An Introduction to Markov Chains and Markov Chain Monte Carlo

A.E. Charman[*]
Department of Physics, University of California, Berkeley

Extensive terminology and theory have developed around the constructs of *Markov chains* and *Markov processes*, but here we will focus on just the basic aspects of the theory of particular relevance for typical applications in statistical physics.

I. BACKGROUND ON MARKOV CHAINS

Markov chains model random processes where successive states are dependent, but with limited memory, such that the conditional probabilities of future states depend only on the *most recent* past. Specifically, a *discrete-time, time-homogeneous, finite Markov chain* is a stochastic process over a finite set of mutually exclusive and exhaustive states, in which a system passes through a sequence of states at what can be regarded as a sequence of discrete time-steps, and where the conditional probability of ending up in a particular state at a particular time-step, given knowledge of the state occupied at the previous time-step, is stochastically independent of the total elapsed time, and of any earlier history prior to the most recent time-step. (Generalizations involving countable or even uncountable state spaces, to continuous time evolution, and to inhomogeneous, or explicitly time-dependent, transition probabilities, are all possible, leading to a broader theory of *Markov processes*, as well as other sorts of *Markov networks* that are prevalent in machine learning).

That is, in a Markov chain, knowing the state in the immediate past makes any knowledge of the state further in the past irrelevant for predicting the future.

Given a set $\mathcal{M}$ of distinct states $a, b, c, \ldots$, let $\mathcal{T}(a \to b)$ denote the transition probability from a to b, defined as the conditional probability that the state b will be occupied at a time-step t_{i+1}, given that the state was known to be a at time-step t_i. If $P(a \,|\, t_i)$ represents the probability distribution over states at time t_i, then the *law of total probability*, combined with the Markovian "memoryless" property, ensures that

$$P(b \,|\, t_{i+1}) = \sum_{a \in \mathcal{M}} P(a \,|\, t_i)\, \mathcal{T}(a \to b),$$

where the sum is over all possible predecessor states in $\mathcal{M}$. For those of a visual bent, such Markov chains may be conveniently visualized as weighted, *directed graphs*, where nodes of the graphs represent states, and edges are labeled by the non-zero transition probabilities between states. If we think of the probability distribution (as to the state occupied at a time-step t) as a vector in $\mathbb{R}^{|\mathcal{M}|}$, with components restricted to satisfy $P(a \,|\, t) \geq 0$ and $\sum\limits_a P(a \,|\, t) = 1$, and interpret the transition probabilities $\mathcal{T}(a \to b)$ as the elements of an $|\mathcal{M}| \times |\mathcal{M}|$ transition matrix $\mathcal{T}$, satisfying the conditions

$$\begin{aligned} \mathcal{T}(a \to b) \geq 0; &\quad \text{for all pairs of states } a \text{ and } b \text{, and} \\ \sum_b \mathcal{T}(a \to b) = 1 &\quad \text{for every starting state } a, \end{aligned}$$

then the Markovian time evolution of the probability distribution is just equivalent to successive matrix-vector multiplications. Any matrix satisfying the former properties is known as a *stochastic* matrix. Whether this is written as multiplication of a row vector by a matrix to its right, or multiplication of a

[*] Electronic address: `acharman@physics.berkeley.edu`

column vector by a matrix on its left, is a matter of convention. Mathematicians tend to use the former convention, and physicists the latter. Either way, it is important not to confuse conventions, because the actual transition matrices will differ by a transposition. Notice that diagonal elements $\mathcal{T}(a \to a)$, if non-zero, correspond to the chance that the system remains (or at least begins and ends) in the same state over one time-step. Time evolution of the probability distribution over n successive time-steps will correspond to multiplication by the corresponding nth power $\mathcal{T}^n = \mathcal{T} \cdots \mathcal{T}$ of the transition matrix.

A probability distribution $\{\pi(a)\colon a \in \mathcal{M}\}$ over states is said to be a *stationary distribution*, a *steady-state distribution*, an *equilibrium distribution*, or an *invariant measure*, if it remains invariant under the time evolution induced by the Markov transitions, meaning

$$\pi(b) = \sum_a \pi(a)\, \mathcal{T}(a \to b) \;\text{ for all } b \in \mathcal{M}.$$

That is, if a state is chosen at random from the stationary distribution, and we update stochastically but according to the Markov chain dynamical rules, subsequent states will also be sampled (albeit not independently) from the stationary distribution. Again, thinking of the $\mathcal{T}(a \to b)$ as elements of a matrix, we see that a stationary distribution $\pi(a)$ (for all $a \in \mathcal{M}$) corresponds to an *eigenvector* of this transition matrix $\mathcal{T}$, with eigenvalue $\lambda = 1$, and with all vector components being nonnegative and summing to unity: $\pi(a) \geq 0$ for all $a \in \mathcal{M}$, and $\sum\limits_a \pi(a) = 1$.

A state $c \in \mathcal{M}$ is said to be *accessible* from a state $a \in \mathcal{M}$ when there is some non-zero probability of transitioning from a to c in some *finite* number of time-steps (not necessarily after just one step). Two states are said to *communicate* if each is accessible from the other. Communicability is an *equivalence relation* (being necessarily reflexive, symmetric, and transitive), and the corresponding equivalence classes are known as *communicating classes* of the Markov chain.

A state is said to be *recurrent* if, starting in that state, the probability of eventually returning to the state is unity. Otherwise the state is *transient*, and there is some chance of never returning to the state. Equivalently, we can say that a recurrent state is one in which the expected total number of future revisits is infinite, while a transient state is one for the system is only expected to re-enter the state at most a finite number of times. This also entials that a state a is recurrent if and only if $\sum\limits_{k=0}^{\infty} \mathcal{T}^k(a \to a) = \infty$, and is transient otherwise.

Furthermore, a recurrent state is said to be *positive recurrent* if the expected waiting time to return is finite, and *null recurrent* otherwise. The *period* $\tau(a)$ of state a has is the greatest common divisor of all positive n for which $\mathcal{T}^n(a \to a) > 0$. If $\tau(a) > 1$, we say that the state is periodic with period $\tau(a)$. If $\tau(a) = 1$, the state is instead said to be *aperiodic.*

All states within a given communication class must be of the same periodicity and the same recurrence/transience properties. An entire communicating class may therefore may be described as periodic or aperiodic, and transient or (positive or null) recurrent.

A finite Markov chain is said to be *irreducible*, or *ergodic*, if every state is accessible from every other state; otherwise it is *reducible.* Equivalently, we can say that an irreducible Markov chain is one which has a single communication class, containing all the states. For any irreducible Markov chain, the entire chain will be either transient or recurrent, and if recurrent, either positive recurrent or null recurrent. An irreducible Markov chain over a finite state space is always recurrent.

Next we consider the question of the possibility of time-reversal of Markov chains. In general, the transition matrix $\mathcal{T}$ need not be invertible, even when the chain is irreducible, or even recurrent. But let $X_1, \ldots, X_n$ be a sample of n successive steps from an irreducible Markov chain with stationary probabilities $\pi(a)$ and transition probabilities $\mathcal{T}(a \to b)$, and suppose X_1 is drawn from the stationary distribution. Then from Bayes' theorem, it can be verified that $Y_1 = X_n, Y_2 = X_{n-1}, \ldots, Y_n = X_1$ is a sample of n successive steps from an irreducible Markov chain also with stationary distribution $\pi(a)$, but transition probabilities $\frac{\pi(b)}{\pi(a)}\, \mathcal{T}(b \to a)$ for going from b to a, and where Y_n is also drawn from the stationary distribution $\pi(a)$.

A Markov chain is then said to be *reversible* if its stationary distribution $\pi(a)$ satisfies *detailed balance*, whereby

$$\pi(a)\,\mathcal{T}(a\to b)=\pi(b)\,\mathcal{T}(b\to a)$$

for every pair of states a and b in $\mathcal{M}$. This says that in equilibrium, the probability of being in a, and then subsequently jumping to b, is equal to the probability of being in b and then jumping to a.

Conversely, if a probability distribution $\pi(a)$ satisfies detailed balance, then it must be a stationary distribution for the chain, since

$$\sum_a \pi(a)\,\mathcal{T}(a\to b)=\sum_a \pi(b)\,\mathcal{T}(b\to a)=\pi(b)\sum_a \mathcal{T}(b\to a)=\pi(b).$$

for all states b.

From the results on time-reversed chains mentioned above, it also follows that a reversible Markov chain is one for which the transition probabilities of the forward and time-reversed chains are the same *in equilibrium*. To see this explicitly, let X_t be the state of the system at time-step t. Then in an equilibrium characterized by detailed balance, it follows from the product rule of probability that

$$P(X_t=b)\,P(X_{t-1}=a\,|\,X_t=b)=P(X_{t-1}=a,X_t=b)=P(X_{t-1}=a)\,P(X_t=b\,|\,X_{t-1}=a)$$

for any pair of states a and b. But the detailed balance and stationarity properties together ensure that

$$\begin{aligned}P(X_{t-1}=a)\,P(X_t=b\,|\,X_{t-1}=a)&=P(X_{t-1}=b)\,P(X_t=a\,|\,X_{t-1}=b)\\&=P(X_t=b)\,P(X_{t+1}=a\,|\,X_t=b),\end{aligned}$$

and after equating these expressions and canceling the common factor of $P(X_t=b)$, we find under detailed balance that

$$P(X_{t-1}=a\,|\,X_t=b)=P(X_{t+1}=a\,|\,X_t=b),$$

for all pairs of states a and b. This can be generalized to say

$$P(X_{t-k}=a\,|\,X_t=b)=P(X_{t+k}=a\,|\,X_t=b),$$

for any integer k. This perhaps offers the most transparent meaning of the reversibility condition: in an equilibrium characterized by detailed balance, the forward-time and time-reversed conditional probabilities are the same, in the sense that starting in equilibrium in any state at the present moment, the probability of finding the system at some future time in some given state is equal to the probability that the system would have been found in this same state if observed the same amount of time in the past.

Note that such time-reversibility does not require *symmetry* of the transition matrix $\mathcal{T}$, although a symmetric transition matrix, satisfying $\mathcal{T}(a\to b)=\mathcal{T}(b\to a)$ for all pairs of states a and b, will be time-reversible, with respect to a uniform equilibrium distribution $\pi(a)=\pi(b)=\pi(c)=\cdots$.

Fulfilling such a detailed balance condition is a sufficient, but not necessary, prerequisite for a finite Markov chain to possess a stationary distribution. But in practice, many Markov chains that arise in physical applications do exhibit detailed balance, often by design, or by virtue of the nature of the underlying microscopic model. (Frequently, we actually end up imposing an even stronger, "intra-channel" form of detailed balance, based on the following property: if, for dynamical "channels" labeled by $j=1,2,\ldots,$ the matrix elements $T_j(a\to b)$ are sets of Markov transition probabilities, each set separately satisfying detailed balance conditions with respect to the *same* stationary distribution $\pi(a)$ over the same space of states, then the mixture transitions

$$\begin{aligned}T(a\to b)&=\sum_j w_j(a,b)\,T_j(a\to b),\quad\text{such that}\\ w_j(a,b)&\geq 0,\quad\text{and}\\ \sum_j w_j(a,b)&=1,\end{aligned}$$

define another reversible Markov chain over the same states, and with the same stationary distribution. That is to say, suppose transitions $a \to b$ can occur in a number of different ways, perhaps depending on the values of other *latent* variables or *nuisance* variables not included explicitly in the Markov state descriptions $a, b, \ldots$ proper. If detailed balance holds separately within each possible pathway, in the sense that

$$\pi(a)\, w_j(a,b)\, \mathcal{T}_j(a \to b) = \pi(b)\, w_j(b,a)\, \mathcal{T}_j(b \to a),$$

then, upon summing over j, we see that detailed balance will hold for the mixture transitions as a whole).

It turns out that any irreducible (i.e., ergodic) Markov chain on a finite state space will have a *unique* stationary probability distribution $\pi(b)$, which will have support (i.e., some non-zero probability) on all states $b \in \mathcal{M}$. The stationary distribution can be determined by solving for the detailed balance conditions, or in principle as a limit of an average,

$$\pi(b) = \lim_{n \to \infty} \tfrac{1}{n} \sum_{k=1}^{n} \mathcal{T}^k(a \to b) \tag{3}$$

starting from any initial state $a \in \mathcal{M}$.

Furthermore, any irreducible, positive recurrent Markov chain also must have a unique stationary distribution, where each probability $\pi(a)$ can be interpreted as the reciprocal of the average waiting time between visits to the state a. While this is a bit involved to prove rigorously, hopefully it makes sense intuitively: in the long run, the system would be anticipated to spend about a fraction $\pi(a)$ of the time in any state a, so,on average, we would expect to have to wait $\frac{1}{\pi(a)}$ steps between visits to a. Also, any irreducible Markov chain over a *finite* state space must automatically be positive recurrent, and hence possesses a unique stationary distribution whose probabilities are just equal to the reciprocal of the average recurrence times.

Aperiodicity is of interest because it entails another convergence result. Specifically, a positive recurrent Markov chain will satisfy

$$\pi(b) = \lim_{n \to \infty} \mathcal{T}^n(a \to b)$$

starting from any initial state a, if and only if the chain is also aperiodic. Mathematically, this is known as weak convergence, although it is a rather strong result. It follows that any weakly convergent chain (starting from any initial state) must also satisfy the *equilibration* condition,

$$\pi(a) = \lim_{n \to \infty} P(a \,|\, t_n)$$

for all states a, starting from *any* initial probability distribution $P(a \,|\, t_0)$. (Warning: in some sources, ergodicity is instead defined as the combination of positive recurrence and aperiodicity, but this does not conform with other used of ergodicity in mathematics, and should be avoided). For a positive recurrent, aperiodic Markov chain, the *Kolmogorov criterion* says that the chain is reversible if and only if the product of transition probabilities over any closed loop of states is the same in both directions of traversal around the loop.

A finite Markov chain is said to be *regular* if some finite, positive power $\mathcal{T}^n = \mathcal{T} \cdots \mathcal{T}$ of the transition matrix $\mathcal{T}$ has all positive matrix elements—that is, if for some positive integer n, the matrix elements satisfy $\mathcal{T}^n(a \to b) > 0$ strictly for every pair of states a and b, meaning that, starting in any state, after some finite number n of time steps, there is some non-zero chance of ending up in any state. (For any finite Markov chain, it is easy to see that if $\mathcal{T}^n$ has all positive matrix elements, then so does $\mathcal{T}^{2n}$, $\mathcal{T}^{3n}$, and so on. What about the matrix $\mathcal{T}^{n+1}$? If $\mathcal{T}^n$ has all positive entries, then because $\mathcal{T}$ has all nonnegative entries, clearly the matrix element $T^{n+1}(a \to b) \geq 0$, and could equal zero if and only if $\mathcal{T}(c \to b) = 0$ for all states c, implying b would never be accessible from any other state after any number of time-steps, which contradicts the assumption of regularity. So we may infer that $\mathcal{T}^{n+1}$, $\mathcal{T}^{n+2}$, etc., have all strictly positive entries if $\mathcal{T}^n$ does.) Any regular, finite Markov chain will be irreducible, but not necessarily vice versa.

It can be shown that any finite, regular Markov chain ls both *positive recurrent*, (meaning that the expected waiting time to return to any state is finite), and *aperiodic* (meaning that the possible return times to any initial state are always relatively prime). For such a chain, it follows that a *unique* stationary measure also exists, satisfying the *equilibration* condition $\pi(a) = \lim_{n\to\infty} P(a\,|\,t_n)$.

However, none of this of course says anything about the *rate* at which the stationary distribution might be approached. (This depends on the other eigenvalues of the transition matrix).

II. MARKOV CHAIN MONTE CARLO

Standard, or simple, Monte Carlo (MC) methods sample directly (at least to some acceptable degree of approximation) from a probability distribution (typically using computer-generated, pseudo-random deviates) in order to perform simulations, numerical quadrature, or statistical estimation.

If it is impossible or impractical to generate samples from the target distribution directly, then Markov Chain Monte Carlo (MCMC) methods can instead sample a Markov chain whose stationary distribution corresponds to the target probability distribution. In fact, such MCMC methods are applicable, and particularly useful, even in situations where we are unable even to normalize the full probability distribution of interest, because calculation of the partition function or normalization constant is prohibitively difficult.

Also note that If we are trying to simulate thermodynamic properties, efficient but good-quality pseudo-random number generators are often essential, since many MCMC methods applied to physical models that exhibit phase transitions or so-called critical behavior behavior turn out to be quite sensitive to the statistical properties of the random deviates. In fact, simulation of the scaling behavior of Ising models has provided some of the most stringent tests of computer-generated pseudo-random numbers.

Most MCMC methods rely on finding *regular* and *reversible* Markov chains with some desired invariant measure $\pi(a)$. After some *burn-in*, *mixing*, or *equilibration* period, it can be expected (or at least hoped!) that subsequent samples are being drawn from something close to the stationary distribution of the chain. Of course, even if draws are described according to the correct *marginal* distribution $\pi(a)$, successive samples may be strongly correlated, so we must wait for some *mixing* or *de-correlation* time interval between simulated samples if we want to acquire samples which can be approximated as stochastically independent. In practice, if statistical averages are of interest, all the equilibrated samples from the chain can be used in the average, but because of stochastic dependencies between successive steps, the precision of such an estimate only improves with the reciprocal square-root of the *effective* number of independent samples, not the total number of samples. Rigorous theoretical bounds on these equilibration and mixing rates are very often elusive, so practical use of MCMC often relies on empirical judgement hopefully backed up by numerical evidence and experience.

To describe many standard MCMC algorithms, it is convenient to first factorize the state transitions and corresponding probabilities into *attempt*, or suggestion sub-steps, where a candidate state transition to some "neighboring" state is proposed, followed by *accept/reject* sub-steps, where a randomized decision is made whether to actually effect the proposed transition or not. (By neighbors, we will mean pairs of states that are accessible in just one transition). This leads to transition probabilities of the form

$$\mathcal{T}(a \to b) = \mathcal{S}(a \to b)\,\mathcal{A}(a \to b),$$

where, starting from state a, $\mathcal{S}(a \to b)$ is the probability of suggesting or proposing a candidate transition to state b, and $\mathcal{S}(a \to b)$ is the probability, given that this particular state transition $a \to b$ has been proposed, that the transition is accepted and thence performed.

Typically, the transition probabilities are carefully engineered so that the resulting Markov chain is (i) ergodic, and moreover, regular (or at least aperiodic), and also (ii) reversible, satisfying the detailed balance condition appropriate to the desired target distribution:

$$\pi(a)\,\mathcal{S}(a \to b)\,\mathcal{A}(a \to b) = \pi(b)\,\mathcal{S}(b \to a)\,\mathcal{A}(b \to a),$$

or equivalently,

$$\frac{\mathcal{A}(a \to b)}{\mathcal{A}(b \to a)} = \frac{\pi(b)}{\pi(a)} \frac{\mathcal{S}(b \to a)}{\mathcal{S}(a \to b)}.$$

For the purposes of statistical averaging and bookkeeping, a rejection of a proposed transition $a \to b$ at any time-step t_i is usually to be interpreted as a "do-nothing" or "idempotent" transition $a \to a$ in the simulated chain, from a back to itself, over the time interval between t_i and t_{i+1}.

In addition to (i) regularity and (ii) reversibility, we also would like the Markov chain to be (iii) efficient, in the sense of leading to reasonably fast equilibration and mixing times. Different Markov chains with the same stationary distribution can have very different rates of equilibration and de-correlation, and different scalings of these rates with respect to problem size. Finding a good set of transition rules is not always easy or obvious.

III. SIMULATED ANNEALING

Simulated Annealing (SA) refers to a class of MCMC techniques that seek solutions to combinatorial (or continuous) optimization problems by way of an analogy to the annealing, or melting and re-solidification, in thermodynamics. Simulated annealing can prove effective for problems where more traditional deterministic optimization algorithms falter because of the shear number of possible states, and/or the existence of many local minima.

The underlying idea is based on recognition that if a thermodynamic system is first heated to a high temperature, then gradually (i.e., quasi-statically) cooled to very low temperature, the system will, with high probability, find its energetic ground state. However, if cooled too fast, the system may be 'quenched" rather than "annealed," and end up effectively trapped near a metastable, local but not global energy minimum. But If cooled slowly enough, the system is afforded an opportunity to jump out of these local minima by virtue of thermal fluctuations, and eventually relax its way toward the global minimum.

The idea of SA is to introduce an "objective function" $H(a)$ which varies as a function of the occupied configuration a in the set $\mathcal{M} = \{a, b, \dots\}$ of possible states, and whose (initially unknown) minimum a^* will correspond to the solution of some optimization problem of interest. For example, SA has been successfully applied to the "traveling salesperson" problem, where the goals is to find the shortest path through a set of cities, or to problems of optimally arranging logic gates on an integrated circuits to minimize writing costs.

This objective function is then treated as the Hamiltonian (i.e., energy function) for an imaginary statistical mechanical system, which at any time can occupy one and only one of the "microstates" in the set $\mathcal{M}$. In thermodynamic equilibrium at fixed temperature T, probabilities over these microstates will be described by a canonical Boltzmann distribution, where the relative odds of finding the system in state a versus state b will be an exponential function of the differences in their respective energies,

$$\tfrac{\pi(a)}{\pi(b)} = e^{-\beta[H(a)-H(b)]}$$

in which $\beta = \frac{1}{k_{\mathrm{B}}T}$ is the usual "coolness" or inverse temperature parameter. So as the temperature drops, the probability of finding the system in a lower-energy state should increases exponentially, as long as the system remains near equilibrium.

The trick is to find an efficient Markov chain with the desired canonical distribution as its stationary distribution, and an annealing schedule for lowering the temperature which keeps the simulated system near equilibrium without wasting too much time on unnecessary iterations.

At any given temperature T, the Markov chain is to be defined by a distribution over the suite of possible candidate transitions out of any given state, and the acceptance/rejection probabilities given a proposed transition, which together should lead to the Boltzmann distribution as the unique stationary distribution for the chain. Different choices of suggestion and acceptance rules, all resulting in the same equilibrium distribution, can still lead to very different equilibration times and overall algorithmic efficiency.

A. Transition Proposals

At least at sufficiently low temperatures, we would typically not want to propose candidate trnasitions to new states completely at random. If the system is already in a lower-energy state, most jumps to arbitrary destination states would make things much worse rather than better energetically speaking, and we would end up wasting a lot of computer time rejecting most of these unhelpful suggestions. But it still must be at least possible to climb out of local energy minima with a series of jumps. This implies that we should skew the suggested transitions towards those effecting small jumps between states of *similar* energy, while still ensuring the possibility of getting between any two states in a finite number of moves. What we mean by "similar" can either be kept fixed for the sake of simplicity, or shrunk in proportion to $k_{\mathrm{B}}T$ as the temperature is lowered.

That is, states that are "neighboring" in the sense of being accessible in a single transition should mostly also be "nearby" in energy space. When choosing the distribution of possible candidate moves, keep in mind that the overarching goal will entail achieving modest, gradual improvements in energy on average as the system is progressively cooled to lower temperatures, while occasionally allowing for jumps to higher energy to avoid getting trapped in local minima.

To refrain from wasting a lot of time attempting then rejecting many bad transitions with large increases in energy, this strategy of proposing energetically similar transitions does however make it unlikely to achieve substantially large drops in energy in any one step. But for most optimization problems of the sort that we would be motivated to tackle with simulated annealing, there will be many more very bad energetic moves than very good moves, so it would be unlikely to randomly stumble upon the latter anyway. So suggesting modest moves to "nearby" states with small changes in energy should be more effective in the end than vainly searching for dramatically good leaps.

This is also closer to how real thermodynamic systems behave. Imagine we put a system in thermal contact with a heat bath consisting of a surrounding gas. In any one collision between a part of the system and one molecule of the heat bath, an energy exchange of order $k_{\mathrm{B}}T$ is typical.

However, when choosing transition proposal rules, it will also be important to try to avoid introducing artificial barriers, where a cluster of neighboring states that excludes the global minimum nevertheless all have substantially lower energy than the immediately accessible neighbors outside of the clusters. These subsets, called "semi-closed catchment basins" could then trap the system for exponentially long stretches of simulation time.

B. Acceptance/Rejection Rules

Once we decide how to stochastically generate candidate transitions, we need to probabilistically accept or reject these transitions so as to guarantee that the desired Boltzmann distribution emerges as the stationary distribution of the Markov chain. Once again, the idea is to jump between "nearby" states in state space, preferring lower energy neighbors overall but sometimes allowing thermal fluctuations to higher-energy states to avoid getting stuck.

The first, and still one of the most widely used procedures for MCMC acceptance/rejection decisions is known as the *Metropolis-Hastings* algorithm, developed at Los Alamos. (Really this should be called the Rosenbluth-Rosenbluth-Hastings algorithm, for, according to reliable accounts, it was apparently physicist Marshall Rosenbluth who had the original idea, and his wife Arianna who developed the details and wrote the original computer program. Metropolis contributed nothing except computer time, but got his name first. Edward Teller's contributions were small at best, but he then insisted that he and his spouse also be added as co-authors to the original paper. Somewhat later, Hastings extended the algorithm in useful ways, to allow for more general transition sets with non-symmetric probabilities of suggested transitions).

We assume that we have settled on a set of possible transitions between each pair of states $\mathcal{M} = \{a, b, c, \dots\}$, along with their corresponding attempt probabilities $\mathcal{S}(a \to b)$, chosen such that the resulting Markov chain will be *regular* assuming only that acceptance probabilities are sometimes non-zero. Starting in some state a, a candidate transition $a \to b$ is first suggested at random according to these probabilities, then accepted with probability

$$\mathcal{A}(a \to b) = \min\Big[\, \frac{\pi(b)}{\pi(a)} \frac{\mathcal{S}(b \to a)}{\mathcal{S}(a \to b)} \, , \, 1 \Big].$$

Otherwise, the suggested transition is rejected, implicitly in favor of the transition $a \to a$ back to the old state. Since for any nonnegative real number $x > 0$, $x \leq 1$ if and only if $x^{-1} \geq 1$, this acceptance/rejection rule ensures that detailed balance holds for the target stationary distribution $\pi(a)$. Because the Markov chain was presumably designed to be regular, this in turn implies that $\pi(a)$ is the unique equilibrium distribution of the chain. Such a choice of acceptance probabilities also tends to increase the rate of acceptances given the actual suggestions, and to foster escape from local minima, compared to some other variants of the MCMC method.

In the case of simulated annealing, the target distribution is chosen to be the Boltzmann distribution, such that $\pi(a) \propto e^{-\beta H(a)}$. Since only ratios of probabilities will appear, this distribution need never be normalized in order to calculate the acceptance probabilities, which is a tremendously useful feature of MCMC approaches when applied to systems with large state spaces. In this case of the Boltzmann distribution, the acceptance probability then becomes

$$\mathcal{A}(a \to b) = \min\Big[\, e^{-\beta[H(b) - H(a)]} \, \tfrac{\mathcal{S}(b \to a)}{\mathcal{S}(a \to b)} \, , \, 1 \Big].$$

In many situations, the forward and reverse suggestion probabilities are chosen to be equal, so that $\frac{\mathcal{S}(b \to a)}{\mathcal{S}(a \to b)} = 1$, and this Metropolis-Hastings rule further simplifies. In these circumstances, we see that the a candidate transition is always accepted if it lowers the energy, but is sometimes accepted even if it raises the energy, where the chance of fluctuating upward in energy decreases with decreasing temperature.

C. Annealing Schedule

In practice, choosing an effective *annealing schedule* that specifies how and how quickly the temperature is lowered is crucial, but often involves more trial-and-error numerical experimentation than reliable theory. Cool too fast, and the system quenches before finding the global minimum. Cool too slow, and we run out of patience before the system finds the minimum. But how slow is sufficiently slow can unfortunately involve a great deal of subtlety depending on the particular details of the system's state space and Hamiltonian.

Ideally, the initial temperature is chosen large enough so that large majority of proposed early transitions will be accepted, and the state becomes effectively randomized. The final temperature reached should be low enough so that the ground state is predcited to enjoy substantially more probability of occupancy than even low-lying excited state(s).

If our goal is to find the minimum of $H(a)$ rather than faithfully simulate thermodynamic properties *per se*, then our efforts are not necessarily invalidated by a cooling a bit faster than we ideally should for true equilibration at each temperature setting. However, if we push our luck in this regard, the system can quench. But it is a good idea to keep track of not only the current state and energy, but the best minimum found so far, in case things do not go to plan.

Here are a few suggestions as to how to adjust the temperature during simulated annealing:

- At the tth time step ($t = 0, 1, 2, 3, \ldots$), use the reciprocal temperature $\beta(t) = B\,\log(\Omega t + 1)$, for some positive constants B and Ω. Some amount of theoretical work has focused on studying mixing and convergence using such a logarithmic cooling schedule, but in practice this tends to be too slow to be workable;

- after every m moves (attempted), reduce the temperature from T to $T' = (1 - \epsilon)T$, for some constant ϵ in the range $0 < \epsilon < \frac{1}{2}$. Ideally we should choose m to be at least a few times the minimum number of individual steps that would be required to move between any pair of states. So for example, if the states consist of all permutations of n lables (with $n!$ states in all), and we generate proposed transitions consisting of pairwise permutations, we might choose m to be a few times n;

- budget a large total of K moves in all, and after every m moves at the current temperature (where $1 \ll m \ll K$), and a total of k time steps so far (throughout the entire simulation), adopt a new temperature $T = T_0(1 - \frac{k}{K})^\alpha$, where α is a positive constant (often chosen as an integer in the range $1 \le \alpha \le 4$), and T_0 is the initial temperature. The optimal value of α would depend on the distribution of minima of various depths relative to the global minimum, which is rarely known ahead of time. Larger values of α correspond to allotting more of the total budget of iterations to lower simulation temperatures;

- After every m moves at temperature T, set the new temperature to $T' = \max\big[(1 - \epsilon)T, \lambda \frac{E_T - E_B}{k_\mathrm{B}}\big]$, where λ is a positive constant of order unity, ϵ is again a constant factor in the range $0 < \epsilon < \frac{1}{2}$, E_T is the *lowest* energy found at the current temperature T, and E_B is the best (lowest) energy encountered throughout the entire simulation so far. This simple adaptive adjustment allows the temperature to go down as long as progress is being made, but go back up if it appears that we have cooled too fast or gotten stuck in a non-optimal part of state space. Other more elaborate adaptive adjustments can also be employed;

- Rather than choosing the number of time steps per temperature to remain fixed, we can also try assess equilibration adaptively based on thermodynamic considerations. For example, we can make use of the Einstein fluctuation formula. Ignore the first b “burn-in”’ steps at any given temperature, then begin to keep a running estimate of the average energy $\bar{E}(T)$ and variance in energy $\sigma_E^2(T)$ at the current simulation temperature, re-calculated after every s steps at temperature T, where $s > 1$. In thermal equilibrium, these should satisfy $\sigma_E^2(T) = k_\mathrm{B} T^2 C(T)$, where the heat capacity $C(T) = \frac{\partial \bar{E}}{\partial T}$ can be estimated from a finite difference with respect to the current temperature T and previous temperature $T + \Delta T$: $C(T) \approx \frac{\bar{E}(T + \delta T) - \bar{E}(T)}{\Delta T}$. If the Einstein relation is approximately satisfied within some reasonable tolerance, then we regard the system as equilibrated at the currently simulated temperature, and therefore ready to have the temperature adjusted downward. How *much* the temperature is lowered can also be informed by the estimated heat capacity. If $C(T)$ is rising precipitously, this may be the signature of a “phase transition” in the system, requiring a more gradual adjustment in the temperature. Also, by numerically integrating $C(T)/T$, starting from the high initial temperature to the current temperature once we predict equilibration, we can also estimate the entropy of the simulated system, and thereby get a (rough) sense for the effective number of microstates that are accessible at a given stage of optimization;

- Sometimes, it is may be possible to estimate self-consistently the expected fraction of proposed moves that should be accepted (versus rejected) if the system really is equilibrated. The observed fraction can be compared to the predicted fraction, and if sufficiently close, we can then lower the temperature, according to some rule like one of those given above.

D. Starting and Restarting

Initially, the state should probably be chosen at random, although in any case it should be “mixed” for a while at higher temperatures before the too much cooling.

Whether converged or seemingly stuck, if one can afford the time, it is generally useful to restart any annealing simulation. Re-starting at the current temperature bur from the lowest-energy state found so far (at any temperature) can sometimes be effective to nudge the system toward greater progress. Re-starting the entire simulation from a different, randomly chosen initial condition can be helpful to assess convergence and the extent of distinct minima.

E. Some Bookkeeping and Other Implementation Details

To avoid keeping track of extraneous factors like Boltzmann's constant k_{B}, it is convenient to work in units where $k_{\mathrm{B}} = 1$, or equivalently, utilize a scaled temperature $\tau = k_{\mathrm{B}}T$, and choose the overall scaling of the Hamiltonian so that this temperature τ and typical energy jumps $\Delta E = H(a) - H(b)$ will fall within convenient numerical ranges.

For debugging and transparency purposes, it is generally s good idea to save the initial state, initial seed for the random number generator, and other details to enable replication.

Store the current state and its energy separately from the proposed state and energy, so it will be quick to reject or accept the proposed transition without much additional effort.

In addition, store the best state and energy found so far, and the best found at the current simulation temperature.

Other statistics, such as average and RMS energy, estimated entropy, Markov chain rejection/acceptance rates, etc., will often be of interest.