computer Simulation Studies in Condensed Matter Physics: Recent Developments*, pages 283–292. Springer-Verlag, Berlin-Heidelberg, 1988.

[28] S. A. Solla, G. B. Sorkin, and S. R. White. Configuration space analysis for optimization problems. In E. Bienenstock, F. F. Soulie, and G. Weisbuch, editors, *Disordered Systems and Biological Organization*, NATO ASI series. Series F, Computer and System Sciences, number 20, pages 283–292. Springer-Verlag, New York, 1986.

[29] G. B. Sorkin. Combinatorial optimization, simulated annealing, and fractals. RC 13674, I.B.M., Apr. 1988.

[30] G. B. Sorkin. Bivariate time series analysis of simulated annealing data. Technical Report UCB/ERL M90/6, Univ. of California at Berkeley, Jan. 1990.

[31] G. B. Sorkin. *Theory and Practice of Simulated Annealing on Fractal Landscapes.* PhD thesis, University of California at Berkeley, 1991. In preparation.

[32] R. F. Voss. Fractals in nature: From characterization to simulation. In H.-O. Peitgen and D. Saupe, editors, *The Science of Fractal Images*, chapter 1, pages 21–70. Springer-Verlag, New York, 1988.