I just wanted to give you a thumbnail sketch of the Metropolis-Hastings algorithm that we went over in class. Let's start with a problem. You have a large set $\mathcal{S}$ that we'll assume is finite for simplicity. To fix notation, suppose $\mathcal{S} = \{1, 2, 3, \ldots, M\}$. Let π be a probability distribution on $\mathcal{S}$, and assume that $\pi_j > 0$ for all $j \in \mathcal{S}$. Your goal is to compute expected values of the form

$$E_\pi(f) = \sum_{j \in \mathcal{S}} f(j)\pi_j$$

for various functions $f : \mathcal{S} \to \mathbb{R}$. To make matters more complicated, you might not know π exactly to begin with. Let's assume you know π up to an arbitrary multiplicative constant, so you have $p_j = c_o\pi_j$ for every $j \in \mathcal{S}$ for some unknown $c_o > 0$. Note that

$$c_o = \sum_{j \in \mathcal{S}} p_j$$

because $\sum_{j \in \mathcal{S}} \pi_j = 1$, but $\mathcal{S}$ might be so large and the formulas for the p_j so complicated that computing c_o is intractable.

To find $E_\pi(f)$, it would suffice to draw an independent sequence of samples $\{Z_m : m > 0\}$ from $\mathcal{S}$ identically distributed according to π. The Strong Law of Large Numbers would then imply that, with probability 1,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} f(Z_m) = E_\pi(f) \ .$$

But how to generate the samples Z_m? And might there be a better way to find $E_\pi(f)$?

Suppose we could construct an irreducible Markov chain with state space $\mathcal{S}$ whose unique stationary distribution π^* was π. If X_n is the state of the Markov chain at time $n > 0$ starting from any initial state in $\mathcal{S}$, the Ergodic Theorem for Markov chains guarantee that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} f(X_m) = E_{\pi^*}(f) = E_\pi(f) \tag{1}$$

with probability 1. Furthermore, the multi-set of states you obtain by recording the various states visited by a typical Markov-chain run $\{X_n : n > 0\}$ will be distributed over $\mathcal{S}$ according to π. Coming up with such a Markov chain would not only generate a population of points from $\mathcal{S}$ distributed according to π, but would also provide a recursive way of approximating $E_\pi(f)$ for any f: simply run the Markov chain and apply equation (1).

Many Markov chains on $\mathcal{S}$ will have π as a stationary distribution. The Metropolis-Hastings algorithm provides an ingenious technique for constructing irreducible such chains that have π as their (necessarily unique) stationary distribution. Before presenting the algorithm itself, I'll need to go over some preliminaries.

First consider an arbitrary homogeneous Markov chain with state space $\mathcal{S}$ and transition probabilities $P(i, j)$. If $\overline{\pi}$ is any stationary distribution for the Markov chain, then

$$\sum_{i\in\mathcal{S}} \overline{\pi}_i P(i,j) = \overline{\pi}_j \ \text{ for every } \ j \in \mathcal{S}\ .$$

People call this set of equations the *balance conditions,* and $\overline{\pi}$ is a stationary distribution if and only if $\overline{\pi}$ satisfies them

Because the $P(i,j)$ are transition probabilities, $\sum_{i\in\mathcal{S}} P(j,i) = 1$ for every $j \in \mathcal{S}$. Accordingly,

$$(2) \qquad \sum_{i\in\mathcal{S}} \overline{\pi}_i P(i,j) = \left(\sum_{i\in\mathcal{S}} P(j,i)\right)\overline{\pi}_j = \sum_{i\in\mathcal{S}} P(j,i)\overline{\pi}_j$$

for every $j \in \mathcal{S}$.

Equations (2) recast the balance conditions as the equality between two sums over $i \in \mathcal{S}$. One way, but not the only way, for a distribution π to satisfy the balance conditions is for the two infinite sums in (2) to be equal term-by-term, i.e.

$$(3) \qquad \pi_i P(i,j) = P(j,i)\pi_j \ \text{ for all } \ i,j \in \mathcal{S}\ .$$

The conditions in (3) are the *detailed balance conditions.* It follows that **if** π satisfies the detailed balance conditions (3), **then** π is a stationary distribution for the Markov chain. The result is even sharper when the Markov chain is irreducible, in which case it has a unique stationary distribution π^*. If a distribution π on $\mathcal{S}$ satisfies (3), then π is stationary and therefore must equal π^*.

Back now to the problem we started with. π is a distribution on $\mathcal{S} = \{1, 2, 3, \ldots, M\}$ satisfying $\pi_j > 0$ for all $j \in \mathcal{S}$. We don't necessarily have access to π_j, but we do have access to $p_j = c_o\pi_j$ for some $c_o > 0$. Let $Q(i,j)$ be any set of transition probabilities of a Markov chain on $\mathcal{S}$ that satisfies the following conditions:

- $Q(i,j) = Q(j,i)$ for every i and j in $\mathcal{S}$.
- The Markov chain with transition probabilities $Q(i,j)$ is irreducible.

Many choices of $Q(i,j)$ are possible. Perhaps the simplest example is

$$Q(i,j) = \begin{cases} q & \text{if } j = i+1 \text{ or } j = i-1 \\ r_i & \text{if } j = i\ , \end{cases}$$

which corresponds to a random walk on $\mathcal{S}$. Note that $r_i = 1 - 2q$ if $2 \le i \le M-1$ while $r_1 = r_M = 1 - q$. Now define $P(i,j)$ for every i and j in $\mathcal{S}$ as follows:

$$P(i,j) = \begin{cases} Q(i,j) & \text{if } i \ne j \text{ and } p_j \ge p_i \\ Q(i,j)p_j/p_i & \text{if } i \ne j \text{ and } p_j < p_i \\ Q(i,i) + \sum_{k\in\mathcal{S}_i} Q(i,k)\,(1 - p_k/p_i) & \text{if } i = j\ , \end{cases}$$

where $\mathcal{S}_i = \{k \in \mathcal{S} : p_k < p_i\}$. You can check easily that $\sum_{j\in\mathcal{S}} P(i,j) = 1$ for every $i \in \mathcal{S}$, so the $P(i,j)$ are transition probabilities for a Markov chain on $\mathcal{S}$. Furthermore, since $P(i,j) > 0$ if $Q(i,j) > 0$, the chain with transition probabilities $P(i,j)$ is irreducible.

It turns out that π is the unique stationary distribution for the P-chain. To see why, just check the detailed-balance conditions (3). Keep in mind that $p_j/p_i = \pi_j/\pi_i$ for all i and j because the c_o-factor cancels. Furthermore, the detailed balance conditions always hold when $i = j$, so it suffices to make sure they hold when $i \ne j$.

If $i \neq j$ and $p_j < p_i$, then

$$\begin{aligned} P(i,j) &= Q(i,j)p_j/p_i \\ &= Q(j,i)p_j/p_i \\ &= P(j,i)p_j/p_i \ , \end{aligned}$$

where the first line holds by definition of $P(i,j)$, the second by symmetry of $Q(i,j)$, and the last again by definition of $P(j,i)$, which must equal $Q(j,i)$ because $p_i > p_j$. It follows that

$$\pi_i P(i,j) = P(j,i)\pi_j$$

for every i and j in $\mathcal{S}$ with $\pi_j < \pi_i$, A similar argument works when $\pi_j \geq \pi_i$.

So we've accomplished our mission, which was to produce an irreducible Markov chain on $\mathcal{S}$ with stationary distribution π. By running the chain starting from an arbitrary initial state we can use (1) to approximate $E_\pi(f)$ for functions $f : \mathcal{S} \to \mathbb{R}$. We can also produce a set $\mathcal{P}$ of samples from $\mathcal{S}$ distributed according to π. Here's an algorithmic description of how to construct $\mathcal{P}$, which is actually a population (i.e. a multi-set) of points in $\mathcal{S}$:

Initialization: Set $\theta_0 = i$, where $i \in \mathcal{S}$ is an arbitrary state. Set $\mathcal{P} = \{\theta_0\}$. Proceed to the Proposal Step.

Proposal Step: Given θ_m, choose $\psi_{m+1} \in \mathcal{S}$ according to transition probabilities $Q(i,j)$; that is, set $\psi_{m+1} = j$ with probability $Q(i,j)$ when $\theta_m = i$. Proceed to the Accept-Reject Step.

Accept-Reject Step: When $\theta_m = i$ and $\psi_{m+1} = j$,

- If $p_j \geq p_i$, set $\theta_{m+1} = \psi_{m+1}$.
- If $p_j < p_i$, set $\theta_{m+1} = \psi_{m+1}$ with probability p_j/p_i and set $\theta_{m+1} = \theta_m$ with probability $1 - p_j/p_i$.

Add θ_{m+1} to $\mathcal{P}$ and return to the Proposal Step.

This description reflects the algorithm's original formulation by Metropolis et al. They were addressing problems in statistical mechanics. For them, the state space $\mathcal{S}$ was a fixed discretization of the phase space of a statistical-mechanical system. The distribution π was the so-called *Boltzmann distribution* or *canonical ensemble* on $\mathcal{S}$. In other words,

$$\pi_j = (1/c_o)p_j = (1/c_o)e^{-\frac{\mathcal{E}(j)}{kT}}$$

for every $j \in \mathcal{S}$, where $\mathcal{E}(j)$ is the energy of state j, T is the temperature of the system, k is Boltzmann's constant, and c_o is an unkown normalization constant. Metropolis et al. initialized their algorithm by starting from some arbitrary initial state $i \in \mathcal{S}$. They generated a population $\mathcal{P}$ of states as follows. First they proposed a next state j by perturbing the current state i according to a uniform distribution centered on i — that was the "$Q(i,j)$-part" of their procedure. If the proposed next state had lower energy than the current state (i.e. $p_j \geq p_i$), they accepted the proposal with probability 1. If the proposed next state had higher energy (i.e. $p_j < p_i$), they accepted it with probability

$$p_j/p_i = e^{-\frac{\mathcal{E}(j)-\mathcal{E}(i)}{kT}}$$

and rejected it (i.e. returned to state i) with probability

$$1 - p_j/p_i = 1 - e^{-\frac{\mathcal{E}(j)-\mathcal{E}(i)}{kT}} \ .$$

Accepting a proposal of j resulted in adding a copy of j to $\mathcal{P}$ whereas rejecting j and returning to i added another copy of i to $\mathcal{P}$.